BMC Medical Imaging

RESEARCH

Open Access

# Comparative analysis of intestinal tumor segmentation in PET CT scans using organ based and whole body deep learning

Mahsa Torkaman[1*], Skander Jemaa[1], Jill Fredrickson[1], Alexandre Fernandez Coimbra[1], Alex De Crespigny[1] and Richard A. D. Carano[1,2]

## Abstract

**Background** 18-Fluoro-deoxyglucose positron emission tomography/computed tomography (FDG-PET/CT) is a valuable imaging tool widely used in the management of cancer patients. Deep learning models excel at segmenting highly metabolic tumors but face challenges in regions with complex anatomy and normal cell uptake, such as the gastro-intestinal tract. Despite these challenges, it remains important to achieve accurate segmentation of gastro-intestinal tumors.

**Methods** Here, we present an international multicenter comparative study between a novel organ-focused approach and a whole-body training method to evaluate the effectiveness of training data homogeneity in accurately identifying gastro-intestinal tumors. In the organ-focused method, the training data is limited to cases with intestinal tumors which makes the network trained with more homogeneous data and with stronger presence of intestinal tumor signals. The whole body approach extracts the intestinal tumors from the results of a model trained on the whole-body scans. Both approaches were trained using diffuse large B cell (DLBCL) patients from a large multicenter clinical trial (NCT01287741).

**Results** We report an improved mean(±std) Dice score of 0.78(±0.21) for the organ-based approach on the hold-out set, compared to 0.63(±0.30) for the whole-body approach, with the *p*-value of less than 0.0001. At the lesion level, the proposed organ-based approach also shows increased precision, recall, and F1-score. An independent trial was used to evaluate the generalizability of the proposed method to non-Hodgkin's lymphoma (NHL) patients with follicular lymphoma (FL).

**Conclusion** Given the variability in structure and metabolism across tissues in the body, our quantitative findings suggest organ-focused training enhances intestinal tumor segmentation by leveraging tissue homogeneity in the training data, contrasting with the whole-body training approach, which, by its very nature, is a more heterogeneous data set.

**Keywords** FDG-PET/CT, Intestinal tumors, Segmentation, Data homogeneity, Deep learning

## Introduction

Convolutional neural networks (CNNs) have shown remarkable efficacy across diverse medical imaging modalities, excelling in different tasks like lesion detection , organ segmentation, disease classification, and predicting treatment outcomes [1–5]. Specifically in tumor

*Correspondence:
Mahsa Torkaman
torkaman.mahsa@gene.com
[1] Genentech, Inc, South San Francisco, CA, USA
[2] F. Hoffman-La Roche Ltd, Basel, Switzerland

identification, they have enabled rapid and reproducible tumor extraction, surpassing manual and semi-automatic methods commonly used in clinical settings [6, 7]. Accurate identification and delineation of lymphoma tumors throughout the entire body are crucial for early detection, evaluating disease progression, and guiding personalized treatment strategies, with CNNs demonstrating notable precision in this field [8–12]. Among various imaging modalities, PET imaging plays a pivotal role in lymphoma diagnosis and management [13, 14].

FDG-PET is a noninvasive functional imaging technique widely used in oncology that enables evaluating tumor stage and metabolic characteristics with high specificity and sensitivity [14, 15]. FDG-PET scans, combined with complementary anatomical information from low-dose CT images, are the key components to assess the metabolic activity of tumors which contributes to disease diagnosis and treatment assessments in lymphoma cancer patients [16, 17]. $^{18}$F-radiolabeled glucose, administered intravenously during PET imaging, concentrates in tumors with high metabolic activity due to increased glucose metabolism in cancer cells. This makes FDG-PET imaging superior to magnetic resonance imaging and CT, providing an effective method for early cancer detection and diagnosis by quantifying metabolically active tumor burden [18]. Many cancer types, such as non-small cell lung cancer, non-Hodgkin's lymphoma and other malignancies, exhibit significant uptake of the FDG tracer by cancer cells, rendering them highly visible in PET images [19, 20]. Overall, FDG-PET/CT provides valuable clinical information, contributing to improved diagnosis, treatment planning, and monitoring of various diseases [14, 21, 22].

