Method for detecting voxelwise changes in fluorodeoxyglucose-positron emission tomography brain images via background adjustment in cancer clinical trials

Lei Qin
Armin Schwartzman
Keisha McCall
Nezamoddin N. Kachouie
Jeffrey T. Yap

Method for detecting voxelwise changes in fluorodeoxyglucose-positron emission tomography brain images via background adjustment in cancer clinical trials

Lei Qin,^{a,*} Armin Schwartzman,^b Keisha McCall,^a Nezamoddin N. Kachouie,^c and Jeffrey T. Yap^d

^aDana-Farber Cancer Institute, Department of Imaging, Boston, Massachusetts, United States

^bUniversity of California, San Diego, Division of Biostatistics, La Jolla, California, United States

^cFlorida Institute of Technology, Department of Mathematical Sciences, Melbourne, Florida, United States

^dUniversity of Utah, Huntsman Cancer Institute, Department of Radiology and Imaging Sciences, Salt Lake City, Utah, United States

Abstract. An important challenge to using fluorodeoxyglucose-positron emission tomography (FDG-PET) in clinical trials of brain tumor patients is to identify malignant regions whose metabolic activity shows significant changes between pretreatment and a posttreatment scans in the presence of high normal brain background metabolism. This paper describes a semiautomated processing and analysis pipeline that is able to detect such changes objectively with a given false detection rate. Image registration and voxelwise comparison of the pre- and posttreatment images were performed. A key step is adjustment of the observed difference by the estimated background change at each voxel, thereby overcoming the confounding effect of spatially heterogeneous metabolic activity in the brain. Components of the proposed method were validated via phantom experiments and computer simulations. It achieves a false response volume accuracy of 0.4% at a significance threshold of 3 standard deviations. It is shown that the proposed methodology can detect lesion response with 100% accuracy with a tumor-to-background-ratio as low as 1.5, and it is not affected by the background brain glucose metabolism change. We also applied the method to FDG-PET patient images from a clinical trial to assess treatment effects of lapatinib, which demonstrated significant changes in metabolism corresponding to tumor regions. © 2017 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.4.2.024006]

Keywords: brain oncology; image analysis; neuroimaging; positron emission tomography.

Paper 16275RR received Dec. 30, 2016; accepted for publication May 10, 2017; published online Jun. 1, 2017.

1 Introduction

Positron emission tomography (PET) imaging can use a tracer dose of injected 18F-fluorodeoxyglucose (FDG) to measure the rate of glucose metabolism in tissue.¹ Quantitative FDG-PET has been used as an imaging biomarker to evaluate response to brain cancer therapy between pretreatment (baseline) and posttreatment (follow-up) PET images. The standard quantitative analysis consists of the selection of tumor regions-of-interest (ROIs) by a nuclear medicine physician on a pre- and posttreatment images. The maximum standardized uptake value (SUVmax) is calculated from the voxels contained within the ROI, and then the change in this metric is compared between the baseline and follow-up scans. The SUVmax is widely accepted as a simplified semiquantitative metric of tumor metabolism. The EORTC criteria² calculate SUVmax in selected tumor ROIs of the baseline and follow-up images and categorize the therapeutic response based on the qualitative and quantitative changes in follow-up images relative to baseline. The more recent PERCIST criteria use similar ROI analysis as well with alternative metrics and thresholds.³ Since the comparison is reduced to a single parameter for each ROI, information about the heterogeneity of tumor response and the spatial distribution of disease is lost. The process of identifying tumor ROIs is also time-consuming and subjective. Furthermore, it does not take

into account the variation in glucose metabolism in healthy brain tissue as a result of normal brain function and pharmacodynamic drug effects.^{4,5}

Voxelwise analysis, on the other hand, offers many advantages, such as providing information about the magnitude and spatial distribution of response everywhere in the brain, and it is more objective since the arbitrary delineation of ROIs is not required. A voxelwise analysis requires accurate registration between pre- and posttreatment images and performing voxelwise hypothesis tests on the difference image to localize statistically significant changes. Depending on whether these changes are positive or negative, these significantly changed voxels are identified as regions where the disease is progressing or responding to therapy, respectively. The voxelwise image analysis method, sometimes called statistical parametric mapping,^{6,7} has been widely used for investigating cognitive mechanisms,^{6,7} neurodegenerative diseases,^{8–10} and anatomical structure,¹¹ but has not been used to evaluate treatment effects on brain tumors. In this paper, we developed a voxelwise analysis algorithm for brain tumor therapeutic response assessment that accounts for background changes based on our previous method.¹²

The evaluation of metastatic brain tumors presents unique challenges in contrast to other voxelwise applications in the

*Address all correspondence to: Lei Qin, E-mail: lei_qin@dfci.harvard.edu

brain. First, the number, location, and size of metastatic lesions are unpredictable and unique to each patient. Thus, while the general anatomical brain structure can be gathered from population atlases, evaluation of voxelwise changes for a particular patient is limited to the two images (pre- and posttreatment) available for that patient. This limited amount of data poses a statistical challenge compared to studies where voxelwise population effects can be inferred from a sizable number of independent subjects. In addition, in the case of FDG-PET, the analysis is further confounded by a large background signal in healthy brain tissue since the brain derives nearly all of its energy from glucose metabolism.^{13,14}

The objective of the voxelwise analysis described above is to generate a map of voxelwise statistics that capture the difference in metabolism between the pre- and posttreatment scans that we are interested in, and which can be uniformly thresholded to determine statistically significant regions. For the inference to be valid, the distribution of the voxelwise statistics should be statistically homogeneous in the regions where there is no real change between the pre- and posttreatment scans. However, as observed in Ref. 12, normal metabolic activity may change heterogeneously over the brain, in particular differentially between the gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). Moreover, noise in PET images is dependent on the underlying brain metabolism resulting in the accumulation of FDG. While the noise in nuclear counting follows the Poisson distribution, there are additional confounding factors that impact the noise distribution of reconstructed PET images including the axial sensitivity profile of the PET scanner as well as processing and data corrections performed during iterative image reconstruction. As documented in this paper, the noise in PET images is not homogeneous in different axial slices along the scanner's z -axis, due to differences in both the brain anatomy and function as well as the scanner sensitivity.

For these reasons, the change in PET intensity at each voxel needs to be normalized by the background change for that voxel, that is, the expected change in metabolism for that voxel that is not associated with treatment effects or disease progression. We call this crucial process background adjustment; it ensures that voxel changes can be evaluated on the same basis throughout the brain. To achieve this, soft segmentation into GM, WM, and CSF tissues is used to estimate each of the backgrounds within each tissue type and within each slice. The specific background estimate for each voxel is obtained by compounding the tissue and slice specific background estimates according to each voxel's specific probabilities of belonging to each tissue, as given by the soft segmentation.

This is the general approach from the previous work.¹² However, our investigations show that the background adjustment and inference method used there are inaccurate. In order to provide a reference distribution for statistical inference, the previous method proposes manual selection of healthy slices. This paper shows that the reference distribution obtained that way is highly variable, and therefore, cannot be used reliably for inference. Moreover, the background adjustment method used there does not take into account the inhomogeneity of the brain metabolism and PET signal in different axial planes, which in itself contributes to the variability in the reference distribution. These effects are shown in both the phantom and patient data (Secs. 3.1 and 5). In the current work, we introduce two very important improvements to the methodology in Ref. 12 that solve the above problems. First, the background adjustment

is performed separately at each PET slice rather than the whole brain, thereby making the estimation of the background parameters adaptive to the metabolic background and noise variability in each transaxial image. Second, we remove the need to manually select presumably healthy slices. Instead, the inference relies on the theoretical assumption that, with proper background adjustment, the appropriate reference distribution is the standard normal distribution. We show that this assumption holds true empirically for our data. To ensure this, the estimation of the background is applied at each slice within an iterative algorithm that excludes the tail of the distribution, thereby reducing the estimation bias produced by the extreme localized changes that we wish to detect.

