Quantitative
Imaging
Biomarkers
Alliance
RSNA

Application for QIBA Round-2 Project Funding

Title of Proposal: : *Evaluation of the Variability in Determination of Quantitative PET Parameters of Treatment Response across Performance Sites and Readers*		
QIBA Committee/Subgroup: FDG PET		
NIBIB Task Number(s) which this project addresses: (multiple, notably variance in quant estimates Rx response by PET)		
Project Coordinator or Lead Investigator Information:		
Last Name: Wahl	First Name: Richard	Degree(s): MD
e-mail:	Tel #:	
Institution/Company: Johns Hopkins University + 14 Additional FDG PET Tech (Analysis) Sites		
Amount Requested:		

Please check the primary category for this proposal from among the following:

☑ 1. Identification of Technical Characteristics and Standards

- ☐ a. Creation and refinement of protocols for image acquisition, analysis, quality control, etc., for specific clinical utility
- ☐ b. Phantom development and testing
- ☑ c. Identification and assessment of intra-reader bias (1) and variance across scanners and centers
- ☑ d. Identification and assessment of inter-reader bias and variance across scanners and centers
- ☐ e. Other

☐ 2. Clinical Performance Groundwork

- ☐ a. Assessment of intra-reader sensitivity and specificity
- ☐ b. Assessment of inter-reader sensitivity and specificity
- ☐ c. Other

☐ 3. Clinical Efficacy Groundwork

- ☐ a. Assessment of correlation between new biomarker and 'accepted-as-standard' method
- ☐ b. Characterization of value in clinical trials
- ☐ c. Characterization of value in clinical practice
- ☐ d. Development/merger of databases from trials in support of qualification
- ☐ e. Other

☑ 4. Resources (money and/or people) committeed from other sources.

Johns Hopkins University, Image Response Assessment Lab will provide digital data from several patient's studies for analysis by QIBA performance sites (at least 15). A targeted two readers/site will allow 30 separate reads to be performed. Image data will be anonymized from JHU clinical patients and therapeutic trials in breast, lung, lymphoma, colon cancer, and other cancers receiving chemotherapy or combination chemo biological therapy. This data has already been collected. Imaging archive and anonymization tools are in place.

Statistical support will be secured from the JHU QIN statistician, Dr. Hao Wang. Dr. Constantine Gatsonis will serve as a statistical consultant. Analysis workstations for PET/CT assessment by sites who will do quantitative reviews of imaging data are available, and no new software would be purchased.

Please provide a one-page summary that includes the following information:

Project Description-

There is very limited data on the performance of varying readers and quantitative imaging workstations in determining cancer treatment response using FDG PET/CT. We propose a study design using well-defined anonymized pre-treatment and post-treatment FDG PET scans of cancer patients as an analysis set. All studies will have been performed at Johns Hopkins, and all will have been done using a 3D PET scanner with LySO crystals and modern iterative image processing. No scans from 2D PET will be used.

We will determine how reproducible quantitative analysis of several major PET parameters are across sites and readers. A detailed statistical plan is included. Our primary metric will be % change in SUV max, determined pre- and post-Rx in the "hottest tumor" as defined by the reader. If the tumor has disappeared fully with treatment, background will be noted in the liver.

Secondary metrics will include absolute SUV peak, SUL max, SUL peak, and TLG (as determined by site), normal liver SUV and SUL in a 3 cm diameter sphere, as well as SD of this region of interest. We will also secure correlative measurements of tumor size when available. In this way, we will determine what component of variability there is in the reader/workstation/lesion selection elements of quantitative assessments of treatment response, when all sites have the same realistic human FDG PET/CT digital data set available. This analysis will target 15 performance sites, and 30 experienced imaging specialists. Such information will inform our field and help us determine if current tools and training are sufficient for deployment in a more general manner of quantitative PET/CT of treatment response, by precisely defining the variability in estimates of % decline in SUV across sites.

Primary Goals and Objectives-

To determine the degree of variance among varying PET centers, using a variety of different workstations and interpreters in measuring the % change in SUV max in the "hottest tumor" before and after therapy. Thirty paired pre- and post-Rx PET data sets (whole body) will be reviewed. Our study will allow us to estimate with considerable certainty the variability in this determination due to measurement differences. We expect an ICC of 0.85 or higher and have powered our study to measure this, with an expectation that the true ICC will be approximately 0.9 (prior single site ICC have been 0.94 to 1.0).

We also will explore the variability among estimates of SUV peak changes, SUL max and peak changes, and TLG changes. We will also determine variance in normal liver and normal blood pool SUV, mean and SD in defined ROIs.

Deliverables-

1) A survey of interested sites, their technical workstation status, and medical qualifications of readers, will be conducted. Data reporting forms will be developed, and interested sites will be identified (3 months).

