Automated measurement of uptake in cerebellum, liver, and aortic arch in full-body FDG PET/CT scans

Christian Bauer and Shanhui Sun

Department of Electrical and Computer Engineering, The University of Iowa, Iowa City, Iowa 52242

Wenqing Sun, Justin Otis, and Audrey Wallace

Department of Radiation Oncology, University of Iowa Hospitals and Clinics, Iowa City, Iowa 52242

Brian J. Smith

Department of Biostatistics, The University of Iowa, Iowa City, Iowa 52242

John J. Sunderland and Michael M. Graham

Department of Radiology, The University of Iowa, Iowa City, Iowa 52242

Milan Sonka

Department of Electrical and Computer Engineering and Department of Radiation Oncology, The University of Iowa, Iowa City, Iowa 52242

John M. Buatti

Department of Radiation Oncology, University of Iowa Hospitals and Clinics, Iowa City, Iowa 52242

Reinhard R. Beichel^{a)}

Department of Electrical and Computer Engineering and Department of Internal Medicine, The University of Iowa, Iowa City, Iowa 52242

(Received 17 February 2012; revised 19 April 2012; accepted for publication 20 April 2012; published 14 May 2012)

Purpose: The purpose of this work was to develop and validate fully automated methods for uptake measurement of cerebellum, liver, and aortic arch in full-body PET/CT scans. Such measurements are of interest in the context of uptake normalization for quantitative assessment of metabolic activity and/or automated image quality control.

Methods: Cerebellum, liver, and aortic arch regions were segmented with different automated approaches. Cerebella were segmented in PET volumes by means of a robust active shape model (ASM) based method. For liver segmentation, a largest possible hyperellipsoid was fitted to the liver in PET scans. The aortic arch was first segmented in CT images of a PET/CT scan by a tubular structure analysis approach, and the segmented result was then mapped to the corresponding PET scan. For each of the segmented structures, the average standardized uptake value (SUV) was calculated. To generate an independent reference standard for method validation, expert image analysts were asked to segment several cross sections of each of the three structures in 134 F-18 fluorodeoxyglucose (FDG) PET/CT scans. For each case, the true average SUV was estimated by utilizing statistical models and served as the independent reference standard.

Results: For automated aorta and liver SUV measurements, no statistically significant scale or shift differences were observed between automated results and the independent standard. In the case of the cerebellum, the scale and shift were not significantly different, if measured in the same cross sections that were utilized for generating the reference. In contrast, automated results were scaled 5% lower on average although not shifted, if FDG uptake was calculated from the whole segmented cerebellum volume. The estimated reduction in total SUV measurement error ranged between 54.7% and 99.2%, and the reduction was found to be statistically significant for cerebellum and aortic arch.

Conclusions: With the proposed methods, the authors have demonstrated that automated SUV uptake measurements in cerebellum, liver, and aortic arch agree with expert-defined independent standards. The proposed methods were found to be accurate and showed less intra- and interobserver variability, compared to manual analysis. The approach provides an alternative to manual uptake quantification, which is time-consuming. Such an approach will be important for application of quantitative PET imaging to large scale clinical trials. © 2012 American Association of Physicists in Medicine. [<http://dx.doi.org/10.1118/1.4711815>]

Key words: automated uptake measurement, cerebellum, liver, aortic arch

I. INTRODUCTION

There are several underlying assumptions in using F-18 fluorodeoxyglucose (FDG) standardized uptake values (SUVs) for characterizing and following metabolic activity in tumors.¹ The failure of some of these assumptions can introduce significant variability in the calculated SUVs. These assumptions include accurate measurement of the injected dose; accurate decay correction of all measurements; complete injection of the dose, i.e., no infiltration; correct entry of all relevant values into the reconstruction program; and similar blood time–activity curves of all subjects. Problems have been seen with all of these assumptions. One approach that has been suggested to avoid these problems and to improve the reproducibility of SUV measurements is to use lesion to normal tissue SUV ratios as the index of metabolic rate rather than the SUV by itself.

Other assumptions are introduced when using the lesion to normal tissue ratio. The fundamental underlying assumption is that a normal tissue can be identified where the uptake is an indication of the availability of FDG to the tissue during the entire uptake period. A second assumption is that the normal tissue activity can be identified in a reproducible fashion. Several authors have examined the utility of ratio for discrimination of benign from malignant nodules. Obrzut *et al.*² found that lung nodule to cerebellum ratio was more accurate than SUV alone. Blake *et al.*³ also found that tissue ratios performed better than absolute SUV measures in diagnosing adrenal nodules with the adrenal to liver ratio being a better discriminator than nodule SUV.

Since the main goal of this approach requires accurate and reproducible measures of average uptake in normal tissue, appropriate methodology that reliably and automatically yields the average FDG uptake in normal tissue regions is required. This will allow the empirical determination of which normal tissue is best under different circumstances, should simplify using the lesion to normal tissue method, and should produce highly reproducible results.

The automated segmentation and/or uptake measurement of structures like the cerebellum, liver, and aortic arch reference regions in full-body FDG PET/CT scans is largely unaddressed in the literature. Li *et al.*⁴ presented a method to segment the liver in low-contrast CT based on a segmentation of the liver in PET datasets, while a variety of methods have been presented in the literature for contrast-enhanced CT scans.^{5,6} For segmentation of brain structures, atlas-based registration approaches are typically utilized (e.g., van der Lijn *et al.*⁷). Xia *et al.*⁸ presented a method which utilizes information from PET and CT images to segment brain structures. An approach for aortic arch segmentation in CT scans without contrast enhancement was presented by Feuerstein *et al.*⁹ Note that all these methods were developed for imaging modalities or protocols that are different from the ones utilized in this work. For example, the resolution of full-body PET/CT images is typically lower than a dedicated brain PET/CT scan.

