Radiotherapy and Oncology 141 (2019) 78–85

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Quantitative imaging: Erring patterns in manual delineation of PET-imaged lung lesions

Fei Yang [a,*], Lori Young [b], Yidong Yang [c]

[a] Department of Radiation Oncology, University of Miami, Miami, FL, USA; [b] Department of Radiation Oncology, University of Washington, Seattle, WA, USA; and [c] The First Affiliated Hospital of University of Science and Technology of China, Hefei, PR China

ARTICLE INFO

Article history:
Received 18 March 2019
Received in revised form 14 August 2019
Accepted 17 August 2019
Available online 5 September 2019

Keywords:
Quantitative PET
Target delineation
Manual contouring
Segmentation accuracy
Lung cancer

ABSTRACT

Background and purpose: Uncertainty and variability in manual contouring of PET-imaged tumor targets are well recognized; however, the error patterns associated with it were little known and rarely investigated. The present study is aimed to quantitatively assess the erring patterns inherent to manual delineation of PET-imaged lung lesions in a setting with complete ground truth.

Materials and methods: Images being used for assessment consisted of 26 synthetic PET datasets created by using the anthropomorphic Zubal phantom in conjunction with the Monte Carlo based SimSET computational package. Each dataset included one PET-positive lesion differing in shape, dimension, uptake heterogeneity, and anatomical location inside the lung. Target contours were provided by 10 raters and the contour accuracy was evaluated using 12 metrics from five categories – spatial overlap, pair counting, information theory, distance, and volume.

Results: In terms of spatial overlap, manual contouring results intersect substantially with the ground truth whereas tend to oversegment the lesions. Shapes of the segmented tumor volumes are in general geometrically consistent with the ground truth but lack sensitivity in characterizing topographical details. No complete consensus could be achieved between manual contours and the ground truth for any of the given lesions being examined when assessing using pair counting- and informatics-based metrics thus indicating an intrinsic stochastic component of manual contouring. Evaluation based on metrics related to distance and volume demonstrated that it is at the borderline areas between the lesions and the normal tissues where the majority part of manual delineation errors occurred and the extent of volume being identified false positively as cancerous by the raters is appreciable.

Conclusion: Quantification of segmentation errors associated with expert manual contouring of PET positive lesion in the lung reveals general patterns in what otherwise might be thought of as randomness. Findings from the current study may allow for the formation of new hypotheses towards improving the accuracy and precision of manual delineation of PET positive tumor targets in the lung.

Worldwide, lung cancer is the most common malignant tumor, accounting for 1.69 million new cases in 2015 [1], and remains the leading cause of cancer-related mortality with a predicted 5-year survival rate of 8–13% [2]. Histopathologically, lung cancer is broadly categorized as small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) with the latter constituting approximately 85% of the cases [3]. ^{18}F-fluorodeoxyglucose positron emission tomography (FDG-PET) with demonstrated high sensitivity and specificity for tumor detection has become the standard of care in baseline staging and restaging for both SCLC and NSCLC, and it has recently also been making inroads into a number of other key essential components of the clinical management of patients with lung cancer ranging from therapy planning to response assessment and surveillance — and even to prognosis and prediction of therapeutic outcomes where the emerging concept of using radiological imaging traits as surrogates for tumoral phenotypic spatial heterogeneity is advocated [4–7]. Exploiting FDG-PET to its full potentials in these areas of lung cancer management lies largely in the extraction or mining, especially with high levels of fidelity and integrity, of quantitative information in relevant to tumoral phenotypic characteristics from the image data, a process that is clinically appealing yet filled with many practical challenges.

Amongst the various challenges encountered by quantitative analysis of PET imaging data, one of the most prominent and largely inevitable obstacles is the accurate and precise determination of the tumor volume [8]. Despite numerous efforts being devoted

[*] Corresponding author at: University of Miami School of Medicine, Department of Radiation Oncology, 1475 NW 12th Ave., Miami, FL 33136, USA.
E-mail address: fei@miami.edu (F. Yang).

https://doi.org/10.1016/j.radonc.2019.08.014
0167-8140/© 2019 Elsevier B.V. All rights reserved.

towards the development of automatic or semi-automatic approaches for the segmentation of tumor volume in PET images, a convincing "solution" has not been reached so far [9,10]. Difficulties posed to these efforts comprise a number of the intrinsic limitations of PET imaging, including but not restricted to the low signal-to-noise ratio, the poor spatial resolution, the partial volume effect, the heterogeneous background, and the uneven uptake of radiotracer within tumors [11,12]. As is in nowadays clinical practices, manual contouring, albeit known to be labor intensive and vulnerable to inter-/intra-observer variability, still remains as the mainstay and is oftentimes regarded to be the "gold standard" in the vast majority, if not all, of the tumor delineation settings within the context of quantitative PET imaging [13]. Manual contouring as a tumor segmentation method for lung FDG-PET has been explored by multiple studies with regard to its inter-/intra-observer variability and test–retest repeatability [14,15]; however, to the best our knowledge little work has been done to date to validate its accuracy as to be a reliable surrogate for volume assessment of PET-imaged tumors.

In the current study, accuracy of manual delineation of PET positive targets in the lung was validated against complete ground truth data for 260 manual drawn contours (10 raters × 26 datasets) in terms of five categories of evaluation metrics related, respectively, to spatial overlap, pair counting, information theory, distance, and volume to illustrate the various underlying aspects of the behavior patterns of manual contouring. Erring patterns associated with manual contouring were assessed individually and collectively and the implications of the results are discussed within the context of easing the deployment of quantitative PET into those aforementioned rapidly advancing domains of the clinical management of lung cancer patients.

