Am J Nucl Med Mol Imaging 2013;3(2):175-181
www.ajnmmi.us /ISSN:2160-8407/ajnmmi1212005

Original Article

In situ study of the impact of inter- and intra-reader variability on region of interest (ROI) analysis in preclinical molecular imaging

Frezghi Habte[1,2], Shradha Budhiraja[1,2*], Shay Keren[1,2**], Timothy C Doyle[1,3], Craig S Levin[1,2], David S Paik[1,2]

[1]Molecular Imaging Program at Stanford (MIPS), [2]Department of Radiology, [3]Department of Pediatrics, Stanford University, Stanford, CA, USA. [*]Current Address: Adobe Systems India Private Limited, City Center, Sector 25-A, Noida 20130, India. [**]Current Address: Nofim School, Haifa, Israel.

Received December 24, 2012; Accepted January 28, 2013; Epub March 8, 2013; Published March 18, 2013

Abstract: We estimated reader-dependent variability of region of interest (ROI) analysis and evaluated its impact on preclinical quantitative molecular imaging. To estimate reader variability, we used five independent image datasets acquired each using microPET and multispectral fluorescence imaging (MSFI). We also selected ten experienced researchers who utilize molecular imaging in the same environment that they typically perform their own studies. Nine investigators blinded to the data type completed the ROI analysis by drawing ROIs manually that delineate the tumor regions to the best of their knowledge and repeated the measurements three times, non-consecutively. Extracted mean intensities of voxels within each ROI are used to compute the coefficient of variation (CV) and characterize the inter- and intra-reader variability. The impact of variability was assessed through random samples iterated from normal distributions for control and experimental groups on hypothesis testing and computing statistical power by varying subject size, measured difference between groups and CV. The results indicate that inter-reader variability was 22.5% for microPET and 72.2% for MSFI. Additionally, mean intra-reader variability was 10.1% for microPET and 26.4% for MSFI. Repeated statistical testing showed that a total variability of CV < 50% may be needed to detect differences < 50% between experimental and control groups when six subjects (n = 6) or more are used and statistical power is adequate (80%). Surprisingly high variability has been observed mainly due to differences in the ROI placement and geometry drawn between readers, which may adversely affect statistical power and erroneously lead to negative study outcomes.

Keywords: Molecular imaging, preclinical, region of interest analysis, variability, microPET, multispectral fluorescence imaging

Introduction

Following image acquisition, quantitative image analysis of complex biological processes is a critical component in molecular imaging studies [1-6]. In particular, radionuclide imaging and optical imaging have emerged as the most utilized approaches to molecular imaging that determine signal uptake in target organs *in vivo*. Positron emission tomography using small-animal models (microPET) provides a powerful technique to explore *in vivo* pathophysiology in a flexible, non-invasive, and potentially highly quantitative fashion [7, 8]. Similarly, optical bioluminescence and fluorescence imaging are also economical methods for pre-clinical tumor imaging studies that are widely used in various molecular imaging research studies including drug development [9-11]. Compared to a simple qualitative analysis (e.g., visual inspection), which is highly subjective, quantitative image analysis characterizes biological processes in an objective and repeatable manner [9]. However, due to the complexity of the imaging instrument and associated analysis tools, there is still significant variability in quantitative imaging, which makes the development of standardized methods difficult [13, 14].

Figure 1 describes the major sources of variability involved in the molecular imaging ana-