In this paper, we introduce and validate a fully automated approach for uptake measurement of cerebellum, liver, and

aortic arch in full-body FDG PET/CT scans to enable the reproducible calculation of tracer uptake in these structures. All acquisitions in this study used an imaging protocol specifically developed for head and neck cancer, but our proposed methodology can be easily adapted to other acquisition protocols. The presented method represents a powerful enabling tool for tumor-to-reference tissue ratio measurement in single- and multicenter studies that could be adopted for clinical application. These methods will be particularly useful in clinical and clinical research applications where physiologic variability associated with direct SUV measurements is unacceptably high. It is also possible that the tumor-to-reference tissue ratio measurements enabled by this work may prove to be a generally more robust response-to-therapy assessment approach than the more common SUV method in both clinical and clinical research applications.

II. METHODS

Methods for automated segmentations of the cerebellum, liver, and aortic arch were developed to automatically quantify FDG uptake in these whole-body FDG PET/CT reference regions. In the case of the cerebellum, a complete 3D segmentation of the structure is generated, whereas a sphere and tube are identified inside the liver and aortic arch, respectively, to deal with uncertainties due to respiratory motion and partial volume effects.

All three reference regions are automatically identified in subparts of the image relative to the brain, which increases robustness of the segmentation and reduces computation time. Note that the approach was specifically tailored to the imaging data described in Sec. III.A, which generally showed the whole cerebellum, but in the majority of cases, the brain was only partially imaged. To segment the brain, each PET scan is thresholded at $2.5 \text{ SUV}_{\text{bw}}$, and morphological closing and opening with a spherical structuring element with the radius of 5 mm is performed to avoid holes inside the brain or leakage in case of adjacent tumors. In the resulting binary image, the brain is identified as the connected component with a volume larger than 1000 mm^3 closest to the top of the scan. Relative to the brain segmentation, ROIs (see Fig. 1) are defined as reference for the different reference region identification methods. ROIs for the cerebellar hemisphere center locations and the aortic arch are aligned with the lower left and right outer bounds of the brain. The ROI for the liver spans the whole lower right area of the image with a minimum distance from the brain center location. There are no restrictions on the ROIs in the sagittal direction.

In this context, segmentations of cerebellum and liver are directly performed in the SUV_{bw} normalized PET scans, while the segmentation of the aortic arch is performed in the CT volume of the PET/CT to increase robustness. The above described ROI generation was designed for and may need adaptation when applied to pediatric PET/CT image data.

II.A. Cerebellum reference region segmentation

Our method for cerebellum segmentation is based on an active shape model (ASM) (Ref. 10) segmentation approach.

FIG. 1. Segmented brain (yellow) and regions of interest for identification of (a) liver center (green), (b) cerebellum hemisphere centers (red), and (c) aortic arch (blue). Distance measures for the ROIs are in centimeters. All PET images are shown with inverted gray-scales.

First, in a training stage, a model of the cerebellum is generated from learning data. Second, the model is automatically initialized at the approximate location of the cerebellum. Finally, segmentation is performed by means of a robust model matching approach.

II.A.1. Cerebellum model generation

To model the shape variability of cerebella as well as the gray-value appearance around its surface along normal profiles, learning data are required. For this purpose, nine normal cerebella were segmented in training PET scans and utilized for model generation, as described below.

II.A.1.a. Statistical shape model. From the binary segmentations, triangle mesh representations were generated with a marching cubes algorithm.¹¹ A set of 642 corresponding vertices (landmark points) were automatically identified on all meshes by means of a minimum description length approach¹² based on shape index and curvedness.¹³ The selected number of landmarks was found to be a good trade-off between processing speed and representation of shape details. Using the landmark correspondences, the learning shape models were aligned in a common coordinate frame using Procrustes analysis.¹⁴ Based on the mean point locations and a principal component analysis of the covariance matrix, a point distribution model¹⁰ was derived, which allows describing a cerebellum shape in terms of a mean shape and variation about the mean in a linear model: $\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b}$, where $\bar{\mathbf{x}}$ denotes the mean shape, $\mathbf{P}$ the shape eigenvector matrix, and $\mathbf{b}$ the shape coefficients. Figure 2 depicts the mean shape and the variation associated with the first eigenvector, which describes the most dominant shape variation of cerebella.

II.A.1.b. Gray-value surface profile model. The gray-value patterns around different cerebellum surface locations

FIG. 2. Visualization of the shape variation of the cerebellum learned from training datasets. The first mode of variation ($\sigma = -2$ to $\sigma = +2$) is shown with the mean shape ($\sigma = 0$) located in the center.

are quite characteristic (Fig. 3). This information is incorporated in our model by means of gray-value profiles as follows. First, landmark points of our statistical shape model are transformed to their related PET training dataset. Second, for each landmark point, a surface profile model $\mathbf{p}_i$ reflecting the typical SUV change along the surfaces normal direction is generated. For this purpose, PET values along landmark surface normals are sampled in a range of ± 1 cm from the surface points with a sampling interval of 0.5 mm. For all profiles, the first derivatives are computed and averaged for each landmark, which results in $\mathbf{p}_i$.

