RIDER (Reference Database to Evaluate Response) Committee Combined Report, 9/25/2008 Sponsored by NIH, NCI, CIP, ITDB

Causes of and Methods for Estimating/Ameliorating Variance in the Evaluation of Tumor Change in Response to Therapy

CT Volumetrics subcommittee members:

Sam Armato
Reinhard Beichel
Luc Bidaut
Larry Clarke
Barbara Croft
Chuck Fenimore
Marios Gavrielides
Hyun (Grace) Kim
Lisa Kinnard
Geoffrey McLennan
Chuck Meyer
Nick Petrick
Anthony Reeves
Binsheng Zhao
Mike McNitt-Gray, Chair

MR subcommittee members:

Daniel Barboriak
Luc Bidaut
Larry Clarke
Barbara Croft
Chuck Meyer
Ed Jackson, Chair

PET-CT subcommittee members:

Sam Armato
Edwin JR van Beek
Luc Bidaut
Larry Clarke
Barbara Croft
Geoffrey McLennan
Paul Kinahan, Chair

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Introduction

Early detection of tumor response to therapy is a key goal. Finding a measurement tool capable of early detection of tumor response could individualize therapy treatment as well as reduce the cost of bringing new drugs to market. On an individual basis the urgency arises from the desire to prevent continued treatment of the patient with a high cost, high risk regimen with no demonstrated individual benefit, as well as the need to rapidly switch the patient to another therapy that may increase treatment efficacy *for that patient*. Regarding the process of bringing new drugs to market such tools could demonstrate efficacy in much smaller populations allowing phase III trials to be have smaller populations and thus arrive at statistically significant decisions in shorter durations. Such studies would be much less costly in time and dollars spent to bring the drug to market.

The emphasis placed on the word "*Early*" implies that we are interested in the measurement regime of zero change, i.e. the detection of truly small changes from whatever tool and parameter set we measure. From detection theory we understand that our ability to do so rests on the ratio of signal to noise, alternatively described as effect size to variance. As this ratio increases we migrate from the condition being able to detect changes in large populations by averaging, to the condition of using fewer subjects until we are able to detect such changes in an individual with clinically useful statistical accuracy.

Given the large task required to implement these measurements across a broad spectrum of tools and measurement parameters in search of optimal tools, this RIDER group has focused on ways of estimating a measurement tool's noise, i.e. variance, under the condition of no change across several modalities and measurement techniques. Arguably the most realistic and useful datasets representing zero change come from patients harboring tumors who are imaged, removed from scanner, allowed to drink, snack, etc, and then are rescanned. We affectionately refer to these interval exams as "coffee-break" exams. These datasets then represent all of the realities of short interval imaging with whatever modality was used, i.e. tissue contrast to noise, patient motion artifacts, repositioning errors, etc., that will be encountered in the real world. Additionally because the time interval between scans is on the order of hours or less, we can safely assume that there are no macroscopic changes to the tumor in the interim. Note that datasets with expert annotations were not used due their demonstrated variability in segmentation and thus lack of certainty in the change assumptions associated with using these datasets[1]. An alternative to collecting these coffee-break experiments which describe all sources of noise in the null hypothesis against which treatment effects can be compared, is the collection of a large database of treatment trials along with clinical endpoints which can be modeled to determine the sources of multiparametric covariance; we suggest that the collection of coffee-break data may be far more efficacious at much lower collection cost.

Most importantly, in addition to the following body of text several groups have provided de-identified coffee-break datasets to NCIA (http://ncia.nci.nih.gov), NCI's Archive web site, for public downloading of this invaluable data, along with descriptions of their measurements, post-processing analyses and results, so that their methodologies and results can be compared with the results of others to follow. While some of the RIDER participants were partially funded for these efforts by NCI sponsored contracts, others volunteered their time and data to participate. We are indebted to all for their expertise, collegial participation and many hours of work.

crm

[1] Meyer, CR, TD Johnson, G McLennan, DR Aberle, EA Kazerooni, H MacMahon, BF Mullan, DF Yankelevitz, EJR van Beek, SG Armato III, MF McNitt-Gray, AP Reeves, D Gur, PH Bland, CI Henschke, EA Hoffman, G Laderach, R Pais, A Starkey, D Qing, C Piker, J Guo, D Max, BY Croft, LP Clarke (2006) *Evaluation of lung MDCT nodule annotations across radiologists and methods*, Acad Radiol **13**(10):1254-1265.

RIDER Project

CT Volumetrics Subcommittee

Summary Report v2

1. Research Questions to be asked regarding CT Volumetrics as a Biomarker for response and how RIDER can contribute to this.

There are several open research questions that will help us establish whether CT Volumetrics can be useful in determining a patient's response to therapy (i.e. can serve as a Biomarker of response). These include:

a. Patient Repeatability – How much variation can be expected when the same acquisition, image analysis and image-derived measurement calculation are repeated on the same patient over a very short time period?
b. Accuracy, Bias and Reproducibility using a test object (Phantom) – How close can one match a known value (size, function, etc) using a specific image acquisition, image analysis and image-derived measurement calculation? What are the effects of changing parameters on the ability to obtain an accurate result? If the same phantom is scanned on the same device, how much variance can we observe? Now move the phantom to another device or another institution and measure variance, etc.
c. Patient Change – When a patient is imaged over time, is the change observed predictive of clinical outcome?

These questions can also be asked in terms of a comparison with the clinically used RECIST criteria.

The RIDER CT Volumetrics subcommittee has focused on the above questions specifically in the context of the use of tumor volumes obtained from CT scans (primarily focusing on thoracic CT and the measurement of lung tumor volumes). This subcommittee has identified both an extended list of potential sources of bias and variance that could be encountered in the measurement of tumor volumes using CT as well as sources of data that could be used to help answer some of these questions. These are described below.

2. Potential Sources of Bias and Variance

There are many potential sources of bias and variance when measuring tumor volumes using CT, especially for lesions in the lung. These have been broken into three major groups, with the understanding that there may be significant interactions between sources of error (e.g. covariance)) linked to various effects. These include:

a. Patient Factors: (controllable factors in a patient or possibly controllable in a phantom)

i. Breathhold (ability to hold breath during a single scan)
    1. breathing motion
ii. Patient motion

   1. voluntary (e.g. just not holding still during scan)
   2. involuntary (e.g. cardiac motion)
   3. cyclic v. non-cyclic motion (e.g. related to the possibility of gating)

iii. Inspiratory level (such as the consistency of breathold for all serial scans)

iv. Patient orientation/positioning
   1. Supine or prone (gravity dependent effects)
   2. Feet first or head first

v. Patient Size (e.g. body habitus)

vi. Implants (e.g. metallic implants, pacemakers, etc.)

vii. Abnormalities
   1. Resected lungs
   2. Scarring
   3. Competing/Concomitant disease (e.g. inflammation)
   4. High-density tissues (see also vi)
   5. Girth of patients (e.g. obesity)

viii. Other Desirable Information
   1. Patient medical history – such as treatment history

b. Lesion Factors (possibly controllable in a phantom)

i. nodule size
ii. nodule shape (e.g. speculation, lobulation)
iii. nodule density
iv. nodule texture
v. nodule margin
vi. nodule internal composition and homogeneity (see also vii)
vii. nodule margin
viii. nodule radiological solidity – solid or ground glass or mixed
ix. location – juxtapleural vs. mediastinal vs. juxtavascular vs. near bone (e.g. rib or vertebra) or soft (e.g. muscle, heart) tissue
x. size and density of attaching vessel

c. Imaging Modality – specific factors that vary within CT

i. Differences between imaging hardware MDCT
   1. different number of detector rows
   2. different number and size of detector elements in each row
   3. different number of sources (e.g. dual source CT)
   4. different detector material
   5. different x-ray spectrum/beam filtration (e.g softer vs. harder x-ray beam)
   6. different bowtie (compensating) filter
   7. different acquisition, correction or reconstruction software principles and versions (e.g. some features enabled)

ii. Differences in imaging protocol
   1. different collimation setting
   2. different mAs
   3. different pitch

4. reconstructed slice thickness and spacing/overlap
5. varying slice thickness/spacing in different regions (e.g 5 mm thickness through upper and lower lungs, 3 mm through hilar region)
6. reconstruction kernel
7. correction and reconstruction algorithms (e.g. different interpolation scheme)
8. different kVp
9. Reconstructed Field of View
10. contrast vs. non-contrast
    a. injection rate
    b. volume of contrast used
        i. including whether a weight based scheme is used to determine total contrast amount
        ii. concentration of contrast material used (there are different concentrations sold commercially)
        iii. type/physiology of contrast material
    c. delay (e.g. delay time used to determine when scan should start following IV injection)

11. Overlap or non-overlapped acquisitions (full chest, then abdomen or through the chest down to the dome of the diaphragm, then delay and then from diaphragm down to abdomen)
12. With larger detector row arrays providing extended z-axis coverage, sequential scans may be performed, and lesions may would span - or be at the interface of - several acquisitions FOVs

iii. Scanner Calibration
1. Phantom QA
2. Calibration to water
3. Calibration to air
4. HU consistency serially and across FOV
5. Contrast Scale

iv. Scanner maintenance schedule/parts replacement (e.g. parts like x-ray tubes, and detectors may degrade or at least change performance over time, which may affect image quality)

d. Image Analysis – image segmentation, registration and calculation

i. Segmentation vs. non-segmentation

ii. Non-segmentation method (such as registration method for change analysis or template-method)

iii. Manual vs. semi-automated vs. fully automated analysis

iv. Region growing factors
1. Thresholding
2. Algorithm (e.g. criterion)

v. Calculating Volume
1. Indirect estimates/surrogates of volume
    a. 1D measurement (RECIST)

b. 2D measurement (Product of Diameters, WHO)