Both modifications to the method make the process more objective and efficient. In addition, they have as a result the convenient property that the marginal distribution of the voxelwise standardized PET intensity change values is very close to a standard normal distribution. As a consequence, the standard normal density can be used as a basis for choosing a formal threshold for statistical inference with a given false change detection rate. From a statistical point of view, being able to use the standard normal as a null distribution is remarkable given that the inference is based on two time points only.

The methods proposed in this paper are validated in three different ways. First, they are applied to a series of images acquired using a PET Hoffman brain phantom,¹⁵ obtained from experiments performed to simulate various effects such as change in position and treatment effects. Test/retest phantom experiments show that a false change detection rate of $\sim 0.4\%$ is achieved at a significance threshold of 3 standard deviations (SDs), while an intensity change of 25% in a simulated lesion can be reliably detected (Sec. 3). Second, the methods are applied to computer simulated images based on the experimental studies in order to cover a wider variety of conditions. It is shown that the proposed methodology can detect lesion response with 100% accuracy with tumor-to-background-ratio (TBR) as low as 1.5, and it is not affected by the background brain glucose metabolism change (Sec. 4). In addition, we applied the method to FDG-PET patient images from a clinical trial to assess treatment effects of lapatinib (Sec. 5). Both the simulation and patient data results were compared to the Guo et al.¹² method and to a simpler method involving a global correction of the background, showing the need to address slice heterogeneity. The results demonstrate the capability of the proposed method to highlight the tumor metabolism change after eliminating background variations.

2 Image Analysis Methods

The proposed methodology consists of the following elements: (1) image preprocessing, including calibration, and registration between baseline and follow-up scans; (2) image segmentation via mixture modeling; (3) background estimation and adjustment; (4) outlier correction; and (5) significance thresholding. The implementation is on MATLAB (Mathworks Inc., Natick, Massachusetts).

2.1 Image Preprocessing

The follow-up scan is spatially coregistered to match the baseline scan via affine transformation, using the "Mutual Information registration of Whole Brain 3D images (BRAINSFit)" module¹⁶ of the 3D Slicer software¹⁷ (Surgical Planning Laboratory, Brigham and Women's Hospital, Boston, Massachusetts).

After registration, the images are resampled into the same reference frame for further voxelwise analysis. But because resampling an image involves interpolation and therefore an implicit smoothing, it may change the voxel intensity distribution if only the follow-up image is resampled. Therefore, we resample both baseline and the registered follow-up PET images to the anatomical reference frame [baseline magnetic resonance images (MRI) were used as anatomical reference frame], so that both images go through the same resampling steps. Finally, the image intensity is calibrated by transforming it to standardized uptake values (SUVs) in order to adjust for the effects of the patient's size, sensitivity of the PET scanner, and the quantity of administered radiopharmaceutical (FDG).¹⁸ The necessary calibration factors are found in the image metadata.

2.2 Soft Image Segmentation

This step aims to separate the brain tissue into GM, WM, and CSF for background estimation, which is necessary because they may have different background mean levels and within-tissue variability. The segmentation was implemented using the software package Statistical Parametric Mapping (SPM5)^{6,19} (Wellcome Trust Centre for Neuroimaging, University College London), which is based on the expectation-maximization (EM) algorithm.²⁰ The SPM segmentation algorithm was applied using the prior probability maps of GM, WM, and CSF available based on PET scans. From SPM5, each voxel v is assigned a probability of belonging to the GM [$p_{GM}(v)$], WM [$p_{WM}(v)$], CSF [$p_{CSF}(v)$], or outside the brain [$p_{OB}(v)$]. These four probabilities add up to 1 at each voxel. In order to include only voxels within the brain, subsequent analysis was restricted to a binary brain mask, composed of voxels for which $p_{GM}(v) + p_{WM}(v) + p_{CSF}(v) > 0.5$. Binary brain masks were generated for both baseline and follow-up images from their soft segmentation maps. The final brain mask was obtained by performing a logical AND to the two brain masks, i.e., $\text{mask} = \text{mask}_1 \text{ and } \text{mask}_2$, where 1 and 2 denote baseline and follow-up scans. The three probabilities of belonging to the GM, WM, and CSF were then renormalized so they add up to 1 at each voxel within the mask.

2.3 Background Estimation and Adjustment

Due to the spatial inhomogeneity of brain metabolism and noise variance in different transaxial images, background parameters were estimated at each slice separately. For clarity, we first describe the background estimation approach ignoring the presence of lesions. The full iterative algorithm addressing the effect of lesions follows in Sec. 2.4.

Let $Y(v) = [Y_1(v), Y_2(v)]^T$ be a 2×1 column vector denoting the pair of measured SUVs for the baseline and follow-up images at voxel v . We assume each such pair is a single realization from a background distribution at voxel v with bivariate mean column vector

$$\mu(v) = E[Y(v)] = [\mu_1(v), \mu_2(v)]^T, \quad (1)$$

and covariance matrix

$$\Gamma(v) = \text{Cov}[Y(v)] = \begin{bmatrix} \Gamma_{11}(v) & \Gamma_{12}(v) \\ \Gamma_{21}(v) & \Gamma_{22}(v) \end{bmatrix}. \quad (2)$$

The mean and covariance constitute the background parameters for that voxel. Following Ref. 12, these parameters are given by the compounded mean

$$\mu(v) = \sum_s p_s(v) \mu_s, \quad (3)$$

and variance

$$\Gamma(v) = \sum_s p_s(v) \{ \Gamma_s + [\mu_s - \mu(v)][\mu_s - \mu(v)]^T \}, \quad (4)$$

where the index s in the sums takes values over the set (GM, WM, CSF). The compounding weights are the probabilities $p_{GM}(v)$, $p_{WM}(v)$, and $p_{CSF}(v)$ of voxel v belonging to each of the three tissue types, given by the soft segmentation of the baseline image. The tissue-specific background means μ_{GM} , μ_{WM} , μ_{CSF} and Γ_{GM} , Γ_{WM} , Γ_{CSF} are estimated by averaging over each tissue type via

$$\hat{\mu}_s = \frac{\sum_v p_s(v) Y(v)}{\sum_v p_s(v)}, \quad (5)$$

and

$$\hat{\Gamma}_s = \frac{\sum_v p_s(v) [Y(v) - \hat{\mu}_s][Y(v) - \hat{\mu}_s]^T}{\sum_v p_s(v)}, \quad (6)$$

where the sums are now over voxels. Plugging these estimates in Eqs. (3) and (4) results in estimated voxelwise background mean and covariance parameters $\hat{\mu}(v)$ and $\hat{\Gamma}(v)$.

