Published in final edited form as:

Stat. 2014 January 1; 3(1): 172–186. doi:10.1002/sta4.53.

Voxelwise single-subject analysis of imaging metabolic response to therapy in neuro-oncology

Mengye Guo^{a,*}, Jeffrey T. Yap^b, Annick D. Van den Abbeele^c, Nancy U. Lin^d, and Armin Schwartzman^e

^aDepartment of Biostatistics, University of Washington, 6200 NE 74th St, Seattle, WA, 98115

^bCenter for Quantitative Cancer Imaging, University of Utah, 2000 Circle of Hope, Salt Lake City, UT 84112

^cDepartment of Imaging, Dana-Farber/Harvard Cancer Center, 450 Brookline Avenue, DL101, Boston, MA 02215

^dDepartment of Medical Oncology, Dana-Farber/Harvard Cancer Center, 450 Brookline Avenue, Boston, MA 02215

^eDepartment of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, NC 27615

Abstract

F-18-Fluorodeoxyglucose positron emission tomography (FDG-PET) has been used to evaluate the metabolic response of metastatic brain tumors to treatment by comparing their tumor glucose metabolism before and after treatment. The standard analysis based on regions-of-interest has the advantage of simplicity. However, it is by definition restricted to those regions and is subject to observer variability. In addition, the observed changes in tumor metabolism are often confounded by normal changes in the tissue background, which can be heterogenous. We propose an analysis pipeline for automatically detecting the change at each voxel in the entire brain of a single subject, while adjusting for changes in the background. The complete analysis includes image registration, segmentation, a hierarchical model for background adjustment and voxelwise statistical comparisons. We demonstrate the method's ability to identify areas of tumor response and/or progression in two subjects enrolled in a clinical trial using FDG-PET to evaluate lapatinib for the treatment of brain metastases in breast cancer patients.

Keywords

Statistical parametric mapping; mixture distribution; multiple testing; false discovery rate

1. Introduction

The development of metastatic brain tumors in cancer patients is a serious condition that most often results in a very poor prognosis for these patients. For example, 10-16% of women with metastatic breast cancer develop brain metastasis and have a median survival of only 5-7 weeks if left untreated (Patanaphan et al., 1988). The severity of the condition demands tools that can both identify brain metastases and indicate response to therapy early in order to guide treatment. FDG-PET is a medical imaging technique that uses positron emission tomography (PET) imaging to measure the rate of glucose metabolism in tissue following intravenous administration of the glucose analog F-18-Fluorodeoxyglucose (FDG). Because cancer cells actively consume glucose (Warburg, 1956; Kim & Dang, 2006), functional changes in tumor glucose metabolism as measured by FDG-PET can be detected earlier than the structural changes in tumor size and shape, typically observed via X-ray Computed tomography (CT) or Magnetic Resonance Imaging (MRI). For this reason, FDG-PET has been used as an early predictor of therapeutic response. FDG-PET has shown a reduction in tumor metabolism as early as one day after the initiation of therapy, and this positive metabolic response has been shown to correlate with survival (Van den Abbeele & Badawi, 2002; Herbst et al., 2007; Demetri et al., 2002; Holdsworth et al., 2007; Van den Abbeele, 2008). Thus FDG-PET shows great promise as an early indicator of response to therapy, both as an imaging biomarker for cancer clinical trials and as a tool for clinical decision making.

Since glucose is also the primary source of energy for brain cells (Sokoloff et al., 1977; Kim & Dang, 2006), analysis and interpretation of the FDG-PET signal in brain cancer poses the particular challenge that it may be confounded by changes in glucose metabolism as a result of normal brain function and pharmacodynamic drug effects to healthy tissue. As an example, described in more detail in Section 2, Figure 1 shows two patients' baseline (BL) PET scans acquired at the beginning of treatment and follow-up scans obtained a week later (W1). For the patient in the first row, the indicated tumor (as in the circle) shows a clear reduction in metabolism. However, the background tissue also appears lighter, indicating a reduction in glucose metabolism over the entire brain, more so in the gray matter (dark areas) than in the white matter (light areas) based on the known physiologic differences in glucose metabolism in these regions. The patient in the second row shows that the background glucose metabolism may also increase from baseline to follow-up.

The standard methods for evaluating FDG-PET images in oncology do not properly account for these confounding factors. The standard method, described in more detail in Section 2, essentially consists of manually delineating regions-of-interest (ROIs), such as the one marked in Figure 1, and comparing simplified measures of these regions, such as the maximum intensity within the region, between the baseline and follow-up scans. This approach is robust but is limited in scope, time-consuming, and subjective. Clinical practice typically involves qualitative comparisons of images and most clinical trials still rely on a single metric for comparison of changes in tumor metabolism.

The goal of this paper is to develop more objective and sophisticated methods for quantifying changes in brain metabolic activity of a single subject as a result of cancer

therapy, that correct for global and regional changes in normal brain glucose metabolism and provide an objective evaluation of significant changes anywhere in the brain as a result of therapy. In contrast to the standard ROI approach, we propose a voxelwise approach, where analysis is performed at all voxels (volumetric pixels) simultaneously. Voxelwise image analysis methods (Friston et al., 1995b; Genovese et al., 2002) have been developed previously for investigating cognitive mechanisms, neurodegenerative diseases, and neuropharmacologic effects, but are not used clinically to evaluate treatment effects on brain tumors.