Segmentation of FDG-avid tumors on FDG-PET has shown high prognostic capabilities in NHL [13, 22–25]. In clinical settings, this segmentation is typically performed manually, with the help of image analysis softwares, by a radiologist or a nuclear medicine specialist. Using specialized softwares, physicians often manually outline the desired region of interest (tumor boundaries) by drawing contours around the target area in the PET images. Metrics such as maximum standardized uptake value (SUV) within a target region of interest (ROI) and tumor volume are typically recorded and tracked over the course of treatment. In advanced-stage malignancies, the presentation of tumors on FDG-PET scans exhibits variability in different anatomical locations. The manual delineation or seeding of target regions of interest (ROIs) is often meticulous, subjective, and prone to high intra- and inter-observer variability [26, 27]. Several methods leveraging the recent advances of Deep Learning have been proposed to automatically delineate metabolically active lesions on eyes to thighs PET/CT scans [8–11].

Segmentation of intestinal tumors is crucial for cancer patients, improving diagnosis, treatment, and monitoring. For instance, in gastric cancer as one of the leading causes of cancer-related deaths, accurate segmentation addresses challenges like late-stage diagnoses and poor treatment outcomes. Similarly, in lymphoma intestinal involvement requires precise segmentation to enable precise tumor assessment and guides effective therapy. While traditional treatments like surgery and chemotherapy are common, emerging approaches in radionanomedicine are demonstrating significant potential for medical advancements. Advances such as integrin receptor-targeted therapies, HER2, Claudin 18, and glutathione-responsive systems improve diagnostic precision and enable personalized treatment, addressing early diagnosis challenges and minimizing systemic toxicity for gastric cancer patients [28]. This work focuses on using PET imaging, a minimally invasive method for effective intestinal tumor segmentation, to enable diagnostic precision and personalized treatment through a more complete disease burden assessment for lymphoma patents.

Intestinal tumors are particularly challenging for both manual and AI-based tumor segmentations due to their intricate anatomical shapes, close adjacency to organs with similar appearances, locally variable presentations, the presence of acute non-malignant uptakes, particularly notable in patients with diabetes or those taking metformin, and the small sizes of the tumors [29–32].

This study specifically focuses on organ-based tumor segmentation in lymphoma, with a particular emphasis on the intestine. While previous research has explored tumor segmentation in lymphoma patients at a holistic level, this work highlights the significance of examining tumors at the organ level for more precise diagnostics and effective treatment planning. The study's focus extends to a comparative analysis, aiming to evaluate the impact of training data homogeneity on the accuracy of intestinal lesion segmentation in FDG-PET/CT images of patients with NHL. The study introduces a segmentation approach that focuses on the intestine organ, employing an organ-focused strategy that minimizes interference from surrounding non-intestinal tissue (Fig. 1). This organ-focused approach is compared to a previously published approach that utilizes whole body scans and the intestinal lesion segmentation is extracted from the results of whole body training [8]. The organ-focused approach is a targeted approach that provides stronger signal during training, resulting in improved quantitative results and more accurate segmentation of intestinal lesions compared to the whole body training approach. We evaluated the generalizability of the organ-focused approach by testing it to two independent, international, and multi-center clinical trials involving subtypes