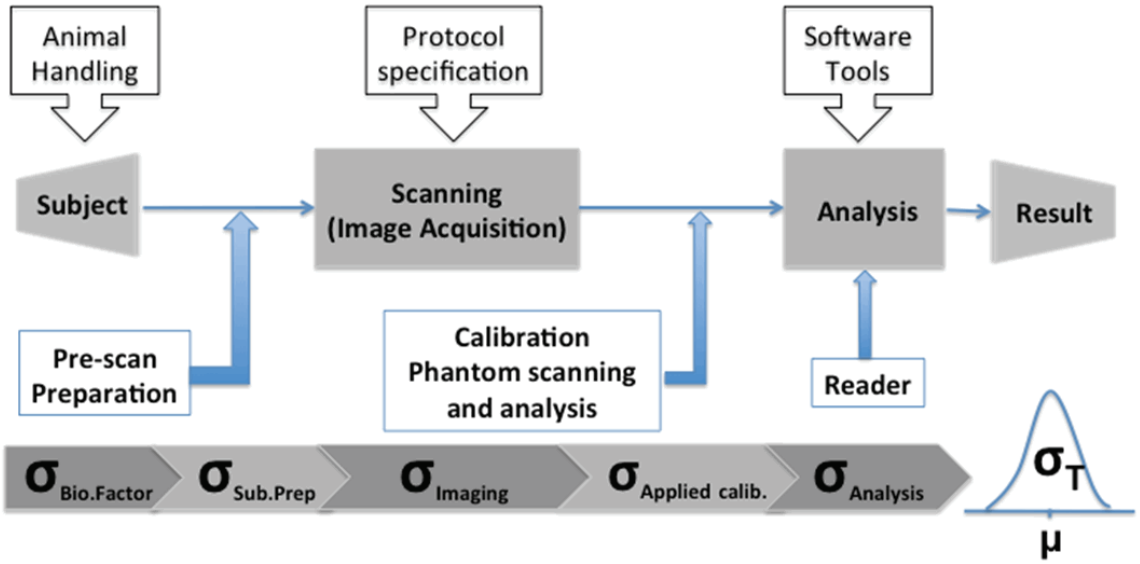

Figure 1. Numerous sources of variability in typical pre-clinical molecular imaging, where the s's are the variation (standard deviations) from each step in the analytic pipeline with respect to the measured mean value (m).

lytic pipeline from subject preparation to scanning and to quantitative analysis. Prior to image acquisition, subject preparation introduces some level of variation. These operator-dependent processes include all animal handling issues, anesthesia, tracer/contrast injection, and placement of animal in the scanner resulting in variation to the final image quantitation. For example, Feuger, et al. [10] demonstrated variations of fluorodeoxyglucose (FDG) uptake ranging between 3- to 15-fold within various tissue types/organs of a mouse model. This resulted from a few basic animal-handling conditions such as fasting, stage warming and the type of anesthesia used.

Other major sources of variability include natural biological variation, even in syngeneic animals, and imaging-specific factors such as adherence to imaging protocols and instrument calibration. To reduce variability, most pre-clinical molecular imaging experiments currently follow highly controlled animal handling procedures and acquire images with the same instrument configurations. Following some of these assumptions, Jan, et al. [11] for example, have estimated the biological variability of FDG uptake using microPET in mouse models. The result showed that inter-animal biological variability varies between 15% and 35% for liver, heart, kidney and leg muscles while intra-animal variability remains below 10%, which was attributed to the variation of the imaging instrument and other associated variability factors. This indicates that the underlying biology is highly variable in nature and reduction of variability in other factors is critical in molecular imaging.

The last contribution to the variability in a typical molecular imaging study occurs post image acquisition during image analysis specifically when performing region of interest (ROI) analysis. Most image analysis software tools require accurate delineation/segmentation of ROI on the images that often is difficult and heavily dependent on the image quality and the application [12]. For practical reasons, interactive manual or semi-manual definition of ROI approaches is commonly used. Thus, the quantitative image analysis process also adds some level of subjectivity both due to the image analysis tool used and the reader (operator) inconsistencies. In this study, we estimated the variability due to image analysis methods and investigated its impact, with emphasis on studying experienced molecular imagers in real-world (i.e., in situ) conditions.

