Combining multiple FDG-PET radiotherapy target segmentation methods to reduce the effect of variable performance of individual segmentation methods

Ross J. McGurk^{a)}

Medical Physics Graduate Program, Duke University, Durham, North Carolina 27705

James Bowsher

Department of Radiation Oncology, Duke University Medical Center, Durham, North Carolina 27705

John A Lee

Molecular Imaging and Experimental Radiotherapy Unit, Université Catholique de Louvain, 1200 Brussels, Belgium

Shiva K. Das

Department of Radiation Oncology, Duke University Medical Center, Durham, North Carolina 27705

(Received 3 December 2012; revised 16 January 2013; accepted for publication 11 February 2013; published 11 March 2013)

Purpose: Many approaches have been proposed to segment high uptake objects in 18F-fluoro-deoxy-glucose positron emission tomography images but none provides consistent performance across the large variety of imaging situations. This study investigates the use of two methods of combining individual segmentation methods to reduce the impact of inconsistent performance of the individual methods: simple majority voting and probabilistic estimation.

Methods: The National Electrical Manufacturers Association image quality phantom containing five glass spheres with diameters 13–37 mm and two irregularly shaped volumes (16 and 32 cc) formed by deforming high-density polyethylene bottles in a hot water bath were filled with 18-fluoro-deoxyglucose and iodine contrast agent. Repeated 5-min positron emission tomography (PET) images were acquired at 4:1 and 8:1 object-to-background contrasts for spherical objects and 4.5:1 and 9:1 for irregular objects. Five individual methods were used to segment each object: 40% thresholding, adaptive thresholding, k-means clustering, seeded region-growing, and a gradient based method. Volumes were combined using a majority vote (MJV) or Simultaneous Truth And Performance Level Estimate (STAPLE) method. Accuracy of segmentations relative to CT ground truth volumes were assessed using the Dice similarity coefficient (DSC) and the symmetric mean absolute surface distances (SMASDs).

Results: MJV had median DSC values of 0.886 and 0.875; and SMASD of 0.52 and 0.71 mm for spheres and irregular shapes, respectively. STAPLE provided similar results with median DSC of 0.886 and 0.871; and median SMASD of 0.50 and 0.72 mm for spheres and irregular shapes, respectively. STAPLE had significantly higher DSC and lower SMASD values than MJV for spheres (DSC, $p < 0.0001$; SMASD, $p = 0.0101$) but MJV had significantly higher DSC and lower SMASD values compared to STAPLE for irregular shapes (DSC, $p < 0.0001$; SMASD, $p = 0.0027$). DSC was not significantly different between 128×128 and 256×256 grid sizes for either method (MJV, $p = 0.0519$; STAPLE, $p = 0.5672$) but was for SMASD values (MJV, $p < 0.0001$; STAPLE, $p = 0.0164$). The best individual method varied depending on object characteristics. However, both MJV and STAPLE provided essentially equivalent accuracy to using the best independent method in every situation, with mean differences in DSC of 0.01–0.03, and 0.05–0.12 mm for SMASD.

Conclusions: Combining segmentations offers a robust approach to object segmentation in PET. Both MJV and STAPLE improved accuracy and were robust against the widely varying performance of individual segmentation methods. Differences between MJV and STAPLE are such that either offers good performance when combining volumes. Neither method requires a training dataset but MJV is simpler to interpret, easy to implement and fast. © 2013 American Association of Physicists in Medicine. [<http://dx.doi.org/10.1118/1.4793721>]

Key words: PET segmentation, majority vote, STAPLE, Dice coefficient

I. INTRODUCTION

Segmentation of metabolically active regions of tumors for radiation therapy treatment planning on functional images,

either for refining the tumor margins or for dose escalation studies is an active area of research.^{1–4} Accurate and reproducible delineation of the tumor volume has several benefits: The potential to reduce local failure in radiotherapy caused by

geographic misses; reduction in intra- and interobserver variation; accurate and reproducible treatment response evaluation. However, the low spatial resolution and noise characteristics of positron emission tomography (PET) images complicate the segmentation process.

Many segmentation approaches have been studied to date. The simplest include thresholding techniques where pixels with more than 40%–50% of the maximum pixel intensity in a region-of-interest (ROI) are included in the image.^{5,6} While simple and fast, they fail to accurately segment small lesions and struggle with low contrast.⁷ Thresholding images based on specific uptake values (SUV) has also been proposed but SUV is influenced by time from injection to imaging, patient size, plasma glucose levels, and the partial volume effect.⁸ Methods that rely on calibration curves iteratively derived from signal to background measurements can increase the segmentation accuracy but are scanner specific.^{6,9} Clustering approaches such as k-means (KM) or fuzzy C-means have also been implemented in segmenting nuclear medicine images.¹⁰ K-means based methods segment PET images into two regions (lesion and background) by initially choosing a cluster center and then iterating to minimize the within-cluster sum-of-squares. This approach tends to be sensitive to the initial choice of cluster center selection and noise in the image. Fuzzy C-mean algorithms replace the hard decision of k-means by allowing voxels to belong to multiple clusters via a membership matrix and similarity criteria, e.g., minimizing the Euclidian distance between each element and the cluster mean. However, they are typically initialized with greater numbers of clusters which are later merged. Thus, the accuracy of the segmentation can be affected by the choice of merging criteria. Seeded region growing methods rely on the initial selection of a seed voxel(s) and searching neighboring voxels.¹¹ Neighboring voxels that satisfy some acceptance criteria are added to the region and the region continues to grow until some termination condition is met. One advantage of region growing includes the flexible acceptance conditions that can be adopted. However, such methods tend to be slow compared to thresholding techniques. Finally, other methods such as Bayesian and Markov chain segmentation or gradient and clustering-type approaches have shown promising results.^{12–15}