Materials and methods

PET simulation

PET image data was generated by aid of an anthropomorphic thoracic digital phantom. The motivation of using a digital phantom over clinical PET data was that the digital phantom would provide image data with a known association to the ground truth and therefore offering the true gold standard for the evaluation of manual contours. Such information, however, would generally be unobtainable for clinical data. The simulation was conducted by using the Zubal anthropomorphic phantom [16] as the attenuation map while employing the Monte Carlo (MC) based Simulation System for Emission Tomography software package (SimSET) [17] for PET event detection process. This method was published previously [18,19] and has been utilized in a broad range of PET segmentation studies [20,21,50]. In brief, for any given location within the phantom an activity index was assigned to approximate the ^{18}F-FDG accumulation of the spatially corresponding anatomy, which is either of metabolically active tumors or of normal tissues. For a complete list of the major organs being incorporated in the thoracic model along with their activity indices and tissue types for attenuation please see Table S1 (Supplementary Material I). The PET system modeled was a Siemens Biograph scanner featuring a pixelated block BGO detector with ring radius of 42.1 cm. With SimSET, positron range was modeled using the empirical algorithm developed by Palmer et al. [22] and photon deviations from collinearity were modeled by a Gaussian random variable with mean of 180 degrees and standard deviation of 0.5 degrees [17]. Compton scatter was simulated by aid of the Klein-Nishina equations [23], while for coherent scatter interaction probabilities and scatter angular distributions were calculated based on the Lawrence Livermore National Laboratories Evaluated Electron Density Library (LLNL EPDL) database [17]. Random events for each detector pair were computed based on separate singles simulation [24]. The emission data produced from the simulations was re-binned into 128 × 128 sinograms by single-slice re-binning, followed with slice-by-slice reconstruction using an ordered subset expectation maximization (OSEM) algorithm (8 iterations, 4 subsets). Attenuation correction was carried out for each organ with using a tissue index that corresponds to an attenuation coefficient specified in the SimSET package. The resulting image slices were further convolved with a 5 mm full width at half maximum (FWHM) 3D Gaussian filter for noise suppression. The various aspects of simulation specifics with regard to detector property, sonogram acquisition, reconstruction algorithm, and post-processing were chosen to emulate those of the PET scanner and the imaging protocol for lung cancer of the local facility. The validity of the described PET imaging simulation process was demonstrated by comparison of the actually acquired and the simulated image data in terms of gray-level intensity histogram, intensity profile, and statistical textures (Supplementary Material I, Fig. S1-2 and Table S2).

Target contouring

Raters participating in the study consisted of a total of 10 radiation oncology physicians equipped with extensive clinical experience in contouring PET-imaged lung lesions as part of the radiation therapy (RT) treatment planning process. Image data provide for contouring included the simulated PET along with a co-registered CT of the digital phantom. MIM Maestro® v6.7.11 (MIM Software, Cleveland, OH) was used as the contouring platform. No specific instructions were given with regard to the display settings such as window/level (W/L), thresholding, and pixel representation, amongst others. Note that only the basic contouring tools of the MIM Maestro program such as 2D/3D brush, pen, interpolation, and smoothing were allowed for use while advanced techniques of the program such as segmentation based on atlas, region growing, and PET edge *etc.* were prohibited. Furthermore, the study was carried out in a double-blind fashion to guard against potential biases, *i.e.*, the raters were not able to view the work from one another and the rater identities were kept anonymous to the investigators.

Quantitative evaluation

Manual contours delineated by the raters (*M*) were evaluated by reference to their respective ground truth data (*G*) in terms of 12 accuracy metrics. These metrics were sub-divided into five groups using spatial overlap, pair counting, information theory, distance, and volume as categorical descriptors (Table 1). Starting with spatial overlap, four metrics including Dice coefficient (*DICE*), false negative Dice (*FND*), false positive Dice (*FPD*), and global consistency error (*GCOERR*) were considered. *DICE* can be seen as the percentage of spatial overlap between *M* and *G* [25]. *FND* lays emphasis on under-segmentation while *FPD* highlights over-segmentation of *M* with respect to *G* [26]. *GCOERR* measures the extent to which *M* can be viewed as a refinement of *G* by quantifying the consistency error at each voxel [27]. The second category consisted of pair counting based metrics featuring the Rand index (*RNDIND*) and the adjusted Rand index (*ADJRIND*). Both *RNDIND* and *ADJRIND* measure the agreement between *M* and *G* by means of counting corrected segmented pairs of voxels while the latter being adjusted for chance [28,29]. *RNDIND* can have a value between 0 to 1, with 0 indicating that *M* and *G* do not agree on labeling any pair of voxels and 1 indicating that the two match identically on all pairs of voxels. *ADJIND* has a range of $[-1, 1]$, with an expected value of 0 for random segmentation and 1 for perfect agreement while negative values indicate worse than blindly

Table 1
Quantitative metrics used for accuracy evaluation of manual contours along with their values for ideal segmentation.

Category	Metric	Value for ideal segmentation
Spatial Overlap	Dice (DICE)	1
	False Negative Dice (FND)	0
	False Positive Dice (FPD)	0
	Global Consistency Error (GCOERR)	0
Pair Counting	Rand Index (RNDIND)	1
	Adjusted Rand Index (ADJRIND)	1
Information Theory	Normalized Mutual Information (NMUTINF)	1
	Normalized Variation of Information (NVARINFO)	0
Distance	Symmetric Mean Absolute Surface Distance (SMASD)	0
	Average Hausdorff Distance (AHDST)	0
	Mahalanobis Distance (MDST)	0
Volume	Absolute volumetric difference (AVD)	0

guessing. Quantitative metrics of the third category were based on information theory, including normalized mutual information (*NMUTINF*) and normalized variation of information (*NVARINFO*). *NMUTINF* quantifies the relative extent to which the uncertainty of *M* is reduced given *G* is known while *NVARINFO* assesses the relative information alteration resulting from the geometric transition from *M* to *G* [30,31]. Together an *NMUTINF* of 1 and *NVARINFO* of 0 indicate a complete agreement. Three metrics containing symmetric mean absolute surface distance (*SMASD*) and average Hausdorff distance (*AHDST*) together with Mahalanobis distance (*MDST*) were employed to quantify spatial distance difference between *M* and *G*. *SMASD* estimates the average distance over which the surfaces of *M* and *G* differ spatially; *AHDST* inspects the average Hausdorff distance over all voxels and takes either the mean Hausdorff distance from *M* to *G* or from *G* to *M*, whichever is greater; and *MDST* gives consideration to the spatial correlation between voxels in *M* and *G* [32,33]. The three distance metrics all evaluate to 0 for perfect segmentation. The final category concerned the extent to which *M* deviates from *G* in volumes through using absolute volumetric difference (*AVD*) that evaluates the absolute fractional volumetric difference between *M* and *G* [12]. For more detailed descriptions of these accuracy measuring metrics, please refer to Supplementary Material II.