Materials and methods

Image acquisition

In order to be able to report on variation as observed in practice, we favored the use of image datasets from ongoing research studies

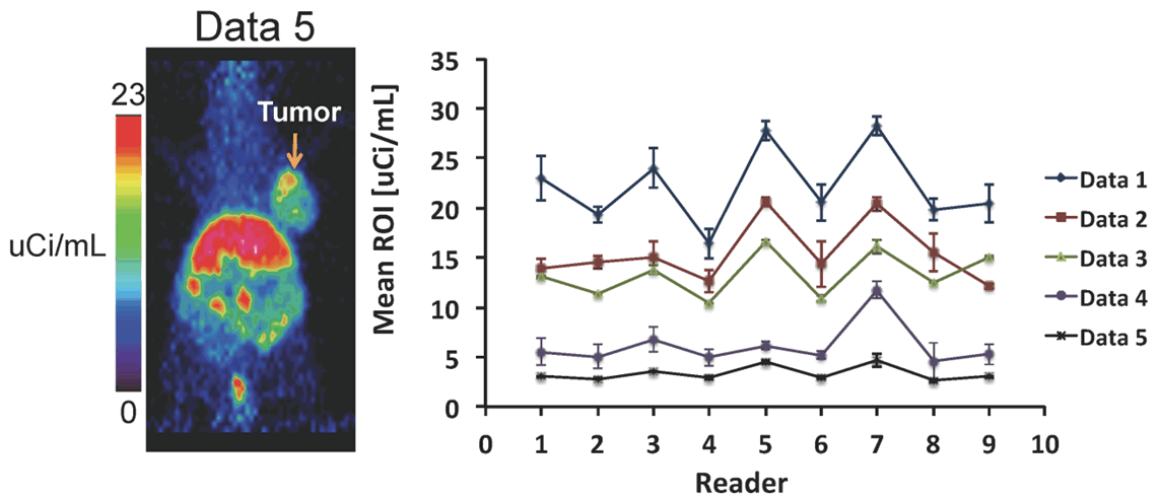

Figure 2. Estimate of inter- and intra-reader variability study using microPET image data. *Left*, a sample of selected microPET image acquired at our imaging facility for other projects. *Right*, comparison of extracted mean ROI value showing significant variation with nearly consistent pattern of either under- or over-estimation of mean ROI value of respective reader.

rather than artificially contrived datasets analyzed with artificially constrained toolsets. We designed this experiment to include two different imaging modalities with a sampling of independent real-world studies being performed within the Stanford Center for Innovation in In-Vivo Imaging (SCI[3]). We selected five images from five separate cancer studies each acquired using the R4 microPET system (Concorde Microsystems, Knoxville, TN), which is amongst the most highly utilized imaging instruments in the facility and five using the CRi Maestro multispectral fluorescence imaging (MSFI) system (Caliper Life Sciences, Hopkinton, MA). Each of the datasets contained exactly one visible xenograft tumor. All microPET images were acquired for 5 min static scans duration in list mode data and reconstructed using a two-dimensional ordered subsets expectation maximum (OSEM) algorithm. In order to avoid bias on the study, we purposely blinded readers from information such us tumor type, tracer used, and specific animal handling issues. Likewise, all MSFI images were acquired with an exposure time of 1000 ms and emission wavelength filter between 500 and 950 nm sampled in 10 nm steps. Again readers were blinded on the type of tumor and fluorophore labeling type used.

Image analysis

Image analysis was performed using vendor-supplied image analysis software tools (ASI Pro VM for microPET and CRi Maestro for MSFI). All readers selected to analyze the data were familiar with these software tools prior to the study. ASI Pro allows the definition of various types of ROIs geometries such as ellipse, rectangle, or free trace for both 2D and 3D ROIs. It also displays the image in the three windows (transverse, coronal and sagittal) making it convenient for locating and delineating the target ROIs. CRi Maestro acquires a cube of image data consisting of multi-spectral 2D images. Before ROI analysis, the MSFI data must be "unmixed" to remove the autofluorescence signal (**Figure 3**). The unmixing tool with manual or auto unmixing options allows the user to either manually specify regions with each pure single spectrum component or to use a Real Component Analysis (RCA) algorithm in order to unmix the spectral components.