2. Summing voxels
3. Other estimates of volumes (e.g. taking into account partial volume effects or tumor models) that may be patient related

vi. Software version (again, depending on which features are implemented in software being used and how they are being implemented).

e. Interactions between Source of Bias and Variance

The CT Volumetrics subcommittee recognizes that there will be some instances of strong interaction between these sources of error. This means that analyses will likely have to take into account covariances and not just consider univariate analyses. A few examples include;

i. A segmentation method that uses a threshold value to determine lesion boundary extent may adapt this value when different slice thicknesses are used due to different degrees of partial volume effects at the tumor’s edge.

ii. Based on the sampling theory, there likely is a strong interaction between the size of a lesion and the slice thickness used to reconstruct images when volume is to be determined

f. Mitigation measures

These were discussed briefly by the CT Volumetrics committee and a few mitigation measures were prospectively identified. These include:

i. When possible, keep all potential sources of variance the same, including:

1. The same device (same scanner)
2. Using the same technical parameter settings as previous exams for any given patient (though some thought to adjusting parameter settings due to change in patient size, such as weight gain or weight loss, may be appropriate).
3. Use the same patient factors (positioning, breathing instructions, etc.)
4. Perform analysis with the same software
5. Perform analysis with the same software settings (threshold, etc.) as previous exams.
6. Perform analysis with the same software version if updates were to add another source of variance (i.e. change absolute measurements) without improving the change analysis (e.g. when seeking relative rather than absolute changes)

3. Data to be made available and how it contributes to each question

There will be several important sets of image data that will be made publicly available, either through the RIDER project directly or through other data collection efforts that are being performed outside of RIDER, but that share the same goals in answering questions related to CT Volumetrics as a Biomarker for response. These are described below:

a. As part of the RIDER Project

i. **The Patient Repeat CTs ("Coffee Break Experiment") performed at Memorial Sloan Kettering**.
   1. In this study, 32 lung cancer (NSCLC) patients were imaged twice on the same scanner in an interval less than 15 minutes apart.
   2. Imaging was done using identical technical parameter settings on the same machine (GE LightSpeed 16 for 28 subjects, GE VCT(64 slice scanner) for 4 subjects). LightSpeed 16 settings (VCT settings in parentheses where different) were:
      a. 120 kVp
      b. detector configuration of 16 x 1.25 mm (64x0.625mm)
      c. pitch of 1.375:1 (0.984)
      d. rotation time of 0.5 second.
      e. The standard-dose thoracic images were obtained without intravenous contrast during a breath hold.
      f. Thin-section images of 1.25 mm slice thickness were reconstructed using the Lung convolution kernel.
   3. Expected image-derived metric would be tumor volume, but this could be expanded to other metrics.
   4. Data will be made available to the public through NCIA by late September.

ii. **Unmarked Repeat CT lung studies at different time intervals from MD Anderson**
   1. In this study, many cases of patients (not sure of the current or ultimate number) with known nodules or masses in the lungs (both primary and metastatic lesions were included) were submitted to NCIA.
   2. Each case had at least 2 image data sets from different time points; many had 3 or more time points (information on time interval between scans will be available) .
   3. No truth about tumor volume was provided – no reader segmentation or volume estimate.
   4. For several cases, two readers at NCIA (C. Jaffe and R. E) made RECIST measurements, but these are subjective ratings and would not be considered as an "externally determined truth".
   5. Typically, scans were done with the same acquisition parameters and/or on similar scanners, but there is no implicit guarantee thereof as this was not one of the inclusion criteria (though technical parameter information will be made available through DICOM headers).

b. **Outside RIDER**

i. **CT lung nodule phantom from FDA (Nick Petrick, Marios Gavrielides, Lisa Kinnard).**
   1. In this study, several synthetic nodules of known size, composition and shape were placed inside an anthropomorphic phantom and attached to lung vasculature.

2. This phantom has been scanned multiple times under different combinations of dose (low and high mAs), slice thickness (thin and thick slices), pitch, and reconstruction filter (smooth and sharp kernels). This includes, effectively, a “phantom coffee break” experiment.
3. These datasets will be available to RIDER working group sites to have them use their own segmentation and volume calculation software to estimate volume of each lesion.
4. The phantom has also been scanned at Mallinckrodt Institute of Radiology (Washington University, St. Louis) and similar imaging has been performed on a Siemens scanner. This phantom is also being scanned on a GE system at the National Institutes of Health. Thus, the variability between scans obtained on different platforms can be investigated.
5. Again, main focus here was on estimating the volume of simulated nodules and analyzing the effects of different technical parameters, especially as a function of nodule size and composition.
6. All of these scans will eventually be archived on the NCIA website for general public use.

ii. **Cornell Coffee Break**

1. In this study, several patients were scanned multiple times during the course of a clinically indicated needle biopsy study. These are multiple scans of the same lesion (the one about to be biopsied).
2. More information at http://www.via.cornell.edu/crpf.html

iii. **Biochange 2008 experiment**

1. In this study, colleagues at NIST performed a nodule volume change pilot study in which they collected several patient studies and some studies from the FDA’s phantom with simulated nodules and had multiple investigative teams perform measurements.
2. More information is available at: http://www.itl.nist.gov/iad/894.05/biochange2008/Biochange2008-webpage.htm

4. **Planned Analyses of Bias and Variance**

a. **Planned Analyses that will be performed as part of RIDER include:**

i. From patient repeatability under no change condition **(“Coffee Break”) experiment, the MSK group** will perform an analysis of variance of the tumor volume measurements made for each patient between the two scans performed. See Appendix 1 for MSKCC resulted submitted for RIDER.

b. **Planned Analyses that will be performed outside of RIDER include:**

i. **From FDA scans of anthropomorphic phantom:**

1. an analysis of the effects of different acquisition parameters (such as slice thickness, mAs level, beam collimation, reconstruction algorithm, pitch and scanner manufacturer) and possibly image analysis parameters on tumor volumes will be performed.
2. Their analysis will also include investigating the effects of lesion related parameters (such as the differences between spherical and

non-spherical nodules as well isolated lesions vs. lesions contacting a simulated vessel) on tumor volumes.

   ii. Analyses similar to the MSK Coffee Break experiment are expected to be performed on the **Cornell Coffee Break** experiment image data as well.
   iii. The **NIST Biochange 2008** experiment will be reporting the results of its Pilot Study at SPIE 2009 in which it will investigate changes in tumor volumes from repeat studies of patients and phantoms using different observers/methods.

5. Roadmap for future data collection and analysis efforts
   a. Other analyses/experiments to perform on RIDER datasets
      i. While tumor volumetrics are the primary subject of these investigations, there are several other metrics that the RIDER CT Volumetrics subcommittee would suggest warrant future investigation. These include:
         1. Other changes in tumor characteristics besides just volume, such as changes in tumor density and composition. Some analysis may be possible with RIDER datasets acquired for volumetric purposes (such as Coffee Break experiment data).
   b. It is recognized that with RIDER and related datasets, a unique opportunity exists both to perform analyses and to provide guidance as to how future analyses should be performed. This may include describing the perceived strengths and weaknesses of various approaches (say, univariate analysis of variance to multivariance analysis of variance, etc.) and describing in detail each analysis method how results may vary depending on the kind of data under scrutiny, and also depending on the question that needs to be answered.
   c. The RIDER group may also consider describing barriers to **either scenario**:
      i. short-term issues (such as barriers to sharing data due to IRB issues) and
      ii. longer terms issues such as
         1. workflow issues about how tumor volumetric issues would be performed in a clinical setting
         2. local cultural or logistical issues such as who would actually perform those measurements
         3. infrastructure issues such as how those results would be obtained, communicated, stored and reported both for regulatory purposes and for clinical patient management issues
         4. economic issues such as:would trial sponsors be willing to bear the cost of the increased effort required to perform tumor volumetric measurements (and reporting, etc.)? would they be willing to do so only if the efficacy of these methods were demonstrated unequivocally?
   d. Other Data sets that would be nice to have and the analyses that would be great to do
      i. The FDA is considering expanding the range of simulated nodules to those that may be more peripherally located and of different size and composition. This would permit a stratified analysis by different kinds of tumors that are seen clinically (albeit under ideal phantom conditions). For example, if

ground glass tumors were to be physically simulated, investigations could be performed into the effects of technical parameters on estimating volume for ground glass tumors. Alternately, a similar setup could also help in assessing the effects of having different software packages to estimate the volume of ground glass tumors

ii. Of course, the obvious extension of RIDER would be to collect image data sets that have corresponding clinical outcomes, such as progression-free survival (PFS) or some other "truth" measure. Ideally, these image datasets would be from the same patient acquired over multiple time points and the timing with regard to therapy received would be known along with the clinical outcome. Other supplementary information could be the ground truth on imaged lesions through biopsies, or through complete surgical samples collected shortly after the last imaging session for this lesion.

Appendix 1: MSKCC Repeat CT Study – Summary on methods and results

Background:

The CT scan is the most widely used assessment of response to treatment for lung tumors. However, little is known about the reproducibility of the CT scan measurements obtained. The purpose of this study was to evaluate the variability of tumor unidimensional, bidimensional and volumetric measurements on same-day repeat CT scans in patients with non-small cell lung cancer (NSCLC).

Methods:

Thirty-two consecutive patients with measureable NSCLC, each underwent two CT scans of the chest within 15 minutes using the same imaging protocol, were evaluated. We applied our home-grown segmentation method to assisting in calculation of the two greatest diameters and the volume of each lung lesion on both scans of thin-section images. An experienced radiologist visually inspected all segmentation results and corrected a result if it was suboptimal. Concordance correlation coefficient (CCC) and Bland-Altman plots were used to assess the agreement between the measurements on the two repeat scans.