Specifically, our proposed approach includes spatial registration between the two images, segmentation into major anatomical regions - gray matter (GM), white matter (WM) and cerebro-spinal fluid (CSF), background adjustment according to each region, and statistical comparison at each voxel. The background mean and covariance for each region are estimated by fitting a bivariate (baseline and follow-up) mixture model to image data from healthy brain slices. The division into the GM, WM and CSF regions is necessary as they typically have different metabolic levels. We adjust for the background by deducting the background mean from each voxel and dividing it by its background covariance. The relative change in the residuals between baseline and follow-up is then calculated and used as the statistic to evaluate the metabolic change. After background adjustment, a map is generated containing a statistic at each voxel, called statistical parametric map (SPM) in the neuroimaging literature (Friston et al., 1995b) (not to be confused with the image analysis software also called SPM). This map is then thresholded to identify voxels of statistically significant change while controlling for the false discovery rate (FDR), where the null distribution is estimated nonparametrically from the healthy brain slices mentioned above. This voxelwise approach provides information about the magnitude and spatial distribution of metabolic response everywhere in the brain, identifying areas of necrosis that may appear as structural lesions on MRI or conversely, and allowing discovery of metabolically progressing brain regions that have not increased in size.

The manuscript is organized as follows: section 2 describes the data and motivating example, including the standard ROI analysis and pre-processing of the data for our method; section 3 describes our proposed voxelwise analysis approach, including the background estimation, background adjustment, generation of the SPM, and the FDR controlling procedure. For illustration purposes, we accompany each step of the method with the corresponding results from one of the subjects from the motivating example. We then give the final results for that subject in Section 4 and apply the approach to the second subject (details shown in Supporting Information). Section 5 briefly describes the performance of the proposed method via simulations (Details given in Supporting Information), followed by a discussion in Section 6.

2. Data and Standard Approach

2.1. Data and registration

Our motivating example comes from a clinical trial designed to evaluate the treatment effect of Lapatinib, a new molecular targeted chemotherapeutic agent, on brain metastases for patients with human epidermal growth factor receptor 2 (HER-2) positive breast cancer (Lin

et al., 2008), conducted at the Dana-Farber/Harvard Cancer Center. In the study, HER-2 positive breast cancer patients received treatment of Lapatinib and 39 of these patients had FDG-PET scans performed at baseline and one week after initiating the treatment. Patients also received MRI scans at baseline and every eight weeks afterward until they withdrew consent, experienced unacceptable toxicity, or had progressive disease. The primary endpoint of the study was response rate based on the MRI measurement of tumor size. Using the Response Evaluation Criteria In Solid Tumors (RECIST) criteria, only one patient had partial response. In this paper, we analyzed two subjects that had FDG-PET and MRI scans demonstrating at least one visible tumor in the brain and visually appeared to have metabolic response (subject 12) and metabolic progression (subject 10), but were classified as stable by RECIST.

The data for each of these subjects consists of two arrays of PET image intensities, one at baseline (BL) and one at week-1 (W1). The arrays are of size $128 \times 128 \times 51$ for subject 12 and $128 \times 128 \times 21$ for subject 10 at each scan. The native FDG-PET image data was stored in the Digital Imaging and Communications in Medicine (DICOM) format and converted to Interfile format to be read into MATLAB (The Mathworks, Inc.). Using image meta-data contained in the DICOM header, the image intensity values were converted to standardized uptake value (SUV) units, obtained as the measured intensity of FDG radioactive concentration divided by the ratio between the administered FDG dose and the patient's body mass (Sugawara et al., 1999).

To compare the BL and W1 images at the voxel level, spatial image registration was performed to account for differences in patient positioning between the two imaging sessions. This was accomplished by applying a linear transformation to the W1 image coordinates that minimizes the mean square error between the BL and the transformed W1 image (Friston et al., 1995a). This step was implemented using commercial software (Hermes Medical Solutions, Stockholm, Sweden).

Selected transverse slices from the spatially registered volumetric scans for the two subjects are shown in Figure 1. Our proposed methods to identify metabolic changes in the brain will be illustrated in these two subjects as examples of responding and non-responding subjects.

2.2. Standard approach: ROI analysis

The standard ROI approach originally used to analyze this type of data was as follows. For the two spatially registered FDG-PET images, multiple masking ROIs that contain lesions are manually drawn on the BL image and the corresponding ROIs are identified on the W1 image. The maximum SUV among the SUVs of all voxels within the ROI, SUV_{max}, was identified within each ROI in the BL and W1 images. Then the relative change of each lesion is calculated as the difference between the SUV_{max} in the W1 image and the SUV_{max} in the BL image divided by the BL SUV_{max}. For each subject, a summed SUV_{max} is calculated by adding the SUV_{max} of all lesions. According to the thresholds in the standard European Organization for Research and Treatment of Cancer (EORTC) criteria (Young et al., 1999), the metabolic response of the individual lesion or patient is categorized by the relative reductions of the lesion SUV_{max} or patient-summed SUV_{max} as follows: partial metabolic response occurs when the relative change consists of a reduction

of more than 25%; progressive metabolic disease occurs when the relative change consists of an increase of more than 25%; and stable metabolic disease occurs when the relative change is less than 25% in either direction.