Besides assessing the delineation performance of each rater with respect to the ground truth data, contour results for a given lesion from the different raters were pooled together via voxel-wise majority voting to establish consensus contour in the interest of assessing if a consensus among the raters would be more accurate. Furthermore, variability in contouring accuracy among the raters was examined for dependency on the shape and heterogeneity of the lesions. Variability in segmentation accuracy metrics among the raters for a given lesion was quantified by coefficient of variation (*CV*). Shape irregularity of the lesions was depicted by the shape complexity index while radiologically revealed heterogeneity of the lesions was characterized with the use of three categories of radiomics parameters related to gray-level intensity histograms (*GLIH*), gray-level size zone matrices (*GLSZM*), and gray-level co-occurrence matrices (*GLCM*) with spatial scale emphasis of each being on global, regional, and local, respectively. A detailed description of the shape complexity index and radiomics parameters being employed is provided in Supplementary Material III. Additionally, manual contours were also measured for agreement in terms of various accuracy metrics through using the Williams' index (*WI*) [34]. For each measuring metric, three agreement scenarios are possible and would result in the value of *WI*: (i) greater than 1 when manual contours are more similar to the ground truth than to each other; (ii) equal or close to 1 when manual contours are as similar to each other as they are to the ground truth; (iii) less than 1 when manual contours are more similar to each other than each is to the ground truth.

Statistical analysis

Statistical analysis was performed using JMP Pro® Version 12 (SAS Institute Inc., Cary, NC) statistical software. Associations, including those between the segmentation accuracy metrics and those between the segmentation accuracy metrics and the lesion volume along with those between the variability in contouring accuracy among the raters and the shape or heterogeneity of the lesion, were examined through the use of Kendall's rank correlation coefficient tau (τ) [35]. Correlations were considered weak if $|\tau| < 0.40$, moderate if $0.40 \le |\tau| < 0.60$, relatively strong if $0.60 \le |\tau| < 0.80$, and strong if $0.80 \le |\tau|$ [19]. *P*-values were determined for each correlation coefficient to assess the statistical significance with the null hypothesis being that the correlation coefficients did not differ significantly from zero. Differences in segmentation accuracy metrics between raters and between lesions were examined by the Kruskal–Wallis H omnibus test — the null hypothesis being that the distribution of the accuracy metric was the same in all raters or for all lesions — and significant results were further subjected to post hoc rank tested using the Dunn's pairwise test with Bonferroni adjustment for multiple comparisons [36,37]. For all statistical analyses, *p*-values of 0.05 or lower were considered statistically significant.

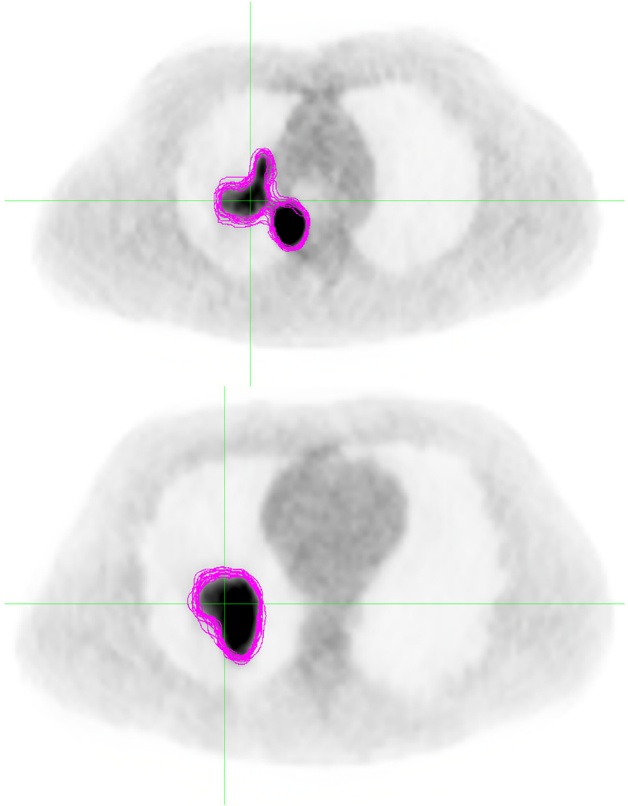

Fig. 1. Synthetic PET images showing simulated lesions and the variability in contouring among raters.

Results

A total of 26 synthetic PET dataset were generated. Each dataset contained a different PET-positive lesion which was situated either within the lungs or adjacent to the mediastinum or to the chest wall. These lesions featured irregular shapes and heterogeneous radiotracer accumulation with volume ranging from 30 cm^3 to 345 cm^3 (mean ± SD = 131 ± 74 cm^3). As examples, transverse view of central slices of two of the simulated lesions located, respectively, immediately adjacent to the mediastinum within the posterior region of the upper lobe of the right lung and within the posterior region of the middle lobe of the right lung are shown in