To date, image segmentation research efforts have attempted to create a single method that performs well over a wide variety of lesion sizes, contrasts, and image noise levels. This challenging goal has yet to be achieved. However, combining several individual methods, based on consensus methods used in other fields^{16–20} may improve the delineation of lesions on PET images. Not only would improvements in segmentation accuracy and robustness aid target delineation for radiation therapy, they may offer improvements in determining treatment response. To date, the most common determination of treatment effectiveness is via volumetric measurements.²¹ However, changes in a patient's SUV values during or following radiation therapy could offer a more complete characterization of treatment efficacy.^{22–24} This application is of interest because tumor sizes may remain constant while SUV values decrease. In addition to this

effect, necrosis or other enhancement in SUV values could be missed using a purely volumetric approach.

The aim of this study is to investigate how segmentation methods can be combined to increase lesion segmentation accuracy. Importantly, this study determines whether combining methods may be used to create a segmentation that is robust to the variable performance of individual segmentation methods (i.e., individual segmentation methods that perform well in some situations but not others). The independent individual segmentation methods used in this study are: 40% thresholding, adaptive thresholding (ADP), k-means clustering, seeded region growing (SRG), and clustering with a watershed transform (WSC). Two techniques for combining the segmentation methods are studied—one developed here using a majority voting concept and another previously used in MRI imaging studies and other computer vision problems.²⁵

II. MATERIALS AND METHODS

II.A. Phantom scans

The National Electronic Manufacturers Association (NEMA) image quality phantom was used to acquire images of five spherical volumes ranging from 1.1 to 26.5 cc. A total of 225 MBq of 18F-fluoro-deoxy (FDG) was used. Each volume was filled with a mix of FDG and 5% Gastrographin® CT contrast agent. The contrast agent allowed the inner surface of each volume to be contoured on a high resolution CT scan (1 mm axial slices using a grid size of 512×512 and 0.975×0.975 mm pixel size) performed prior to the emission scan. For the emission scan, two contrasts (8:1 and 4:1) were used with the background activity concentration set to approximately 5.7 MBq/ml to mimic that previously observed in head-and-neck patients.²⁶ A Siemens Biograph mCT PET/CT scanner was used to acquire 30 min of list-mode data. The list-mode data were then rebinned such that five-image series were obtained for 5-min scan durations. Image volumes were reconstructed on a grid size of 128×128 ($4.2 \times 4.2 \times 5$ mm) voxel dimensions and 256×256 ($2.1 \times 2.1 \times 2$ mm voxel dimensions). The Ordered Subsets Expectation Maximization (OSEM) algorithm with four iterations and 21 subsets was used with the time-of-flight (TOF) information.

We repeated the above experiment with the spherical inserts replaced with two irregularly shaped high-density polyethylene (HDPE) plastic vessels (Fig. 1). The irregular shape was created by deforming the vessels in a hot water bath. The plastic vessels were glued to long plastic rods mounted on a Styrofoam™ base, which was positioned within the phantom. The volumes of the vessels were determined to be 16 and 31 cc using volume displacement. The procedure used to fill the spherical volumes described above was repeated this time with 185 MBq of FDG of total activity. The filling procedure resulted in average signal to background (S:B) ratios of 9:1 and 4.5:1 at similar background activity concentrations. Images were reconstructed on the same 256×256 grid size as above.

FIG. 1. (a) The two irregular HDPE plastic volumes mounted on polystyrene base prior to inserting into NEMA IQ phantom. (b) In place within the NEMA image quality phantom.

II.B. ROI definition

The PET and high-resolution CT image volumes were imported into the VelocityAI image registration software (Version 2.8.1, Velocity Medial Solutions, Atlanta, GA). The inside surfaces of each of the spherical and irregular volumes were contoured on the high-resolution CT images using the smart brush tools and refinements made by hand. A larger 3D spherical ROI was also constructed to encompass each volume and used to define the region on the images within which the segmentation methods were confined. The high-resolution CT contours were resampled to match the resolution of the PET images because each segmentation method works on the PET images directly. Down-sampling the contours from the high-resolution CT to the PET means there is no interpolation of PET image values, and allows for faster segmentation. The 10 mm (0.52 ml) sphere was excluded from analysis as it was not visible in many of the images. Further to this, there is extensive ongoing discussion in the literature about whether objects with glass walls are suitable for segmentation studies.^{27,28} Given that the spherical inserts have wall thicknesses of 1.2 mm, the wall thickness for the 10 mm sphere represents almost 50% of the total insert volume. The inclusion of this sphere would further increase bias via the production of smaller volumes by some segmentation methods.²⁹ Also observed was the incomplete filling of the 28 mm (11.5 ml) sphere. The incomplete filling of the 28 mm sphere should not cause any issues with our analysis because its measured volume (11 cc) was still close to the theoretical value (11.5 cc) and distinctly different from the next smallest and next largest objects which had volumes of 3.6 and 26.5 cc, respectively.