The above ROI approach is simple and relatively easy to implement on standard imaging workstations by nuclear medicine physicians. It is also robust, as small changes in the contour of any ROI are unlikely to affect the SUVmax value for that ROI. However, it has important limitations. First, by definition, the analysis is restricted to the ROIs drawn on target lesions, so the changes in brain regions outside these ROIs may be missed. Second, the distribution of potentially heterogeneous metabolic response of voxels within the ROI is ignored in favor of the single SUVmax summary statistic. Third, the identification of target lesions and manual delineation of ROIs varies according to the radiologist's experience, which results in observer variability, and is also time consuming as it requires inspection of the image volume slice by slice. Finally, the standard analysis might be confounded by changes in the two scans' background SUV, either due to changes in normal brain metabolism or as a result of pharmacodynamic effects of the therapy on the brain (Kuhl et al., 1977; Sokoloff et al., 1977). Moreover, different anatomical brain regions, such as GM, WM and CSF, have different levels of physiologic glucose metabolism and therefore may change in a different way as a result of treatment between the BL and W1 scans. The standard analysis does not account for the heterogeneous background in these three components.

The problem is illustrated in Figure 1. For the subject in the first row, the ROI marked has a SUVmax reduction of 29%. Since this exceeds the 25% EORTC cutoff, the lesion would be classified as a partial metabolic response. However, since the background of the W1 scan is reduced relative to the BL scan, the reduction in the tumor might have resulted from the reduced background.

A variation of the standard method to adjust for the change in the background is to compute for each ROI and each image the tumor-to-background ratio, obtained as the ratio between the SUVmax in the ROI and the SUVmax in another corresponding ROI drawn on the contralateral side to the tumor, representing the background. The relative change is then computed as the ratio between the tumor-to-background ratio in the W1 and that in the BL image.

Doing this for the ROI marked in Figure 1 (contralateral ROI not shown) resulted in a change in tumor-to-background ratio of 7.8%. According to the EORTC criterion, the lesion would no longer be classified as a partial metabolic response to treatment, contradicting the previous assessment. This example highlights the importance of adjusting for changes in the background. The contralateral ROI method, while it approaches this problem partially, still suffers from the basic limitations of ROIs, mainly subjectivity and limited scope.

In the proposed method, we use a voxelwise approach that measures the change in glucose metabolism for the whole brain, while adjusting for the tissue-specific background. Besides the registration procedure that is also standard for ROI analysis, we also need segmentation, background estimation, calculating the statistics of metabolic change for each voxel, and

FDR thresholding for identifying the regions of change. The segmentation step is described next, and the statistical methods are described in Section 3.

2.3. Pre-processing: segmentation

As explained above, the proposed method requires separation of the tissue into GM, WM and CSF components. As shown in a transverse view in Figure 1, the full volumetric image is composed of both the brain and other structures outside the brain including skull and scalp. The analysis is therefore restricted to voxels inside the brain. This is achieved by first constructing a soft segmentation map that assigns to each voxel a probability of belonging to the GM, WM, CSF, or none of those (i.e. outside the brain). A binary brain mask is then generated where each voxel is assigned the value 1 if it has a high probability of belonging to the brain.

Specifically, the soft segmentation is implemented using the software package Statistical Parametric Mapping (SPM) (Wellcome Trust Centre for Neuroimaging, University College London) for Matlab, which is based on the Expectation-Maximization (EM) algorithm (Ashburner & Friston, 2005). In our analysis, we obtained the segmentation probabilities for both the BL and W1 images.

A binary mask for the brain and masks for each tissue type are then generated using the segmentation probabilities from both images, as follows. First, an initial brain mask is generated including voxels whose probability of belonging to either the GM, WM or CSF is greater than 0.5, separately for the BL and W1 images. The cutoff of 0.5 was chosen so that the brain mask includes as much brain tissue as possible while excluding the non-brain tissues. The final binary mask is then obtained as the intersection of the two initial masks for the BL and W1 images. Within this mask, the probabilities of belonging to the GM, WM and CSF are reweighted so that they sum to 1 at each voxel.

Figure 2 shows a slice of the 3D map of the BL segmentation probabilities for GM, WM and CSF in the upper row. In the GM probability map (left), the voxels with darker colors are more likely to belong to GM than the voxels with lighter color, since they have higher probabilities of belonging to GM. Here the full brain mask has already been applied, so that the three maps sum to 1 when added voxelwise. For this patient, the number of voxels considered inside the brain for analysis is 142,964, smaller than $128 \times 128 \times 51 = 835,584$ voxels in the original images. The CSF probability map (right) shows some edge effects, which might be due to a slight mismatch in the spatial registration.

The bottom row of Figure 2 shows hard segmentation maps obtained from the soft segmentation maps by maximum likelihood, i.e. by assigning each voxel to the tissue type to which it belongs with the highest probability. Note that the full brain mask may be obtained by adding these hard segmentation maps voxelwise. While the statistical analysis will be based on the soft segmentation maps, the hard segmentation maps will be used for display purposes below.

Unlike MRI segmentation, PET segmentation does need additional registration and has a direct interpretation in terms of metabolism: the PET segmentation is not exactly anatomical

but rather functional, where the regions labeled “GM”, “WM” and “CSF” correspond to regions of high, medium and low metabolism, respectively.

3. Statistical Methods

3.1. A mixture model

For the purposes of our analysis, we think of the spatially registered BL and W1 images as a single bivariate image, where each voxel is attached to a pair of SUV values, one for each of the scans.