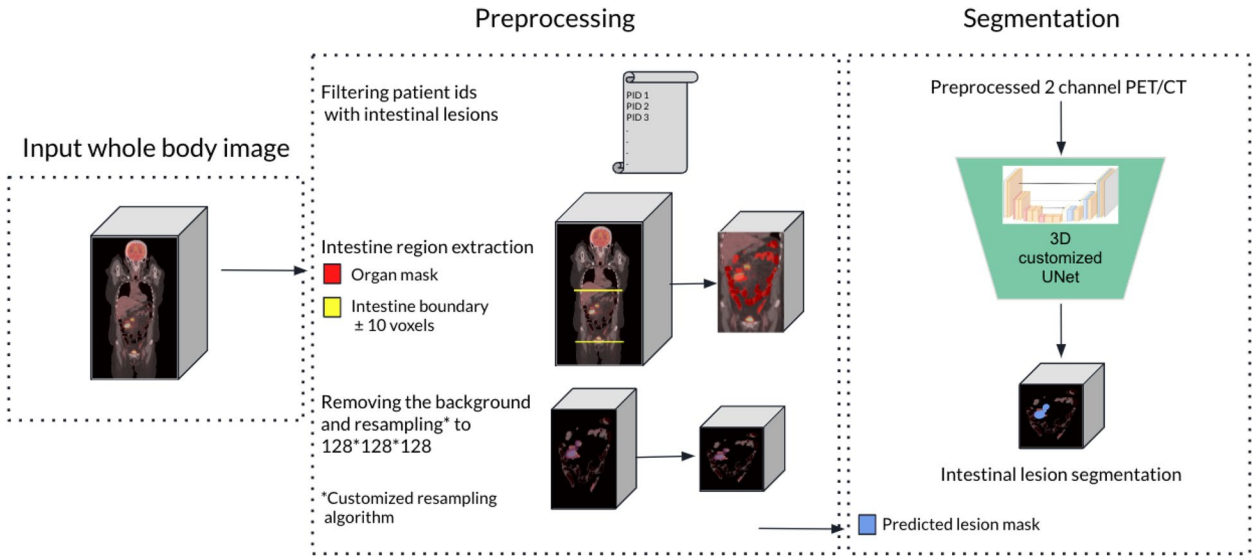



**Fig. 1** Organ-based model workflow diagram

of NHL: DLBCL and FL. This assessment allowed us to determine the effectiveness and applicability of organ-focused approach across different types of lymphomas and ascertain its potential for broader use in medical imaging segmentation tasks.

## Methodology

To investigate the impact of training data homogeneity on the segmentation results of intestinal tumors, we propose an organ-focused 3D CNN approach where the training data is limited to the small intestine region and compare the results to the previously published whole body training approach [8]. By focusing on the intestine, we can provide the CNN with more homogenous data and allow the network to be better guided during the training process.

### Data and preprocessing

Goya (NCT01287741 [33]) and Gallium (NCT01332968 [34]) are randomized open-label phase III clinical trials for previously untreated patients with DLBCL and FL respectively. All scans were taken following standardized image acquisition procedures and were reviewed by an independent review committee. Three board-certified nuclear medicine physicians segmented and assessed the tumors semi-automatically (using the MIM Software Inc, OH, USA) at an independent review facility [24]. These annotations served as the "ground truth" for our models' development and assessment.