The voxelwise background-adjusted intensity for both baseline and follow-up images is obtained by the standardizations

$$Z_1(v) = \frac{[Y_1(v) - \hat{\mu}_1(v)]}{\hat{\Gamma}_{11}(v)}$$

and

$$Z_2(v) = \frac{[Y_2(v) - \hat{\mu}_2(v)]}{\hat{\Gamma}_{22}(v)}, \quad (7)$$

which now have mean zero and SD one. Once these have been obtained, the background-adjusted metabolic change is measured by the standardized difference

$$\Delta(v) = \frac{[Z_1(v) - Z_2(v)]}{\sqrt{2[1 - \hat{\rho}(v)]}}, \quad (8)$$

where the denominator is the estimated SD of the numerator so that the normalized score $\Delta(v)$ has an SD of 1. Here, $\hat{\rho}(v)$ is the estimated correlation between $Z_1(v)$ and $Z_2(v)$:

$$\hat{\rho}(v) = \frac{\hat{\Gamma}_{12}(v)}{\sqrt{\hat{\Gamma}_{11}(v)\hat{\Gamma}_{22}(v)}}. \quad (9)$$

2.4 Outlier Correction

The estimation of the background parameters in Eqs. (5) and (6) over all voxels within the mask at each slice is based on the assumption that most voxels indeed represent the background. This assumption is true if the lesions in the brain are sparse and occupy a relatively small volume. However, lesions can affect the estimation in the slices where they exist. Thus, for the purposes of background estimation, lesions are considered outliers and should be excluded from the estimation. This is challenging

because identifying such lesions objectively without human intervention is the very goal of this paper.

To address this problem, we apply the background adjustment algorithm of Sec. 2.3 above within each slice iteratively, as follows:

1. Estimate the background parameters Eqs. (5) and (6) over all voxels within the mask and slice. Use these parameters to obtain the standardized differences $\Delta(v)$ in Eq. (8).
2. Re-estimate the background parameters Eqs. (5) and (6) over the voxels v within the mask and slice for which $|\Delta(v)| \leq c$, where c is a fixed constant.
3. Calculate the new standardized differences $\Delta(v)$ using Eq. (8) times the constant factor $\sqrt{1 - 2c\varphi(c)}/\sqrt{2\Phi(c) - 1}$, where φ and Φ are the standard normal density and cumulative distribution function, respectively.
4. Go back to step 2.

The restriction $|\Delta(v)| \leq c$ in step 2 above excludes outliers from the estimation of the background parameters. In this paper, we use $c = 2$, being the smallest threshold at which any voxel would be declared significant, and thus an outlier, using pointwise testing at the 0.05 level according to the normal distribution. This number can be adjusted (see Sec. 6). While the restriction in step 2 helps reduce the influence of outliers, it results in underestimation of the background variance. The factor in step 3 compensates so that the new standardized differences $\Delta(v)$ will have an SD of about 1. This factor is equal to the inverse of the SD of a standard normal variable conditioned on it having absolute value less than c and is based on the assumption that the background standardized differences $\Delta(v)$, as shown in Sec. 3, approximately follow a standard normal distribution. In the situations analyzed in this paper, it was found that three iterations were sufficient in the sense that, by the third iteration, the standardized differences $\Delta(v)$ did not change in value by more than 1.5% in at least 95% of the brain voxels. To ensure convergence, it was verified that the estimated values remained stable for at least 10 iterations. However, a smaller number of iterations are preferred for computational speed.

2.5 Significance Thresholding

As it will be shown in the phantom experiments, in the case of no change in lesions between the two images, the distribution of the standardized differences $\Delta(v)$ over the brain mask is very close to a standard normal distribution. Using the standard normal as the reference null distribution, a voxel v is considered to contain a statistically significant change if its absolute standardized difference score $|\Delta(v)|$ is higher than a significance threshold. Thanks to the background adjustment process, all voxels may be compared with the same reference null distribution and assessed with the same significance threshold. In Sec. 3, we use a significance threshold of ± 3 , targeting a false change detection rate of 0.27%.

3 Experiments and Results

Images were acquired on GE Discovery ST PET-CT scanner, which was used in three-dimensional (3-D) acquisition mode

and has transverse spatial resolution of 7-mm full-width half-maximum (FWHM) and axial spatial resolution of 6-mm FWHM. The Hoffman 3-D brain phantom¹⁵ was scanned to test the robustness and accuracy of the proposed algorithm. FDG radioactivity concentration in the Hoffman 3-D phantom was chosen to simulate a PET emission count rate within the GM of a typical brain study by assuming a 70-kg patient injected with ~ 10 mCi of FDG and imaged by a standard 15-min PET/CT scan of the brain acquired 60-min postinjection. In each experiment, the brain phantom was filled with activity concentrations of FDG to simulate GM mean SUV of 10 to 17 and TBR ranging from 1.0 to 2.0. PET scans of the brain phantom were acquired for 15 min per scan over a single bed position. The PET signal, of detected coincident events, was corrected for random events during acquisition, while model-based corrections for scattered events and photon attenuation were applied during the iterative reconstruction. Images were reconstructed with 3-D Fourier rebinning (3D FORE)²¹ followed by weighted least-squares ordered subset EM using 10 iterations and 35 subsets with a 4-mm noise reducing filter applied postreconstruction. To magnify the resolved details within images of the brain and to minimize reconstruction time, only the 30-cm central portion of the transverse FOV was reconstructed at 1.17-mm pixel spacing. The 47 transaxial slices of each PET image were reconstructed with a matrix size of 256×256 and a slice thickness of 3.27 mm. Low-dose CT scans were acquired prior to PET acquisition and used for attenuation correction of the PET data. This spatially matched CT was also used for image registration between repeated PET-CT scans of the phantom.

Three types of experiments were performed to evaluate the effect of noise on the robustness of the algorithm and to validate the modeling assumptions. A test-retest study was performed where the phantom was not moved between the two scans to represent optimal conditions with perfect spatial registration. A second test-retest study was performed by repositioning the phantom in between scans to include the process of spatial registration. In addition, a third experiment was conducted where a lesion was artificially added to the phantom and radioactive levels were changed between the two scans to evaluate the ability of the methodology to detect real localized changes.

3.1 "Test-Retest" Phantom Study with No Movement

Five sets of PET test-retest scans of the Hoffman 3-D brain phantom with various noise levels were analyzed (10 scans in total). The noise within images was modified by decreasing the acquisitions of the full-count images from 525 million coincident counts in an initial reference 15-min 3-D study, down to 263 M counts (1/2 counts), 132 M counts (1/4 counts), 52 M counts (1/10 counts), and finally 18 M counts (1/30 counts). The phantom position was kept exactly the same between all imaging acquisitions of this experiment and was not moved or refilled. The phantom was prepared with a TBR of 1.25 and the background SUV of 17 on delayed imaging at ~ 100 min. The activity concentration in the phantom background was $1.3 \mu\text{Ci/ml}$ (47 kBq/ml) at the start of the initial acquisition, but decayed to $0.7 \mu\text{Ci/ml}$ (29 kBq/ml) at the end of the final scan of the test-retest study. In order to obtain statistically comparable images, the PET acquisition was performed to achieve a predefined number of coincident events (counts). Radioactive decay was thus accounted for by increasing the acquisition times as the activity in the phantom decay over

Fig. 1 One slice of the 525 M count PET image of the Hoffman phantom. (a). Original PET image in SUV units. (b) Segmented GM probability map. (c) Brain mask generated from the tissue map.

the multiple acquisitions. Different image noise levels were obtained by acquiring 525 M counts, 263 M counts, down through 18 M counts.

Figure 1 shows the soft segmentation and generation of the brain mask. One slice from the 525 M count PET image in the SUV unit is shown in Fig. 1(a). Using SPM5, the PET image was segmented to WM, GM, and CSF probability maps. For illustration, Fig. 1(b) shows the GM map, where each voxel value indicates the probability of belonging to GM. A binary brain mask was generated based on the segmented probability map as described in Sec. 2.2 above. The generated binary brain mask, Fig. 1(c), dramatically reduced the number of voxels to be analyzed from $256 \times 256 \times 47 = 3,080,192$ voxels in the original PET matrix to only 380,456 voxels within the mask, an 88% reduction. The consequent analysis was based on the voxels inside the brain mask only.