Assessment of variability

To estimate inter- and intra-reader ROI analysis variability, we selected ten experienced readers that had not previously seen the source images. Each reader extracted the mean ROI value after delineation of suitable ROI around the tumor entirely based on his or her experience and judgment. Readers were allowed to select the ROI tool of their choice according to the habits they routinely use to analyze their own data. We provided raw image data and it was up to the reader to adjust the threshold and image contrast based on their experience

Table 1. microPET ROI mean CV for inter- and intra-reader variability

	Data 1	Data 2	Data 3	Data 4	Data 5
Inter-reader CV	17.8	19.9	16.8	36.0	22.1
Mean Intra-reader CV [%]	10.2	11.7	6.8	16.5	5.1

Table 2. MSFI ROI mean CV for inter- and intra-reader variability

	Data 1	Data 2	Data 3	Data 4	Data 5
Inter-reader CV [%]	35.6	62.6	77.3	100.3	85.2
Mean Intra-reader CV [%]	20.3	21.1	29.4	17.1	44.0

that would allow them draw ROI around the tumor optimally. The mean ROI generated by each user was used to estimate inter-reader variability using the coefficient of variation (CV). Readers repeated the analysis three times for each data set to estimate the intra-reader variability. Readers were aware of the repeat measurements, but they were spaced as far apart as possible to minimize memory recall effects. Nine readers completed the entire analysis (one reader was unable to finish the study due to time constraints).

Assessment of the impact of variability using statistical hypothesis testing

To help provide context in the assessment of the impact of variability on statistical hypothesis testing, we performed numerous iterations of statistical hypothesis tests on randomly sampled data. We assumed a control group and an experimental group with an equal number of subjects in each group over several values (n=3, 6, 10 and 15). We also assumed normal distributions with variance determined by CV ($\sigma^2=CV^2\mu^2$) and the difference in means under the alternative hypothesis was expressed as a percentage. For each data point, t-tests ($\alpha=0.05$) were performed 5,000 times and Type II error rate (β) was reported as the statistical power ($1-\beta$).

Results

Inter- and intra- reader variability

Estimates of inter- and intra-reader variability were obtained by computing CV from the mean ROI value extracted from the five data sets for each modality and from the three repeated measurements by each reader, respectively. **Figure 2** shows the mean microPET ROI values for each reader/dataset combination and error bars show the range across three repeat measurements. A comparison of extracted ROI mean values between readers (**Table 1**, **Figure 2**, right) indicates that there is notable inter-reader variability with an average CV of 22.5% (**Table 1**). In addition, a pattern where each reader either consistently underestimated or overestimated the extracted mean ROI value (relative to the group mean) was observed. Connected line segments between data points shown in **Figure 2** (right) clearly demonstrate this pattern. The observed mean intra-reader variability (indicated as an error bar in **Figure 2**, right) was significantly lower compared to inter-reader variability ($p=0.038$) with an average CV of 10.1%.

For MSFI data, a much higher ROI mean variability was observed with an average CV of 72.2% (**Table 2**, **Figure 3**) compared to microPET. This is mainly due to higher noise and less specificity of optical imaging due to high scatter of photons propagating through biological tissue [13]. A similar pattern to microPET was observed where each reader either underestimated or overestimated (relative to the group mean), though not as consistently as in microPET. The estimated intra-reader average CV computed from repeated measurements of each reader was 26.4%, significantly ($p=0.0395$) higher than microPET data.

Quantitative estimates of the impact of variability on tumor quantification

To assess the impact of variability, we performed t-tests over the randomly sampled data under different conditions. **Figure 4** shows the computed probability of true detection of biological responses under the alternative hypothesis (i.e., statistical power) as function of variability for different conditions (different number of subjects, magnitude of differences, and total