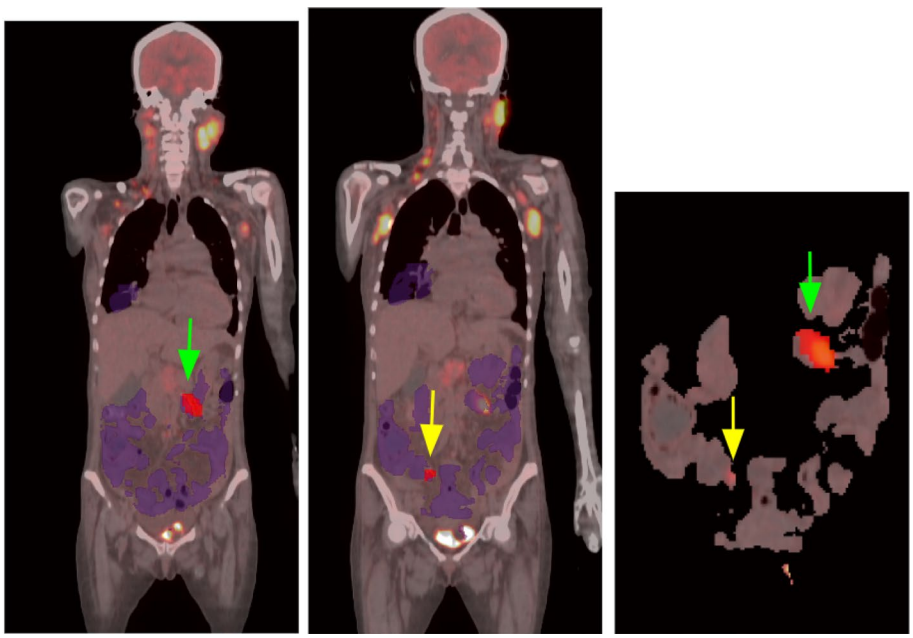
The training data set in the whole body approach is sourced from the Goya clinical trial. There are 1166 FDG-PET/CT scans in this trial which 80 percent of it used for the training purpose, whereas the remaining 20 percent was included in the test set. The whole body model was tested on both Goya and Gallium clinical trials to show the generalizability of the model to different NHL subtypes. To prevent any cross-contamination of data between training and test sets, we adopted a patient-level data division strategy in the whole body approach. Whole body scans underwent resampling to have a uniform isotropic voxel size of $2 \times 2 \times 2$ mm.

The same Goya trial dataset was utilized in training the organ-focused approach. Figure 1 illustrates the high-level workflow of the organ-based model, providing an overview of the entire process which is explained in more detail in this section, Data and preprocessing and the subsequent section, Whole body and organ-focused training. Both whole-body and organ-focused approaches were trained and evaluated using PET and CT as dual-channel inputs. While FDG-PET is the primary modality for detecting tumors by highlighting areas of increased glucose metabolism, CT enhances tumor localization and boundary delineation by providing detailed anatomical context. This combination improves segmentation accuracy [35].

Samples that contained intestinal lesions were chosen and underwent a preprocessing procedure for the organ-focused model development. From the complete set of whole body scans in the Goya trial, 310 scans were detected with presence of at least one tumor with an overlap of no less than 20 percent with the intestine (defined by an intestine segmentation mask described below). The training and test sets in the organ-focused approach were chosen from the training and test sets

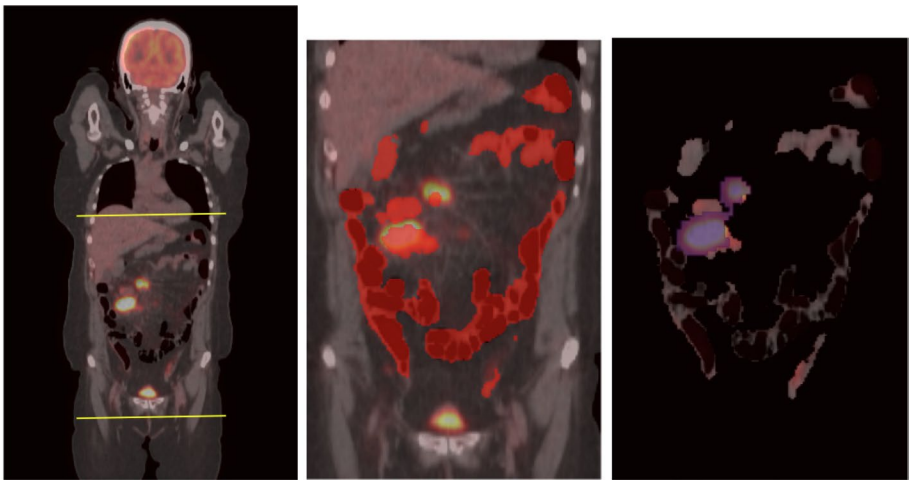
of the whole body approach respectively to ensure fair comparison between the two models. Based on this, we partitioned the 310 scans from the Goya trial into train (215 scans), validation (26 scans), and test sets (69 scans). The test set specifically was used to compare the two approaches on patients with DLBCL subtype. Gallium trial was used as an additional test set to show that the model is generalizable to the NHL follicular lymphoma subtype as well. 116 scans with intestinal lesions were chosen from Gallium trial to evaluate and compare the two approaches on patients with follicular lymphoma. Patients characteristics in the test sets can be seen in supplementary Tab. 1.