While nuclear counting statistics follow the Poisson distribution, the resulting noise in reconstructed PET images is complex and is not strictly Poisson noise, since the count data are subtracted from, added to, and scaled before and during the iterative reconstruction processes. Variance in the acquired projection data is reduced during the iterative reconstruction process as well as during postreconstruction smoothing. The relative sensitivity at each slice of the PET image is demonstrated in Fig. 2(a). A uniform 20 cm $\times$ 20 cm cylinder phantom was scanned with the same acquisition parameters as the 525 million counts PET acquisition using the Hoffman brain phantom. For each slice, the mean of the signal and the SD of the background noise inside the cylinder were measured. Figure 2(a) shows the slice dependent noise, which is higher at the end of the 3-D volume due to fewer detected counts: while central planes have high sensitivity and count-rates, lower total counts are acquired at the end-planes. Note that simply removing the end slices, as is often done in PET studies, would not be sufficient because this would not remedy the variable axial sensitivity and noise profile in the remaining slices. More importantly, the mean and SD of the signal from the Hoffman phantom changes across slices [Figs. 2(b) and 2(c)], justifying the need to perform the background correction within each slice separately.

The test-retest experiment allowed us to evaluate the effect of background adjustment and the validity of the normal

distribution assumption. As shown in Fig. 2(d), the distribution of voxelwise changes $Y_1(v) - Y_2(v)$ between scan 1 and scan 2 of the 525 M counts acquisition is observed to be slice dependent (the distribution from the other noise levels scans is similar and is not shown here). The voxel-wise change distributions here and elsewhere in the paper were generated using the function `ksdensity` in MATLAB. This function estimates the probability density using a Gaussian kernel function with an automatically chosen bandwidth that is optimized for Gaussian densities.

Performing the proposed per-slice background adjustment described in Sec. 2.3 results in the displayed distributions of $\Delta(v)$ in Eq. (8) within each slice, which are close to the standard normal distribution [Figs. 2(e) and 2(f)]. The similarity with the standard normal is even greater when the distribution is taken over all slices (after per-slice background adjustment). Mismatches are observed more than three SDs away from the mean, where the number of voxels exhibiting such large changes is small. The standardized difference image between scan 1 and scan 2, Fig. 2(g), shows that most of the differences are in the range of ± 3 SDs. More importantly, it shows that the background adjustment process is successful in that it is able to remove all of the anatomical structure inside the mask, making the GM, WM, and CSF regions truly comparable in the same scale.

By thresholding the difference image, voxels with values beyond the threshold can be regarded as statistically significant changes. Because there are no true changes in this experiment, the percentage of detected voxels in the brain mask serves as an estimate of the false response volume accuracy (FRVA) as a function of threshold. As shown in Table 1, with increasing the threshold, the number of detected voxels as a percentage of the total brain volume decreases. It can be seen that the FRVA is close to that predicted by the standard normal distribution with mean = 0 and variance = 1, regardless of the signal count used in the PET acquisition. We emphasize that this analysis does not account for the spatial clustering of voxels, which could be used to further decrease FRVA. We emphasize again that the error rate is measured in terms of detected voxels and should not be confused with the global probability of detecting a lesion change.

Fig. 2 (a) Mean and SD measured in a uniform cylinder phantom, showing slice dependent noise. (b) and (c) Estimated mean and SDs measured inside the brain mask of the first scan in a test/retest experiment using the Hoffman phantom. WM, white matter; GM, gray matter; CSF, cerebro-spinal fluid. (d) Distribution of voxelwise changes between the first and second scans in the test/retest experiment for each slice (gray) and all slices (blue). Red curve is the standard normal density. Without background adjustment, it can be observed that the various slices exhibit different distributions. (e) Per-slice background adjustment brings the distributions within each slice (gray) closer to the standard normal (red), and more so for the distribution of all slices together (blue). The difference between the two curves is very small, as shown in logarithmic scale in (f). Mismatches are observed more than three SDs away from the center. (g) The standardized difference image between scan 1 and scan 2. The color bar shows the units of the display in SD.

3.2 “Test–Retest” Repeatability Phantom Study with Movement

A “coffee-break” experiment was conducted by changing the position of the phantom between sequential PET-CT scans in order to simulate a longitudinal brain study that is acquired in different positions. Two 15-min PET image acquisitions were acquired of the phantom. At the start of the first PET-CT scan, the phantom contained an activity concentration of

0.7 $\mu\text{Ci/ml}$ (26 kBq/ml), which decayed to 0.6 $\mu\text{Ci/ml}$ (20 kBq/ml) at the start of the second scan. This experiment was conducted to evaluate the combined effect of Poisson statistical noise due to repeated counting experiments, as well as errors in image registration and interpolation. Image registration was performed between the test and retest PET scans using the method described in Sec. 2.1. An MRI scan was performed using the phantom (images not shown) to mimic similar data

Table 1 Percentage voxels (%) with significant changes between scan 1 and scan 2 to the total volume of brain mask as a function of threshold.

Counts (million)	Threshold				
	± 1 SD	± 2 SD	± 3 SD	± 4 SD	± 5 SD
525	31.75	4.73	0.37	0.02	<0.01
262.5	31.72	4.73	0.37	0.02	<0.01
131.25	31.59	4.92	0.42	0.02	<0.01
52.5	31.49	4.76	0.38	0.02	<0.01
17.5	30.99	4.45	0.32	0.02	<0.01
Normal distribution	31.73	4.55	0.27	0.01	<0.01

sets as in the clinical experiment (Sec. 5) and after registration of PET to baseline MRI, both test and retest PET were resampled to the baseline MRI reference frame.

Figures 3(a) and 3(b) show the distribution of voxelwise changes before and after background adjustment. As in Fig. 2,

the background adjustment brings the distribution of the voxels over the brain very close to a normal distribution within each slice and over the entire brain. In contrast to the FRVA of 0.37% from the previous test/retest repeatability study without movement, setting the threshold to 3 SDs in this study resulted in a larger FRVA of 0.67%. Closer inspection of the standardized difference map [Fig. 3(c)], as well as the higher resolution CT images from scan 1 and scan 2, [Fig. 3(d)], revealed that the apparent “false” response detections were caused by air bubbles. These air bubbles were trapped between the polycarbonate layers of the Hoffman brain phantom during filling, which displaced the radioactive fluid and decreased uptake-intensities. Unlike the experiment in Sec. 3.1 where no movement was performed between the test–retest scans and air bubbles did not move, in this experiment, the air bubbles did move within the phantom between the two scans creating a true difference in regional activity. Comparing scan 2 to scan 1, one bubble became larger, which led to a decrease in PET SUV, while the other one became smaller and resulted in an increase in PET SUV. When the 2 air bubble regions reflecting real changes were excluded, the FRVA dropped to 0.43%, comparable to Sec. 3.1. The two air bubbles, though not intended, could be regarded as virtual lesions and demonstrated not just the

Fig. 3 (a) Distribution of voxelwise changes between the first and second scans in a test/retest experiment for each slice (gray) and all slices (blue). Red curve is the standard normal density. Without background adjustment, it can be observed that the various slices exhibit different distributions. (b) Per-slice background adjustment brings the distributions within each slice (gray) closer to the standard normal (red), and more so for the distribution of all slices together (blue). (c) The standardized difference image between scan 1 and scan 2. The color bar shows the units of the display in SDs. (d) The corresponding CT image from scan 1. Two air bubbles can be clearly seen. The red square in the full brain CT images was amplified and showed on the right of the image, with scans 1 and 2 displayed individually. The size of the bubble on the left of the image increased, leading to a decrease of PET SUV. On the contrary, the bubble on the right decreased in size, leading to an increase of PET SUV. L, left; R, right.

robustness of the algorithm, but also the sensitivity in lesion change detection.