II.A.2. Automated model initialization

Before the model can be matched to PET volume data, it needs to be initialized and placed in rough proximity of the target structure. We utilize the mean cerebellum shape parameters ($\mathbf{b} = \mathbf{0}$) and assume no rotation and scaling. The approximate location of the cerebellum is additionally calculated in an automated fashion. For this purpose, the bounding box of the brain segmentation serves as a reference to specify subvolumes containing the right/left hemispheres of the cerebellum (see Fig. 1). In these subregions, the cerebellum hemisphere centers are identified based on the fact that metabolism is lower at the centers compared to the surrounding tissue. Thus, all gradient vectors point away from this location. To identify this location, the corresponding Gaussian gradient vector field is computed at a scale of $\sigma = 3$ mm, individual vectors normalized to unit-length, and the average outward flux¹⁵ (AOF) is determined. The AOF measures the divergence of a vector field at a certain location. The centers of the cerebellum hemispheres are identified as the locations with the maximum AOF in the search regions and a minimum distance of 3 mm to the surface of the brain segmentation. The average of the left and right cerebellum hemisphere centers is used as the initial location for our model.

FIG. 3. Axial (a), coronal (b), and sagittal (c) image slices of an FDG PET scan showing the brain with the segmented cerebellum. All PET images are shown with inverted gray-scales.

II.A.3. Active shape model matching

Matching of an active shape model to the target structure in the PET volume is an iterative process. In each iteration, two processing steps are performed. First, by utilizing the present model shape parameters $\mathbf{b}$ and transformation parameters $\mathbf{T}$ (translation, rotation, and scaling), shape points are updated by utilizing the gray-value profile models (Sec. II.A.1). For each landmark point, the gray-value information along a line normal to the model surface is extracted from the image and the first derivative is computed. The update point along this line is found by matching against the template $\mathbf{p}_i$ of the corresponding landmark i . For the matching function, the cross correlation is used. The utilized range to search for update points was set to ± 10 mm, and gray-values in the PET scan < 1 SUV were set to 1 to avoid mismatches at the transitions between head and air. Second, the update shape points are utilized to recalculate model ($\mathbf{b}$) and transformation ($\mathbf{T}$) parameters. Usually, this is achieved by utilizing Procrustes alignment and least squares optimization.¹⁰ However, this approach does not perform robustly in the case some of the identified update points are wrong (outliers). In addition, we have to potentially deal with truncated PET data, where parts of the brain (or even of the cerebellum) were not imaged and with nontypical adjacent image structures (e.g., tumors). To avoid such issues, we utilize the robust ASM matching approach that was recently presented by Sun *et al.*¹⁶ The basic idea behind this approach is to identify outliers and to exclude them from the model parameter recalculation. For segmentation, we perform 100 model matching iterations, which were found to be sufficient in combination with the proposed model initialization to achieve convergence of the model.

II.B. Liver reference region segmentation

The liver is located in the lower right part of the image relative to the brain (see VOI in Fig. 1). In this area, the liver can be identified as the largest organ above a certain uptake value, because it is mostly surrounded by large structures with low uptake (lungs, muscles, fat tissue, etc.) and smaller organs with similar or higher uptake (heart, kidneys, etc.). The processing steps of our 3D liver reference region identification method are illustrated in Fig. 4 on an axial slice of the 3D PET scan. After thresholding the PET scan with a threshold of $1.0 \text{ SUV}_{\text{bw}}$ in the liver ROI [Fig. 4(b)], holes in the binary image are closed using morphological operations [Fig. 4(c)], and a distance transformation¹⁷ is calculated [Fig. 4(d)]. The distance transformation assigns the distance value to the closest surface point to each location of the binary segmentation. The maximal hyperellipsoid inside the liver is determined in two steps. First, the center is found by finding the maximum point location in this distance transformed image [Fig. 4(d)]. Second, the axes lengths of the hyperellipsoid in x -, y -, and z -direction are calculated. All axes lengths are reduced by 1 cm to avoid the influence of partial volume effects and respiratory motion close to the boundary of the liver [Fig. 4(e)]. In this context note that a direct utilization of the thresholding result [Fig. 4(b) or 4(c)] for uptake measurement is not suitable,

FIG. 4. Processing steps of the 3D liver reference region identification approach depicted on a single axial PET image. (a) PET image slice showing the liver. (b) Thresholding. (c) Hole closing. (d) Distance transformation with maximum distance point marked. (e) 2D contour of the maximal hyperellipsoid with 1 cm distance to liver boundary. Note that all PET images are shown with inverted gray-scales.

because frequently kidneys or other structures would be included in the segmentation. In comparison, the proposed method avoids such problems, since only small portions of the liver surface are typically affected by adjoined structures.

II.C. Aortic arch reference region segmentation

Our approach for segmentation of a reference region in the aortic arch is based on a method for the extraction of tubular structures in CT data.^{18,19} First, tubular structures are identified in the CT scan of PET/CT dataset. Second, the carina is identified which then serves as a reference point for the aorta identification similar to the ideas of Feuerstein *et al.*⁹ The individual processing steps are outlined in Fig. 5 and will be described in detail in the following paragraphs.

II.C.1. Extraction of tubular structures

Inside the VOI of the aortic arch (Fig. 1) of the CT image [Fig. 5(a)], edge information is obtained by utilizing a Gaussian gradient magnitude filter at scale $\sigma = 1$ mm. Subsequently, a tube detection filter (TDF) (Refs. 18 and 19) is applied. This filter generates a tube-likeness measure together with a radius estimate at the centers of elongated objects with approximately circular cross sections. To avoid artifacts from calcifications in the aorta, the gray-value range in the CT dataset is truncated to a maximum value of 200

FIG. 5. Processing steps for the aortic arch reference region segmentation. (a) Volume rendering of the subregion in the CT dataset. (b) Identified tubular structures with trachea and main bronchi candidates (blue) and aortic arch candidates (red). (c) Identified trachea and main bronchi with carina and detected aortic arch. (d) Sagittal image slice with identified aortic arch.