Results:

Lesions had a mean diameter of 3.8 +/- 2.0 cm. The computer method successfully segmented 19 (59%) lesions on both scans. The remaining 13 (41%) lesions required radiologist's manual corrections. The CCCs and relatively narrow 95% limits of agreement have demonstrated that CT scan measurements are highly reproducible (Table 1). Taking the volume as an example, differences measured on the two repeat scans falling outside the range of -12.1% and 13.4% could be considered a true change in tumor volume.

	Concordance correlation coefficient		**Mean % relative difference**	**95% Limits of agreement**
	ρ_c	**95% CI**		
Uni-dimensional	1.00	(1.00, 1.00)	-0.6%	-7.3 %, 6.2 %
Bi-dimensional	1.00	(0.99, 1.00)	1.1%	-17.6 %, 19.8 %
Volume	1.00	(1.00, 1.00)	0.7%	-12.1 %, 13.4 %

Table 1. Computer-generated measures of reproducibility on repeat CT scans

RIDER Project

MR SUBCOMMITTEE SUMMARY REPORT

1. **Types of MR imaging biomarkers that would be useful in providing change/response assessment.**
   Of the large number of morphological and functional MR measures that can be obtained and might be useful in assessment of response to therapy, the MR Subcommittee focused on the following aspects:
   - Phantom data obtained on four scanners from two vendors and at two field strengths to address 1) repeatability (short-term and long term) of T_1 measurements, 2) contrast-to-noise ratio and signal intensity stability, 3) limits of agreement of T_1 measurements using two acquisition techniques, and 4) between vendor contrast differences and effect on computed contrast agent concentration results (M.D. Anderson Cancer Center).
   - Repeat (assumed zero pathological change) dynamic contrast enhanced MRI (DCE-MRI) (Duke University).
   - Repeat (assumed zero pathological change) diffusion MRI, including diffusion-weighted imaging and diffusion tensor imaging (University of Michigan and Duke University).

   These three foci of effort provided deidentified source data, transferred to the NCIA, to be used by other researchers interested in change analysis of DCE-MRI and diffusion imaging biomarkers, as well as data analyses and derived parametric maps, also transferred to NCIA. The individual contract reports address the specific data analyses performed on each of the data sets outlined above. This summary document only briefly reviews the general goals of the subcommittee's efforts, data acquisition and submission, perceived limitations of the current data and analyses, difficulties encountered in the efforts to share such clinical research data, and suggestions for future studies.

2. **Specific goals of the MR subcommittee efforts.**
   The initial studies described herein were undertaken to begin the process of addressing the following general goals identified by the full RIDER committee:
   - Accuracy, Bias and Reproducibility Using a Test Object (Phantom) – How closely can one match known value (contrast response, relaxation times, *etc*.) using a specific image acquisition technique and image analysis methodology? If the same phantom is scanned on the same device, how much variance is observed? If the phantom is scanned on a different device (same vendor as well as different vendor) how much variance is observed? How closely can acquisitions on one vendor's scanner be matched on a scanner from a different vendor?
   - Patient Repeatability – If repeated scans are obtained on a given patient, how much variation is observed when the same acquisition, image analysis, and image-derived measurement calculation are used?
   - Patient Change – When a patient is imaged over time, how much change can reliably be observed, and is such change observed predictive of clinical outcome? (This ultimate goal of the RIDER efforts is not addressable within the limited RIDER initiative lifetime.)

3. **Identification of sources of bias and variance**

   Although MR provides a potential wealth of functional and morphological information, there are many sources of variance when using MR techniques to measure imaging biomarker response. These have been broken into three major groups, with the understanding that there may be significant interactions between sources of variance (e.g. covariances) between effects. These include:

   a. **Patient Factors:**

      i. Patient motion
         1. voluntary (*e.g.*, not holding still during scanning, swallowing)
         2. involuntary (*e.g.*, respiratory and cardiac motion, coughing, swallowing, spasm)
         3. cyclic (e.g., "gatable") vs. non-cyclic

      ii. Physiological variation independent of disease
         1. variation in cardiac output
         2. modulation of blood volume and flow due to, for example, prandial status and caffeine consumption.

      iii. Lesion size and location
         1. lesion size vs. spatial resolution provided by scanner hardware and acquisition protocol (partial volume)
         2. lesion location proximal to sources of signal variation (*e.g.*, mediastinal lesions in DCE-MRI applications)

      iv. Patient positioning / repositioning / immobilization

      v. Implants (e.g., metallic implants, pacemakers, *etc*.)

      vi. Effects due to therapies other than the target therapy
         1. Prior therapies
         2. Concomitant therapies (*e.g.,* steroids, antivascular/antiangiogenic therapies, *etc*.)

   b. **Acquisition-Dependent Factors:**

      i. Differences due to imaging hardware design and performance characteristics
         1. differences in gradient subsystem performance and image acquisition rates
         2. differences in B_0 and B_1 homogeneity
         3. phased array coil sensitivity characteristics

      ii. Differences in imaging protocols (extensive list!)
         1. differences in pulse sequence implementation, across vendors and within vendors, which can strongly affect contrast response
         2. differences in acquisition parameter selections available for a given pulse sequence across vendors and within vendors
         3. corrections for image intensity non-uniformity
         4. parallel imaging implementation and acceleration factors
         5. the "upgrade dilemma" – software upgrades that can, in unanticipated and unadvertised ways, modify pulse sequences and affect contrast response, temporal resolution, *etc*.

iii. Scanner characteristics and calibration
   1. magnetic field homogeneity
   2. RF subsystem calibration and stability
   3. gradient subsystem calibration and stability
   4. spatial accuracy (over specific volume or volumes) - affected by gradient nonlinearity corrections (in-plane and, for some scanners and pulse sequences, through-plane)
   5. contrast response
   6. signal stability (signal intensity, SNR, CNR, and ghosting)
   7. B_1 homogeneity and receiver coil sensitivity characteristics

c. Image analysis factors

i. Strongly application-dependent.
   1. Correction for motion, if required, including deformable registration.
   2. Correction for image intensity non-uniformity (B_1 transmit / RF coil sensitivity characteristics)
   3. DCE-MRI requirements:
      a. vascular input function assessment
      b. native tissue T_1 measurements required if contrast agent concentration assessment required
      c. temporal sampling characteristics

ii. General
   1. repeatability of analysis technique
      a. interobserver
      b. intraobserver
      c. (semi-)automatic
   2. sensitivity to CNR and other input data characteristics

Interactions between sources of variance

As noted by the other RIDER subcommittees, there are known (and, undoubtedly, as yet unknown) interactions between some of these sources of variance and these will need to be addressed in an application specific manner.

Basic steps for mitigation of patient-dependent variance

While such techniques were not discussed specifically by the MR Subcommittee, common concepts for mitigation include:

- Use the same device (same scanner) for each patient, when possible, and have rigorous QC processes in place to assure consistent performance for a given scanner and, if applicable, across scanners.
- Use stored, and “user locked”, if possible, protocols on scanners to make sure the same technical parameter settings are used.
- Use consistent patient positioning and preparation procedures (IV placement, breathing instructions and/or feedback, *etc*.).
- Use consistent contrast agent doses and rates of delivery (and consistent use of saline flush) when techniques call for intravascular contrast agents.

- Carefully monitor the scanners for any software and/or hardware upgrades. If either occurs, completely reassess baseline scanner performance characteristics.
- Perform all data analyses with the same equipment and software release versions, if possible. If upgrades or changes occur, verify consistency of results from one or more test cases with those obtained using prior versions of the hardware and/or software.

4. **Data made available by MR Subcommittee members and how they contribute to each question above**

   Several sets of image data will be made publicly available the RIDER project and can be used to partially address one or more of the three major categories of variance identified above. These are described below:

   a. Phantom data to assess system bias and variance (E. Jackson, M.D. Anderson Cancer Center).

      i. In this study, data were obtained using an 18-compartment EuroSpin TO5 contrast response phantom (Diagnostic SONAR, Ltd, Livingston, West Lothian, Scotland). Data were obtained in "coffee break" fashion on the same day as well as one week later. In the coffee break measurement setting, the RF coil and phantom were positioned and data obtained, the phantom was then removed and positioned and scanned on a second nearby scanner. The phantom was subsequently returned to the first scanner, the RF coil and phantom repositioned, and data obtained using a new exam ID but using the same stored data acquisition protocol. Data were obtained at all three time points from three 1.5T scanners (two, with differing gradient subsystems, from a single vendor and one from a second vendor) and from one 3.0T scanner. Data acquired included T_1 measurements using a 2D inversion recovery (IR) sequence (once at 1.5T and twice at 3.0T) and a 3D multiple flip angle fast spoiled gradient recalled echo (MF-FSPGR) sequence (each time on each scanner). An approximately 7 min DCE-MRI data set was also acquired each time on each scanner using a 3D FSPGR sequences with sections matched those acquired for the MF-FPSGR T_1 measurements..