II.C. Segmentation methods and combination

Five segmentation methods were used to segment each object: 40% of the maximum voxel intensity within the ROI (40%); adaptive thresholding using the method described in Tyłski *et al.*³⁰ (ADP); the standard k-means algorithm implemented in MATLAB® (Natick, MA) to partition the image into tumor and background (KM); a seeded region-growing method modified from the work of Li *et al.*³¹ (SRG); and the gradient, watershed transform and clustering method of Geets *et al.*¹⁴ (WSC). The modifications to the seeded region-growing of Li *et al.* include (1) a sorting step where the voxel intensities within the current region are sorted before under-

going the acceptance criteria testing and (2) depending on the grid size, once the region growing process has stopped, all uncovered voxels (i.e., the outer shell of the volume) are removed instead of the dual-front contouring method presented in their work. The sorting step acts to prevent the region mean from decreasing quickly and leading to the premature termination of region-growing process when segmenting low contrast objects. The removal of uncovered voxels is based on our observations that the initial region-growing step generally overestimates the boundaries of objects. This phenomenon was also present in results of Li *et al.* Further, their dual front contouring refinement led to an underestimation of the object size, particularly for small objects of low contrast on a 128×128 images. For our 128×128 images, the removal of the outer shell also resulted in volume underestimation, sometimes even completely removing the smallest objects, lowering Dice similarity coefficient (DSC) and increasing symmetric mean absolute surface distance (SMASD) values and so was not used. However, for objects reconstructed on a 256×256 matrix size, the removal of the outer shell results in higher DSC and lower SMASD values for the segmented objects.

For the 40%, ADP, KM, and SRG methods, a 5-mm Gaussian filter commonly seen in the literature was applied to the reconstructed images.^{32–34} For the gradient-based method a 7-mm bilateral filter was applied and the deblurring step was accomplished with 30 iterations of Landweber's algorithm with a symmetrical 6 mm FWHM point spread function (PSF) in accordance with Geets *et al.*¹⁴ The choice of a 7-mm bilateral filter was based on the method outlined by Hofheinz *et al.*³⁵ as it provides comparable signal-to-noise in background regions to the 5 mm Gaussian filter used for the other segmentation methods.

The five individual segmentation methods were combined using two strategies: (1) a majority vote combination where if a voxel is segmented by a majority of the methods, i.e., three or more, it is included in the final volume and (2) the Simultaneous Truth And Performance Level Estimate (STAPLE) algorithm.²⁵ STAPLE takes a collection of segmentations as input and produces a continuous probability map which estimates the likelihood of each voxel belonging to a given class (in our case tumor or background). A maximum likelihood segmentation can be produced for each voxel by selecting the label that maximizes the probability for that voxel based on this probability map. Alternatively a fixed threshold (e.g., 75%) can also be specified. The second approach results in segmentations where only voxels with a probability of belonging to the tumor label greater than the specified threshold are included. In addition, STAPLE provides a performance estimate for segmentations in terms of sensitivity and specificity. It also allows for spatial homogeneity constraints to be incorporated in order to refine the estimate of the probabilistic estimate of the true segmentation using Markov Random Fields. This is done via an interaction neighborhood and strength parameter β . Our implementation of STAPLE used the voxel-wise maximum likelihood estimation method and a value of $\beta = 1.0$ as a spatial homogeneity constraint chosen based on performance testing with our data.

FIG. 2. (a). Individual and consensus segmentations for the 13 mm sphere imaged at 4:1 contrast. (b) Segmentation results for the 16 cc irregular volume imaged at 8:1 contrast. Contours are overlaid on subimages extracted from the 256×256 grid sized images filtered using 5 mm FWHM Gaussian filter.

II.D. Quantitative metrics

Two metrics were selected due to their common use in the literature. The first was the DSC which provides a measure of spatial overlap between a segmented volume, S , and the defined ground truth, T .

$$\text{DSC} = 2 \frac{|S \cap T|}{|S| + |T|}.$$

The second metric is the SMASD which is an average distance of agreement between the surfaces of two volumes.

$$\text{SMASD [mm]} = \frac{1}{n_S + n_T} \left(\sum_{i=1}^{n_S} |d_i^{ST}| + \sum_{j=1}^{n_T} |d_j^{TS}| \right),$$

where n_S and n_T are the number of surface voxels on the segmented and true volumes, respectively and d_i^{ST} is the distance to the closest voxel on T for the i th surface voxel on S and similarly, d_j^{TS} is the closest voxel on S for the j th surface voxel on T .¹⁵ MATLAB was used for the calculation of DSC and SMASD values which were saved for statistical analysis.

II.E. Statistical analysis

Results for each metric were pooled according to the object type (spherical and irregular). Due to non-normality and heteroscedasticity, the nonparametric Wilcoxon signed-rank test was used to determine if DSC and SMASD were significantly different between the MJV and STAPLE consensus volumes across all object sizes, contrasts, and scan durations. The Wilcoxon rank-sum test was used to test whether differences in image size were significant using DSC and SMASD values from the spherical volumes. All statistical tests were

carried out using JMP (Version 10.0, SAS Institute Inc., Cary, NC).