The data underwent the following preprocessing steps in preparation for the organ-based training. First, the intestine organ mask from the organ segmentation results was extracted for each of the scans. In this method, skilled imaging technologists initially segmented various organs, including the small intestine, followed by review and validation from a radiologist. The resulting masks were used to train a VNet model exclusively on low dose CT data for organ segmentation. This model underwent training on a subset of the training data from the Goya trial and was subsequently tested on both the Goya and Gallium hold-out sets. Additional information about the organ segmentation process can be found in [36]. Second, we iterated through all the tumors in the whole-body images and included the ones that overlapped by at least 20 percent or more with the intestine organ and excluded other non-intestinal tumors to make sure our lesion masks contained only intestinal lesions. The chosen threshold is adaptable and can be tailored according to the specific organ under investigation, the quantity of available data, and the density of lesions surrounding the organ of interest. We chose a 20 % threshold to ensure an adequate inclusion of intestinal lesions for both training and analysis. Regions that correspond to non-intestinal tumors are removed from both PET and CT scans to avoid including misleading signals in the training. Figure 2 shows an example with an intestinal (green arrow) and a non-intestinal lesion (yellow arrow) on a sample PET/CT scan. Through these preprocessing steps, the whole body scans were transformed to exclusively encompass the intestine and intestinal lesions overlapping with it. Figure 3 shows a sample PET/CT 2D slice from a whole body scan on the left where the yellow bars show the region of the intestine. The middle figure shows a 2D slice from the restricted volume to the intestine region ± 10 voxels and the red label shows the intestine organ mask. In the figure on the right, the background is removed and only the intestine organ in PET/CT and tumors overlapped with it shown in purple were kept. Third, small lesions with volume less than 0.8 ml (100 voxels) which were distant from any other lesion by at least 10 voxels were removed. This step is implemented to prevent the unintentional removal of small lesions that may arise as a result of tumors with weak connections breaking apart during the resampling procedure.



(a) intestinal lesion (b) non-intestinal lesion (c) extracted intestine

**Fig. 2** An example of intestinal (green arrow) and non-intestinal (yellow arrow) lesions



(a) intestine region (b) intestine organ mask (c) intestinal lesions

**Fig. 3** An example of a whole-body scan transformed to exclusively encompass the intestine and intestinal lesions

In other words, the objective was to ensure the exclusion of only small lesions that were not generated as a direct consequence of the resampling process. In alignment with the manual TMTV annotations protocols, lesions below 0.8 mL were excluded. Minimal size and metabolic activity thresholds are often used in TMTV reading protocols [24, 37, 38]. These thresholds for size and metabolic activity help minimize the inclusion of false positive uptakes on the PET and increase the medical experts' confidence in their measurement of the total metabolic tumor burden. In addition, very small objects might not have clinical significance and can be considered as noise or artifacts. Removing them can help ensure that the model focuses on clinically relevant structures. Fourth, processed data from the previous steps was resampled to the target size of $128 \times 128 \times 128$ in preparation for the training phase. We performed a customized resampling to prioritize the preservation of information and minimize data loss. Therefore, for each processed whole body scan, the maximum dimension was adjusted to 128, while other dimensions were padded with zero in case of PET/mask and $-1024$ in case of CT scans to achieve the desired target size. The predicted masks from the whole body model underwent the same preprocessing step as described above to ensure an unbiased assessment of the models' performance on unseen data.