In what follows, we assume a mixture bivariate distribution with three components for each voxel, each voxel being assigned a mixing probability of belonging to the GM, WM or CSF. The deviation of each voxel's intensity from its background can be described by the difference between its observed value and the mean of its distribution, standardizing by its covariance. In this sense, the background for each voxel is captured (at least partially) by the marginal mean and covariance for that voxel. The marginal mean and covariance for each voxel are computed from the conditional mean and covariance for each of the tissue types, weighted according to the mixing probabilities at that voxel. For the distribution of each component, the mean and covariance are estimated by maximizing the marginal likelihood of the data, where the data for this step is chosen in healthy regions of the brain so that only contains the background.

More specifically, assuming each voxel in $\mathbb{R}^3$ has coordinate $\underline{l}$, it belongs to one of the three tissue types with probabilities $P_{GM}(\underline{l})$, $P_{WM}(\underline{l})$, $P_{CSF}(\underline{l})$. These probabilities were obtained for all voxels through the soft segmentation procedure described earlier. Here we use the BL segmentation probabilities instead of the W1 segmentation probabilities because baseline scan has no pharmacodynamic effects induced by the therapy. Let $S(\underline{l})$ be a categorical random variable indicating the assignment of voxel $\underline{l}$ to one of the three tissue types:

$$S(\underline{l}) = \begin{cases} GM & \text{with prob. } p_{GM}(\underline{l}) \\ WM & \text{with prob. } p_{WM}(\underline{l}) \\ CSF & \text{with prob. } p_{CSF}(\underline{l}) \end{cases}, \quad (1)$$

Since the change in SUV is a multiplicative factor of the background SUV, we first apply a log transformation to the data. We assume the pair of BL and W1 log-SUV values at location $\underline{l}$ is a single realization from a distribution with mean $\mu(\underline{l})$ and covariance matrix $\Sigma(\underline{l})$, or equivalently,

$$Y(\underline{l}) = \begin{pmatrix} Y_{BL}(\underline{l}) \\ Y_{FU}(\underline{l}) \end{pmatrix} \sim (\mu(\underline{l}), \Sigma(\underline{l})). \quad (2)$$

Note that no particular distributional form is assumed.

For each voxel $\underline{l}$, the conditional mean and covariance of the bivariate log-SUV value of the voxel given the assignment are defined as

$$Y(\underline{l}) | S(\underline{l}) \sim \begin{cases} (\mu_{GM}, \Sigma_{GM}) & \text{if } S(\underline{l}) = GM \\ (\mu_{WM}, \Sigma_{WM}) & \text{if } S(\underline{l}) = WM \\ (\mu_{CSF}, \Sigma_{CSF}) & \text{if } S(\underline{l}) = CSF \end{cases} \quad (3)$$

Each pair would be the mean and covariance of each voxel in the GM, WM and CSF if the true anatomical assignment were known. However, the assignment is known only up to the soft segmentation. Thus the marginal distribution for the log SUV at each voxel $\underline{l}$ has mean

$$\mu(\underline{l}) = E[Y(\underline{l})] = E\{E[Y(\underline{l}) | S(\underline{l})]\} = \sum_s p_s(\underline{l}) \mu_s, \quad (4)$$

where $s \in \{GM, WM, CSF\}$, and has covariance structure

$$\begin{aligned} Var[Y(\underline{l})] &= E_s(Var[Y(\underline{l}) | S(\underline{l})]) + Var_s(E[Y(\underline{l}) | S(\underline{l})]) \\ &= \sum_s p_s \left(\Sigma_s + [\mu_s - \mu(\underline{l})][\mu_s - \mu(\underline{l})]' \right), \end{aligned} \quad (5)$$

where we have used the standard formula for conditional variance, conditioning on the voxelwise region assignment $S(\underline{l})$.

The marginal mean for each voxel could be viewed as the weighted average of the means from the three tissue types, with weights being the probability of belonging to the three tissue types. For example, a voxel with high probability of belonging to GM has more weight on the GM-specific mean and hence the marginal mean is close to the GM-specific mean than the WM- or CSF-specific means.

For illustration, Figure 3 shows a contour plot of the mixture bivariate distribution of the BL and W1 log-SUV, overall for the whole brain and separately for each tissue type, for patient 12. In this figure, voxels were grouped according to the tissue-specific hard segmentations (binary masks) of Figure 2, and then mixed to create the scatterplot for the whole brain. In the plot for the whole brain (upper left), the center of the contour plot deviates from the 45° line, which indicates a difference in the background mean of the baseline and follow-up images and hence a need for background adjustment. The three tissue-specific plots (the rest of the figures) show that the extent of the deviation from the 45° varies for GM, WM and CSF, indicating that the degree of background adjustment might be different for the three tissue types.

3.2. Background estimation

To properly adjust for the background, we need estimates of the marginal mean and covariance at each voxel, obtained via Equations 4 and 5. To obtain estimates for the means and covariance matrixes of the three tissue types, we use marginal likelihood estimates. These estimates are based on data from a subset of transverse slices in the brain containing only healthy tissue identified visually by radiologist. We exclude the slices with tumor because the estimates for the background should not be contaminated by the tumor.