III. RESULTS

Figure 2 shows both the individual segmentation methods, together with the MJV and STAPLE consensus volumes for a small (13 mm diameter), low-contrast (4:1), sphere [Fig. 2(a)] and the 16 cc, high-contrast (8:1), irregular shape [Fig. 2(b)]. The ground truth volumes defined using the high-resolution CT are also shown and have been down sampled to match the grid size. All contours are overlaid on images filtered with the 5 mm FWHM Gaussian filter for display purposes only. The adaptive thresholding (ADP), k-means, seeded region growing (SRG), and gradient-based watershed (WSC) methods all produce plausible segmentations of the sphere, while the 40% thresholding method results in numerous false positive regions outside of the known object. For the high-contrast irregular object, all segmentation methods provide volumes that closely resemble the contour generated from the high-resolution CT. The exception is the SRG method which noticeably overestimates the object edge compared to the CT contour. Importantly, both majority vote (MJV) and STAPLE segmentations appear robust to the false positives generated the 40%, and overestimation of the edges by other methods.

Figure 3 shows the overall performance of the independent segmentation methods, and the two consensus methods in terms of their DSC [Fig. 3(b)] and SMASD [Fig. 3(b)]. For both spherical and irregular shapes, all methods either exceed or are close to a mean DSC of 0.7 which indicates good quality segmentations.^{36–38} In addition, all mean SMASD values are either below or close to below the 2 mm

FIG. 3. (a) Performance metrics of Dice similarity coefficient (DSC) and (b) symmetric mean absolute surface distance (SMASD) for the independent and consensus segmentations for both spheres and irregular shapes.

pixel size of the highest resolution images indicating they are achieving subpixel accuracy on average. Figure 3 also shows that both MJV and STAPLE display greatly improved DSC and SMASD values compared to the worst performing methods and are close to the best performing methods in all cases.

The median and interquartile range for DSC of the spherical objects were 0.866 [0.807, 0.924] (with mean and standard deviation of 0.860 ± 0.083) for MJV contours and 0.886 [0.819, 0.927] (with mean and standard deviation 0.872 ± 0.068) for the STAPLE contours. However, the minimum DSC values for the spheres was 0.594 for MJV compared to 0.718 for STAPLE. While the difference was small, STAPLE had significantly higher DSC values compared to MJV ($p < 0.0001$). For the irregular shapes, the median and interquartile ranges were 0.875 [0.870, 0.889] (with mean and standard deviation of 0.875 ± 0.015) for MJV and 0.871 [0.868, 0.889] (with mean and standard deviation 0.872 ± 0.016) for STAPLE. The minimum DSC was 0.842 for MJV and 0.843 for STAPLE. Again, while differences were small, MJV had significantly higher DSC values compared to STAPLE ($p < 0.0001$).

The median and interquartile range for SMASD of the spherical objects were 0.520 mm [0.434 mm, 0.658 mm] (with mean and standard deviation of 0.629 ± 0.305 mm) for MJV and 0.503 mm [0.426 mm, 0.686 mm] (with mean and standard deviation of 0.571 ± 0.209 mm) for STAPLE. Here STAPLE had significantly smaller SMASD values compared to MJV ($p = 0.0101$). For the irregular shapes the median and interquartile ranges were 0.708 [0.692, 0.763] (with mean and standard deviation of 0.748 ± 0.117 mm) for MJV and 0.718 [0.693, 0.762] (with mean and standard deviation of 0.767 ± 0.129 mm) for STAPLE. SMASD values for MJV were now significantly smaller than STAPLE ($p = 0.0027$).

The achievable segmentation accuracy can be affected by the reconstructed voxel size. We grouped the DSC and SMASD values corresponding to the spheres according to the reconstructed grid size. For MJV, mean and standard deviations of DSC for the 128×128 and 256×256 grid sizes were 0.847 ± 0.080 and 0.873 ± 0.085 , respectively. This difference was not significant ($p = 0.0519$). For STAPLE, mean DSC values were 0.871 ± 0.065 and 0.874 ± 0.071 for 128×128 and 256×256 , respectively. This difference was also not significant ($p = 0.5672$). For MJV, mean values for SMASD were 0.723 ± 0.287 and 0.534 ± 0.295 mm between the 128×128 and 256×256 grid sizes, respectively. This difference was significant ($p < 0.0001$). Finally, the mean SMASDs for STAPLE were 0.621 ± 0.228 and 0.522 ± 0.176 mm which was also significant ($p = 0.0164$). Interesting, the mean DSC and SMASD values both indicate that for the 128×128 grid size STAPLE offers better performance compared to MJV but at the 256×256 grid size, both offer equally good performance.

Given that both MJV and STAPLE display much better performance compared to the worst performing method and comparable performance relative to the best performing method, an obvious question is why not use the best performing method all the time? The answer is that in the absence of a ground truth we simply do not know *a priori* which method will be the best. Certain techniques perform favorably under different circumstances and if a given voxel is segmented by the majority of techniques, the likelihood of it being a false positive decreases. One way of investigating whether or not it is viable to simply use the best-performing method all the time is to rank the DSC and SMASD values for the five methods when they are presented different circumstances. Therefore, if one individual method was observed to consistently outrank the others, then using a consensus

FIG. 4. Proportion of #1 DSC and SMASD rankings of the individual segmentation methods across multiple imaging and object characteristics.

approach may not offer any improvement over using this “best” method.