#### Whole body and organ-focused training

Tumor segmentation from whole-body FDG-PET/CT scans was performed using cascaded 2D and 3D CNNs [8]. This method employs a two-step approach involving 2D and 3D tumor segmentation using modified UNet and VNet networks. The process begins with 2D tumor segmentation on individual slices using an adapted UNet. Subsequently, a refined 3D VNet architecture is employed to enhance the 2D segmentation. The final mask is then derived by averaging the tumor masks from both 2D and 3D segmentations. This method has undergone training on the Goya trial dataset and was tested on two distinct NHL subtypes: DLBCL and FL. Further information regarding the whole body approach is available in [8].

After preprocessing of the data based on the steps explained in the Data and preprocessing section, a UNet like network was used to train an organ-based model for segmentation of intestinal lesions. Supplementary Fig. S4 shows the architecture used to train the organ-focused model. The network follows an encoder-decoder architecture with symmetrical skip connections between various stages, inspired by the original UNet model [39].

The hyperparameters were chosen empirically by adjusting the network based on its performance on the validation set. The network's initial weights of the layers were selected randomly from normal distribution and Adam optimizer [40] was utilized to optimize the weights during the training. The initial learning rate was set to 0.0005, and we trained the model for 300 epochs with a batch size of 4. We monitored the learning curve throughout to ensure convergence. Additionally, we employed a learning rate scheduling strategy using ReduceLROnPlateau from the Keras library, which monitored the validation loss. If the validation loss did not improve for 10 consecutive epochs, the learning rate was reduced by a factor of 0.1, with a minimum learning rate of 0.0. The main software and libraries used in this study, along with their respective versions, are detailed in

Supplementary Material, Section "List of Softwares and Libraries". We combined both the cross entropy and dice losses (Eq. 1) to address the class imbalance issue inherent in the images, where there is a significantly higher proportion of negative voxels (no lesion) compared to positive ones (lesions). Combining both losses helps mitigate the effect of the class imbalance and encourages the network to produce accurate pixel-wise probabilities and well-defined object boundaries [41].

$$L = \left[1 - \frac{2|P \cap T| + 1}{|P| + |T| + 1}\right] - \left[\sum_{v \in V} \frac{|V|}{\sum_{v \in V} y_v} \left(y_v \log\left(\hat{y_v}\right) + \left(1 - \frac{|V|}{\sum_{v \in V} y_v}\right)\left(1 - y_v\right) \log\left(1 - \hat{y_v}\right)\right)\right] \tag{1}$$

In the above equation, $V$ represents individual voxels within an image. $T$ represents the collection of positive voxels in the ground truth, while $P$ denotes the set of voxels predicted as positive by the model. $y_v$ represents the value of the voxel $v$ in the tumor mask, and $\hat{y_v}$ is the value of the voxel $v$ in the predicted tumor mask.

#### Evaluation metrics and statistical analysis

To assess the accuracy of the organ-focused approach and contrast it with the whole-body approach, we employed the Dice score and conducted Wilcoxon test on this metric to demonstrate statistical significance. Additionally, we conducted lesion-level analysis, calculating precision, recall, and F1 score to further evaluate performance of the intestinal tumor segmentation at the lesion level. Total metabolic tumor volume (TMTV) was also computed for the predicted masks of the two approaches and compared to the ground truth to show that the organ-based method has a higher correlation with the ground truth in computing the burden of metabolically active disease in patients with lymphomas. The Spearman correlation coefficient was computed for both approaches, serving as a statistical measure to further evaluate the strength of the relationship between predicted results and the ground truth in each case.

To calculate the precision, recall, and F1 score for the lesion-level analysis, we assessed lesions in the ground truth against the predicted masks and vice versa. We classified a lesion in the predicted mask as a true positive (TP) if it had an overlap of over 20 percent with the ground truth mask. Conversely, if the overlap was less than 20 percent, we categorized it as a false positive (FP). Similarly, for false negatives (FN), we identified lesions in the ground truth mask with less than 20 percent overlap to the prediction mask as false negatives. These metrics were used to compute the precision, recall, and F1 score (Eq. 2). $p$-values at the lesion level were generated using 1000 bootstrap random samples with replacement [42]. The bootstrapped random samples were of the same size as the original dataset.