Obtained by marginal maximum likelihood (see supporting information), the estimates for the tissue specific means and covariance matrixes are:

$$\hat{\mu}_s = \sum_{\underline{I}} p_s(\underline{I}) Y(\underline{I}) / \sum_{\underline{I}} p_s(\underline{I}) \quad (6)$$

and

$$\hat{\Sigma}_s = \sum_{\underline{I}} p(\underline{I}) (Y(\underline{I}) - \hat{\mu}_s) (Y(\underline{I}) - \hat{\mu}_s)' / \sum_{\underline{I}} p(\underline{I}). \quad (7)$$

which are the weighted sample means and sample covariances. These estimates are biased, but the bias is minor, as illustrated in the simulations in Section 5. We also describe the other possible approaches for estimation in the Discussion section.

In Figure 3, the superimposed ellipses are equidensity contours containing 95% of the probability corresponding to a Gaussian distribution with the estimated tissue-specific mean and covariance matrix. As shown in the individual panels, the centers of the ellipses closely match the modes of the data, indicating a good fit of the mixture model. The upper left figure shows that the three ellipses do not coincide, which indicates the heterogenous estimates for the mean and covariance of the three tissue types and the need to treat them separately.

3.3. Background adjustment and statistics for metabolic change

For each voxel at location $\underline{I}$, we consider two approaches to adjust the background: adjusting only the background mean and adjusting both the background mean and covariance.

The first approach is to center Y , that is, to take the difference between the observed intensity (log-SUV) and the marginal mean for each voxel: $\tilde{Y}(\underline{I}) = Y(\underline{I}) - \hat{\mu}(\underline{I})$, where the voxelwise mean estimate $\hat{\mu}$, is obtained by substituting (6) in (4). We then calculate the relative change of $\tilde{Y}$ and compare it to the EORTC 25% threshold. Any voxels with relative change greater than 25% is classified as a response. This is similar to the tumor-to-background ratio approach mentioned in Section 2, except that the estimation of the background considers all voxels in the brain instead of only the voxels in the contralateral ROI.

The second approach is to standardize $Y(\underline{I})$, that is, take the difference between the observed intensity and the marginal mean at each voxel and divide it by the root marginal covariance at the same voxel, that is,

$$Z(\underline{I}) = \hat{\Sigma}(\underline{I})^{-1/2} [Y(\underline{I}) - \hat{\mu}(\underline{I})] \sim (0, I), \quad (8)$$

where the marginal mean is estimated as before and the marginal covariance is estimated replacing (6) and (7) in (5).

The notation on the right of (8) indicates that the distribution of the standardized Y has zero mean and identity covariance matrix. This approach accounts not only for the heterogenous

background mean, as the centered approach also does, but also for the different background variance. Since we estimated the background mean/covariance based on the mixture model, each voxel has its unique background mean/covariance and hence the heterogeneous background for the three tissue types is automatically incorporated.

For each voxel at location $\underline{l}$, we want to conduct a test that can identify positive and negative changes between the baseline and follow-up SUV. Specifically, we want to test the hypotheses

$$H_0: \mu_{W1}(\underline{l}) - \mu_{BL}(\underline{l}) = 0 \text{ vs. } H_1: \mu_{W1}(\underline{l}) - \mu_{BL}(\underline{l}) > 0 \quad (9)$$

and

$$H_0: \mu_{W1}(\underline{l}) - \mu_{BL}(\underline{l}) = 0 \text{ vs. } H_1: \mu_{W1}(\underline{l}) - \mu_{BL}(\underline{l}) < 0 \quad (10)$$

for all voxels. These tests are treated separately as they lead to different clinical conclusions. Given the standardized scores $Z(\underline{l})$, an appropriate test statistic for both tests at each voxel is given by the contrast

$$\Delta(\underline{l}) = \frac{Z_{W1}(\underline{l}) - Z_{BL}(\underline{l})}{\sqrt{2}} \sim (0, 1),$$

which has mean zero and standard deviation one. In the testing procedure, we will account for multiple testing over voxels.

3.4. False discovery rate inference

Recall that we want to test (9) and (10) at each voxel, where $\underline{l}$ varies within the brain. This is a large-scale multiple testing problem over thousands of voxels. We decide to control the false discovery rate (FDR), defined as the expected proportion of falsely rejected tests among all the significant tests. For this we follow the empirical FDR approach of Storey et al. (2004), where the FDR is estimated for a range of thresholds and the significance threshold is chosen for which the estimated FDR falls below the desired level.

Assuming we have a total of M voxels/hypothesis in the full brain mask, among them an unknown number M_0 are nulls (i.e. the null hypothesis holds) and $M_1 = M - M_0$ are non-nulls (i.e. the null hypothesis does not hold). In general, let $T_1, \dots, T_M$ denote the corresponding test statistics (in our case $T_1, \dots, T_M$).