3.3 Lesion Change Detection Study

A malignant lesion was simulated within the central GM at a location superior and anterior within the brain, by placing a 1.5-cm diameter sphere of FDG. The TBR of activity concentrations of the lesion to GM used for these experiments was 2:1 and 1.5:1. Due to the physical construction of the phantom, the two different TBR levels were achieved by increasing the activity in the phantom background, i.e., injecting more radio-tracer to the background, while keeping the activity concentration in the lesion constant. Specifically, after scan 1 at a TBR of 2:1, the phantom was taken out from the scanner and more FDG was injected into the background compartment of the phantom to make TBR 1.5:1. For the 2:1 TBR experiments, the activity concentration in the lesion was 3.0 $\mu\text{Ci/ml}$ (111 kBq/ml) at the time of filling the lesion compartment, whereas 1.5 $\mu\text{Ci/ml}$ (54 kBq/ml) was added to the background. The activity concentration in the background was then increased to 2.0 $\mu\text{Ci/ml}$ (74 kBq/ml) for a TBR of 1.5. At the time of acquisition, background activity concentrations were 1.2 $\mu\text{Ci/ml}$ (43 kBq/ml) and 1.3 $\mu\text{Ci/ml}$ (46 kBq/ml) for the 2:1 and 1.5:1 TBR experiments, respectively. To meet the TBR objectives for this experiment, all activity assays and activity concentration calculations were corrected for the radioactive decay. A single reference time was used for the calculations, which corresponded to the time that the activity in the lesion compartment of the phantom was assayed. Image registration was performed between the two scans.

Figures 4(a) and 4(b) show one slice of the registered scan 1 and scan 2 images. The lesion is seen in the anterior region within the brain. Figure 4(c) shows the difference image between registered scan 1 and scan 2, in which no lesion difference is obvious, but only background. If the conventional ROI-based SUVmax were used as a biometric, no changes would be detected between the two scans. Using the proposed algorithm, the background adjusted images are shown in Figs. 4(d) and 4(e) where background differences are eliminated and the lesion appears to be going down in intensity with respect to the background, as intended by the experiment. The standardized difference image with significance threshold of 3 SDs, Fig. 4(f), clearly shows the lesion, but also some false response voxels. By setting the significance threshold as 4 SDs, those false response voxels disappear [Fig. 4(g)]. Note that Figs. 4(f) and 4(g) are shown with the same color scale to allow a direct comparison. Figures 4(h) and 4(i) show the distribution of voxelwise changes between scan 1 and scan 2 for each slice and all slices averaged, before and after background adjustment, respectively. The background adjusted distribution of the relative changes matched nicely with the standard normal, except at the left tail, where the simulated tumor signal is.

4 Computer Simulations

The variation of glucose uptake measured with FDG-PET in healthy brain tissue can result in changes in normal brain background SUV levels between baseline and follow-up scans. Computer simulations were performed to evaluate the performance of the algorithm in detecting change in lesions under more general and extreme cases of changes in background than those available in the phantom experiments. Note that the simulations were not intended to model all physical factors

and reconstruction parameters of a specific PET scanner. Rather, the goal was to generate objects that reasonably represent healthy brain structure and tumors while using count rates that would represent the typical clinical scenario. The digital Hoffman brain phantom was used to provide a realistic model of normal brain anatomy that has been segmented into gray matter, WM, and ventricles. Geometric objects were superimposed on the normal anatomic structures to represent lesions. Various background and tumor changes were simulated to test the lesion response detection accuracy based on the 525 million counts PET brain phantom image pair acquired in Sec. 3.1. Scan 1 of the 525 million counts PET image was not modified and was representative of normal anatomy in the absence of disease, whereas the background level for scan 2 was changed by multiplication of a new background level and a simulated hypermetabolic lesion was inserted into the volume of scan 2 at various TBR values with respect to the new background level. This allowed the evaluation of a progressing tumor in the presence of varying normal background metabolism (modified scan 2) compared to a normal baseline (scan 1). Two types of background change were simulated, one with a homogeneous global change, and the other with an inhomogeneous change.

The proposed algorithm was applied on scan 1 and modified scan 2 images to test the capability of detecting the simulated lesion. To compare with the proposed algorithm, a global normalization was also performed where scan 2 was scaled by the ratio of the average intensities of scan 2 and scan 1 in the brain mask, i.e., $Y_2^G(v) = Y_2(v)/E[Y_2(v)] * E[Y_1(v)]$, and then the difference was taken between the global corrected image $\Delta^G(v) = Y_2^G(v) - Y_1(v)$ normalized to the SD of $\Delta^G(v)$ (i.e., $\Delta^G(v)/\sigma[\Delta^G(v)]$).

Our prior algorithm proposed by Guo et al.¹² was implemented as well for comparison (referred to as the method of Guo et al.). According to that method, background parameters were estimated from healthy slices selected manually, and then applied to the entire volume for background adjustment. Significance thresholds were used to evaluate the true response volume accuracy (TRVA), defined as the number of detected true tumor voxels divided by the simulated entire tumor volume and FRVA.

4.1 Homogeneous Background Change

Global background change was simulated with a global multiplication constant ranging from 1 to 3. This provides a range of variability in physiologic uptake as well as scanner calibration errors that would render conventional SUV analysis useless in the absence of a valid reference tissue. A $12 \times 9 \times 2$ cuboid lesion (total volume 216 voxels) was simulated and inserted to slice 25 and 26 of the 525 million count scan 2 PET image. The TBR of the lesion was simulated to be between 1.2 and 3.

Figures 5(a) and 5(b) show slice 26 of the PET images, where the background amplitude of scan 2 was twice that of scan 1, and TBR was 1.5. While the lesion is barely visible in panel (b), it is visible in the difference image, panel (c), on the right side of WM. However, the background change is also obvious in the difference image. Panels (d)–(f) show the global corrected images and the difference after global correction. Panels (h) and (i) show the background adjusted images using the proposed method from scans 1 and 2. Compared to (a) and (b), the background difference is clearly removed in (d) and (e), and (h) and (i). The difference images (g) and (k) detect the change between the two scans with threshold set at 3 SDs for both methods.

Fig. 4 Phantom experiment simulating a lesion and a change in the background. (a) One slice from the original scan 1. (b) The corresponding slice from registered scan 2. (c) Difference image of the same slice. (d) Background adjusted image from scan 1. (e) Background adjusted image from scan 2. (f) and (g) Standardized difference image from background adjusted scan 1 and 2, when 3 and 4 SDs are used as threshold respectively. (h) and (i) Distribution of voxelwise changes between scan 1 and scan 2 for each slice (gray) and all slices (blue) without and with background adjustment. Red curve is the standard normal density. Background adjustment eliminates the axial variability between slices.

The TRVA is 100% for both global correction and our proposed methods. There appear to be more false positive voxels in Fig. 5(k) with the proposed algorithm compared to Fig. 5(g) with the global correction method. However, at the top and bottom slices of the phantom, a large number of false positive voxels were produced using the global correction method, Fig. 6(a), as compared to our proposed method, Fig. 6(b). This demonstrates the advantage of using slice-by-slice background

correction in our method, achieving a homogeneous error rate and a better overall accuracy (see also Fig. 8). Figure 7(a) shows the volume distribution of the difference image for our method and the global correction method compared with the normal distribution. Our method shows better agreement with normal distribution.