HU. After applying the TDF, centerline-based representations are extracted for tubular structures by utilizing a height ridge traversal with hysteresis thresholding (upper threshold of 0.5 and a lower threshold of 0.1).¹⁸ The resulting extracted tubular structures are depicted in Fig. 5(b).

II.C.2. Aortic arch detection

From the detected tubular structures, candidate centerlines for the trachea, main bronchi, and the aorta are selected. Candidate centerlines for the trachea and main bronchi must have a maximum average radiodensity of -800 HU and a minimum average radius of 2 mm [blue centerlines in Fig. 5(b)]. The candidates for the aorta must have a minimum average radiodensity of 100 HU and a minimum average radius of 7 mm [red centerlines in Fig. 5(b)]. Among the trachea and main bronchi candidate centerlines, the trachea is identified as the tubular structure with the largest extent in z direction, ending at a bifurcation. This bifurcation point provides the location of the carina which we use as a landmark for the identification of the aortic arch. The spatial relation between aortic arch and carina is biologically very stable. The ascending aorta passes through an axial plane anterior to the carina and the descending aorta passes through an axial plane posterior to the carina, with the aortic arch in between connecting these parts. Thus, a predefined part of the aortic arch is selected as the centerline part passing through two planes with a predefined location relative to the carina. An example of the identified trachea/main bronchi and the resulting aortic arch are shown in Fig. 5(c). Figure 5(d) depicts the aorta with the estimated radius together with a sagittal image slice of the CT dataset.

In few datasets, the above described rule-based approach for aortic arch detection is not able to identify a tubular object following all those rules (e.g., due to too low image contrast or resolution). In these cases, no segmentation is performed and an error reported instead.

III. VALIDATION METHODOLOGY

III.A. Image data

One hundred and thirty four PET/CT scans from 49 subjects with head and neck cancer were acquired between one

and six times on either a Siemens Biograph Duo or Siemens Biograph 40 PET/CT scanner (Siemens Healthcare, Hoffman Estates, IL) between 2003 and 2010. All subjects were injected with $370 \text{ MBq} \pm 10\%$ of $[\text{F-18}] \text{ FDG}$ with an uptake time of $90 \text{ min} \pm 10\%$. In all cases, subjects were fasted for >4 h and had blood glucose <200 mg/dl. Patients were imaged with arms down, because of the interest in the head and neck region in this particular image dataset. The CT-based attenuation correction works equally well with arms up/arms down orientations, thus arm positioning should neither impact nor bias the methodology. All reconstructions were performed with 2D OSEM iterative algorithms. Reconstructions on the Biograph 40 were performed with 4 iterations, 8 subsets, and a 7 mm Gaussian filter and on the Biograph Duo with 2 iterations, 8 subsets, and a 5 mm Gaussian filter. Biograph Duo PET images were reconstructed onto a 128×128 pixel image matrix ($3.5 \text{ mm} \times 3.5 \text{ mm} \times 3.4 \text{ mm}$), while Biograph 40 PET images were reconstructed onto a 168×168 pixel image matrix ($3.4 \text{ mm} \times 3.4 \text{ mm} \times 2.0 \text{ mm}$). The reconstructed CT images from the Emotion Duo/Siemens Sensation 40 CT units associated with the Biograph Duo/Biograph 40 were reconstructed onto 512×512 pixel CT image matrices ($0.98 \text{ mm} \times 0.98 \text{ mm} \times 3.4/3.0 \text{ mm}$).

III.B. Independent reference standard and statistical analysis

III.B.1. Manual SUV measurement

To allow for a quantitative evaluation of the uptake measurement methods, human reviewers traced reference regions for cerebellum, liver, and aortic arch in axial, coronal, and sagittal slices of the datasets. For tracing, the human reviewers used a drawing tool, which allows users to specify polygonal lines to indicate the outline of the measurement region. The reviewers were able to navigate through the three-dimensional PET and CT datasets in all directions and to see the contours they drew in the other image slices. The default window/level used for visualization of the PET scan was $0\text{--}10$ SUV for the cerebellum and $0\text{--}6$ SUV for liver and aortic arch, but the users were able to adjust window/level as desired. The default window/level used for visualization of the CT scan was $-100\text{--}900$ HU.

In case of the cerebellum where our method obtains full 3D segmentations, the human reviewers were instructed to trace the exact outline of the cerebellum in four predefined slices per dataset: one axial, one coronal, and two sagittal slices (one in each hemisphere of the cerebellum). These slices were selected randomly inside the bounding box of the segmented cerebellum with a margin of 1 cm to the boundary to avoid problems due to partial volume effects. In case of the aortic arch and the liver, the human reviewers traced a reference region in one axial, one coronal, and one sagittal slice they selected for each of the measurement regions. The human reviewers were instructed to choose appropriate slices showing larger areas of the measurement region of interest and to trace the regions in such a way that boundary artifacts are avoided.

Two expert reviewers manually traced reference regions for the cerebellum, liver, and aortic arch in all 134 datasets. In the case of the cerebellum, two additional reviewers traced 44 and 20 datasets, respectively. The fewer cerebellum regions traced by the latter two reviewers were a random selection from the larger set and representative of those traced by the other reviewers. Statistical analysis methods were employed to accommodate differences in the numbers of regions traced.

On average, the segmentation of all three structures took 16.9 min (median: 10.35 min) per PET/CT scan. Average SUVs were computed for each manually segmented structure (i.e., over all traced slices) and utilized for comparison with averages calculated on automatically segmented volumetric regions. Note that for cerebellum segmentations, the average SUV was calculated over the four predefined slices (four-slice) described above, in addition to the average over the whole 3D structure (volume).