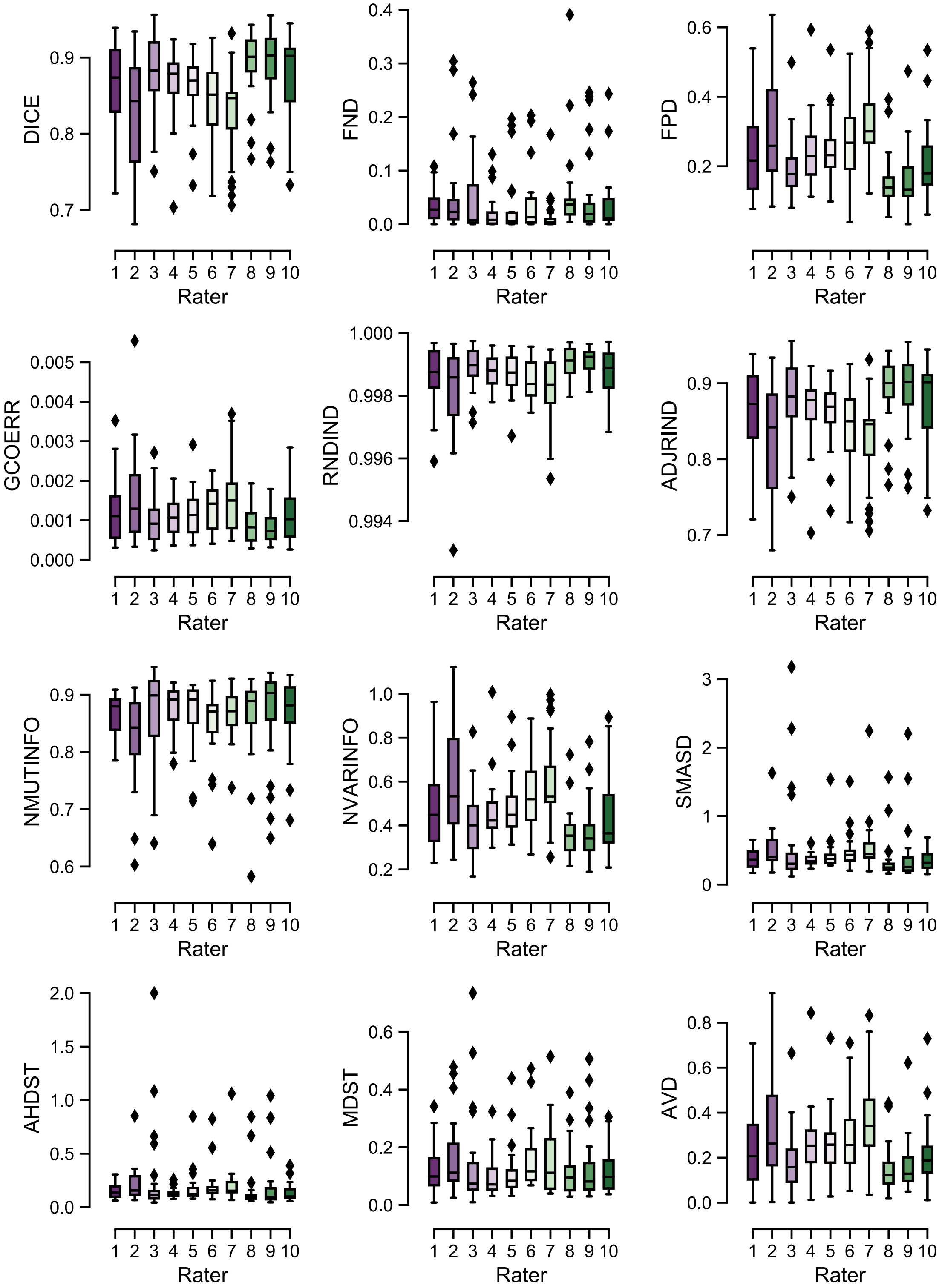

Fig. 2. Boxplots comparing segmentation accuracy metrics between raters for the simulated PET lesions. On each box, the central mark indicates the median, and the top and bottom edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers are 1.5× the inter-quartile range, and values outside these are plotted individually.

Fig. 1 with the manual contours from the raters overlaid. The simulated image data were appraised by nuclear medicine and radiation oncology professionals that read and interpret lung PET images on routine basis as being visually indistinguishable from clinical PET data.

Boxplots illustrating contouring performance of the raters as measured by the accuracy metrics are shown in Fig. 2, from which it can be readily observed that the distribution patterns of the examined accuracy metrics differed greatly between the raters in terms of central tendency, spread, skewness, and tail length among other distribution statistics. *DICE* between *M* and *G* of the 10 raters had means ranging from 0.828 to 0.895, indicating partial while substantial overlap between contours by the raters and the ground truth data for the lesions being examined. Mean values of *FND* between *M* and *G* ranged from 0.014 to 0.055, showing that there was only very limited amount of lesion volume being missed out by manual contouring. The range of mean *FPD* between *M* and *G* of the raters spanned from 0.156 to 0.328, demonstrating there existed nonnegligible amounts of abutting normal tissues were perceived by manual contouring as cancerous and revealing the tendency of the raters' inclination towards oversegmentation of the targets. With maximum *GCOERR* of all raters being less than 0.006, it would be rational to say that *M* and *G* are overall geometrically consistent except that the latter features finer levels of topographical details. Considering minimum *RNDIND* of the raters being no less than 0.993 while mean *ADJIND* of all raters ranging from 0.826 to 0.894 for the examined lesions, it would be reasonable to infer that, although manual contours agreed in general with the ground truth, no complete consensus could be achieved between the two for any of the given lesions being investigated and also suggesting that the manual contouring process is somewhat stochastic in nature. With *NMUTINF* of the raters averaging from 0.830 to 0.881 and *NVARINFO* averaging from 0.368 to 0.593 for the examined lesions, there appeared to be an evident gap between *M* and *G* in terms of information measures. This echoes what pair counting based metrics implied, which is the presence of a certain amount of randomness being involved in the manual contouring. Given *SMASD* of the raters averaging from 0.342 to 0.570 voxel while worst case scenario approaching as much as 3.177 voxels for the examined lesions, it would be fair to say that the spatial extent to which manually defined tumoral surfaces differed from their corresponding ground truth surfaces are on average less than one voxel size whereas could reach multi-voxel extent in certain cases. *AHDST* exhibited results highly similar to those of *SMASD*, with averages between 0.129 to 0.260 voxel and worst case scenario as large as 2.001 voxel. *MDST* of the raters had means ranging from 0.102 to 0.162 and maxima ranging from 0.305 to 0.734 for the examined lesions, showing that the distance difference between *M* and *G* was mitigated with spatial correlation of the two being taken into consideration. Volume differences between *M* and *G* as measured by *AVD* were associated with means ranging from 0.148 to 0.392 among the raters whereas could be as great as 0.930 in some cases, resonating with what the results of *FPD* metric demonstrated, *i.e.*, the amounts of abutting normal tissues being false positively identified as cancerous is non-trivial.