         For the fast spin echo IR T1 measurements, 10 inversion times were used (50, 100, 250, 500, 750, 1000, 1500, 2000, 2500, and 3000 ms). For the MF-FSPGR T_1 measurements, 7 flip angles were used (2, 5, 10, 15, 20, 25, 30 degrees). For the DCE-MRI acquisitions, the total scan time was approximately 7 min with a temporal resolution of approximately 9 sec (dependent on specific scanner). Echo and repetition times were matched as closely as possible across scanners, with TE ranging from 0.90 to 1.35 ms and TR ranging from 4.09 to 5.10 ms. The flip angle was 30 degrees for all DCE-MRI scans.

      ii. All source data will be transferred to NCIA by 9/25/2008.

   b. Repeated DCE-MRI and Diffusion Data in Brain Tumor Patients (D. Barboriak, Duke University Medical Center)

      i. In this study, repeat DCE-MRI datasets in 19 patients with recurrent glioblastoma multiforme and repeat diffusion tensor imaging (DTI) datasets in 17 of the 19 patients were performed. The interval between scans was 2 days or less. The technical parameters are summarized in the Appendix. All images were obtained on the same 1.5T imager (Siemens Avanto). In brief, for DCE-MRI, T_1 mapping using a multi-flip angle approach and 6 flip angles was performed. Dynamic imaging was performed every 4.8 seconds during the intravenous infusion of 0.1mmol/kg of Gd-DTPA at 3ccs/second. DTI imaging was performed using a 12 direction sequence, TR=6000ms, TE=100 ms, 90 degree flip angle, 4 signal averages, 128 x 128 acquisition matrix, 1.72 x 1.72 x 5 mm voxel size, and a *b*-value of 1000 sec/mm^2.

ii. All source data have been transferred to NCIA. In addition, parameter maps for fractional plasma volume, volume of the extracellular extravascular space, and K^{trans} (for DCE-MRI data) and apparent diffusion coefficient and fractional anisotropy (for DTI data) have been transferred to NCIA. Segmentations of tumor-related enhancement and FLAIR signal abnormality volumes have also been transferred.

c. Repeated ADC Diffusion of Breast Cancer (C. Meyer, University of Michigan)

i. In this study, during the first cycle of neoadjuvant chemotherapy the following imaging protocol was implemented:

- Two pre-initiation (pre-therapy) baseline MRI scans were typically obtained within 15 minutes, where initiation of neoadjuvant chemotherapy was no later than one week from the last baseline MRI, and
- One post-initiation MRI was obtained within 8-11 days after initiation of neoadjuvant chemotherapy

ii. The second cycle of neoadjuvant chemotherapy was implemented with no further imaging studies.

iii. Axial DWI parameters on our Philips 3T scanner: FOV=350mm; acquisition matrix = 196x86; SENSE factor 2; 30 slices; 4mm thick; 2D SE-EPI 3-axis DWI at *b*=0 and 800s/mm^2; STIR (fat suppression), TR/TE/TI = 12000 / 62 / 150 ms; 2 signal average; acquisition time = 4:41 min.

iv. All source data will be transferred to NCIA by 9/25/2008.

5. Analyses of bias and variance

a. Phantom data collection (M.D. Anderson Cancer Center):

Limits of agreement and coefficients of repeatability, as proposed by Bland and Altman (Lancet, 1986), were computed for the T_1 measurements obtained from each scanner using the IR-based and multiple flip angle-based data (limits of agreement) or multiple flip angle measurements at each time point (coefficients of repeatability). T_1 measurement correlation analysis was also performed using the data obtained at each time point on each scanner. Contrast response was assessed at each time point on each scanner using the multiple compartment phantom data, and stability of the CNR and signal intensity was assessed from each of the DCE scans obtained at each time point on each scanner. Simulated DCE uptake curves were also generated for each scanner using measured data from the multiple compartment phantom and commonly assumed signal intensity response for an ideal fast spoiled gradient echo sequence.

b. Repeated DCE-MRI and Diffusion Data in Brain Tumor Patients (Duke University Medical Center):

A similar analysis of repeatability was performed on both the DCE-MRI and DTI datasets from brain tumor patients. For DCE-MRI parameters, mean values were obtained at both time points in areas of tumor-related enhanced segmented from 3D volumetric isotropic T_1-weighted contrast-enhanced FLASH images. The coefficient of repeatability (Bland and Altman) and the 95% confidence interval for percent change in parameter (Roberts, Roberts C, Issa B, Stone A, Jackson A, Waterton JC, Parker GJ *Comparative study into the robustness of compartmental modeling and model-free analysis in DCE-MRI studies*. J Magn Reson Imaging. 2006;23(4):554-63). For DTI parameters, mean values were obtained at both time points in both the areas of tumor-related enhanced, as well as in areas FLAIR signal abnormality segmented from 3D volumetric isotropic contrast-enhanced FLAIR images.

c. Repeated ADC Diffusion of Breast Cancer (University of Michigan)

Registration of the tumor in the interval exams was implemented using MIAMI Fuse©. Tumor volumes of interest (VOI) were drawn on the anatomical image volume and were warped from the anatomical volume onto one of the pre-therapy diffusion volumes denoted as the reference; warping is necessary due to the susceptibility artifacts in the diffusion acquisitions not present in the anatomical volumes. Subsequent registrations, either between the two pre-therapy exams or the two pre- and post-therapy scans, are also warped to account for repositioning deformations to the breast as well as any small compartmental changes to the tumor. Warping is accomplished using thin plate splines where the degrees of freedom (DOF) of the warp are related to the volume of the tumor. The user only needs to pick the location of three control points in the homologous tumor volume that approximate their loci in the reference tumor volume. The multiscale registration first implements rigid body registration, then low DOF warping, and finally full DOF warping.

Apparent diffusion coefficient (ADC) volumes are computed from the interleaved b=0 s/mm^2 and b=800 s/mm^2 acquisitions. For each pair of registered ADC images, a 128x128 joint density histogram (JDH) is constructed by incrementing the count of the 2D histogram defined by the two ADC values of the registered tumors**.** For the JDH of the two pre-therapy exams, bias is removed from this realization and variance is generalized, *i.e.*, increased, by adding the transpose of its JDH to itself. Then linear regression is performed after rotating the JDH onto its principal component axes. The resulting linear estimate (green), the estimates of its 95% confidence limits (red) and the 95% confidence limits of the histogram (yellow) are computed and plotted on the modified joint density histogram. The means of both joint density histograms are computed and plotted. Note that pre-therapy/pre-therapy JDF represents estimates of sources of measurement covariance variance associated with the null hypothesis, *i.e.*, the presence of all sources of noise, but no change in the tumor.

6. Recommendations for future data collection and analysis efforts and guidance to users of MR-based imaging biomarkers

a. Once the MR data analyses are fully available on NCIA, the MR Subcommittee and parent RIDER Committee should carefully evaluate the complete data set and describe perceived strengths and weaknesses of various approaches utilized. Once this process is complete, the RIDER Committee should be in a unique position, one based on multiple sources of commonly shared data, to make recommendations to the imaging biomarker community at large regarding source of variance and bias as well as means of mitigating such sources.

b. The RIDER Committee should also, in communication of its findings and recommendations for future efforts, consider describing barriers to the implementation of quantitative imaging biomarkers, including:

    i. barriers to sharing of data, such as those due to IRB compliance and review and sponsor-specific concerns of the proprietary nature of clinical trial data

    ii. workflow-related issues, including 1) the volume of data to be stored and analyzed for single or multiple time points, 2) the time-commitment requirement for not just the data analysis, but also the associated QC required to assure optimal data collection and analysis, 3) the funding required for such data acquisition and analyses and sources of such funding (or lack thereof), and 4) Structured Reporting and further extensions of current DICOM standards to allow for the appropriate data storage and reconciliation, the communication of such quantitative imaging biomarker findings in a consistent manner, and to allow further data mining and meta analyses.

c. Patient outcome studies should be strongly encouraged in any RIDER Committee report. The data obtained during the very limited lifetime of the RIDER initiative is clearly only the initial phase of a process that begins with a cultural change in the very nature of the data acquired (which currently are obtained using techniques optimized for qualitative image quality, not quantitative image data integrity) and progresses through the identification and minimization of sources of bias and variance at the device, patient preparation and scanning, and data analysis levels to come to the stage where imaging biomarkers can be utilized as true surrogate markers of therapy response. The ultimate proof of the validity and utility of such non-invasive imaging biomarkers, of course, lies in the ability of such measures to predict patient outcome earlier than currently possible. A key future effort must focus on such outcome measure assessments as well as continued efforts to identify and quantify sources of bias and variance in order to be able to properly determine sample size requirements for single subject response and group response assessments to novel therapies.

RIDER MR Subcommittee members:
Daniel Barboriak, MD – Duke University Medical Center
Luc Bidaut – M.D. Anderson Cancer Center
Edward Jackson, Chair – M.D. Anderson Cancer Center
Charles Meyer – University of Michigan

Submitted to the RIDER Committee and Laurence Clarke on 9/24/2008.

Appendix 2

Guide to Duke submissions to NCIA

Imaging data on 19 patients with recurrent glioblastoma who underwent repeat imaging sets was submitted to NCIA. These images were obtained approximately 2 days apart (with the exception of patient 786, whose images were obtained one day apart).

DICOM images

DCE-MRI:

All 19 patients had repeat dynamic contrast-enhanced MRI (DCE-MRI) datasets on the same 1.5T imaging magnet. On the basis of T2-weighted images, technologists chose 16 image locations using 5mm thick contiguous slices for the imaging. For T1 mapping, multi-flip 3D FLASH images were obtained using flip angles of 5, 10, 15, 20, 25 and 30 degrees, TR of 4.43 ms, TE of 2.1 ms, 2 signal averages. Dynamic images were obtained during the intravenous injection of 0.1mmol/kg of Magnevist intravenous at 3ccs/second, started 24 seconds after the scan had begun. The dynamic images were acquired using a 3D FLASH technique, using a flip angle of 25 degrees, TR of 3.8 ms, TE of 1.8 ms using a 1 x1 x 5mm voxel size. The 16 slice imaging set was obtained every 4.8 sec.

DTI:

Seventeen of the 19 patients also obtained repeat diffusion tensor imaging (DTI) sets. Whole brain DTI were obtained using TR 6000ms, TE 100 ms, 90 degree flip angle, 4 signal averages, matrix 128 x 128, 1.72 x 1.72 x 5 mm voxel size, 12 tensor directions, iPAT 2, *b* value of 1000 sec/mm^2.

Contrast-enhanced 3D FLASH:

All 19 patients underwent whole brain 3D FLASH imaging in the sagittal plane after the administration of Magnevist. For this sequence, TR was 8.6 ms, TE 4.1 ms, 20 degree flip angle, 1 signal average, matrix 256 x 256; 1mm isotropic voxel size.