In the lesion-level analysis, we incorporated a binary closing operation with the kernel size of (3,3,3) as a post processing step to counteract the potential issue of lesion fragmentation caused by weak connections, as previously discussed in the preprocessing step. This binary closing procedure involves sequentially applying dilation and erosion operations to the masks, effectively reconnecting or "closing" lesions that may have been split. By doing so, we aimed to enhance the accuracy of the lesion analysis, ensuring that fragmented or loosely connected lesions do not affect the overall assessment. Following this, we repeated the process of removing small and isolated lesions, similar to the preprocessing step. These steps ensure that the lesion-level analysis focuses on clinically important lesions with coherent structures.

$$\begin{aligned} precision &= \frac{TP}{TP + FP} \\ recall &= \frac{TP}{TP + FN} \\ F1 &= \frac{2 \times (precision \times recall)}{(precision + recall)} \end{aligned} \tag{2}$$

Higher F1 score suggests a better balance between precision and recall, which is often desired in many applications. In the statistical analysis we employed Wilcoxon test and computed p-value to assess the significance of differences between the two models based on the computed Dice scores. Results with p-values less than 0.05 considered as statistical significance. We also used a 95 percent confidence interval measure on computed Dice scores. Confidence interval is a statistical measure which quantifies the uncertainty or variability on the computed measures, helping in assessing the precision of the estimates.

## Results

The quantitative results on both Goya and Gallium test sets can be seen in Tables 1 and 2. Both the Dice score and the lesion level metric F1 demonstrate that the organ-focused method outperforms the whole body approach quantitatively. Figures 4 and 5 shows the graphs of the average Dice score with standard deviations and

**Table 1** Goya trial test results

| Goya | Dice[a] | Precision | Recall | F1 |
|---|---|---|---|---|
| Organ-focused | **0.78±0.21** | **0.91±0.04** | **0.85±0.05** | **0.88±0.04** |
| Whole-body | 0.63±0.30 | 0.85±0.3 | 0.82±0.31 | 0.82±0.3 |

Quantitative results on Goya trial test cases show that the organ-focused approach outperforms the whole-body approach

[a]Dice score is shown in mean±standard deviation format

**Table 2** Gallium trial test results

| Gallium | Dice[a] | Precision | Recall | F1 |
|---|---|---|---|---|
| Organ-focused | **0.70±0.25** | 0.67±0.07 | **0.75±0.07** | **0.70±0.04** |
| Whole-body | 0.58±0.31 | **0.73±0.4** | 0.70±0.4 | 0.69±0.4 |

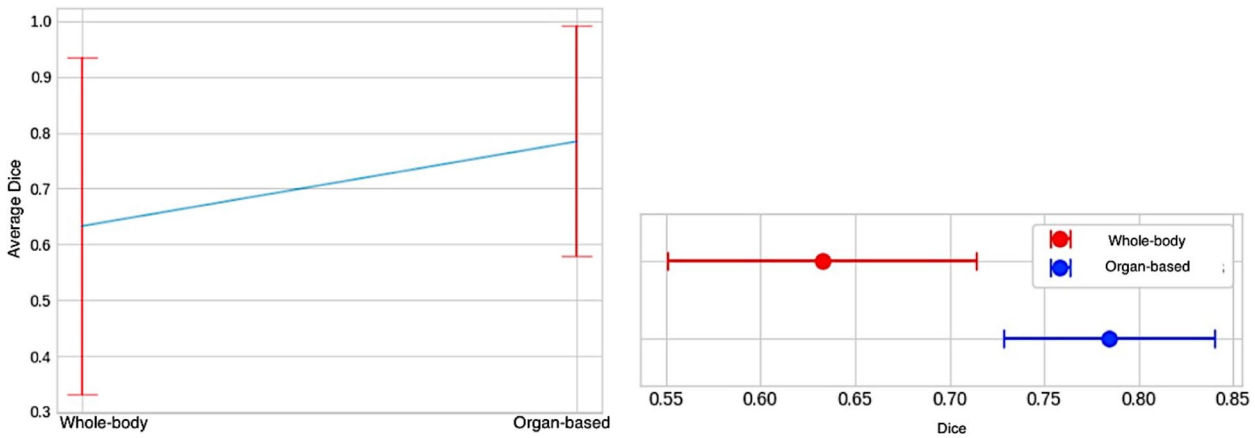
Quantitative results on Gallium trial test cases show that the organ-focused approach outperforms the whole-body approach in terms of the majority of the metrics