Boxplots comparing segmentation accuracy between consensus and individual contours of the raters as a function of the lesions are shown in Fig. S3 (Supplementary Material IV), from which it can be readily appreciated that consensus contour in overall outperformed the majority of the individual contours, even though to varying degrees dependent on the lesions and accuracy metrics. Fig. 3 presents a heatmap that indicates the degree of correspondence between the segmentation accuracy metrics along with their association with the lesion volume as being examined by the Kendall's rank correlation coefficients (τ). Across all the accuracy metric pairs, strong correlations ($\tau \geq 0.80$) emerged within 9 pairs while 7 pairs born relatively strong correlations ($0.60 \leq \tau < 0.80$). As to the association with lesion volume, all accuracy metrics showed weak correlations ($|\tau| < 0.40$) except for *GCOERR* and *RNDIND* that demonstrated moderate correlation with the lesion volume with τ of 0.58 and -0.55, respectively. Kendall's rank correlation coefficients (τ) between *CV* of the accuracy metrics among the raters and image features depicting shape irregularity and activity heterogeneity of the lesions are summarized in Table S3 (Supplementary Material V), from which it can be observed that variability of the accuracy metrics, in general, depended only weakly ($|\tau| < 0.40$) on image features except for the *CVs* of *RNDIND* and *FND* that correlated moderately to 8 and 2 of the features ($0.40 \leq |\tau| < 0.60$), respectively.

Kruskal–Wallis H omnibus test demonstrated that 9 out of the 12 investigated segmentation accuracy metrics distributed differently across the raters while all of them distributed differently across the lesions. Rater pairs and lesion pairs revealed by the Dunn's pairwise test as being significantly different in their distribution of segmentation accuracy metrics are illustrated in Fig. 4 and Fig. S4 (Supplementary Material VI), respectively. As can be seen from these figures, either the rater pair or the lesion pairs that exhibited significant difference in contouring accuracy is largely measuring metric dependent and may vary greatly from one metric to another. Whether manual contours of the raters are more similar to each other or more similar to the ground truth was examined in terms of the investigated accuracy metrics by aid of *WI*. Estimated distribution of *WI* with respect to the ground truth for each of the metrics is presented in Fig. 5, from which it can be appreciated that manual contours of the raters are as similar to each other as they are to the ground truth across all the examined metrics except for *FPD* and *AVD* ($WI > 1$) as regards which manual contours are more similar to the ground truth and for *FND* ($WI < 1$) as to which they are more similar to each other.

Discussion

Accurate tumor target identification is of essential significance for quantitative analysis of oncological FDG-PET imaging data for lung cancer. It influences multifarious key aspects of RT clinical

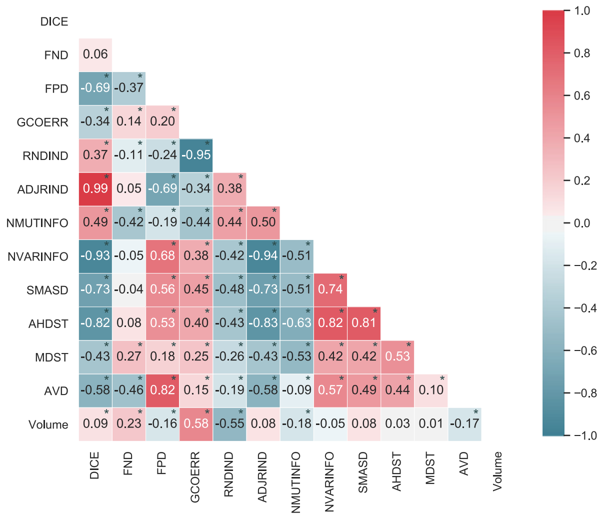

Fig. 3. Heatmap of Kendall's rank correlation coefficients (τ) between segmentation accuracy metrics and those between the segmentation accuracy metrics and lesion volume. The color key represents the τ values of Kendall's correlations. Numerical values of correlations coefficients are presented in the corresponding boxes with statistically significant values being highlighted by asterisks.