Once two or more manual SUV measurements were established by the reviewers for all three structures and datasets, two-stage regression models were used to compare results from the human reviewers to those from the automated methods, as described in Secs. III.B.2–III.B.4 below.

III.B.2. Consensus-true SUV model

Reviewer-based measurements were used to estimate (1) a consensus-true SUV for each image, (2) systematic differences between reviewers, and (3) within- and between-reviewer variability. Letting y_{ij} denote the measured value from image i and reviewer j , the first stage of the regression model can be expressed mathematically as

$$\begin{aligned} y_{ij} &= \tau_i + \gamma_j + \epsilon'_{ij} \\ \tau_i &\sim N(\mu, \sigma^2) \\ \gamma_j &\sim N(0, \sigma_{\text{reviewer}}^2) \\ \epsilon'_{ij} &\sim N(0, \sigma_{\text{manual}}^2), \end{aligned}$$

where τ_i are (normally distributed) image-specific true SUVs; μ and σ^2 are the SUV mean and variance in the patient population; γ_j are reviewer-specific mean effects; $\sigma_{\text{reviewer}}^2$ is the between-reviewer variance; and the ϵ'_{ij} are within-reviewer errors. Variability in manual SUV measurements is due to

both between- and within-reviewer variability. Thus, the sum of the two associated variances ($\sigma_{\text{reviewer}}^2 + \sigma_{\text{manual}}^2$) represents the total variability for the reviewer-based method.

III.B.3. Automated SUV model

In the second stage, agreement between automated SUV measurements and consensus-true SUV was assessed, and variability due to the automated method was estimated. Both were accomplished with the regression model

$$\begin{aligned} z_i &= \beta_0 + \beta_1 \tau_i + \epsilon''_i \\ \epsilon''_i &\sim N(0, \sigma_{\text{auto}}^2), \end{aligned}$$

where z_i is the automated measurement from image i ; β_0 and β_1 are intercept (shift) and slope (scale) parameters that define a linear relationship between automated and true SUV; and the ϵ''_i are residual errors associated with the automated calculation of SUV. Since automated results are deterministic, the residual error variance (σ_{auto}^2) represents the total variability for the method. The relative total variances in the automated and manual procedures can thus be estimated as

$$\lambda = \left(1 - \frac{\sigma_{\text{auto}}^2}{\sigma_{\text{reviewer}}^2 + \sigma_{\text{auto}}^2} \right) \times 100\%$$

and are interpreted as the percent by which the total variability in measured SUVs would be decreased if the automated method was used instead of the manual method.

III.B.4. Bayesian analysis

In the analysis, the stage-one and stage-two models were jointly specified as a Bayesian hierarchical model, and their parameters were estimated simultaneously. Vague prior distributions of $N(0,1000)$ for mean parameters and $\text{Gamma}(0.001,0.001)$ for variance parameters were specified and confirmed, via sensitivity analysis, to have negligible impact on posterior inference. OpenBUGS (Ref. 20) was used to generate 9000 Markov chain Monte Carlo simulated samples from the posterior distributions of model parameters. Three parallel chains of the sampler were generated by varying sampler starting values and were checked for convergence with the diagnostic of Gelman and Rubin.²¹ For inference, posterior means are reported along with 95% credible intervals (CrI) computed as the highest posterior density intervals.²² Credible intervals may be interpreted as containing associate model parameters with 95% probability.

IV. RESULTS

Examples of automated segmentation results are depicted in Fig. 6. The segmentation of all three reference regions in a PET/CT scan took about 7 min, on average. The automated aortic arch segmentation failed in ten cases and delivered wrong results (outliers) in three cases of patients with breathing tubes, which were excluded from further statistical analysis. Table I shows the average measured FDG uptake and the measurement volume in the three reference regions over all datasets. Figures 7–10 show the measured SUVs plotted

FIG. 6. Results of automatically identified reference measurement regions. Coronal (top row) and sagittal (bottom row) maximum intensity projections of PET scans with overlaid outlines of projected cerebellum, aortic arch, and liver regions. All PET scans are depicted as inverted maximum intensity projections using a gray-value range of 0–6 SUV. Note that in the majority of cases, only parts of the brain are imaged, which does not affect the ability of our method to locate all three structures. In the example shown in (d), even parts of the cerebellum are missing.

against the consensus-true SUVs estimated from the statistical analysis. In the case of cerebellum, Figs. 7 and 8 depict the plot for average SUVs calculated over all four slices, which were also traced manually by reviewers (Sec. III.B), and the whole volume, respectively. Different plotting symbols are used to identify measurements by reviewer and method (manual or automated). The solid lines represent the estimated relationships between automated measurements and the consensus truth, and the dashed lines the relationships that would exist between the two if there were no shift (intercept $\beta_0 = 0$) or scale (slope $\beta_1 = 1$) differences. Noteworthy, features of the plots include the relationships between the two lines and the amount of scatter about each. Differences between the lines would be indicative of systematic differences between manual and automated measurements. The relative scattering of each type of measurement about its line can help inform about the relative amounts of variability in the methods. More scattering in the manual measurements, than in the automated measurements, would indicate greater variability, and vice versa. The estimated percent changes in the slopes and total variability for the

automatic methods, relative to manual method, are presented in Tables II and III and correspond directly to the data presented in the figures.

In Figs. 7, 9, and 10, there do not appear to be substantial systematic differences between the methods. The solid and dashed lines are similar, and all associated 95% credible intervals include 0 for the slopes and 1 for the intercepts. The greatest systematic difference is seen in the comparison

TABLE I. Average FDG uptake and measurement volume with proposed methods.