Consider first the right-sided test (9). Based on a fixed threshold u , $R(u)$ tests are rejected, among which $V(u)$ are falsely rejected, that is,

$$R(u) = \sum_{all \underline{l}} 1(T(\underline{l}) \geq u), \quad V(u) = \sum_{null \underline{l}} 1(T(\underline{l}) \geq u).$$

Then the right-sided FDR is defined as

$$FDR_R(u) = E \left[\frac{V(u)}{R(u) \vee 1} \right],$$

where the maximal operator $\vee$ ensures that $FDR = 0$ if $R(u) = 0$. The empirical estimate of FDR_R is

$$F\hat{D}R_R(u) = \frac{\hat{M}_0/M \cdot E[V(u)/M_0]}{R(u)/M} = \frac{\hat{p}_0 \hat{P}_{H_0}(T(\underline{I}) \geq u)}{\hat{P}(T(\underline{I}) \geq u)}. \quad (11)$$

The FDR for the left-sided test (10) is obtained similarly as

$$F\hat{D}R_L(u) = \frac{\hat{p}_0 \hat{P}_{H_0}(T(\underline{I}) \leq u)}{\hat{P}(T(\underline{I}) \leq u)}. \quad (12)$$

In both (11) and (12), $\hat{p}_0$ is an estimate of the unknown null proportion $p_0 = M_0/M$, $\hat{P}$ represents the empirical distribution of all M test statistics in the full brain mask and $\hat{P}_{H_0}$ is the empirical distribution of a subset of null test statistics among M_0 . Both distributions are unknown in our case and hence are estimated from the data. Since the null distribution should be composed of background voxels, we use the data from a set of selected healthy slices with no tumor to estimate that empirical null distribution. Note that this nonparametric empirical null approach is unlike that of Efron (2004); Schwartzman et al. (2008). This is because no parametric form may be assumed for the null distribution of $(\underline{I})$, being obtained from only two time measurements. The estimate of the null proportion $\hat{p}_0$ is the same for both the right- and left-sided FDR and is obtained, following the mixture model structure in Efron (2004) and Schwartzman et al. (2008) as the ratio of the peaks of the two empirical densities (obtained by standard kernel estimation) corresponding to the two distributions $\hat{P}$ and $\hat{P}_{H_0}$.

Figure 4 shows the distribution for all test statistics $\hat{P}$ and the empirical null distribution $\hat{P}_{H_0}$ (scaled by $\hat{p}_0$) for patient 12 on the left panel. The null distribution was obtained from the data in six healthy slices selected out of 51 total slices of the PET image for Patient 12. The empirical null approximately matches the bulk of the data in the center, which are more likely to be the voxels in the background. When choosing the null distribution, we also compared $N(0,1)$ and t to the shape of the empirical null distribution and we found the empirical nulls fits the data better.

Following Storey et al. (2004), for a given FDR level α , the right-sided threshold for the test statistic is the minimum value among all u for which $F\hat{D}R(u)$ is no greater than α , that is, $u_R = \inf \{u: F\hat{D}R_R(u) \leq \alpha\}$. Similarly, the left-sided threshold is given by $u_L = \sup \{u: F\hat{D}R_L(u) \leq \alpha\}$. Storey et al. (2004) showed that this procedure provides strong control of FDR for identical and independent tests. They also showed that under weak

independence assumption, the procedure still provides strong control of FDR asymptotically for large M . The brain image data is usually large and the voxels have only local dependence, and hence it meets the weak independence assumption.

Figure 4 summarizes the FDR estimates as a function of threshold on the right panel. We calculated the left- and right-sided FDRs according to varying thresholds. The right-sided FDR is decreasing with increasing threshold and the left-sided FDR is increasing with increasing threshold. The asymptotic values of both FDR curves correspond to the estimated null proportion $\hat{p}_0=0.92$. As commonly encountered in large-scale multiple testing problems, this value is close to 1 (Efron, 2004; Schwartzman et al., 2008). The intersection of the two FDR curves and the pre-specified FDR level of 0.2 yields the threshold values $u_L = -2.97$ and $u_R = 2.60$, which are the thresholds for left-sided and right-sided tests for Delta.

4. Analysis Results

4.1. Identification of metabolic changes for Patient 12

Figure 5 shows the SPM of relative change for the unadjusted approach, the centered approach and standardized approach using the test statistic t_{Δ} for Patient 12. The top row shows the relative change SPM, while the bottom row shows the thresholded relative change, superimposed on a baseline MRI image of the same patient. Here we used the EORTC 25% threshold for the unadjusted and the centered approaches, and used the thresholds from the FDR controlling procedure for the standardized approach. In the bottom left and middle figures, the voxels with relative changes less than -25% or more than 25% are highlighted in light blue and orange. In the bottom right figure, the voxels with t_{Δ} less than -2.97 or greater than 2.60 are highlighted in light blue and orange respectively.

The unadjusted approach (bottom left) identified too many regions with significant change, most of them likely to be false positives. The centered approach (bottom middle) identified the two tumor regions (upper left and lower right) with reduced metabolic activity, after adjusting for the background, but the EORTC criterion is still too liberal. The standardized approach (bottom right) using t_{Δ} eliminated more noise and also identified the two tumor regions with suppressed metabolic activity by Lapatinib (upper left and lower right regions of the brain).

In the standard ROI analysis for Patient 12, only the tumor region on the left frontal lobe (upper right in Figure 5) was identified and a -7.8% reduction in the tumor-to-background ratio was reported. As this relative change is less than the 25% cutoff, this patient was classified as Stable Metabolic Disease. Our analysis, however, confirmed it as Partial Metabolic Response, even after adjusting for the background. Our analysis also identified one more region in the right parietal lobe (lower left in Figure 5) with suppressed metabolic activity, which had not been found by the standard analysis. This finding resulted from properly adjusting for the background and performing statistical testing over the entire brain.

4.2. Main results for Patient 10

Due to space limits, the detailed analysis results for Patient 10 are given in the Supporting Information.