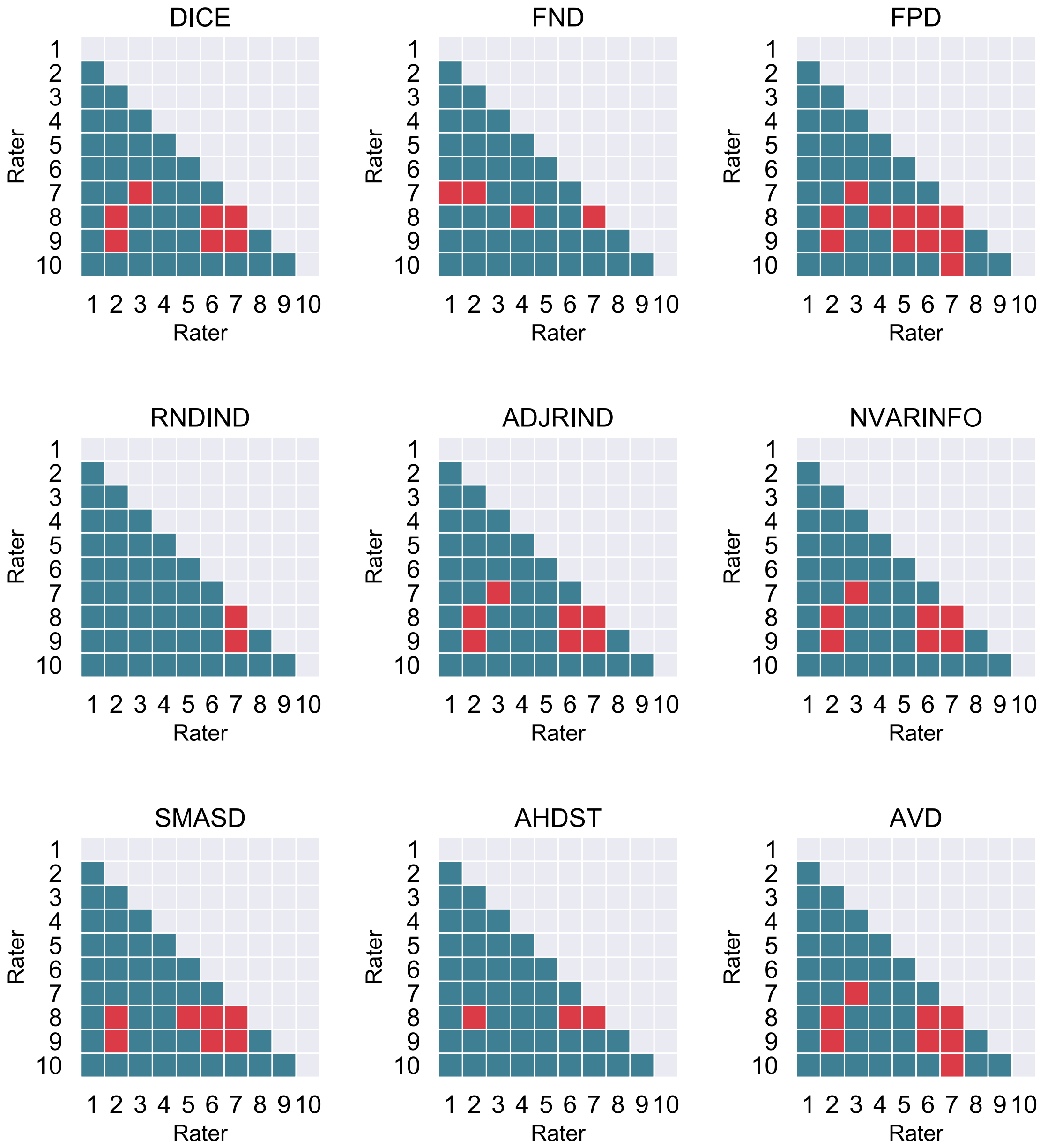

Fig. 4. Rater pairs with significant (in red) or nonsignificant (in teal) difference per Bonferroni adjusted Dunn's multiple comparison test for segmentation accuracy metrics rendering statistically significant results in regard to Kruskal–Wallis H omnibus test.

management of lung cancer. As to treatment planning, accurate and precise treatment target volume definition is of paramount importance in achieving dose escalation to viable malignant tumors while sparing a maximum amount of surrounding health tissue from undesirable radiation [38–40]. With respect to treatment response assessment and surveillance, metabolic tumor burden quantification using total lesion glycolysis makes it imperative for accurate and reliable delineation of FDG-PET imaged tumor lesions [41–43]. Furthermore, accurate knowledge on tumor boundaries is also of crucial relevance and cannot be overstated enough for quantitative PET imaging and radiomics analysis in lung cancer, wherein image content within tumor lesions is converted to a mineable high-dimensional form that aims at characterizing intratumoral heterogeneity and may potentially hold promise in the development of prognostic and predictive models seeking to relate radiological imaging ques to tumoral phenotypic or even to genotypic signatures thus allowing personalized and adaptive care of lung cancer in RT [44–46].

The present study aimed to validate the accuracy of manual contouring for PET positive tumor targets in the lung within the context of complete known ground truth. Although different raters demonstrated distinct conceptions of the ideal boundaries separating cancerous lesions from adjacent normal tissues during manually contouring of PET positive target volumes, there are several clear and consistent trends that emerged from the accuracy evaluation of manual contouring results with respect to the ground truth data: firstly, the extent to which manual contours overlap with their respective ground truth is substantial on average whereas not sufficient enough to ensure manual contouring a reliable and accurate surrogate for the volume determination of PET-imaged tumor targets. Secondly, manual contouring in general tends to inadvertently identify an appreciable amount of adjacent

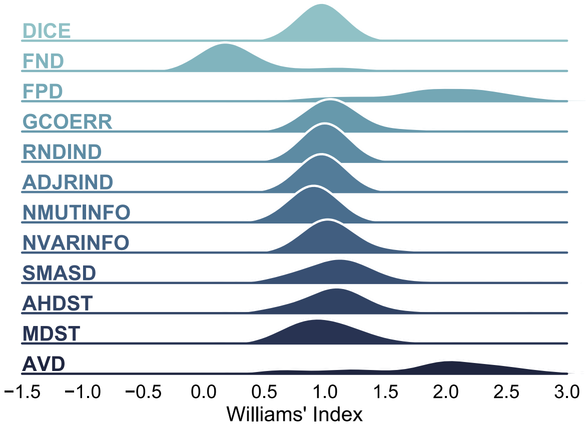

Fig. 5. Kernel density estimations of distribution of Williams' index with respect to ground truth for all segmentation accuracy metrics with pooling of all lesions and all raters.

normal tissue to be cancerous and this oftentimes dominates over the magnitude of the false negatives it brings about, *i.e.*, regions where diseases do exist are erroneously ignored, hence resulting in overestimated tumor volumes as compared to the actual ones. Thirdly, though being capable of retaining the gross dimensions and overall geometrical shapes of the actual tumor lesions manual contouring is evidently associated with the incompetence of capturing fine morphological details of the actual tumors. Furthermore, alongside clinical judgement and reasoning there is also a certain amount of "guesswork" being involved in the manual contouring process, as is evidenced by the pair counting and informatics measures, and this occurs most likely primarily over weighing up the belongings of obscured borderline areas between the normal tissues and the cancerous lesions.