Site	FDG uptake (SUV)	Measurement volume (ml)
Cerebellum	4.96 ± 1.01	160.27 ± 21.90
Aorta	1.70 ± 0.26	4.17 ± 1.02
Liver	2.01 ± 0.31	314.07 ± 123.40

FIG. 7. Scatterplot of measured vs reviewer consensus-true SUV in the four measurement slices for the cerebellum.

Fig. 8. Scatterplot of measured vs reviewer consensus-true SUV for the cerebellum.

of manual SUV measurements to automated volumetric SUV measurements in the cerebellum, where automated measurements are attenuated at the high end (Fig. 8). Estimates indicate that the automated measurements slope is 0.95 (95% CrI: 0.93–0.96) and is significantly different from 1 (that of the dashed line).

With respect to variability, the automated method produces the least amount of scatter about the lines in the cerebellum analyses. This is confirmed by the estimates in Table III showing total variability to be decreased by 99.2% (95% CrI: 98.0%–100%) and 89.8% (95% CrI: 80.6%–99.3%) in the two respective analyses. Variability is also decreased, although less so, in the aorta analysis by 76.7% (95% CrI:

Fig. 9. Scatterplot of measured vs reviewer consensus-true SUV for the aorta.

Fig. 10. Scatterplot of measured vs reviewer consensus-true SUV for the liver.

29.5%–100%). In the liver analysis, there is an estimated mean decrease of 54.7%, but the decrease is not significantly different from 0 (95% CrI: −9.1%–100%). Finally, estimates of the variances due to the automated method, between-reviewer mean differences, and within-reviewer measurement differences are summarized in the last three columns of the table.

V. DISCUSSION

The developed reference region segmentation methods are fully automated and showed good agreement with the independent standard generated from manual segmentations. There was no significant systematic shift or scaling difference, with exception of the slope, which occurred when the average cerebellum SUV was calculated for the whole volume, instead of the four specified cross sections. This can be explained as follows. The cerebellum is an inhomogeneous structure with different normal FDG uptake in its substructures (Fig. 3). These substructures have a characteristic, non-uniform distribution of FDG uptake that is volume averaged in the whole 3D cerebellum region, but are not proportionally sampled by the four cross sections. Because the independent reference was traced in the same four cross sections, it appears as if the automatic volumetric measurements are

TABLE II. Estimated intercept and slope for the linear relationship between automated SUV measurements and reviewer consensus-true SUV. Estimates are posterior means with 95% credible intervals given in parentheses.

Site	Auto	Intercept β_0 (SUV)	Slope β_1 (–)
Cerebellum	4-Slice	−0.03 (−0.34, 0.20)	0.99 (0.98, 1.01)
	Volume	0.04 (−0.24, 0.28)	0.95 (0.93, 0.96)
Aorta	Volume	−0.06 (−0.29, 0.14)	1.03 (0.97, 1.09)
Liver	Volume	0.02 (−0.23, 0.15)	0.97 (0.94, 1.00)

TABLE III. Estimated percent difference in total SUV measurement error variance (λ), as well as automated measurement error variance (σ_{auto}^2), variability between reviewers ($\sigma_{\text{reviewer}}^2$), and variability within reviewers (σ_{manual}^2). Estimates are posterior means with 95% credible intervals given in parentheses.

Site	Auto	λ (%)	σ_{auto}^2 (SUV ²)	$\sigma_{\text{reviewer}}^2$ (SUV ²)	σ_{manual}^2 (SUV ²)
Cerebellum	4-Slice	99.2 (98.0, 100)	0.0005 (0.0004, 0.0011)	0.0828 (0.0043, 0.2597)	0.0244 (0.0206, 0.0282)
	Volume	89.8 (80.6, 99.3)	0.0067 (0.0050, 0.0086)	0.0828 (0.0043, 0.2597)	0.0244 (0.0206, 0.0282)
Aorta	Volume	76.7 (29.5, 100)	0.0051 (0.0036, 0.0068)	0.1721 (0.0013, 0.4536)	0.0025 (0.0019, 0.0032)
Liver	Volume	54.7 (−9.1, 100)	0.0026 (0.0018, 0.0035)	0.0436 (0.0002, 0.1090)	0.0020 (0.0016, 0.0025)

minorly off in scale, even though the average SUV calculated on the volume is more representative of the whole cerebellum. The variability of measured FDG uptake is reported in Table I. In comparison to the cerebellum, liver and aortic arch show less intrinsic variation (lower standard deviation) in the scans utilized by this study.

An important factor in measuring uptake of an organ or structure is the size of the utilized region. This is demonstrated by an experiment on one PET/CT scan reported in Figs. 11 and 12. Eleven different regions (Fig. 11) were generated and utilized to measure/estimate uptake of the liver

(Fig. 12). In this context note that regions A, E1, and E2 directly relate to the proposed fully automated method, manual tracings of reviewer 1, and manual tracings of reviewer 2, respectively. All the other regions were manually generated for comparison. In general, larger reference regions allow a more robust estimate of the true mean. However, in practice (e.g., clinical trials), small circular or elliptic regions (e.g., regions C1–C4 in Fig. 12) are commonly utilized, because they are less time-consuming to generate manually. Small regions may lead to differing estimates, depending on where the region was placed. In contrast, the proposed method for

FIG. 11. 3D rendering of reference regions utilized for liver FDG PET uptake measurement comparison in Fig. 12. (PET) Coronal and sagittal images representing a volumetric PET scan in the liver region. (L) Volumetric liver segmentation. (L*) Same as L, but the volume was eroded by a 1 cm margin. (A) Proposed automated method. (E1) Combination of axial, coronal, and sagittal liver slices segmented by reviewer 1. (E2) Combination of axial, coronal, and sagittal liver slices segmented by reviewer 2. (E1ax) Single axial liver slice segmented by reviewer 1. (E2ax) Single axial liver slice segmented by reviewer 2. (C1) Cylindrical liver region over five slices in axial direction. (C2) Circular region located in the middle of C1. (C3) Circular region located at the bottom of C1. (C4) Circular region located at the top of C1.