The next set of simulations investigates the effect of TBR in scan 2, while keeping the background change as a factor of 2.

Fig. 5 Computer simulation of a lesion with homogeneous background change. (a) Slice 26 from the original scan 1. (b) The corresponding slice of simulated scan 2, with background changed to twice the original amplitude, and an inserted lesion with TBR 1.5. (c) Difference image of the same slice, showing both the change of background and the lesion. (d) and (e) Global background corrected images, (f) the difference between them, and (g) with a threshold set at 3 SDs. (h) and (i) Background adjusted image from scan 1 and scan 2. (j) Standardized difference image between background-adjusted scan 1 and 2, (k) with a threshold set at 3 SDs. (f), (g), (j), (k) use the same range to display the difference between the two methods.

Because the lesion is only added in scan 2, increasing TBR in scan 2 leads to a higher intensity change in the lesion region with respect to scan 1 and should be easier to detect. Figure 8(a) shows the TRVA as a function of TBR from 1.2 to 3. As expected, the higher the TBR, the higher the TRVA, reaching 100% once the TBR is greater than 1.5. In comparison with the Guo et al. method, even though TRVA is similar, the FRVA is much lower with the new method. At threshold 3 SDs, the FRVA for the Guo et al. method is 1.84%, which is ~5 times higher than the new method. The global correction method provides a lower TRVA and a higher FRVA compared to the new proposed method. The Guo et al. method and the proposed method have a constant FRVA regardless of TBR, while the global correction exhibits a decreasing FRVA as

Fig. 6 (a) Slice 1 of the phantom thresholded at 3 SDs using global correction method and (b) the proposed method.

Fig. 7 Distribution of the difference image for (a) homogeneous and (b) inhomogeneous background change simulations, using global corrected (black) and proposed (blue) method; normal distribution is in red.

Fig. 8 Tumor response volume accuracy simulation for homogeneous background change. (a) TRVA as a function of TBR. When TBR increases, tumor detection accuracy increases. (b) FRVA as a function of TBR is constant regardless of TBR. Black is the current method, red is the Guo et al. method, blue is the global correction method. Threshold was set at 3 SDs for all three methods.

a function of TBR. This is because in the Guo et al. and the proposed method, by design, the estimated background SD depends only on the background noise. In the global correction method, on the other hand, it increases because it is biased by the effect of TBR on the global background.

As mentioned above, a set of simulations was also performed to evaluate the effect of background change for a fixed TBR of 1.5 in scan 2. Changing the background of scan 2 did not produce any difference in TRVA or FRVA, even when the background was multiplied by a factor of three (data not shown here because they are similar to Fig. 8).

4.2 Inhomogeneous Background Change

Background change was simulated with a different factor for GM and WM, respectively, to reflect that some drugs may introduce different levels of change in GM and WM. Here, the WM intensity was not changed, but GM intensity was multiplied by a factor of 1.5. The same simulated cuboid lesion as in the above simulation section was inserted to slice 25 and 26 of the 525 million counts scan 2 PET image as well. The TBR of the lesion was simulated to be between 1.2 and 3.

Figures 9(a) and 9(b) show slice 26 of the PET images, where the GM amplitude of scan 2 was 1.5 times that of scan 1, WM amplitude was the same for both scans, and TBR was 1.5.

The difference image in Fig. 9(c) clearly shows the GM and the lesion on the right side of WM. Figures 9(d) and 9(e) show that global background adjustment fails to remove the differences in background between scan 1 and scan 2, which remain in the difference image (f). The thresholded image (g) fails to detect the simulated lesion. The differences in background, however, are removed using our proposed method in (h)–(j). When thresholding was applied at 3 SDs, as shown in (k), the lesion was identified. Figure 7(b) shows that the distribution of the difference image for our method is close to a standard normal, while the distribution for the global correction is far from it.

Figure 10 shows the TRVA and FRVA when TBR was simulated to change from 1.2 to 3, with WM and GM change similar to the above simulation as shown in Fig 9. Three SDs were used for thresholding. We can see in Fig 10(a) that when TBR was 1.2 or 1.3, almost no tumor response voxels could be detected. But if TBR increased to 1.5, as the case shown in Fig 9, 93.1% of tumor voxels could be detected using the proposed method, compared to a slightly lower TRVA for the method of Guo et al. and zero detection of the tumor voxels using the global correction method. The FRVA for the Guo et al. method is 0.46%, which is ~ 3 times higher than the new method. As in Fig. 6, even though TRVA is similar, the FRVA is much lower with the new method. The global correction method fails to

Fig. 9 Computer simulation of a lesion with inhomogeneous background change. (a) Slice 26 from the original scan 1. (b) The corresponding slice of simulated scan 2, with GM changed to 1.5 times the original amplitude, no change in WM and an inserted lesion with TBR 1.5. (c) Difference image of the same slice, showing both the change of GM and the lesion. (d) and (e) Global-background corrected images, (f) the difference between them, which shows a large GM change, and (g) thresholded at 3 SDs. (h) and (i) Background adjusted image from scan 1 and scan 2. (j) Standardized difference image between background-adjusted scan 1 and scan 2, (k) thresholded at 3 SDs. (f), (g), (j), (k) uses the same range to display the difference between the two methods.

detect any tumor voxels when TBR is less than 2 and has a much higher FRVA.

5 Clinical Application

The clinical data are from a Dana-Farber/Harvard Cancer Center clinical trial, conducted to assess the treatment effect of Lapatinib on brain metastases for patients with human epidermal growth factor receptor 2 (HER-2) positive breast cancer.²² The original clinical trial was approved by the institutional review board and our secondary analysis was approved with a waiver of informed consent. PET scans were acquired prior to (baseline) and four weeks after (follow-up) treatment initialization with standardized acquisition parameters conforming with the NCI consensus guidelines for FDG-PET in clinical trials.²³ MRI scans were acquired at baseline as well. PET images were spatially coregistered to MRI and resampled using the MRI reference frame.

Figures 11 and 12 show two patients' cases. For both cases, original baseline and follow-up PET images (a)–(c) show an obvious background change. Therefore, using the conventional ROI method for PET analysis to calculate SUVmax would be biased. Global normalization (d)–(f) and the proposed method (h)–(j) are able to adjust the background and eliminate background variation. However, the difference image (g), (k) after thresholding at 3 SDs, shows that the proposed method better highlights the metabolism changes for the two patients. In the proposed method, the slice dependent variability of voxelwise changes between baseline and follow-up scans (l) is removed after background adjustment and the resulting distribution is close to a standard normal distribution (m) just as the phantom experiments demonstrated.

For comparison, the method of Guo et al. was also applied to one of the patients. Figure 13 shows that the standardized difference maps are influenced by the selection of healthy

Fig. 10 Tumor response detection accuracy simulation for inhomogeneous background change. (a) TRVA as a function of TBR. When TBR increases, tumor detection accuracy increases. (b) FRVA as a function of TBR is constant regardless of TBR. Black is the current method, red is the Guo et al. method. Threshold was set at 3 SDs for all three methods.

slices (a)–(c). The reason is the inhomogeneous background mean and SD among slices (d)–(e), again justifying the need to perform background adjustment slice by slice. Note that, in comparison to Figs. 2(a)–2(c), clinical data show more variation between slices than the Hoffman phantom.

In comparison with MRI, it is observed that in the two cases, areas of abnormality on MRI did not correspond to increased metabolism with FDG-PET and conversely areas of presumed tumor growth with hypermetabolic lesions on FDG-PET appeared normal on MRI. This discordance is consistent with the known differences between the anatomical and functional imaging, particularly with previously treated lesions.