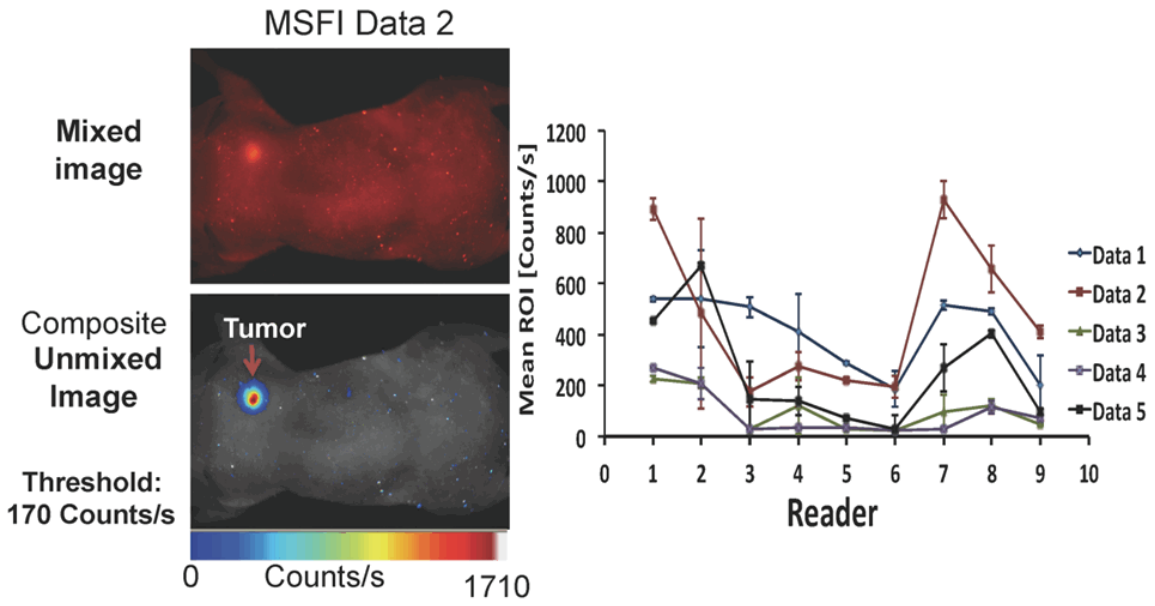

Figure 3. Inter- and intra-reader variability study performed using image data acquired from MSFI. *Left top*, selected raw mixed image data, *Left bottom*, Composite unmixed image sample with tumor (rainbow) and tissue auto fluorescence (grey). *Right*, comparison of extracted mean ROI value of each reader for the randomly selected five data sets.

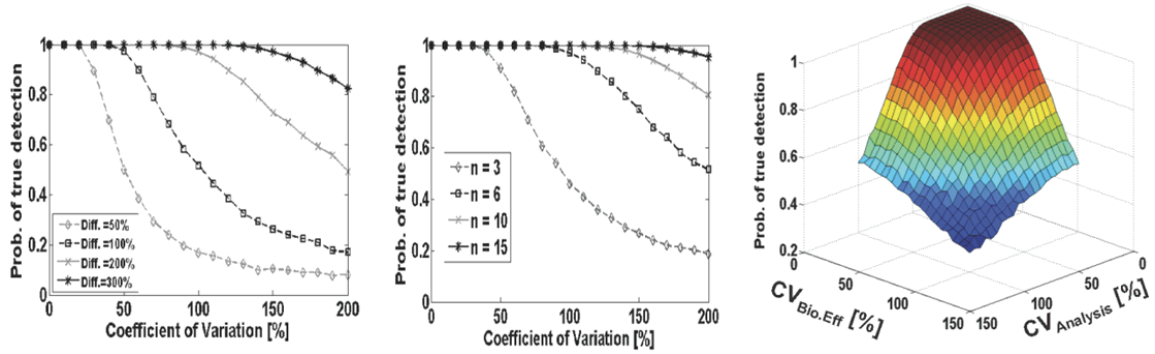

Figure 4. Computed probability of statistically significant biological response under the alternative hypothesis (i.e., statistical power) as function of CV assuming different mean relevant difference in biological response with n=6 (left) and different number of subjects with diff=50% (middle). Right, Illustrating combined variability effects due to biological factors and image analysis with n=6 and diff=50%.

CV). Note that here, total CV includes ROI analysis as well as all the other steps shown in **Figure 1**. The result indicates that a total variability estimates of 50% or less may be needed to detect relatively small difference (< 50%) in biological response for n=6 to ensure reproducibility with an adequate statistical power of 80% (**Figure 4**, left). As expected, statistical power is limited if a small number of subjects is used in the experiment. This is true even if there is a large difference (> 100%) in biological response between control and experimental groups (**Figure 4**, middle). **Figure 4** right, illustrates the combined variability effect both due to the underlying biology and human factors during image analysis assuming n=6. A better accuracy and precision may be obtained when the effective variability remains below 50%. Note that even small reductions in the total CV on the order of 10-15% can produce large jumps in statistical power on the order of 30%.