Quantitative metrics used for segmentation accuracy assessment in the current study consisted of a total of 12 well-established agreement measures, with categories related respectively to spatial overlap, pair counting, information theory, distance, and volume. Pairwise Kendall's rank correlation analysis shows that out of the 66 pairs of accuracy metrics 9 pairs were associated with strong correlation at statistically significant levels while the rest 57 pairs were correlated with varying degrees in the strength ranging from relatively strong through moderate to weak (Fig. 3 first to second-to-last rows from the top). Metrics such as *GCOERR* and *RNDIND* considering true negatives as part of their agreement evaluation are biased against the ratio between the segmented volume and the complement volume used for evaluation, thus giving rise to segmentation with large volume being penalized whereas those with small volume being rewarded. This is evident from the bottom portion of the heatmap as shown in Fig. 3, where it can be seen that moderate correlation existed between these metrics and the tumor volume. With regards measuring accuracy, although the majority of the metrics indicated manual contours are as similar to each other as they are to the ground truth, there do exist exceptions, such as *FND* in terms of which manual contouring results are more similar to each other together with *FPD* and *AVD* in terms of which manual results are more similar to the ground truth than to each other. These observations demonstrate that different measuring metrics may well lead to different impressions about the accuracy of the exact same delineation results.

There are several immediate implications from the current study for the RT clinical management of patients with lung cancer that are worthy of mention. Firstly, the observed general trend of raters toward overestimation of tumor extent in PET when added on top of various margins applied during RT treatment planning would lead to further increase in treatment volume and thereby potentially posing an undue risk of augmented toxicity to the adjacent normal tissues. Reduction of the treatment margins to account for this biased tendency of manual contouring would therefore be advisable when formulating treatment plans for lung cancer from PET-based target volumes. Additionally, the current study calls for prudence with respect to using manual contouring results, even consensus of multiple experts, as the surrogate gold standards for volume determination of PET positive lesions in the lung at least. Automatic segmentation of PET imaged lesions has been of continuous interest since the deployment of PET scanners for clinical use especially along with recent advances in image processing technics. To date there has been a wide array of algorithms being proposed toward this end. In regard to performance evaluation and result validation, a certain amount of them referred to manual contouring data as "gold standard" [9,10,13,47,48]. Caution is therefore advised when adopting any of these methods towards clinical use. Furthermore, the current study showed that different accuracy measuring metrics evaluate the goodness of the segmentation from different perspectives and may likely lead to different conclusions regarding how well manual derived contours are relative to their ground truth. The correct use of measuring metrics for segmentation accuracy evaluation have been recognized previously and led to several studies investigating how to choose appropriate accuracy assessment metrics for a given segmentation task [49]. The current study, besides in line with previous studies on the selection of evaluating metrics for general segmentation tasks, further unveils what is peculiar to accuracy assessment of PET-based target delineation, namely, the imperative of considering measuring metrics capable of characterizing the distinct bias patterns of manual contouring as is identified by the study. Finally, but probably most importantly, it is the consideration of whether there is scope to improve the accuracy of manual contouring of PET positive targets. One possibility would be to utilize the synthetic PET datasets to construct error-compensating models with taking into consideration the erring patterns of manual contouring as is highlighted in the study. Specifically on how to develop and implement models of such types warrants further investigation. One limitation of the current study is that the co-registered CT data provided to the raters during contouring corresponded to the digital phantom being used for PET simulation and hence was radiographic lesion-free. This, though offering the functionality of anatomical visualization and localization, differed to some extent from the actual clinical contouring settings as in which presentation and characteristics of the lesions on CT are available and accessible to the raters. The lack of furnishing this information to the raters for contouring of the current study may affect the clinical translation of the presented findings. To determine specifically in what way and exactly to what extent this limitation of the study plays out for the various erring patterns being identified warrants further investigation. However, for a considerable number of clinical scenarios in lung cancer management wherein PET plays a direct role in determining disease extent such as those with lesions adjacent to the mediastinum, atelectasis, *etc.* or like cases with free-breathing CT the results of the current study are clinically relevant and immediately applicable.

Conclusion

Quantification of segmentation errors associated with manual contouring of PET positive tumor targets in the lung reveals general patterns in what otherwise might be thought of as randomness. It calls for caution in assuming expert manual contouring a reliable surrogate for volumetric assessment of PET-imaged tumor lesions. Findings from the current study may allow for the formation of new hypotheses towards improving the accuracy and precision of manual delineation of PET positive tumor targets in the lung.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors are thankful to the raters who provided manual contour data to this work, although they may not agree with all the interpretation appeared in the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2019.08.014.

References

[1] Organization WH. Cancer Fact Sheet. 2018.

[2] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin 2011;61():69–90.

[3] Siegel R, DeSantis C, Virgo K, Stein K, Mariotto A, Smith T, et al. Cancer treatment and survivorship statistics, 2012. CA Cancer J Clin 2012;62:220–41.

[4] Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 2012;366:883–92.

[5] Fletcher JW, Djulbegovic B, Soares HP, Siegel BA, Lowe VJ, Lyman GH, et al. Recommendations on the use of 18F-FDG PET in oncology. J Nucl Med 2008;49:480–508.

[6] Yang F, Thomas MA, Dehdashti F, Grigsby PW. Temporal analysis of intratumoral metabolic heterogeneity characterized by textural features in cervical cancer. Eur J Nucl Med Mol Imaging 2013;40:716–27.

[7] Johnson PB, Young LA, Lamichhane N, Patel V, Chinea FM, Yang F. Quantitative imaging: correlating image features with the segmentation accuracy of PET based tumor contours in the lung. Radiother Oncol 2017;123:257–62.

[8] Yankeelov TE, Mankoff DA, Schwartz LH, Lieberman FS, Buatti JM, Mountz JM, et al. Quantitative imaging in Cancer Clinical Trials. Clin Cancer Res 2016;22:284–90.

[9] Foster B, Bagci U, Mansoor A, Xu Z, Mollura DJ. A review on segmentation of positron emission tomography images. Comput Biol Med 2014;50:76–96.

[10] Hatt M, Lee JA, Schmidtlein CR, El Naqa I, Caldwell C, De Bernardi E, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: report of AAPM task group No. 211. Med Phys 2017;44:E1–E42.

[11] Schinagl DA, Vogel WV, Hoffmann AL, van Dalen JA, Oyen WJ, Kaanders JH. Comparison of five segmentation tools for 18F-fluoro-deoxy-glucose-positron emission tomography-based target volume definition in head and neck cancer. Int J Radiat Oncol Biol Phys 2007;69:1282–9.