Contrast-enhanced 3D FLAIR:

All 17 patients who had repeat DTI sets also had 3D FLAIR sequences in the sagittal plane after the administration of Magnevist. For this sequence, the TR was 6000 ms, TE 353 ms, and TI 2200ms; 180 degree flip angle, 1 signal average, matrix 256 x 216; 1 mm isotropic voxel size.

Note: before transmission to NCIA, all image sets with 1mm isotropic voxel size were "defaced" using MIPAV software or manually.

Non-DICOM images

A set of images which are not in DICOM format were submitted to provide an example or instance of an analysis of the submitted imaging datasets. The steps used to obtain the imaging and guide to the sample images are given below. All image analysis steps, unless specified, were performed using in house software created using as plugins to the ImageJ platform.

DCE-MRI analysis:

The first step in DCE-MRI analysis was to obtain maps of T1 and equilibrium magnetization from the multi-flip 3D FLASH images. This was accomplished using a non-linear simplex fitting technique of the data to the standard signal intensity equation.

The second step was using motion correction algorithms to minimize the effects of patient motion during the dynamic acquisition. In brief, an algorithm based on measurements of correlation ratios were applied to groups of 5 time points that were averaged and edge filtered to determine to which image sets motion correction would be applied. Sets that were motion corrected were subjected to two algorithms. First, image registration with itk (www.itk.org) using MultiResMIReg (a multi-resolution mutual information algorithm) were performed. These registrations were kept if the correlation ratios measured as described above were improved by the technique. If there was no improvement, a Lucas-Kanade image stabilizing algorithm was performed. Again, the registrations were kept if the correlation ratios improved.

The third step was to use the T1 and equilibrium magnetization images to convert signal intensities to estimates of gadolinium concentrations. For this purpose, the standard equation as formulated in Li, et al.[1] was used. This resulted in contrast agent concentration-time curves generated for each pixel location.

The fourth step was to derive a vascular region-of-interest from the imaging data. Candidate vascular voxels were selected using an automated technique from the signal intensity data on the basis of timing and heights of peaks. An experienced neuroradiologist then selected subsets of these voxels on a single slice on the basis of resulting height and plateau of the concentration agent-time curves.

The fifth step was to derive parameter maps using the extended Tofts model from the data above. To derive K^{trans} (the volumetric transfer constant for contrast agent transit from the plasma space to the extracellular extravascular space), fPV (the fractional plasma volume, or percent of each voxel thought to represent plasma space) and the v_e (the percent of each voxel thought to represent extracellular extravascular space), the matrix-based technique described by Murase[2] was used.

> Of note, these maps were submitted to NCIA. They are under the non-DICOM datasets directory designated by patient number. The maps are then designated Ktrans-mmddyyyy.tif , FPV-mmddyyyy and Ve-mmddyyy, respectively, where mmddyyyy is the anonymized date. These images are 32-bit TIFF stacks of 16 images.

The sixth step was to determine the location of tumor that was repetitively imaged. The contrast-enhanced 3D FLASH images obtained at the first imaging session were filtered using anisotropic diffusion filtering, then candidate pixels were determined by thresholding all pixels whose signal intensity was more than two standard deviations above the mean obtain in a large VOI obtained in normal appearing areas in the corpus callosum. An experienced neuroradiologist then selected those pixels in this pixel set that appeared to be related to tumor contrast enhancement rather than non-tumor objects such as fat, vessels, dura or choroid plexus.

> Of note, these segmentations of tumor-related enhancement (TRE) were submitted to NCIA. They are under the non-DICOM datasets directory designated by patient number. The segmentations are then designated TRE_segmentation-mmddyyyy.tif , , where mmddyyyy is the anonymized date. These images are 8-bit TIFF stacks. These images were originally processed in the axial plane, but have been reformatted into the sagittal plane in order to correspond exactly with the sagittal source DICOM images.

The seventh step was to place the parameter maps into a single parameter space by registering a summation of the dynamic image stack to the isotropic whole brain 3D FLASH images obtained at the same time point. By registering the 3D FLASH obtained at the second imaging session to that obtained in the first, both sets of parameter maps could be registered to the isotropic whole brain 3D FLASH images obtained at the first imaging session. MultiResMIReg was used to perform these registrations.

The eighth step was to derive mean parameters in areas of tumor. The first and last two images in the parameter maps were discarded due to wrap-around artifact. The resultant parameter maps registered to the isotropic whole brain 3D FLASH images obtained at the first imaging session were interpolated to 1 mm isotropic images using a tri-linear technique. Mean parameter values were obtained for points that were segmented as tumor-related enhancement in the sixth step and were imaged in the middle 12 slices on both imaging occasions. Main repeatability indices derived were the repeatability coefficient and the 95% confidence interval for percent change[3].

DTI analysis:

The first step in the DTI analysis was to create pixel-by-pixel maps of apparent diffusion coefficient (ADC) and fractional anisotropy (FA) from the DTI image sets. These images were obtained using JDTI, a Java-based plugin for ImageJ based on the JAMA matrix package (this software is available for free download from http://dblab.duhs.duke.edu). Of note, the fractional anisotropy equation used in this software package is:

$$FA = \frac{\sqrt{(\lambda_1 - MD)^2 + (\lambda_2 - MD)^2 + (\lambda_3 - MD)^2}}{\sqrt{\lambda_1{}^2 + \lambda_2{}^2 + \lambda_3{}^2}}$$

Where MD is the mean diffusivity and λ_1, λ_2, and λ_3 are the three eigenvalues from the diffusion tensor. (ADC is set equal to the mean diffusivity). The values obtained by this plugin may need to be multiplied by $\sqrt{\frac{3}{2}}$ in order to match other formulations of FA in the literature.

> Of note, these maps were submitted to NCIA. They are under the non-DICOM datasets directory designated by patient number. The maps are then designated ADC-mmddyyyy.tif and FA-mmddyyy, respectively, where mmddyyyy is the anonymized date. These images are 32-bit TIFF stacks.

The second step was to create segmentations of TRE and FLAIR signal abnormality (FSA) from the corresponding isotropic 3D image sets. The TRE segmentations obtained in step 6 of the DCE-MRI analysis above were used for DTI analysis. To obtain FSA segmentations, the contrast-enhanced 3D FLAIR images obtained at the first imaging session were filtered using anisotropic diffusion filtering, then candidate pixels were determined by thresholding all pixels whose signal intensity was more than two standard deviations above the mean obtain in a VOI obtained in normal appearing areas in the caudate head contralateral to the bulk of tumor. An experienced neuroradiologist then selected those contiguous pixels in this pixel set that appeared to be involved with or surrounding the tumor-related contrast enhancement.

> Of note, these segmentations of FLAIR signal abnormality (FSA) were submitted to NCIA. They are under the non-DICOM datasets directory designated by patient number. The segmentations are then designated FSA_segmentation-mmddyyyy.tif , , where mmddyyyy is the anonymized date. These images are 8-bit TIFF stacks. These images were originally processed in the axial plane, but have been reformatted into the sagittal plane in order to correspond exactly with the sagittal source DICOM images.

The third step was to place the parameter maps into a single parameter space by registering a summation of the dynamic image stack to the B0 images to the isotropic whole brain 3D FLASH images obtained at the same time point. By registering the 3D FLASH obtained at the second imaging session to that obtained in the first, both sets of parameter maps could be registered to the isotropic whole brain 3D FLASH images obtained at the first imaging session. Similarly, registrations of the 3D FLAIR sequences to the isotropic whole brain 3D FLASH images obtained at the first time point allowed these images and their corresponding segmentations to be placed in the same anatomical frame of reference. MultiResMIReg was used to perform these registrations.

The fourth step was to derive mean parameters in areas of tumor. The parameter maps registered to the isotropic whole brain 3D FLASH images obtained at the first imaging session were interpolated to 1 mm isotropic images using a tri-linear technique. Mean parameter values were obtained for points that were segmented as TRE and FSA on the first study. Main repeatability indices derived were the repeatability coefficient and the 95% confidence interval for percent change[3].

References

1. Li KL, Zhu XP, Waterton J, Jackson A. Improved 3D quantitative mapping of blood volume and endothelial permeability in brain tumors. J Magn Reson Imaging 2000;12(2):347-357.
2. Murase K. Efficient method for calculating kinetic parameters using T1-weighted dynamic contrast-enhanced magnetic resonance imaging. Magn Reson Med 2004;51(4):858-862.
3. Galbraith SM, Lodge MA, Taylor NJ, et al. Reproducibility of dynamic contrast-enhanced MRI in human muscle and tumours: comparison of quantitative and semi-quantitative analysis. NMR Biomed 2002;15(2):132-142.

Summary of RIDER PET/CT Subcommittee Results and Recommendations

September 21, 2008

PET/CT RIDER subcommittee members:
Luc Bidaut (MD Anderson Cancer Center)
Larry Clarke (NCI)
Barbara Croft (NCI)
Robert Doot (U. Washington)
Paul Kinahan (chair) (U. Washington)
Geoffrey McLennan (U. Iowa)
Charles Meyer (U. Michigan)
Edwin van Beek (U. Iowa)
Brian Zimmerman (NIST)

The following reports the results and recommendations of the RIDER PET/CT subcommittee. The PET/CT subgroup was responsible for: (1) archiving de-identified DICOM serial PET/CT phantom and lung cancer patient data in the National Cancer Imaging Archive (NCIA) to provide a resource for the testing and development of algorithms and imaging tools used for assessing response to therapy, (2) conducting multiple serial imaging studies of a long half-life phantom to assess systemic variance in serial PET/CT scans that is unrelated to response, and (3) identifying and recommending methods for quantifying sources of variance in PET/CT imaging with the goal of defining the change in PET measurements that may be unrelated to response to therapy, thus defining the absolute minimum effect size that should be used in the design of clinical trials using PET measurements as end points.