6 Discussion

In this paper, we have proposed an algorithm to detect the local changes in brain lesion(s) of a single subject from a pre- to a posttreatment image at the voxel level. The proposed method accounts for global changes in brain metabolism, in order to identify the regional changes in tumor response. Compared to the standard manual ROI analysis method used in clinical trials, this approach provides an objective assessment of changes in metabolism at each voxel.

The Hoffman brain phantom experiments showed a slice-dependent voxel change distribution. Therefore, the background adjustment was based and applied on each slice. The background of the standardized difference image was found to closely follow the standard normal distribution in both the phantom and clinical datasets.

The choice of applying the background adjustment at each slice separately instead of over the entire brain proved crucial in addressing the noise heterogeneity over slices. However, it still assumes that the background is homogeneous within each tissue type at each slice. This represents a trade-off between bias and variance. While the background adjustment could be applied to smaller regions (e.g., left and right hemispheres within each slice) thereby reducing bias, doing so would reduce the number of voxels available for estimation of the background parameters, thereby increasing variance. One of the limitations of the previous work (Guo et al.¹²) is that the number of voxels within the mask in a typical slice (about 10,000 in a 256×256 slice) is not enough to accurately estimate the intensity distribution nonparametrically. However, it is enough to estimate the

background parameters, as it is done in this paper, solving the problem.

On the other hand, to estimate the background parameters from all the voxels in each slice requires the exclusion of tumor voxels to reduce bias. In Sec. 2.4, we introduced an iterative procedure to include only nonoutliers in the estimation and readjust the variance. Nonoutliers were defined as voxels satisfying $|\Delta(v)| \leq c$. A larger value of c includes more background voxels and includes more outliers, whereas a smaller value of c does the opposite. In our analysis of both the phantom and clinical data, a voxelwise detection threshold of 3 was used to find significant changes, effectively defining voxels with $|\Delta(v)| \geq 3$ as outliers. The choice $c = 2$ used in this paper is conservatively low enough to exclude most outliers but large enough that it allows estimation of the background parameters using $2\Phi(c) - 1 = 95\%$ of the background voxels.

The computer simulation of homogeneous versus inhomogeneous background change revealed that a global background adjustment is not sufficient to adjust for inhomogeneous changes in background. Compared to our previous method (Guo et al.), the TRVA was similar, but FRVA was 3 to 5 times (Figs. 8 and 10) lower. This is because, by estimating the background within each slice, the method adapts to the inhomogeneity between slices in noise levels and background change, and is, therefore, more accurate. This was demonstrated in Fig. 6 as well. When not considering slice-to-slice differences, global correction fails in the top and bottom slices. It may also be noticed that the inhomogeneous background change produces a smaller FRVA than the homogeneous background change. This is the result of an overestimation of the background variance, which leads to a narrower distribution of the standardized difference scores with respect to the standard normal [as can be seen in Fig. 7(b)] and thus less probability in the tails. This overestimation of the variance is the result of bias incurred in the transition regions between the tissue types, which is exacerbated when there is a larger difference between tissues. Fortunately, this leads to a reduction rather than increase in the false positive rate, with almost no loss in sensitivity, so the inference is still valid.

When the algorithm was applied to clinical data, the number of detected “tumor change” voxels was different when comparing the global method and our proposed method. The position of

Fig. 11 Patient treated with lapatinib. (a) and (b) One slice from baseline and coregistered follow-up PET, and the difference image (c). (d) and (e) Global normalization of the baseline and follow-up PET images. (f) Normalized difference image after global correction, and (g) with 3 SDs used as threshold overlaid on MRI. (h) and (i) Standardized background-adjusted image from baseline and follow-up images. (j) Standardized difference image from background adjusted baseline and follow-up, (k) with 3 SDs used as threshold overlaid on MRI. Distribution of voxelwise changes between baseline and follow-up for each slice (gray) and all slices (blue) (l) without and (m) with background adjustment. Red curve is the standard normal density. Background adjustment eliminates the axial variability between slices. (n) Comparison of distribution between global normalization (black) and proposed normalization method (blue) and normal distribution (red).

those voxels seems to be in accordance with MRI contrast enhanced images. However, since contrast enhanced MRI is not specific in showing tumor,²⁴ we cannot conclude if the detected voxels are real tumor changes or not. Because we do not have a ground truth in clinical cases (i.e., to know which voxel is real tumor and whether it is responding to treatment or

progressing), we were not able to determine TRVA as in the phantom experiments and computer simulations. As another indication of the quality of the background adjustment, we can see in the distribution of the background corrected difference image that the proposed method agreed much better with normal distribution than the global correction method.

Fig. 12 Patient treated with lapatinib. (a) and (b) One slice from baseline and coregistered follow-up PET, and the difference image (c). (d) and (e) Global normalization of the baseline and follow-up PET images. (f) Normalized difference image after global correction, and (g) with 3 SDs used as threshold overlaid on MRI. (h) and (i) Standardized background-adjusted image from baseline and follow-up images. (j) Standardized difference image from background adjusted baseline and follow-up, and with (k) 3 SDs used as threshold overlaid on MRI. Distribution of voxelwise changes between baseline and follow-up for each slice (gray) and all slices (blue) (l) without and (m) with background adjustment. Red curve is the standard normal density. Background adjustment eliminates the axial variability between slices. (n) Comparison of distribution between global normalization (black) and proposed normalization method (blue) and normal distribution (red).

The current algorithm was applied to FDG-PET acquisitions on the same scanner using the same imaging protocol. The current consensus in clinical trials using FDG-PET is that it is not appropriate to compare two scans (pre and post) from different scanners.^{23,25} Hence, prospective clinical trials standardize

the acquisition of PET scans to use the same scanner and reconstruction parameters for all the time points of a given subject. However, there are ongoing efforts to harmonize image reconstruction methods across different PET scanners and vendors. If these approaches succeed, our methods for analyzing

Fig. 13 Healthy slice selection in method of Guo results in unstable standardized difference image (3 SDs were used as threshold): (a) slices #37 to #42 used as healthy slices; (b) slices #40 to #42 used as healthy slices; (c) slice #42 used as healthy slice. (d) and (e) Estimated mean and SD inside the brain mask of PET brain baseline and follow-up images. Solid and dashed lines are baseline and follow-up scans, respectively.

differences across different time points would have broader applicability in not requiring the same scanner as may be the case in routine clinical practice.

One limitation of the current implementation is that the segmentation was performed using a relatively old version of SPM. A newer version SPM12, with improved segmentation, might be able to improve the accuracy of our algorithm, and should be incorporated in future development. Another limitation of is that only the segmentation map from the baseline reference image was used for estimation of the background parameters. Thus, the results are to be interpreted as changes with respect to the background as defined by the baseline image. The follow-up image may have a slightly different segmentation. Taking the segmentation of the follow-up image into account may help improve the accuracy, but it requires developing a new algorithm that can combine both segmentations into the bivariate background estimation simultaneously. This is a research direction left for future work. While the method has been shown using glucose as a radiotracer, it may or may not be applicable to other radiotracers depending on whether they also exhibit global, inhomogeneous, and local changes in intensity as does glucose.