To do this the DSC and SMASD values were grouped according to different contrast and object sizes for the spheres and irregular shapes, separately. The independent segmentation methods were ranked according to highest DSC and lowest SMASD. The number of times each method was ranked first in a comparison was recorded and this number was normalized by the total number of comparisons made. The distribution of method rankings is shown in Fig. 4. Here we see that while the SRG or WSC do not receive a top ranking, the best method in any given comparison changes. Therefore, no one method is consistently performing better than another.

To further explore and emphasize why using a consensus method is a good option, we created a new distribution consisting of the highest DSC values from the five independent methods. In other words, this is the distribution of DSC values you would observe for all object sizes and contrasts if you knew *a priori* which method was going to be best out of the five before applying it. For spheres, the mean DSC for this new distribution was 0.885 ± 0.056 , compared to 0.860 ± 0.082 for MJV (mean difference = 0.025) and 0.872 ± 0.068 for STAPLE (mean difference = 0.013), respectively. For irregular shapes, the mean DSC of the best distribution was 0.885 ± 0.010 compared to 0.875 ± 0.015 for MJV (mean difference = 0.010) and 0.872 ± 0.016 for STAPLE (mean difference = 0.013).

For spheres, the mean SMASD for the best distribution (i.e., lowest SMASD values) was 0.52 ± 0.20 mm compared to 0.63 ± 0.30 mm for MJV (mean difference = 0.11 mm) and 0.57 ± 0.21 mm for STAPLE (mean difference = 0.05 mm). For irregular shapes, the mean of the best independent distribution was 0.64 ± 0.04 mm compared to 0.75 ± 0.12 mm for MJV (mean difference = 0.11 mm) and 0.77 ± 0.13 mm for STAPLE (mean difference = 0.12 mm). The mean differences are below the size of even the smallest pixel used in this study. This, together with mean DSC differences approximately 1% of the mean value of the best independent method indicate that both MJV and STAPLE provide performance essentially equivalent to picking the best method for every situa-

FIG. 5. The percentage of MJV volume made up of from each independent method across all object sizes, contrasts, and image grid size.

tion in this study. However, with Fig. 4 showing that knowing *a priori* which method was best under every circumstance in this study was impossible; the use of a consensus method is the best possible choice.

Finally, Fig. 5 shows the percentage of the MJV consensus volume segmented by each individual method in order to show that each method is contributing to the final MJV consensus volume. We see for both spheres and irregular shapes, all methods are contributing to more than 80% of the final MJV volume over all object sizes and contrasts used in this study. However, given the proportions only represents the makeup of volumes within the MJV volume, without reference to the ground truth, they cannot be directly compared with the results of Fig. 3 which does use the ground truth to compute the DSC and SMASD. In general though, we found that the MJV volumes were slightly larger than the ground truth and therefore higher proportions in Fig. 5 would represent lower DSC and higher SMASD values, consistent with Fig. 3.

IV. DISCUSSION

The work presented here outlines the performance of two consensus methods which combine a collection of individual segmentation methods used in segmenting objects from PET images under a variety of object sizes, shapes, contrasts, and image noise levels. In addition to helping define radiation therapy targets, the use of consensus volumes potentially provides a more robust approach to quantifying treatment response. Issues surrounding the use of volumetric measures to characterize treatment response were outlined in a recent report outlining the use of PET for evaluating solid tumor treatment response.³⁹ Also in this report were observations of how SUV measures may provide more relevant information for this task. In fact, SUV_{max} was seen to be the most common metric of evaluating tumor response. However, as the authors noted, SUV_{max} strongly depends on the statistical quality of the images and size of the maximum voxel.³⁹ Approaches using volumes to derive SUV values correlated with treatment outcomes have been explored^{22,24} and it is in this context that a consensus approach may provide researchers and clinicians a more comprehensive picture of SUV values within a tumor region.

The background activity was set equal to that observed in head-and-neck patients. However, it is important to note that consensus volumes may have use for defining volumes in other body regions, such as the liver, lung, brain, and pelvis. They may also be useful in defining volumes for other nuclear medicine modalities such as SPECT with the potential to provide a robustness reflecting that observed in our work.

Importantly, both MJV and STAPLE were consistently the top-performing methods. This was indicated by their DSC and SMASD values being better or comparable to those of the best individual segmentation method. Further, both MJV and STAPLE offer better results than the worst performing individual method. The argument for using either MJV or STAPLE extends to situation when we are presented with a previously unseen activity distribution and do not know *a priori* which method will provide superior segmentation accuracy. The fact that the ranking of an individual segmentation method changes depending on the situation and that both MJV and STAPLE were shown to provide essentially equal performance compared to using the best performing individual segmentation method in every case is compelling evidence that combining multiple segmentation methods is superior in providing robustness against false positives from one individual method. This is especially true with the potentially limitless variety of activity distributions among patient populations for which no ground truth is known.

Consensus methods also offer some mitigation of the individual segmentation method weaknesses. For example, the 40% and SRG methods suffer from image noise because if the maximum value represents a noise spike, rather than a true intensity, all voxels included in the final volume are now relative to noise. Absolute thresholding can also struggle to accurately segment smaller objects and tumor boundaries due to the blurring from the partial volume effect.² ADP provides some level of protection against the effects of noise by using both tumor and background intensities. The traditional KM algorithm implemented in this work could have difficulty with noisy images due to it grouping image intensities without regard to the physical distance between voxels. Extensions to the fuzzy C-means clustering algorithm that does account for distances are available and have been used in PET.⁴⁰ Finally, the gradient based method does require both a filtering and deblurring step with the filtering being carried out using a bilateral filter to preserve edges. Without sufficient filtering, noise in the image can be amplified by this deblurring step and subsequently included as tumor artificially. The noise levels in images acquired for this work were not sufficiently high for this to occur, however. The deblurring step also requires knowledge of the point spread function of the PET scanner that images were acquired on.