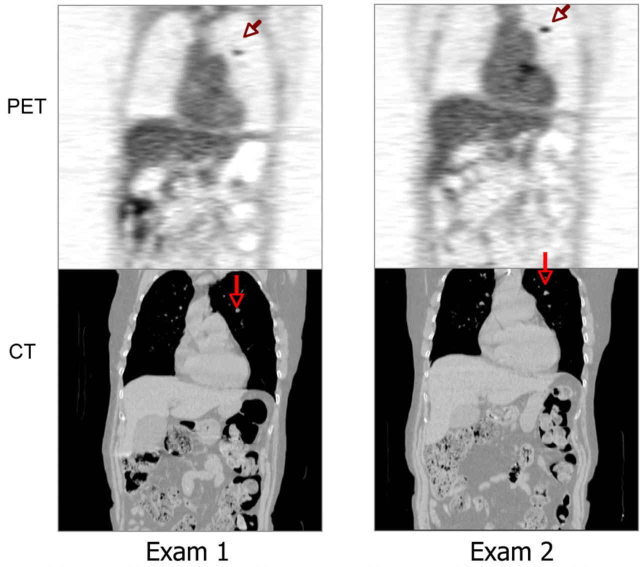


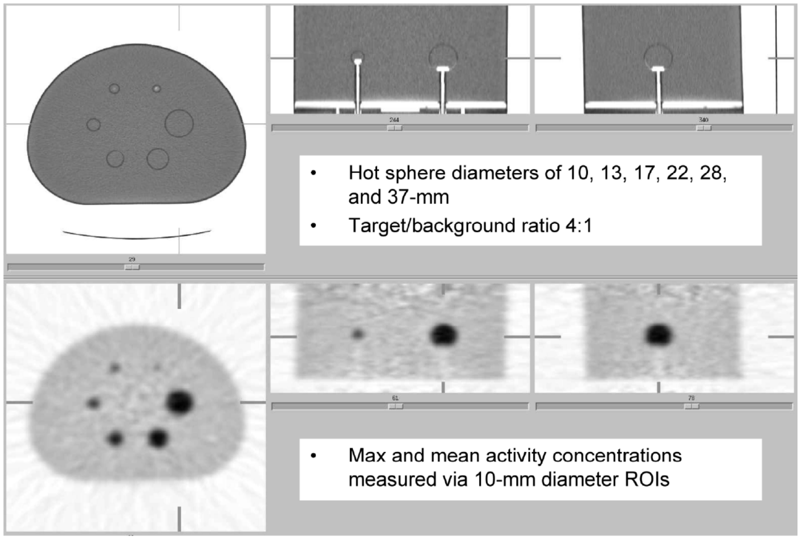
Figure 1: Serial PET/CT images of an lung cancer patient before and after therapy. Black and white arrow heads respectively indicate the tumor in the PET and CT images.

1. Archived PET/CT Images

De-identified DICOM images of serial scans of 28 lung cancer patients (with a total of 65 scans) of patients were uploaded to the National Cancer Imaging Archive (NCIA) using the RSNA's Medical Imaging Resource Center (MIRC) open source clinical trials software. DICOM images from repeat studies of a long half-life 68Ge calibration phantom were also uploaded. These image sets can now be downloaded by users for the testing and development of algorithms and imaging tools for assessing response to therapy. Sample PET/CT images of a lung cancer patient and the phantom included in the NCI archive are in Figures 1 and 2 and derived results are discussed in the next section.

2. Serial imaging studies of patients and a phantom filled with long half-life 68Ge-epoxy compound

De-identified DICOM files from serial PET and PET/CT imaging of 10 lung cancer patients were retrieved from the National Cancer Imaging Archive and analyzed to verify the ability to process the archived files. Sample results from an analysis of the retrieved files using GE PET VCAR software are shown in Figure 3.



- Hot sphere diameters of 10, 13, 17, 22, 28, and 37-mm
- Target/background ratio 4:1

- Max and mean activity concentrations measured via 10-mm diameter ROIs

Figure 2: NEMA NU-2 IQ phantom filled with solid, 9-month half-life 68Ge-epoxy compound. Air gaps (white regions) in the main chamber and stems, and to some extent in the spheres themselves, are evident on the CT images from PET/CT scans at the University of Washington.

Reproducibility studies of the phantom shown in Figure 2 assessed the instrumentation variability of measurements in PET/CT scans by repeatedly measuring the activity concentration of phantom hot spheres with fixed spatial locations and a true target/background ratio of 4:1. Phantom spheres and background volumes were filled with long half-life 68Ge-germanium epoxy to allow repeated measurements.

Several sets of 20 scans were acquired serially and reconstructed using various PET/CT scanners (e.g., GE's Discovery STE - DSTE, Siemens' Biograph Hi-REZ and Philips' Gemini TF) to determine the variability in stationary and 'coffee-break' repeat scans of the same 9-month half-life calibration phantom. Standard deviations for maximum and mean values of absolute recovery coefficients (equal to measured activity divided by true activity) ranged from 0.9% to 4.3% for repeat GE DSTE scans of a stationary phantom, depending on acquisition and reconstruction methods, after averaging the standard deviations for all 6 sphere diameters as shown in Table 1. Sample results in Figure 4 show similar findings for ‘coffee-break’ imaging studies using one PET/CT from each of the 3 major vendors. The GE, Philips, and Siemens PET/CTs were respectively located at the University of Washington, at the University of Pennsylvania, and at the Huntsman Cancer Institute at the University of Utah.

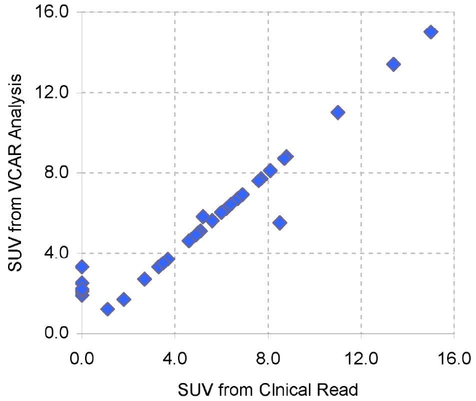


Figure 3: Maximum SUV from analysis of retrieved DICOM files using commercial analysis software versus maximum SUV from clinical reports from serial PET scans of 10 lung cancer patients archived in the NCIA.

The standard deviations of the absolute recovery coefficients for the coffee-break sets in Figure 4 are shown in Table 2. The standard deviations from the first set of GE DSTE coffee-break scans were significantly lower than the determined standard deviations from the Siemens Biograph Hi-REZ and Philips Gemini TF PET/CTs with the exception of the standard deviation of maximum absolute recovery coefficient for the Philips Gemini TF PET/CT whose maximum value of 4.3% exhibited only a trend towards statistical significance (p = 0.06). After adjusting the maximum and mean absolute recovery coefficients to compensate for differences in total counts, the adjusted absolute recovery coefficients for the subsequent coffee-break experiments were no longer significantly different from the DSTE PET/CT set of standard deviations with the exception of the adjusted standard deviation of the mean absolute recovery coefficient for the Gemini TF PET/CT whose average value of 1.8% was significantly lower that the corresponding DSTE PET/CT value of 2.4 (p = 0.04) as found in Table 2.

Table 1: Standard deviations of recovery coefficients for 2D & 3D reconstructions. Average standard deviations (SD) of absolute maximum and mean recovery coefficients (RC) for 5 minute GE DSTE PET/CT scans acquired in 2D and 3D modes. The 3D-FBP method was 3D-ReProjection (3DRP) (n = 20).

Algorithm:	2D-FBP*	2D-OSEM*	3D-FBP*	3D-OSEM*
SD of Max RC for 7-mm Resolution	4.3%	3.8%	2.4%	2.4%
SD of Max RC for 10-mm Resolution	3.2%	2.7%	1.7%	1.6%
SD of Max RC for 13-mm Resolution	2.3%	2.0%	1.2%	1.1%
SD of Mean RC for 7-mm Resolution	2.8%	2.5%	1.4%	1.4%
SD of Mean RC for 10-mm Resolution	2.3%	1.9%	1.1%	1.1%
SD of Mean RC for 13-mm Resolution	1.8%	1.6%	0.9%	0.9%

*Standard deviation (SD) value is mean of SDs of recovery coefficients (RC) for all six spheres.

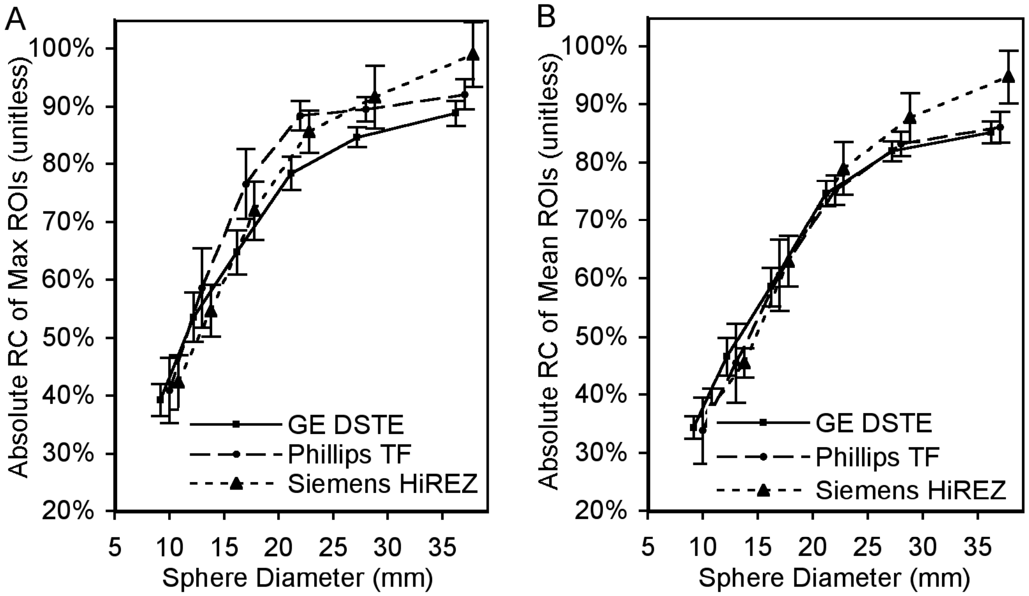


Figure 4: Recovery coefficients vs. size for PET/CT scans of repositioned phantom. Maximum (A) and mean (B) absolute recovery coefficients - with standard deviation error bars - versus sphere diameter for the same phantom repositioned and imaged 20 times using PET/CTs from the 3 major vendors and reconstructed via 3D iterative algorithms with 7-mm post-reconstruction Gaussian smoothing for the GE and Siemens scanners, and with no post-reconstruction smoothing for the Philips scanner with standard deviation error bars.