FIG. 12. Example of SUV variation in dependence of the reference regions depicted in Fig. 11. Bars and lines represent the mean and standard deviation of SUVs in the segmented region, respectively.

liver FDG uptake measurement is fully automated and utilizes a large volume (Table I) to estimate the FDG uptake in the liver. Also, when large regions are utilized, partial volume effects at the object boundary need to be considered to avoid estimation errors (region L in Fig. 12).

All three segmentation methods are fully automated and generate reproducible results (i.e., zero intraobserver variability). Compared to the manual approach, the estimated average SUV measurement error variance was reduced between 54.7% and 95.2%. While this reduction was statistically significant for cerebellum and aortic arch, it was not found to be statistically significant in the case of the liver. In this context, it is interesting to note that the variability between reviewers ($\sigma_{\text{reviewer}}^2$) and variability within reviewers (σ_{manual}^2) are the lowest for the liver. Reasons for both observations might be that large regions inside the liver can be quite easily defined and the low variability of SUVs in the liver.

The human reviewers had detailed instruction on how to generate the manual segmentations and were required to adhere to certain constraints (e.g., preselected cross sections to outline the cerebellum). Thus, it seems very likely that inter- and intraobserver variability will be even higher without such instructions and constraints.

Manual segmentation of all three reference regions took on average 17 min of reviewer tracing time per dataset. In contrast, the proposed fully automated approach completes on average in less than 7 min and can be further reduced by means of code optimization and parallel computing, if needed.

The proposed method for segmentation of the aortic arch failed to produce correct results in 3 cases of 134 (2%). An investigation showed that in all these cases, patients had breathing tubes, which lead to an incorrect identification of the bifurcation of the trachea into main bronchi that is required by the algorithm to correctly locate the aortic arch. We plan to address this issue by incorporating a whole-body segmentation method into the algorithm, which will allow identifying the neck region and, therefore, excluding this area as a possible location for airway bifurcations. Also, the algorithm did not succeed in detecting the aortic arch in ten cases and terminated due to low contrast of aorta to surrounding tissue and low image resolution of the CT data of the PET/CT scan.

A potential application of the proposed methods is the normalization of tumor uptake with respect to a reference region in the context of treatment response assessment. This is an important step for practical application of quantitative imaging for response assessment in oncology clinical trials. In the presented study, we have demonstrated the feasibility of automated and accurate uptake measurement in cerebellum, liver, and aortic arch, all of which are potentially promising candidates. However, at this stage it is not clear which of the three regions, if at all, is best suited for this task. The pairwise correlation analysis shown in Fig. 13 suggests that measured average SUVs for aortic arch and liver correlate more ($r = 0.7$) than cerebellum and aortic arch ($r = 0.52$) or cerebellum and liver ($r = 0.5$). Thus, future research will focus on investigating this issue.

FIG. 13. Pairwise correlation between automated volumetric SUV measurements.

FIG. 14. Examples of segmentation results produced with our algorithm on PET/CT scan with an imaging protocol requiring arms up.

In the current implementation, fixed (generous) SUV_{bw} thresholds for liver and brain (Sec. II) segmentation are utilized, because the approach is geared toward automated reference region segmentation for (tumor) uptake normalization under the assumption of limited image acquisition and SUV_{bw} normalization errors. Additional processing will be required to handle major errors in SUV_{bw} calculation or in quality control applications, where PET image values can be off by a large margin. For example, a histogram analysis could be utilized to adapt all threshold values utilized by our method to a particular scan. Also, plausibility checks could be added to automatically identify completely failed PET/CT image acquisitions.

The presented approach was developed for a head and neck cancer specific imaging protocol where patients were imaged with arms down. We have also successfully utilized the presented algorithm to segment reference regions in PET/CT scans where patients had their arms up during scanning (Fig. 14), even though the algorithm was not developed

for such an imaging protocol. Also, many existing clinical oncology related PET/CT imaging protocols image the cerebellum only partially, if at all. The selected robust ASM matching approach (Sec. II.A.3) is well suited to segment partially imaged target objects, as demonstrated by Sun *et al.*¹⁶ and the example is depicted in Fig. 6(d). In addition, the developed method could be adapted to PET imaging studies of the brain where the cerebellum is often used as a reference region. In the case the brain/cerebellum is not imaged at all, the ROI generation described in Sec. II can be easily adapted accordingly, because the ROIs are quite large and only provide rough information regarding the location of an organ/structure of interest.

VI. CONCLUSION

We have presented a fully automated approach for FDG uptake measurement of cerebellum, liver, and aortic arch in full-body FDG PET/CT scans. Average SUVs were derived from volumetric regions inside all three structures, which were identified by means of automatic segmentation methods. The validation of our approach on 134 FDG PET/CT scans of head and neck cancer patients showed good agreement with the manually generated independent standard. In addition, the proposed method was found to have lower estimated variability compared to manual segmentation. Our fully automated approach requires no user interaction, completes on average in less than 7 min of unattended computing time, and offers reproducible definition of reference regions for PET image normalization. In contrast, manual segmentation of all three reference regions took approximately 17 min of reviewer tracing time per dataset and had greater variability. In the near future, we plan to evaluate several lesion-to-background ratios based on these reference regions to see if they provide a meaningful alternative for uptake normalization in the context of assessing response to therapy in conjunction with an established clinical outcomes database.