Accuracy is guaranteed to improve in a MJV approach if the individual methods have accuracies greater than 0.5 themselves.⁴¹ For example, if individual segmentation methods all have accuracies of 0.6, then combining three of them leads to an accuracy improvement to 0.6480 (8%), five of them to 0.6826 (13.2%), and combining nine leads to an accuracy improvement of 0.7334 (22.2%). With individual ac-

curacies of 0.9, the improvement is to 0.9720 (8%), 0.9914 (10%), and 0.9991 (11%) with three, five, and nine methods, respectively.⁴¹ The flattening of percent improvement implies that combining fewer individual methods may provide sufficient improvement in accuracy, but only if the accuracy of these methods is sufficiently high. With our segmentation methods having accuracies lying between these ranges, five methods provide a tradeoff in accuracy improvement and ease of implementation.

While the MJV and STAPLE were shown to be significantly different based on their DSC and SMASD values, the actual small differences in their median values are probably not convincing enough to sway the decision of which consensus method is better. MJV is also simple, intuitive and while other combination strategies such as the sum rule, or neural network approaches may also be employed, majority vote combination has been shown to provide performance on par with these methods and importantly, requires no training dataset.¹⁹ However, MJV is not robust to voxels that possess labeling errors²⁵ prior to combination. In other words, if an individual method or methods happened to assign an incorrect label to a given voxel or collection of voxels, then that method(s) may not contribute to the final consensus volume. With only two labels and relatively simple activity distributions, this limitation of MJV was less concerning for our study. However, with radiation therapy treatment planning utilizing boost volumes and even dose-painting which can utilize multiple levels of FDG uptake to prescribe the dose, any mislabeling of the boost subvolume(s) by an individual segmentation method would bias the final consensus volume, in that no contribution from that method(s) would be registered. In the extreme case no voxel would exhibit consensus leaving the boost volume undefined, whereas STAPLE would be robust to these sorts of effects.

A further point relevant to radiation therapy treatment planning is the issue of voxel size. We found that DSC values between grid sizes of 128×128 and 256×256 were not significantly different between MJV or STAPLE. However, SMASD values indicated that the average surface-to-surface difference was significantly smaller for the 256×256 grid size compared to the 128×128 grid size for both consensus methods. This implies that (1) SMASD is a more sensitive metric to changes in voxel size compared to DSC, and (2) given that the median SMASD is significantly lower for the higher resolution images, segmentation of target volumes using 256×256 images or images with smaller pixel sizes could lead to more accurate targets for treatment planning purposes.

All objects in this study had homogeneous activity levels. Realistically, patient tumors are more likely to possess heterogeneous FDG uptake. How the combination methods perform when applied to more realistic tumors or actual patient data is an interesting question but lies outside the scope of this study. The lack of a ground truth with patient data complicates the calculation of the accuracy of segmented volumes but it is not infeasible that up to a point, consensus volumes could mitigate certain individual methods that do not adequately deal with the presence of necrosis or hypoxia, for example. This is

especially important with the large variety of activity distributions in a patient population. In order to test this, simulation studies can be undertaken. However, they are restricted in not being able to fully model the biological effect and physical PET system. Another, potentially more useful study may be to recruit expert observers to produce a collection of segmentations which are subsequently combined into a ground truth using STAPLE. While labor-intensive, a pseudo ground truth could be generated this way for accuracy assessment over a more realistic collection of tumor sizes, shapes, and scanner and image acquisition characteristics. Importantly, efforts are currently underway toward the development of benchmark phantoms for oncology applications which should prove invaluable in future segmentation research.⁴² We are also currently applying our consensus volumes to patient data, in order to investigate any correlations between histograms of patient SUV values and patient outcomes. We are also looking at whether consensus volumes can be used as an initial GTV for oncologists during treatment planning. This application potentially offers reductions in interobserver variation between physicians and improvements in clinical workflow.

V. CONCLUSION

The combination of multiple, independent segmentation methods using either a majority vote strategy (MJV), or STAPLE probabilistic estimate, resulted in improved segmentation accuracy relative to any one individual method across all object shapes, sizes, contrasts, and scan durations. MJV and STAPLE provided comparable performance. Generally, for a particular object and imaging characteristics, both MJV and STAPLE were as good as the best performing individual method. Most importantly, because no individual method consistently rated higher than the others across multiple imaging and object characteristics, a combination method provides a level of robustness against the inconsistent performance of individual segmentation methods.

ACKNOWLEDGMENTS

Ross McGurk is supported by a New Zealand Bright Futures Top Achiever Doctoral Scholarship, Grant Number DKUX09001. This research was supported in part by NIH R01 RR021885 from the National Center for Research Resources, and by an award from the Neuroscience Blueprint I/C through R01 EB008015 from the National Institute of Biomedical Imaging and Bioengineering.

^{a)} Author to whom correspondence should be addressed. Electronic mail: ross.mcgurk@duke.edu; Telephone: 919 668 0293; Fax: 919 681 7183.