Table 2: Reproducibility of PET/CT 'coffee-break' scans of repositioned phantom. Reproducibility of 20 PET/CT scans of repositioned phantom with 3D iterative reconstruction and 7-mm Gaussian smoothing for GE DSTE and Siemens Biograph Hi-REZ, and a "sharp" Philips filter for the Gemini TF PET/CTs (n = 20).

	SD of Max RC* (unitless)	SD of Mean RC* (unitless)	PET Slice Thickness (mm)	Time since DSTE Scan (d)	Scan Time (min)
GE DSTE-1 (1/28/07)†	2.9%	2.4%	3.3	0	5
Siemens Biograph Hi-REZ (10/13/07)†	4.8%‡	3.9%‡	2.0	258	6
Philips Gemini TF (12/20/07)†	4.3%	3.2%‡	4.1	326	3
Adjusted Siemens Biograph Hi-REZ	2.9%§	2.4%§	3.3§	0§	5§
Adjusted Philips Gemini TF	2.4%§	1.8%‡§	3.3§	0§	5§

* Standard deviation (SD) value is mean of SDs of recovery coefficients (RC) for all six spheres.

† Date the same phantom filled with solid 68Ge-germanium epoxy was scanned.

‡ These SDs were significantly different from corresponding DSTE-1 PET/CT SDs ($p \leq 0.04$).

§ SDs adjusted via Equation 4 to compensate for differences in total counts.

3. Recommended methods to quantify variance in PET/CT imaging

Response of disease to treatment can be quantified using serial PET scans to measure changes in concentrations of tracer uptake throughout diseased tissue [1]. FDG-PET estimates of therapy response have been reported to predict outcome in non-small cell lung cancer [2, 3], esophageal squamous cell carcinoma [4], ovarian cancer [5], metastatic breast cancer [6], locally advanced adenocarcinoma [7], and neoadjuvant locally advanced breast cancer [8]. Changes are due to response to therapy, natural (i.e.

biologic) variability including lesion size, variability in patient preparation [1], and systemic variability measurement from differences in hardware, reconstruction, ROI analysis, and subject positioning or movement (i.e. respiration).

A growing number of clinical trials are considering using quantitative FDG-PET measurements as endpoints [1, 9] or to guide treatment [7], which increases the importance of understanding the bias and variability intrinsic to serial PET scans and their impact on multi-center clinical trials. The reproducibility of PET quantification must be assessed to enable determination of a threshold change required for classifying patient response to therapy and to aid estimation of variance for clinical trial design. While reproducibility of patient imaging [10-13] and relative merits of different reconstruction methods [12, 14-16] and ROI analysis techniques [12, 15, 17] have been published, the impact of PET/CT hardware and subject positioning on the reproducibility of tracer uptake quantification have not been as well studied.

3.1. What measures would be useful in measuring meaningful change/response?

Currently most clinically relevant: Relative change in SUV [standardized uptake value = (measured ROI value) / ((injected amount)/(volume of distribution))], i.e. % change in SUV - typically maximum SUV - in a ROI. In particular due to physical limitations of FDG-PET, lesions less than 3 cm in diameter are prone to significant partial volume effects (PVE). This leads to errors in the initial value for such lesions, which can then propagate to even larger errors in the % change calculation.

What about dynamic imaging (c.f. pharmacokinetics)? Analyses of dynamic data sets may provide better accuracy and more relevant information [8, 18, 19]. For metabolic changes measured with FDG-PET, the use of SUV may suffice, but, it will likely not be so for new tracers, or if the initial SUV is low (e.g. tumor to background ratio less than 5 [20]), and any subsequent response is producing only a small absolute change . In addition to such considerations, recent publications suggest that a change in blood flow may be a better metric than tracer metabolic rates to predict patient outcome [8].

There are many barriers to dynamic imaging in the clinic: arterial blood sampling, scanner time plus cost, technical issues for dynamic imaging at multiple axial FOVs (e.g. for lesions that require various bed positions).

Solutions for arterial sampling may include: (1) image-based measures of input function from heart or aorta in FOV etc.; (2) population based curves w/ venous sample, Solutions for scanner time plus cost may include: (1) plan into trial from beginning; (2) only use for phase 0 or first in human (FIH) studies to determine potential differences in measured effect sizes between SUV and PET measures determined from dynamic imaging.

3.2. What is the basic variability in these measures, and what are the sources of these variations?

Quantitation in PET is relatively well understood, but a major component thereof remains dose calibrators and long term stability or compensation. In other words, the scanners are stable [22] unless they are modified in some way such as hardware or software updates or recalibration or sensor drift.
There are three major categories of sources that might be modality (and task) specific:

a. Patient Factors – motion (e.g. respiration), physiologic function, etc.
b. Imaging Modality - specific factors that vary for PET and PET/CT equipment and design
c. Image Analysis - image segmentation, registration and calculation methods

Some of the related variable effects are summarized in recent presentations and conference records (e.g. [22,23]).

Physical sources of error (real, estimated and their correction)

a. Attenuation
b. Scattered and Random coincidences
c. Detector efficiency variations, scanner dead-time

Sources of variability

1. Patent specific
   a. Biological changes
   b. Lesion size (partial volume effect), location and environment
   c. Patient and/or lesion motion
2. Imaging protocol
   a. Patient preparation
   b. Uptake period and environment
   c. Scan protocol: amount of FDG (or other tracer) injected, scan duration, 2D or 3D mode,
3. Processing specific
   a. Accuracy of corrections for physical sources of error
   b. Reconstruction method
   c. Image smoothness vs. resolution tradeoff
   d. ROI definition method
   e. Standardized Uptake Value (SUV) calculation

One solution to the variability introduced by image reconstruction protocols is to prescribe extra reconstructions with specific and controlled parameters (e.g. FBP) whenever quantitation is required.

3.3. Are there any mitigating measures to reduce variation?

a. Attention should be paid to the positioning - i.e. centering in the FOV – of the lesion to minimize truncation effects
b. Dose calibrator standard – one should be available next month from NIST – that is cross calibrated with 18F's different branching ratios and additional X-rays from 68Ge.
c. What about dose calibrator reproducibility? NIST has not seen significant variability (i.e. < 2%) but there have been reports of up to 5% variability and results of internal studies will be presented at ENM and SNM.

3.4. Can we separate out actual biological/physiological/pathophysiological change from variability introduced by the measurement process i.e. perform an analysis of the components of variance?

Likely yes: it is generally easier to separate out variability than bias. BUT depends on the details:

1. Patient characteristics (e.g. diabetes, other functional states)

2. Type of scan
    a. single PET scan
    b. serial PET scans (on same equipment and with same protocol) will tend to cancel out the effect of reproducible bias on change analysis
    c. mixing various scanner types and/or generations (e.g. at large PET centers or see also d)
    d. multi-center studies (clinical trials versus clinical care, esp. at tertiary centers)
3. Calibration phantoms are important and have a real role for ensuring that machine and results are maintained within their workable – and reliable range
4. Can we scan a calibration phantom alongside each patient ("a la" QCT)?
5. In general, 3-4 may be more workable with NIST-traceable calibrated sources

3.5. Given the variability in image acquisition and image analysis, what is the minimum change we can detect using a given modality and image analysis method, i.e. how large must the "effect" size ultimately be in order to detect a change in a single individual with significant statistical accuracy?

What we don't know in general:

1. We do not know how the measured changes relate to defined standards such as the EORTC standards [21]. This can be answered with entirely controlled virtual simulations and through physical phantom studies. Hybrid simulation (patient+phantom) methods may also help in this regard.

2. We don't know about coffee-break type variability (e.g. due to a combination of patient and scanner issues) with extended time between imaging sessions. Could we use variability of normal tissue variability from patients? Recall Weber 1999 [11] recommended a minimum limits of 20% change in FDG parameter based on a measured standard deviation of approximately 10% since changes of more than 20% were outside the observed 95% range for spontaneous fluctuations for the fifty tumors examined in his study (within restriction of study, e.g. no Rx therapy and also no change16 patients with less than 10 days between serial scans). It is important to note that figure 3 in the Weber 1999 article [11] implies study of lesions with different average values of initial FDG uptake would require different minimum effect sizes.

3.6. How do the answers to question 3.5 change for multicenter clinical trials?

1. Will need a larger effect size or sample size to compensate for the increased variation
2. Important to identify sources of variation and prospectively control them
3. Introduce calibration phantom as daily QC (could replace doing this with each patient)
4. Important to repeat with a calibration scan to understand bias, then other sources of variability will dominate

3.7. Can we identify software tools that can adequately quantify treatment-induced changes?

1. Scanner vendors have generally still limited tools, e.g. for visualization/comparison, SUV and ROIs: GE PET VCAR, Philips (?), Siemens True3D
2. Similar tools are also available from vendors not directly associated with a specific scanner: Hermes, MimVista, Mirada, Vital
3. Other tools are available through other vendors that may better fit the quantitation niche: Pmod (kinetic analysis)
4. And finally, there are open source tools that may - or generally may not - fit the requirements for a full clinical exploitation: e.g. from caBIG (XIP, AIM)

This list is not exhaustive.