ACKNOWLEDGMENTS

This work was supported in part by NIH/NCI Grant No. U01CA140206, NIH/NIBIB Grant No. R01EB004640, and NIH/NHLBI Grant No. R01HL111453.

^{a)}Author to whom correspondence should be addressed. Electronic mail: reinhard-beichel@uiowa.edu

¹M. M. Graham, L. M. Peterson, and R. M. Hayward, "Comparison of simplified quantitative analyses of FDG uptake," *Nucl. Med. Biol.* **27**(7), 647–655 (2000).

²S. Obrzut, R. H. Pham, D. R. Vera, and C. K. Hoha, "Comparison of lesion-to-cerebellum uptake ratios and standardized uptake values in the evaluation of lung nodules with 18F-FDG PET," *Nucl. Med. Commun.* **28**(1), 7–13 (2007).

³M. A. Blake, J. M. Slattery, E. F. Halpern, A. J. Fischman, P. R. Mueller, and G. W. Boland, "Adrenal lesions: Characterization with fused pet/ct image in patients with proved or suspected malignancy-initial experience," *Radiology* **238**(3), 970–977 (2006).

⁴C. Li, X. Wang, Y. Xia, S. Eberl, Y. Yin, and D. D. Feng, "Automated PET-guided liver segmentation from low-contrast CT volumes using probabilistic atlas," *Comput. Methods Programs Biomed.* 3565–3568 (2011).

⁵T. Heimann, B. van Ginneken, M. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. Binnig, H. Bischof, A. Bornik, P. Cashman, Y. Chi, A. Cordova, B. Dawant, M.

- Fidrich, J. Furst, D. Furukawa, L. Grenacher, J. Hornegger, D. Kainmuller, R. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, R. Li, S. Li, H. P. Meinzer, G. Nemeth, D. Raicu, A. M. Rau, E. van Rikxoort, M. Rousson, L. Rusko, K. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J. Waite, A. Wimmer, and I. Wolf, "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Trans. Med. Imaging* **28**(8), 1251–1265 (2009).
- ⁶H. Ling, S. Zhou, Y. Zheng, B. Georgescu, M. Suehling, and D. Comaniciu, "Hierarchical, learning-based automatic liver segmentation," in *Computer Vision and Pattern Recognition (CVPR)* (IEEE Computer Society, Anchorage, Alaska, 2008), pp. 1–8.
- ⁷F. van der Lijn, M. de Bruijne, S. Klein, T. den Heijer, Y. Hoogendam, A. van der Lugt, M. Breteler, and W. Niessen, "Automated brain structure segmentation based on atlas registration and appearance models," *IEEE Trans. Med. Imaging* **31**, 276–286 (2011).
- ⁸Y. Xia, S. Eberl, M. Fulham, and D. D. Feng, "Dual-modality brain PET-CT image segmentation based on adaptive use of functional and anatomical information," *Comput. Med. Imaging Graph.* **36**(1), 47–53 (2012).
- ⁹M. Feuerstein, T. Kitasaka, and K. Mori, "Automated anatomical likelihood driven extraction and branching detection of aortic arch in 3-D chest CT," in *International Workshop on Pulmonary Image Analysis. Medical Image Computing and Computer Assisted Intervention* (London, UK, 2009), pp. 49–60.
- ¹⁰T. F. Cootes, D. Cooper, C. J. Taylor, and J. Graham, "Active shape models—Their training and application," *Comput. Vis. Image Underst.* **61**(1), 38–59 (1995).
- ¹¹W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *ACM SIGGRAPH Comput. Graph.* **21**(4), 163–169 (1987).
- ¹²T. Heimann, I. Wolf, T. Williams, and H. P. Meinzer, "3D active shape models using gradient descent optimization of description length," in *Proceedings of IP MI*, Springer, Heidelberg (Springer, Glenwood Springs, CO, 2005), Vol. 3565, pp. 566–577.
- ¹³J. J. Koenderink and A. J. van Doorn, "Surface shape and curvature scales," *Image Vis. Comput.* **10**, 557–565 (1992).
- ¹⁴*Statistical Shape Analysis*, edited by I. L. Dryden and K. V. Mardia (Wiley, New York, 1998).
- ¹⁵S. Bouix, K. Siddiqi, and A. Tannenbaum, "Flux driven automatic center-line extraction," *Med. Image Anal.* **9**(3), 209–221 (2005).
- ¹⁶S. Sun, C. Bauer, and R. Beichel, "Automated 3-D segmentation of lungs with lung cancer in CT data using a novel robust active shape model approach," *IEEE Trans. Med. Imaging* **31**(2), 449–460 (2012).
- ¹⁷C. R. Maurer, R. Qi, and V. Raghavan, "A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions," *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(2), 265–270 (2003).
- ¹⁸C. Bauer and H. Bischof, "Extracting curve skeletons from gray value images for virtual endoscopy," in *Workshop on Medical Imaging and Augmented Reality* (Springer, Tokyo, Japan, 2008), pp. 393–402.
- ¹⁹C. Bauer, H. Bischof, and R. Beichel, "Segmentation of airways based on gradient vector flow," in *International Workshop on Pulmonary Image Analysis. Medical Image Computing and Computer Assisted Intervention* (London, UK, 2009), pp. 191–201.
- ²⁰D. J. Lunnon, A. Thomas, N. Best, and D. Spiegelhalter, "WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility," *Stat. Comput.* **10**, 325–337 (2000).
- ²¹A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Stat. Sci.* **7**, 457–511 (1992).
- ²²M. H. Chen and Q. M. Shao, "Monte Carlo estimation of Bayesian credible and HPD intervals," *J. Comput. Graph. Stat.* **8**(1), 69–92 (1999).