¹I. F. Ciernik, E. Dizendorf, B. G. Baumert, B. Reiner, C. Burger, J. B. Davis, U. M. Lutolf, H. C. Steinert, and G. K. Von Schulthess, "Radiation treatment planning with an integrated positron emission and computer tomography (PET/CT): A feasibility study," *Int. J. Radiat. Oncol. Biol. Phys.* **57**, 853–863 (2003).

²E. C. Ford, J. Herman, E. Yorke, and R. L. Wahl, "18F-FDG PET/CT for image-guided and intensity-modulated radiotherapy," *J. Nucl. Med.* **50**, 1655–1665 (2009).

- ³H. Zaidi and I. El Naqa, "PET-guided delineation of radiation therapy treatment volumes: A survey of image segmentation techniques," *Eur. J. Nucl. Med. Mol. Imaging* **37**, 2165–2187 (2010).
- ⁴U. Nestle, S. Kremp, A. Schaefer-Schuler, C. Sebastian-Welsch, D. Hellwig, C. Rube, and C. M. Kirsch, "Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-Small cell lung cancer," *J. Nucl. Med.* **46**, 1342–1348 (2005).
- ⁵D. A. Schinagl, W. V. Vogel, A. L. Hoffmann, J. A. van Dalen, W. J. Oyen, and J. H. Kaanders, "Comparison of five segmentation tools for 18F-fluorodeoxy-glucose-positron emission tomography-based target volume definition in head and neck cancer," *Int. J. Radiat. Oncol. Biol. Phys.* **69**, 1282–1289 (2007).
- ⁶Y. E. Erdi, O. Mawlawi, S. M. Larson, M. Imbriaco, H. Yeung, R. Finn, and J. L. Humm, "Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding," *Cancer* **80**, 2505–2509 (1997).
- ⁷E. C. Ford, P. E. Kinahan, L. Hanlon, A. Alessio, J. Rajendran, D. L. Schwartz, and M. Phillips, "Tumor delineation using PET in head and neck cancers: Threshold contouring and lesion volumes," *Med. Phys.* **33**, 4280–4288 (2006).
- ⁸J. W. Keyes, Jr., "SUV: Standard uptake or silly useless value?," *J. Nucl. Med.* **36**, 1836–1839 (1995).
- ⁹W. Jentzen, L. Freudenberg, E. G. Eising, M. Heinze, W. Brandau, and A. Bockisch, "Segmentation of PET volumes by iterative image thresholding," *J. Nucl. Med.* **48**, 108–114 (2007).
- ¹⁰A. E. Boudraa, J. Champier, L. Cinotti, J. C. Bordet, F. Lavenne, and J. J. Mallet, "Delineation and quantitation of brain lesions by fuzzy clustering in positron emission tomography," *Comput. Med. Imaging Graph.* **20**, 31–41 (1996).
- ¹¹R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans. Pattern Anal. Mach. Intell.* **16**, 641–647 (1994).
- ¹²M. Hatt, C. Cheze le Rest, A. Turzo, C. Roux, and D. Visvikis, "A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET," *IEEE Trans. Med. Imaging* **28**, 881–893 (2009).
- ¹³M. Hatt, F. Lamare, N. Boussion, A. Turzo, C. Collet, F. Salzenstein, C. Roux, P. Jarritt, K. Carson, C. Cheze-Le Rest, and D. Visvikis, "Fuzzy hidden Markov chains segmentation for volume determination and quantitation in PET," *Phys. Med. Biol.* **52**, 3467–3491 (2007).
- ¹⁴X. Geets, J. A. Lee, A. Bol, M. Lonneux, and V. Gregoire, "A gradient-based method for segmenting FDG-PET images: Methodology and validation," *Eur. J. Nucl. Med. Mol. Imaging* **34**, 1427–1438 (2007).
- ¹⁵F. Yang and P. W. Grigsby, "Delineation of FDG-PET tumors from heterogeneous background using spectral clustering," *Eur. J. Radiol.* **81**(11), 3535–3541 (2012).
- ¹⁶X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de-Solorzano, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," *IEEE Trans. Med. Imaging* **28**, 1266–1277 (2009).
- ¹⁷F. Kimura and M. Shridhar, "Handwritten numerical recognition based on multiple algorithms," *Pattern Recognit.* **24**, 969–983 (1991).
- ¹⁸J. Kittler and F. M. Alkoot, "Sum versus vote fusion in multiple classifier systems," *IEEE Trans. Pattern Anal. Mach. Intell.* **25**, 110–115 (2003).
- ¹⁹L. Lam and C. Y. Suen, "Application of majority voting to pattern recognition: An analysis of its behavior and performance," *IEEE Trans. Syst. Man Cybern., Part A Syst. Humans* **27**, 553–568 (1997).
- ²⁰L. Østergaard and O. Larsen, "Applying voting to segmentation of MR images," in *Advances in Pattern Recognition*, edited by A. Amin, D. Dori, P. Pudil, and H. Freeman (Springer, Berlin, 1998), Vol. 1451, pp. 795–804.
- ²¹P. Therasse, S. G. Arbuck, E. A. Eisenhauer, J. Wanders, R. S. Kaplan, L. Rubinstein, J. Verweij, M. Van Glabbeke, A. T. van Oosterom, M. C. Christian, and S. G. Gwyther, "New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada," *J. Natl. Cancer Inst.* **92**, 205–216 (2000).
- ²²E. Schreibmann, A. F. Waller, I. Crocker, W. Curran, and T. Fox, "Voxel clustering for quantifying PET-based treatment response assessment," *Med. Phys.* **40**, 012401 (12pp.) (2013).
- ²³C. Nahmias, W. T. Hanna, L. M. Wahl, M. J. Long, K. F. Hubner, and D. W. Townsend, "Time course of early response to chemotherapy in non-small cell lung cancer patients with 18F-FDG PET/CT," *J. Nucl. Med.* **48**, 744–751 (2007).
- ²⁴S. M. Larson, Y. Erdi, T. Akhurst, M. Mazumdar, H. A. Macapinlac, R. D. Finn, C. Casilla, M. Fazzari, N. Srivastava, H. W. Yeung, J. L. Humm,