Discussion

In molecular imaging, quantitative image analysis is an essential component of molecular imaging commonly used to capture or measure small differences in biological response, where simple qualitative visualization of images is insufficient [14]. Clinically, well-validated quan-

titative techniques play a key role in medical decision-making [15-19]. As a result, those quantitative techniques that are in clinical use are more standardized and robust compared to those in pre-clinical research [20]. Comparably, however, pre-clinical research dealing with the development of new biomarkers and evaluation of new drug responses to therapy entertains entirely much larger variety and volume of image data. Complex analyses of image data are usually involved to measure and compare various biological responses [21]. Yet, there is generally little standardization of tools or methods—in our experience tools become popular mainly due to their simplicity and availability. Many researchers still perform their respective analysis using their favorite tool or in-house software tools specific for their application. Thus, such traditional methods of quantitative image analysis in addition to the complexity of biological process in pre-clinical research introduce several sources of variability.

In this study, we assessed the source of variability and its impact when performing region of interest (ROI) analysis of tumors in mouse models. The study indicates that due to the use of manually defined ROIs, high inter- and intra-reader variability is observed limiting the precision of molecular imaging data analysis. Since ROI definition is generally done manually, the current methods suffer from a high degree of subjectivity introducing significant variability in the subsequent analysis. The results in **Figures 3** and **4** demonstrated this issue with a pattern observed where each user consistently either under- or over-estimates the value relative to the whole group, thus limiting any quantitative study-to-study comparisons. Most importantly, the impact of this variability as we have demonstrated "in situ" is on statistical power. Small increases in variability can lead to precipitous drops in power, which ultimately leads to wasted studies with falsely negative results and missed opportunities.

Progress in biology is increasingly relies on high-throughput data including imaging [22], which makes the task of image analysis more tedious. At the same time, there is potentially more biologically relevant information waiting to be extracted from the large volumes of images. Due to this and other factors, there is a growing need to replace subjective manual evaluation of images with computerized automatic or semi-automatic analysis methods. Automated methods may offer the possibility of improving the sensitivity, precision, reproducibility and objectivity of data analysis methods when especially similar input image quality is utilized. Otherwise, a cautiously performed analysis can significantly improve variability to limit the biological effect and/or other non-human factors with much less impact.

Acknowledgments

The authors acknowledge the use of imaging instruments and image analysis support in the Stanford Center for Innovation in In-Vivo Imaging (SCI^3). This work was supported by grants NCI ICMIC P50-CA114747, NIH U54 CA119367 and Stanford Cancer Center. No other potential conflict of interest relevant to this article was reported

Address correspondence to: Dr. Frezghi Habte, Molecular Imaging Program at Stanford (MIPS), Department of Radiology, Stanford University, Stanford, CA, USA. E-mail: fhabte@stanford.edu

References

[1] Su H, Forbes A, Gambhir SS and Braun J. Quantitation of cell number by a positron emission tomography reporter gene strategy. Mol Imaging Biol 2004; 6: 139-148.

[2] Carlson SK, Classic KL, Hadac EM, Dingli D, Bender CE, Kemp BJ and Russell SJ. Quantitative Molecular Imaging of Viral Therapy for Pancreatic Cancer Using an Engineered Measles Virus Expressing the Sodium-Iodide Symporter Reporter Gene. Am J Roentgenol 2009; 192: 279-287.

[3] Jon NM, Kathryn CP, Dana RA, Michael JS, Gregory ML and Samuel AW. Molecular Imaging With Targeted Perfluorocarbon Nanoparticles: Quantification of the Concentration Dependence of Contrast Enhancement for Binding to Sparse Cellular Epitopes. Ultrasound med biol 2007; 33: 950-958.