5. Simulations

As shown in detail in the supporting information, we conducted synthetic data simulations to evaluate the accuracy of the estimates for the tissue-specific mean and covariance, as well as the error rate control and power of our approach in detecting the area of SUV change from baseline to follow-up scans. The estimates yielded small bias when the segmentation was accurate, with increased bias for worse segmentation. The detection FDR was controlled at the nominal level under various scenarios and the power was shown to increase as a function of the true change in tumor intensity from baseline to follow-up.

6. Discussion

In this manuscript, we have proposed an analysis pipeline to detect the local changes in brain metabolism in a single patient from a baseline to a follow-up image at the voxel level, while adjusting for changes in the background. After the two images were spatially registered and segmented, we assumed for the log SUV at each location a bivariate mixture model with three components, each component representing a tissue type. The background mean and covariance for each voxel were estimated from the background mean and covariance estimated belonging to three components, with the weights corresponding to the mixing probabilities (estimated by the segmentation probabilities). We defined the background-adjusted score Z by deducting the voxelwise background mean from the observed log-SUV and dividing it by the background covariance. We used the standardized difference, the difference in Z divided by $\sqrt{2}$, to measure the voxelwise metabolic change between two scans, adjusting for the different background mean and covariance. We then generated an SPM of the brain, with each voxel attached to a statistic. To test for the change at each voxel, we used an FDR-controlling procedure. Voxels with significant change were marked as having a metabolic change in the map. This approach successfully detected the voxels with metabolic change inside and outside the tumor regions, while controlling the error rate at pre-specified level. We discuss some of the main issues and alternative approaches

In pre-processing the data, the spatial registration between the baseline and the follow-up scan is a key issue for voxelwise analysis. We observed some degree of mis-registration, evidenced for instance by the edge effects in Figure 2 and Figure 5. More accurate registration could possibly be obtained using nonlinear or fluid registration algorithms (Friston et al., 1995b; Lepore et al., 2008).

Since the segmentation procedure involves a spatial prior that accounts for values in the neighborhood, the use of these segmentation probabilities in our approach has allowed us to incorporate the spatial correlation to some extent. However, a more comprehensive approach that fully incorporates the spatial correlation in both the model and the FDR controlling procedure would be an interesting subject.

Because of the inherent bias of the maximum marginal likelihood estimates for the tissue-specific background mean and covariance, a more stringent estimation for the means and covariance structure is needed. However, this is a difficult problem. We have considered

several other approaches. One approach is to give a hard assignment to each voxel by maximum likelihood, as in Figure 2 (bottom row), and calculate the mean and covariance matrix using the voxels in each tissue type. The main problem with this approach is the ambiguity and associated error in the assignment of the voxels on the boundary of the tissue types. Another approach is to apply an EM algorithm, using the baseline soft segmentation probabilities as the prior. However, the gain in accuracy is offset by the great increase in computational time. In addition, the data are used twice in the estimates: once in the prior and a second time in the EM algorithm. To reduce the impact of extreme values (such as those in the tumor) on the estimation, we considered robust estimates such as truncated means and covariance. However, these estimates turned out to be more biased than the maximum marginal likelihood estimates reported here.

Given that there is no formal model for the distribution of log-SUV, we did not assume any parametric distribution for the data except for the normality assumption in estimating the background mean and covariance. Instead, we used an empirical null distribution that was derived from actual null data, obtained from slices of healthy brain, to be used in the FDR controlling procedure. We emphasize that, because no two patients have tumors at the same location within the brain and the selection of healthy slices are not identical for any two patients, our method is intended to be applied to one individual subject at a time.

Although we generated a map to describe the voxelwise metabolic change throughout the whole brain, it would be useful to develop a global statistic that summarizes the results for a single patient and compare it to the EORTC criteria. We will consider in the future global statistics such as the maximal change within the detected regions, the total number of detected lesions (defined as the number of connected detected regions in the map), or the integral of the change within the detected lesions in the map. These global statistics could potentially be compared across subjects and used as biomarkers in predicting response to therapy and survival outcome in clinical trials and population studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by the Dana Foundation and NIH grant R21 EB013795.

References

- Ashburner J, Friston K. Unified segmentation. *Neuroimage*. 2005; 26:839–51. [PubMed: 15955494]
- Demetri GD, von Mehren M, Blanke CD, Van den Abbeele AD, Eisenberg B, Roberts PJ, Heinrich MC, Tuveson DA, Singer S, Janicek M, Fletcher JA, Silverman SG, Silberman SL, Capdeville R, Kiese B, Peng B, Dimitrijevic S, Druker BJ, Corless C, Fletcher CD, Joensuu H. Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. *New England Journal of Medicine*. 2002; 347(7):472–480. [PubMed: 12181401]
- Efron B. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J Amer Statist Assoc*. 2004; 99(465):96–104.
- Friston K, Ashburner J, Frith C, Poline J, Heather J, Frackowiak R. Spatial registration and normalization of images. *Human brain mapping*. 1995a; 3(3):165–189.