2) The clinical data set will be identified and developed using anonymized patient data, including cancer of the breast, lung, colon, lymphoma and elsewhere acquired using suitable methods. This will be anonymized. Sites solicited and selected, will substantially be from interested members of the PET/CT technical committee. This will likely include academic sites, CRO sites, and possibly a site or more from a private practice setting doing PET/CT frequently. Ideally, 2 experienced radiologists (NM specialists) at each site will independently assess the PET/CT scans for SUV max, SUV peak, SUL max, SUL peak (1 and 5 lesions), TLG and normal liver/blood values. Tumor size measures will be determined when feasible.

3) The inter-observer variance will be determined for changes in the SUV parameters between studies, as well as for absolute values. Kappa statistics, and likely Bland Altman plots will be determined to assess change metrics and absolute SUV values Publication will be prepared of results for presentation at national meeting.

Timeline [must include intermediate measureable milestones.]-

Aim 1: (Months 1-3): Site selection, form design, case selection, non-study test cases sent to performance sites for analysis and success of importation of digital data. Selection of relevant cases and anonymization.
Aim 2: (Months 4-6): Final selection of performance sites. Data distribution to sites for analysis and start of analysis.
Aim 3: (Months 7-9): Completion of analysis by sites and reporting to JHU statistics group.
Aim 4: (Months 10-12): Analysis of data centrally, reporting and preparation of first draft of results for publication and to QIBA technical committee and QIBA leadership.

Statistical considerations-

The primary quantitative measurement will be SUV max, and it will be determined based on pre- and post-treatment PET imaging. The % decline in SUV max will be calculated for the single hottest tumor as the percentage difference between the pre- and post-treatment values. Our primary goal is to assess the repeatability of the parameter -- % decline in SUV max across varying workstations or readers. Restricted maximum likelihood estimation of variance components in a random effects model will be used to estimate the intra- and inter-subject variance components for each measure. The model will include tumor ID as a random classification factor. The variance component estimates will be used to compute the intra-class correlation coefficient (ICC) and the intra-subject coefficient of variation (CV) as measures of repeatability for each parameter. In addition, we will assess the agreement between workstations using Bland-Altman plots.

To evaluate the systematic difference between workstations/sites in the quantitative parameter measurement, we will compare the sites using mixed effect models where sites will be included as a fixed effect.

We will also explore additional PET imaging parameters, including SUV (and SUL) mean in liver, SUV mean (and SUL mean) in descending aorta, SUL max tumor , SUV peak tumor , SUL peak tumor , and total lesion glycolysis and total tumor volume. While our major focus is the single hottest tumor identified by the reader, the hottest up to 5 tumors will be assessed if more than one tumor lesion is present and identified. The analysis methods are the same as described for SUV max. We will correlate the imaging parameters with the tumor size measures using regression analysis. Tumor size will be noted in 1D in cases when tumors can be measured.

Sample size and power calculation-

Statistical power is based on the primary objective to evaluate the reproducibility of the % decline measures between sites. Our goal is to assess how much of total variance in the % decline measure is attributable to the variance between sites. We will estimate inter-site/inter-rater correlation coefficient (ICC), which is a measure of between site-agreement on the % decline metric.

Table 1 summarizes the precisions on the ICC estimate under various scenarios of the ICC. Our preliminary data from a single institution/workstation study estimates an ICC of 0.94, which indicates an excellent reliability of the % decline metric. We expect a similar ICC in this multiple site study. Assuming that each site will have two readers and each reader will evaluate 30 hottest tumors to determine their % decline, 15 sites would allow the ICC to be estimated with precision of ± 3% (i.e., half width of the 95% confidence interval is 3%), and if we observe that the ICC is 0.94, its 95% confidence interval (CI) will be [0.91, 0.97]. That is, if the variance between sites accounts for 6% of total variance, its 95% confidence interval will be [3.0%, 9.0%]. If we observe an ICC of 0.90 resulting from greater variation across sites, (which we expect given multiple readers and workstations) the precision of the ICC estimate will be ± 4.8%, and the corresponding 95% CI is [0.852, 0.948]. This estimate will assure the ICC is in excess of 0.85, a figure indicating extremely strong reproducibility across sites.

Table 1. Precision of the ICC estimate.

The observed ICC	Number of sites	Precision (half width of the 95% CI of ICC)	95% confidence interval
0.95	15	2.5%	[0.925, 0.975]
0.94	15	3%	[0.91, 0.97]
0.90	10	5%	[0.85, 0.95]
0.90	15	4.8%	[0.852, 0.948]

We believe 15 sites is the best choice for sample size, especially as ICC may be less than our single center 0.94 estimate. Expecting 0.90 ICC, the range of 0.85 to 0.95 would indicate the user aspects of the process are small contributors to overall variance. Thus, 15 sites with two readers/site is proposed based on the desired 95% confidence intervals.