[4] Li W, Li F, Huang Q, Frederick B, Bao S and Li CY. Noninvasive Imaging and Quantification of Epidermal Growth Factor Receptor Kinase Activation In vivo. Cancer Res 2008; 68: 4990-4997.

[5] Yang D, Han L and Kundra V. Exogenous Gene Expression in Tumors: Noninvasive Quantification with Functional and Anatomic Imaging in a Mouse Model[1]. Radiology 2005; 235: 950-958.

[6] Erdi YE, Humm JL, Imbriaco M and Larson SM. Quantitative bone metastases analysis based on image segmentation. J Nucl Med 1997; 38: 1401-1406.
[7] Vaska P, Rubins DJ, Alexoff DL, Schiffer WK. Quantitative Imaging with the Micro-PET Small-Animal PET Tomograph. Int Rev Neurobiology 2006; 73: 191-218.
[8] Yeh HH, Ogawa K, Balatoni J, Mukhapadhyay U, Pal A, Gonzalez-Lepera C, Shavrin A, Soghomonyan S, Flores L, Young D, Volgin AY, Najjar AM, Krasnykh V, Tong W, Alauddin MM and Gelovani JG. Molecular imaging of active mutant L858R EGF receptor (EGFR) kinase-expressing nonsmall cell lung carcinomas using PET/CT. Proc Nati Acad Sci 2011; 108: 1603-1608.
[9] Wang YX and Ng CK. The impact of quantitative imaging in medicine and surgery: Charting our course for the future. Quant Imaging Med Surg 2011; 1: 1-3.
[10] Fueger BJ, Czemin J, Hildebradt I, Tran C, Halpern BS, Stout D, Phelps ME and Weber WA. Impact of Animal Handling on the Results of 18F-FDG PET studies in Mice. J Nucl Med 2006; 47: 999-1006.
[11] Jan S, Boisgard R, Fontyn Y, Eroukhmanoff C, Comtat C and Trebossen R. Accuracy and variability of quantitative values obtained for mouse imaging using the microPET FOCUS. IEEE Nucl Sci Symp Conf Rec 2004; 5: 2934-2937.
[12] Article R. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. J Med Phys 2008; 33: 136-140.
[13] Qin C, Zhu S and Tian J. New Optical Molecular Imaging Systems. Curr Pharm Biotechnol 2010; 11: 620-627.
[14] Jannin P, Krupinski E and Warfled S. Validation in medical image processing. IEEE Trans Med Imaging 2006; 25: 1405-1409.
[15] Lordick F. PET to assess early metabolic response and to guide treatment of adenocarcinoma of the oesophagogastric junction: the MUNICON phase II trial. Lancet Oncol 2007; 8: 797-805.
[16] Oyen WJ, van der Graaf WT. Molecular imaging of solid tumors: exploiting the potential. Nat Rev Clin Oncol 2009; 6: 609-611.
[17] Prior JO. Early prediction of response to sunitinib after imatinib failure by 18F-fluorodeoxyglucose positron emission tomography in patients with gastrointestinal stromal tumor. J Clin Oncol 2009; 27: 439-445.
[18] Pio BS. Usefulness of 3'-[F-18]fluoro-3'-deoxythymidine with positron emission tomography in predicting breast cancer response to therapy. Mol Imaging Biol 2006; 8: 36-42.
[19] Ott K. Metabolic imaging predicts response, survival, and recurrence in adenocarcinomas of the esophagogastric junction. J Clin Oncol 2006; 24: 4692-4698.
[20] O'Connor JP, Jackson A, Asselin MC, Buckley DL, Parker GJ, Jayson GC. Quantitative imaging biomarkers in the clinical development of targeted therapeutics: current and future perspectives. Lancet Oncol 2008; 9: 766-776.
[21] Swedlow JR, Goldberg I, Brauner E and Sorger PK. Informatics and Quantitative Analysis in Biological Imaging. Science 2003; 300: 100-102.
[22] Kherlopian A, Song T, Duan Q, Neimark M, Po M, Gohagan J and Laine A. A review of imaging techniques for systems biology. BMC Syst Biol 2008; 2: 74.