- Friston KJ, Ashburner J, Poline JB, Frith CD, Heather JD, Frackowiak RSJ. Spatial registration and normalization of images. *Human Brain Mapping*. 1995b; 2:165–189.
- Genovese C, Lazar N, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*. 2002; 15(4):870–878. [PubMed: 11906227]
- Herbst R, Davies A, R.B. N, Dang T, Schiller J, Garland L, Miller V, Mendelson M, Van den Abbeele A, Melenevsky Y, de Vries D, Eberhard D, Lyons B, Lutzker S, Johnson B. Efficacy and safety of single-agent pertuzumab, a human epidermal receptor dimerization inhibitor, in patients with non small cell lung cancer. *Clinical Cancer Research*. 2007; 13(20):6175–81. [PubMed: 17947484]
- Holdsworth CH, Badawi RD, Manola JB, Kijewski MF, Israel DA, Demetri GD, Van den Abbeele AD. Ct and pet: early prognostic indicators of response to imatinib mesylate in patients with gastrointestinal stromal tumor. *American Journal of Roentgenology*. 2007; 189(6):W324–W330. [PubMed: 18029844]
- Kim J, Dang C. Cancer's molecular sweet tooth and the warburg effect. *Cancer Research*. 2006; 66(18):8927–30. [PubMed: 16982728]
- Kuhl D, Phelps M, Hoffman E, Robinson G Jr, MacDonald N. Initial clinical experience with 18f-2-fluoro-2-deoxy-d-glucose for determination of local cerebral glucose utilization by emission computed tomography. *Acta neurologica Scandinavica. Supplementum*. 1977; 64:192. [PubMed: 268784]
- Lepore N, Brun C, Chou Y, Chiang M, Dutton R, Hayashi K, Luders E, Lopez O, Aizenstein H, Toga A, et al. Generalized tensor-based morphometry of hiv/aids using multivariate statistics on deformation tensors. *Medical Imaging, IEEE Transactions on*. 2008; 27(1):129–141.
- Lin N, Carey L, Liu M, Younger J, Come S, Ewend M, Harris G, Bullitt E, Van den Abbeele A, Henson J, et al. Phase ii trial of lapatinib for brain metastases in patients with human epidermal growth factor receptor 2-positive breast cancer. *Journal of Clinical Oncology*. 2008; 26(12):1993–1999. [PubMed: 18421051]
- Patanaphan V, Salazar OM, Risco R. Breast cancer: metastatic patterns and their prognosis. *Southern medical journal*. 1988; 81(9):1109–1112. [PubMed: 3420442]
- Schwartzman A, Dougherty R, Taylor J. False discovery rate analysis of brain diffusion direction maps. *Annals of Applied Statistics*. 2008; 2(1):153–175.
- Sokoloff L, Reivich M, Kennedy C, Rosiers M, Patlak C, Pettigrew K, Sakurada O, Shinohara M. The [14c] deoxyglucose method for the measurement of local cerebral glucose utilization: theory, procedure, and normal values in the conscious and anesthetized albino rati. *Journal of neurochemistry*. 1977; 28(5):897–916. [PubMed: 864466]
- Storey J, Taylor J, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2004; 66(1):187–205.
- Sugawara Y, Zasadny K, Neuhoff A, Wahl R. Reevaluation of the standardized uptake value for fdg: variations with body weight and methods for correctionl. *Radiology*. 1999; 213(2):521–525. [PubMed: 10551235]
- Van den Abbeele A, Badawi R. Use of positron emission tomography in oncology and its potential role to assess response to imatinib mesylate therapy in gastrointestinal stromal tumors (gists). *European Journal of Cancer*. 2002; 38:S60–S65. [PubMed: 12528774]
- Van den Abbeele AD. The lessons of gistOpet and pet/ct: a new paradigm for imaging. *The oncologist*. 2008; 13(Supplement 2):8–13. [PubMed: 18434632]
- Warburg O. On the origin of cancer cells. *Science*. 1956; 123(3191):309–14. [PubMed: 13298683]
- Young H, Baur R, Cernerius U, Herholz K, Hoekstra O, Lammertsma A, Pruin J, Price P, et al. Measurement of clinical and subclinical tumour response using [18f]-fluorodeoxyglucose and positron emission tomography: review and 1999 eortc recommendations. *European Journal of Cancer*. 1999; 35(13):1773–1782. [PubMed: 10673991]

Figure 1.

Baseline (first column) and week-1 (second column) PET images in SUV units for Subject 12 (upper row) and 10 (bottom row) at a particular transverse slice. A darker shade of gray indicates a higher metabolic activity. BL indicates baseline. W1 indicates week 1 after initiating treatment.

Figure 2. Patient 12: Soft segmentation probabilities (upper row) and hard segmentation masks (bottom row) for GM, WM and CSF. Black represents probability 1, white represents probability 0. Same transverse slice as in Figure 1.

Figure 3.

Patient 12: contour plot of the bivariate distribution of the baseline and follow-up log-SUV for the whole brain, GM, WM and CSF. The diagonal lines indicate no change between BL and W1 (red) and 25% change according to the EORTC criterion (green). For visual reference, superimposed are the 95% probability contour of the bivariate distribution component of each tissue type based on the estimated tissue-specific mean and covariance assuming a Gaussian density. Note that the Gaussianity assumption is used here only for illustration purposes but not in the actual data analysis.

Figure 4.

Patient 12: (left panel) the distribution for all test statistics (blue) and the null component estimated from healthy slices scaled by p_0 (green); (right panel) right-tail (green) and left-tail (blue) FDR estimates as a function of threshold.

Figure 5.

Patient 12: Relative change (top row) and thresholded relative change superimposed on baseline MRI (bottom row) for the three approaches: unadjusted, centered and standardized. Same transverse slice as in Figure 1, top row. In the bottom row, regions above the positive thresholds marked in orange; regions below the negative threshold marked in blue.