The rigorous and objective statistical approach presented in this paper provides an objective quantitative approach to complement the qualitative clinical interpretation of FDG-PET that is typically performed. Unlike the standard semiquantitative

approach utilizing the SUVmax, this approach has the ability to identify and characterize the known physical phenomena of tumor heterogeneity, differential treatment response, and changes in normal tissue background. While the ability to identify statistical heterogeneity in metabolic changes following treatment has been demonstrated, further validation and clinical trials are needed to establish whether these changes truly reflect tumor response/progression and can be used to predictive drug efficacy and/or clinical outcome. These methods are not intended to be used immediately as a clinical tool or make treatment decisions. Rather, it is our intention that the objective lesion change detection method presented in this paper will provide additional information that may help assess the treatment response of novel therapeutics in future clinical trials. While the phantom and simulation studies demonstrate the basic principles of this approach, further refinement and testing is needed using prospective clinical trials to evaluation the performance of this methodology under the diverse and complex scenarios due to patient heterogeneity. There are many other PET radiopharmaceuticals (^{11}C -methionine, ^{18}F -FET, ^{18}F -FLT, etc.) and imaging modalities (dynamic contrast enhanced MRI, diffusion-weighted MRI, etc.) being investigated to measured therapeutic response. The approach presented in this paper could be applied to any of these provided they result in comparable quantitative longitudinal measurements at the voxel level.

Disclosures

The authors have no relevant conflicts of interest to disclose.

Acknowledgments

This work was partially supported by the NIH Grant R21 EB013795.

References

1. M. Reivich et al., "The [^{18}F]fluorodeoxyglucose method for the measurement of local cerebral glucose utilization in man," *Circ. Res.* **44**(1), 127–137 (1979).
2. H. Young et al., "Measurement of clinical and subclinical tumour response using [^{18}F]fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. European Organization for Research and Treatment of Cancer (EORTC) PET study group," *Eur. J. Cancer* **35**(13), 1773–1782 (1999).
3. R. L. Wahl et al., "From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors," *J. Nucl. Med.* **50**(Suppl 1), 122S–150S (2009).
4. L. Mosconi et al., "FDG-PET changes in brain glucose metabolism from normal cognition to pathologically verified Alzheimer's disease," *Eur. J. Nucl. Med. Mol. Imaging* **36**(5), 811–822 (2009).
5. A. G. Vlassenko, M. M. Rundle, and M. A. Mintun, "Human brain glucose metabolism may evolve during activation: findings from a modified FDG PET paradigm," *NeuroImage* **33**(4), 1036–1041 (2006).
6. K. J. Friston et al., "Statistical parametric maps in functional imaging: a general linear approach," *Hum. Brain Mapp.* **2**(4), 189–210 (1994).
7. K. J. Worsley et al., "A unified statistical approach for determining significant signals in images of cerebral activation," *Hum. Brain Mapp.* **4**(1), 58–73 (1996).
8. R. Bammer et al., "Magnetic resonance diffusion tensor imaging for characterizing diffuse and focal white matter abnormalities in multiple sclerosis," *Magn. Reson. Med.* **44**(4), 583–591 (2000).
9. N. M. Kemppainen et al., "Voxel-based analysis of PET amyloid ligand [^{11}C]PIB uptake in Alzheimer disease," *Neurology* **67**(9), 1575–1580 (2006).
10. M. G. Naylor et al., "Voxelwise multivariate analysis of multimodality magnetic resonance imaging," *Hum. Brain Mapp.* **35**(3), 831–846 (2014).
11. A. Schwartzman, R. F. Dougherty, and J. E. Taylor, "Cross-subject comparison of principal diffusion direction maps," *Magn. Reson. Med.* **53**(6), 1423–1431 (2005).
12. M. Guo et al., "Voxelwise single-subject analysis of imaging metabolic response to therapy in neuro-oncology," *Stat.* **3**(1), 172–186 (2014).
13. J. W. Kim and C. V. Dang, "Cancer's molecular sweet tooth and the Warburg effect," *Cancer Res.* **66**(18), 8927–8930 (2006).
14. O. Warburg, "On the origin of cancer cells," *Science* **123**(3191), 309–314 (1956).
15. E. J. Hoffman et al., "Assessment of accuracy of PET utilizing a 3-D phantom to simulate the activity distribution of [^{18}F]fluorodeoxyglucose uptake in the human brain," *J. Cereb. Blood Flow Metab.* **11**(S1), A17–A25 (1991).
16. H. J. Johnson, G. Harris, and K. Williams, "BRAINSFit: mutual information registrations of whole-brain 3D images, using the insight toolkit," *Insight J.* (2007).
17. S. Pieper et al., "The NA-MIC Kit: ITK, VTK, pipelines, grids and 3D slicer as an open platform for the medical image computing community," in *Proc. IEEE Int. Symp. Biomed. Imaging*, pp. 698–701 (2006).
18. Y. Sugawara et al., "Reevaluation of the standardized uptake value for FDG: variations with body weight and methods for correction," *Radiology* **213**(2), 521–525 (1999).
19. J. Ashburner, "SPM: a history," *NeuroImage* **62**(2), 791–800 (2012).
20. J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage* **26**(3), 839–851 (2005).
21. O. Mawlawi et al., "Performance characteristics of a newly developed PET/CT scanner using NEMA standards in 2D and 3D modes," *J. Nucl. Med.* **45**(10), 1734–1742 (2004).
22. N. U. Lin et al., "Phase II trial of lapatinib for brain metastases in patients with human epidermal growth factor receptor 2-positive breast cancer," *J. Clin. Oncol.* **26**(12), 1993–1999 (2008).
23. L. K. Shankar et al., "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.* **47**(6), 1059–1066 (2006).
24. L. C. H. da Cruz et al., "Pseudoprogression and pseudoresponse: imaging challenges in the assessment of posttreatment Glioma," *Am. J. Neuroradiol.* **32**(11), 1978–1985 (2011).
25. M. M. Graham et al., "Summary of the UPICT protocol for 18F-FDG PET/CT imaging in oncology clinical trials," *J. Nucl. Med.* **56**(6), 955–961 (2015).

Lei Qin received her PhD in biological resource engineering from the University of Maryland, and completed her postdoctoral research fellowship at Brigham and Women's Hospital, Harvard Medical School, Boston. She is currently a medical physicist in the Department of Imaging at Dana-Farber Cancer Institute and an instructor in the Department of Radiology at Harvard Medical School. Her research interests are in the development of quantitative imaging biomarkers for cancer treatment assessment.

Armin Schwartzman received his BS and MS degrees in electrical engineering from the Technion – Israel Institute of Technology and the California Institute of Technology, and his PhD in statistics from Stanford University. He is currently an associate professor of biostatistics at the University of California, San Diego. His research interests are in signal and image analysis, with applications to biomedicine and the environment.

Keisha McCall received her PhD in medical physics from the University of Wisconsin in Madison and completed her postdoctoral research fellowship at Memorial Sloan-Kettering Cancer Center, New York. She is the nuclear medicine/PET-CT physicist in the Department of Imaging at Dana-Farber Cancer Institute, Boston and is an instructor in radiology at Harvard Medical School. Her research focus is medical imaging and quantitative image analysis in nuclear medicine and oncology.

Nezamoddin N. Kachouie received his BS and MS degrees in electrical and computer engineering. He received his PhD in systems design engineering from University of Waterloo, Canada. He was a research fellow with Harvard Medical School, Harvard-MIT Health Sciences and Technology, and Harvard School of Public Health. He is currently an assistant professor of mathematical sciences at Florida Institute of Technology. His research interests are statistical modeling and machine learning with applications to public health.

Jeffrey T. Yap received his BS degree in physics from the University of Dallas and his PhD in medical physics from the University of Chicago. He is currently a research associate professor in the Department of Radiology and Imaging Sciences at the University of Utah. He is also the codirector of the Center for Quantitative Cancer Imaging at the Huntsman Cancer Institute. His research focus is on the development of multimodality imaging in translational cancer research.