3.8. Can we establish standards that will eventually lead to the acceptance of PET/CT image analysis software by regulatory agencies as surrogate end-points in new drug applications?

1. Establish benchmarks with databases (RIDER) and known phantoms.
2. Would like to separate acquisition from image analysis, so there is an important role for digital reference objects and simulation paradigms.
3. RIDER/NCIA could be a source of data but standards definition/certification needs to be taken up by a standards organization such as MITA/NEMA/ACR/ABSNM
4. RIDER may not be a large enough dataset.

3.9. Datasets

1. What data is needed in general?
    a. Simulation of scanners and phantoms for multiple vendors and images supporting multiple parameters
    b. Serial patient studies with meta-data (disease, therapy, outcomes, ground truth)
2. What data does RIDER have?
    a. Multi-vendor and multi-parameter calibration phantoms
    b. Serial patient studies (but without meta-data).
3. What data is in the literature and/or in progress from other efforts?
    a. Minn 1995, Weber 1999, Young 1999, Shankar 2006, and Boellaard 2008
4. What additional data is needed?
    a. NIST 68Ge calibration data
    b. Access to current trials data with meta-data and outcomes. These need to be accessible publicly and in a useful format.

References

[1] L. K. Shankar, J. M. Hoffman, S. Bacharach, M. M. Graham, J. Karp, A. A. Lammertsma, S. Larson, D. A. Mankoff, B. A. Siegel, A. Van den Abbeele, J. Yap, and D. Sullivan, "Consensus recommendations for the use of 18F-FDG PET as an indicator of therapeutic response in patients in National Cancer Institute Trials," J Nucl Med, vol. 47, pp. 1059-66, Jun 2006.

[2] M. R. MacManus, R. Hicks, R. Fisher, D. Rischin, M. Michael, A. Wirth, and D. L. Ball, "FDG-PET-detected extracranial metastasis in patients with non-small cell lung cancer

undergoing staging for surgery or radical radiotherapy--survival correlates with metastatic disease burden," Acta Oncol, vol. 42, pp. 48-54, 2003.

[3] W. A. Weber, V. Petersen, B. Schmidt, L. Tyndale-Hines, T. Link, C. Peschel, and M. Schwaiger, "Positron emission tomography in non-small-cell lung cancer: prediction of response to chemotherapy by quantitative assessment of glucose use," J Clin Oncol, vol. 21, pp. 2651-7, Jul 15 2003.

[4] H. A. Wieder, B. L. Brucher, F. Zimmermann, K. Becker, F. Lordick, A. Beer, M. Schwaiger, U. Fink, J. R. Siewert, H. J. Stein, and W. A. Weber, "Time course of tumor metabolic activity during chemoradiotherapy of esophageal squamous cell carcinoma and response to treatment," J Clin Oncol, vol. 22, pp. 900-8, Mar 1 2004.

[5] N. Avril, S. Sassen, B. Schmalfeldt, J. Naehrig, S. Rutke, W. A. Weber, M. Werner, H. Graeff, M. Schwaiger, and W. Kuhn, "Prediction of response to neoadjuvant chemotherapy by sequential F-18-fluorodeoxyglucose positron emission tomography in patients with advanced-stage ovarian cancer," J Clin Oncol, vol. 23, pp. 7445-53, Oct 20 2005.

[6] F. Cachin, H. M. Prince, A. Hogg, R. E. Ware, and R. J. Hicks, "Powerful prognostic stratification by [18F]fluorodeoxyglucose positron emission tomography in patients with metastatic breast cancer treated with high-dose chemotherapy," J Clin Oncol, vol. 24, pp. 3026-31, Jul 1 2006.

[7] F. Lordick, K. Ott, B. J. Krause, W. A. Weber, K. Becker, H. J. Stein, S. Lorenzen, T. Schuster, H. Wieder, K. Herrmann, R. Bredenkamp, H. Hofler, U. Fink, C. Peschel, M. Schwaiger, and J. R. Siewert, "PET to assess early metabolic response and to guide treatment of adenocarcinoma of the oesophagogastric junction: the MUNICON phase II trial," Lancet Oncol, vol. 8, pp. 797-805, Sep 2007.

[8] L. K. Dunnwald, J. R. Gralow, G. K. Ellis, R. B. Livingston, H. M. Linden, J. M. Specht, R. K. Doot, T. J. Lawton, W. E. Barlow, B. F. Kurland, E. K. Schubert, and D. A. Mankoff, "Tumor Metabolism and Blood Flow Changes by Positron Emission Tomography: Relation to Survival in Patients Treated With Neoadjuvant Chemotherapy for Locally Advanced Breast Cancer," J Clin Oncol, Jul 14 2008.

[9] M. E. Juweid and B. D. Cheson, "Positron-emission tomography and assessment of cancer therapy," N Engl J Med, vol. 354, pp. 496-507, Feb 2 2006.

[10] H. Minn, K. R. Zasadny, L. E. Quint, and R. L. Wahl, "Lung cancer: reproducibility of quantitative measurements for evaluating 2-[F-18]-fluoro-2-deoxy-D-glucose uptake at PET," Radiology, vol. 196, pp. 167-73, Jul 1995.

[11] W. A. Weber, S. I. Ziegler, R. Thodtmann, A. R. Hanauske, and M. Schwaiger, "Reproducibility of metabolic measurements in malignant tumors using FDG PET," J Nucl Med, vol. 40, pp. 1771-7, Nov 1999.

[12] M. Westerterp, J. Pruim, W. Oyen, O. Hoekstra, A. Paans, E. Visser, J. van Lanschot, G. Sloof, and R. Boellaard, "Quantification of FDG PET studies using standardised uptake values in multi-centre trials: effects of image reconstruction, resolution and ROI definition parameters," Eur J Nucl Med Mol Imaging, vol. 34, pp. 392-404, Mar 2007.

[13] R. Boellaard, W. J. Oyen, C. J. Hoekstra, O. S. Hoekstra, E. P. Visser, A. T. Willemsen, B. Arends, F. J. Verzijlbergen, J. Zijlstra, A. M. Paans, E. F. Comans, and J. Pruim, "The Netherlands protocol for standardisation and quantification of FDG whole body PET studies in multi-centre trials," Eur J Nucl Med Mol Imaging, Aug 15 2008.

[14] R. Boellaard, N. C. Krak, O. S. Hoekstra, and A. A. Lammertsma, "Effects of noise, image resolution, and ROI definition on the accuracy of standard uptake values: a simulation study," J Nucl Med, vol. 45, pp. 1519-27, Sep 2004.

[15] N. C. Krak, R. Boellaard, O. S. Hoekstra, J. W. Twisk, C. J. Hoekstra, and A. A. Lammertsma, "Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial," Eur J Nucl Med Mol Imaging, vol. 32, pp. 294-301, Mar 2005.

[16] V. Bettinardi, P. Mancosu, M. Danna, G. Giovacchini, C. Landoni, M. Picchio, M. C. Gilardi, A. Savi, I. Castiglioni, M. Lecchi, and F. Fazio, "Two-dimensional vs three-dimensional imaging in whole body oncologic PET/CT: a Discovery-STE phantom and patient study," Q J Nucl Med Mol Imaging, vol. 51, pp. 214-23, Sep 2007.

[17] S. M. Larson, Y. Erdi, T. Akhurst, M. Mazumdar, H. A. Macapinlac, R. D. Finn, C. Casilla, M. Fazzari, N. Srivastava, H. W. Yeung, J. L. Humm, J. Guillem, R. Downey, M. Karpeh, A. E. Cohen, and R. Ginsberg, "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, vol. 2, pp. 159-171, May 1999.

[18] N. M. Freedman, S. K. Sundaram, K. Kurdziel, J. A. Carrasquillo, M. Whatley, J. M. Carson, D. Sellers, S. K. Libutti, J. C. Yang, and S. L. Bacharach, "Comparison of SUV and Patlak slope for monitoring of cancer therapy using serial PET scans," Eur J Nucl Med Mol Imaging, vol. 30, pp. 46-53, Jan 2003.

[19] R. K. Doot, L. K. Dunnwald, E. K. Schubert, M. Muzi, L. M. Peterson, P. E. Kinahan, B. F. Kurland, and D. A. Mankoff, "Dynamic and static approaches to quantifying 18F-FDG uptake for measuring cancer response to therapy, including the effect of granulocyte CSF," J Nucl Med, vol. 48, pp. 920-5, Jun 2007.

[20] G. M. McDermott, A. Welch, R. T. Staff, F. J. Gilbert, L. Schweiger, S. I. Semple, T. A. Smith, A. W. Hutcheon, I. D. Miller, I. C. Smith, and S. D. Heys, "Monitoring primary breast cancer throughout chemotherapy using FDG-PET," Breast Cancer Res Treat, vol. 102, pp. 75-84, Mar 2007.

[21] H. Young, R. Baum, U. Cremerius, K. Herholz, O. Hoekstra, A. A. Lammertsma, J. Pruim, and P. Price, "Measurement of clinical and subclinical tumour response using [18F]-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, vol. 35, pp. 1773-82, Dec 1999.

[22] Doot RK, Christian PE, Mankoff DA, Kinahan PK. Reproducibility of Quantifying Tracer Uptake with PET/CT for Evaluation of Response to Therapy. In: 2007 IEEE Nuclear Science Symposium and Medical Imaging Conference, Honolulu, Hawaii Oct 27 - Nov 3, Vol. , pp 2880-2884, 2007.

[23] Kinahan PE, Doot RK, Christian PE, Karp JS, Scheuermann JS, Zimmerman RE, Saffer JR, McEwan AJ. Multi-center comparison of a PET/CT calibration phantom for imaging trials. Journal of Nuclear Medicine, vol. 49, pp P. (abstract), 2008