[12] Yang F, Grigsby PW. Delineation of FDG-PET tumors from heterogeneous background using spectral clustering. Eur J Radiol 2012;81:3535–41.

[13] Hatt M, Laurent B, Ouahabi A, Fayad H, Tan S, Li L, et al. The first MICCAI challenge on PET tumor segmentation. Med Image Anal 2018;44:177–95.

[14] Hofheinz F, Apostolova I, Oehme L, Kotzerke J, van den Hoff J. Test-retest variability in lesion SUV and lesion SUR in (18)F-FDG PET: an analysis of data from two prospective multicenter trials. J Nucl Med 2017;58:1770–5.

[15] Nahmias C, Wahl LM. Reproducibility of standardized uptake value measurements determined by 18F-FDG PET in malignant tumors. J Nucl Med 2008;49:1804–8.

[16] Zubal IG, Harrell CR, Smith EO, Rattner Z, Gindi G, Hoffer PB. Computerized three-dimensional segmented human anatomy. Med Phys 1994;21:299–302.

[17] Lewellen TK, Harrison RL. The simset Program. Monte Carlo Calculations in Nuclear Medicine. CRC Press; 2012. p. 106–29.

[18] Aristophanous M, Penney BC, Pelizzari CA. The development and testing of a digital PET phantom for the evaluation of tumor volume segmentation techniques. Med Phys 2008;35:3331–42.

[19] Yang F, Young LA, Johnson PB. Quantitative radiomics: Validating image textural features for oncological PET in lung cancer. Radiother Oncol 2018;129:209–17.

[20] Werner-Wasik M, Nelson AD, Choi W, Arai Y, Faulhaber PF, Kang P, et al. What is the best way to contour lung tumors on PET scans? Multiobserver validation of a gradient-based method using a NSCLC digital PET phantom. Int J Radiat Oncol Biol Phys 2012;82:1164–71.

[21] Zaidi H, El Naqa I. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. Eur J Nucl Med Mol Imaging 2010;37:2165–87.

[22] Palmer MR, Brownell GL. Annihilation density distribution calculations for medically important positron emitters. IEEE Trans Med Imaging 1992;11:373–8.

[23] Klein O, Nishina Y. The scattering of light by free electrons according to Dirac's new relativistic dynamics. Nature 1928;122:398.

[24] MacDonald L, Schmitz R, Alessio A, Wollenweber S, Stearns C, Ganin A, et al. Measured count-rate performance of the Discovery STE PET/CT scanner in 2D, 3D and partial collimation acquisition modes. Phys Med Biol 2008;53:3723.

[25] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.

[26] Babalola KO, Patenaude B, Aljabar P, Schnabel J, Kennedy D, Crum W, et al. An evaluation of four automatic methods of segmenting the subcortical structures in the brain. Neuroimage 2009;47:1435–47.

[27] Martin D, Fowlkes C, Tal D, Malik J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Eighth Ieee International Conference on Computer Vision, Vol Ii, Proceedings. p. 416–23.

[28] Hubert L, Arabie P. Comparing partitions. J Classif 1985;2:193–218.

[29] Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 1971;66:846–50.

[30] Meila M. Comparing Clusterings by the Variation of Information. Learning Theory and Kernel Machines. Springer; 2003. p. 173–87.

[31] Russakoff DB, Tomasi C, Rohlfing T, Maurer CR. Image Similarity Using Mutual Information of Regions. Lecture Notes in Computer Science. Berlin Heidelberg: Springer; 2004. p. 596–607.

[32] Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Mach Intell 1993;15:850–63.

[33] McLachlan GJ. Mahalanobis distance. Resonance 1999;4:20–6.

[34] Williams GW. Comparing the joint agreement of several raters with another rater. Biometrics 1976:619–27.

[35] Kendall MG. A new measure of rank correlation. Biometrika 1938;30:81–93.

[36] Dunn OJ. Multiple comparisons among means. J Am Stat Assoc 1961;56:52–64.

[37] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47:583–621.

[38] Feng M, Kong FM, Gross M, Fernando S, Hayman JA, Ten Haken RK. Using fluorodeoxyglucose positron emission tomography to assess tumor volume during radiotherapy for non-small-cell lung cancer and its potential impact on adaptive dose escalation and normal tissue sparing. Int J Radiat Oncol 2009;73:1228–34.

[39] Ford EC, Herman J, Yorke E, Wahl RL. 18F-FDG PET/CT for image-guided and intensity-modulated radiotherapy. J Nucl Med 2009;50:1655–65.

[40] Lavrenkov K, Partridge M, Cook G, Brada M. Positron emission tomography for target volume definition in the treatment of non-small cell lung cancer. Radiother Oncol 2005;77:1–4.

[41] Larson SM, Erdi Y, Akhurst T, Mazumdar M, Macapinlac HA, Finn RD, et al. 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 1999;2:159–71.

[42] Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. J Nucl Med 2009;50:122S–50S.

[43] Chen HHW, Chiu NT, Su WC, Guo HR, Lee BF. Prognostic value of whole-body total lesion glycolysis at pretreatment FDG PET/CT in non-small cell lung cancer. Radiology 2012;264:559–66.

[44] Yang F, Young L, Grigsby P. Predictive value of standardized intratumoral metabolic heterogeneity in locally advanced cervical cancer treated with chemoradiation. Int J Gynecol Cancer 2016;26:777–84.

[45] Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48:441–6.

[46] Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imaging 2012;30:1234–48.

[47] Kim H, Monroe JI, Lo S, Yao M, Harari PM, Machtay M, et al. Quantitative evaluation of image segmentation incorporating medical consideration functions. Med Phys 2015;42:3013–23.

[48] Berthon B, Spezi E, Galavis P, Shepherd T, Apte A, Hatt M, et al. Toward a standard for the evaluation of PET-Auto-Segmentation methods following the recommendations of AAPM task group No. 211: requirements and implementation. Med Phys 2017;44:4098–111.

[49] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15:29.

[50] Yang F, Grigsby P. A segmentation framework towards automatic generation of boost subvolumes for FDG-PET tumors: a digital phantom study. Eur J Radiol 2012;81(12):4123–30.