- J. Guillem, R. Downey, M. Karpeh, A. E. Cohen, and R. Ginsberg, "Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging. The visual response score and the change in total lesion glycolysis," *Clin. Positron Imaging* **2**, 159–171 (1999).
- ²⁵S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imaging* **23**, 903–921 (2004).
- ²⁶Y. Yamamoto, T. Z. Wong, T. G. Turkington, T. C. Hawk, and R. E. Coleman, "Head and neck cancer: Dedicated FDG PET/CT protocol for detection—phantom and initial clinical studies," *Radiology* **244**, 263–272 (2007).
- ²⁷T. Shepherd, "Response to "Phantom measurements with glass inserts in a hot background are not suitable for performance assessment of volume delineation algorithms in PET,"" *IEEE Trans. Med. Imaging* **PP**(99) (2012).
- ²⁸J. van den Hoff and F. Hofheinz, "Phantom measurements with glass inserts in a hot background are not suitable for performance assessment of volume delineation algorithms in PET," *IEEE Trans. Med. Imaging* **PP**(99) (2012).
- ²⁹F. Hofheinz, S. Ditttrich, C. Potzsch, and J. Hoff, "Effects of cold sphere walls in PET phantom measurements on the volume reproducing threshold," *Phys. Med. Biol.* **55**, 1099–1113 (2010).
- ³⁰P. Tylski, S. Stute, N. Grotus, K. Doyeux, S. Hapdey, I. Gardin, B. Vanderlinden, and I. Buvat, "Comparative assessment of methods for estimating tumor volume and standardized uptake value in (18)F-FDG PET," *J. Nucl. Med.* **51**, 268–276 (2010).
- ³¹H. Li, W. L. Thorstad, K. J. Biehl, R. Laforest, Y. Su, K. I. Shoghi, E. D. Donnelly, D. A. Low, and W. Lu, "A novel PET tumor delineation method based on adaptive region-growing and dual-front active contours," *Med. Phys.* **35**, 3711–3721 (2008).
- ³²P. T. Meyer, D. Elmenhorst, A. Matusch, O. Winz, K. Zilles, and A. Bauer, "18F-CPPFX PET: On the generation of parametric images and the effect of scan duration," *J. Nucl. Med.* **47**, 200–207 (2006).
- ³³E. Day, J. Betler, D. Parda, B. Reitz, A. Kirichenko, S. Mohammadi, and M. Miften, "A region growing method for tumor volume segmentation on PET images for rectal and anal cancer patients," *Med. Phys.* **36**, 4349–4358 (2009).
- ³⁴M. Hatt, C. Cheze Le Rest, N. Albarghach, O. Pradier, and D. Visvikis, "PET functional volume delineation: A robustness and repeatability study," *Eur. J. Nucl. Med. Mol. Imaging* **38**, 663–672 (2011).
- ³⁵F. Hofheinz, J. Langner, B. Beuthien-Baumann, L. Oehme, J. Steinbach, J. Kotzerke, and J. van den Hoff, "Suitability of bilateral filtering for edge-preserving noise reduction in PET," *EJNMMI Res.* **1**, 23 (2011).
- ³⁶K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells 3rd, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index," *Acad. Radiol.* **11**, 178–189 (2004).
- ³⁷J. J. Bartko, "Measurement and reliability: Statistical thinking considerations," *Schizophr. Bull.* **17**, 483–489 (1991).
- ³⁸A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: Method and validation," *IEEE Trans. Med. Imaging* **13**, 716–724 (1994).
- ³⁹R. L. Wahl, H. Jacene, Y. Kasamon, and M. A. Lodge, "From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors," *J. Nucl. Med.* **50**(Suppl 1), 122S–150S (2009).
- ⁴⁰S. Belhassen and H. Zaidi, "A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET," *Med. Phys.* **37**, 1309–1324 (2010).
- ⁴¹L. I. Kuncheva, *Combining Pattern Classifiers : Methods and Algorithms* (Wiley, Hoboken, NJ, 2004).
- ⁴²T. Shepherd, B. Berthon, P. Galavis, E. Spezi, A. Apte, J. A. Lee, D. Visvikis, M. Hatt, E. de Bernardi, S. Das, I. El Naqa, U. Nestle, C. Schmidlein, H. Zaidi, and A. Kirov, "Design of a benchmark platform for evaluating PET-based contouring accuracy in oncology applications," *Eur. J. Nucl. Med. Mol. Imaging* **39**, 155–303 (2012).