[a]Dice score is shown in mean±standard deviation format

confidence intervals for both the organ-focused and the whole-body approaches. The organ-focused approach achieved better Dice scores at a statistically significant level for both test sets ($p < 10^{-5}$ in Goya and $p < 10^{-3}$ in Gallium). Figures 4 and 5 also show the organ-focused approach can generate intestinal tumor segmentation results with less variability and greater consistency among different cases (smaller standard deviation than the whole body approach). The $p$-values for precision, recall, and F1 score in the Goya test sets after bootstrapping are 0.38, 0.47, and 0.48 respectively. For the Gallium test sets are 0.3, 0.5, and 0.4 respectively. These p-values indicate that there is no statistically significant difference between the methods at the lesion level.

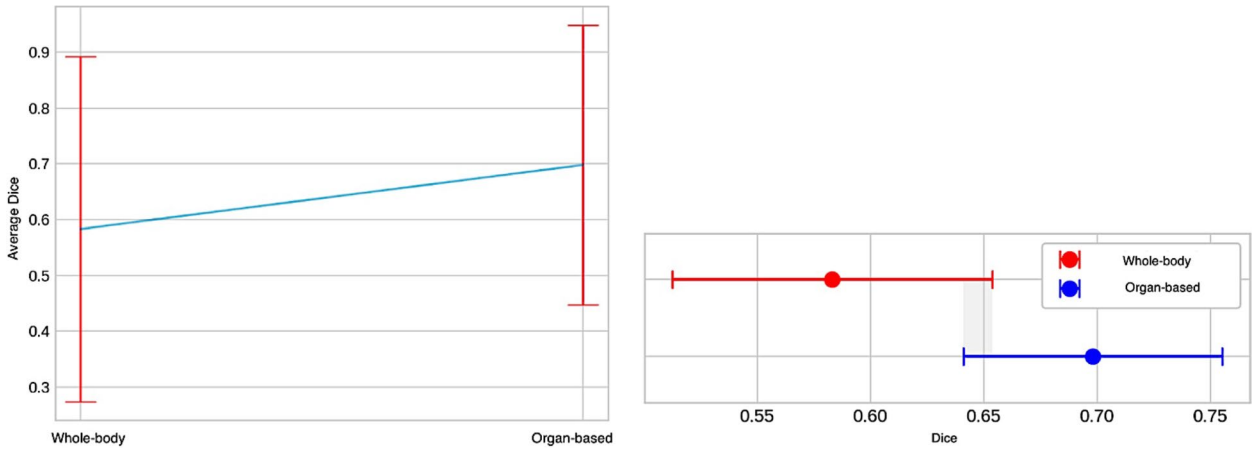
The organ-focused approach also showed better compatibility in terms of the Dice score with the ground truth masks than the whole body approach in the Gallium test set (Table 2, 0.70 vs 0.58, $p < 10^{-3}$). Higher Dice score indicates better overall overlap and similarity between the predicted and ground truth masks. This is valuable in medical images because we prioritize accurately capturing the true positive regions and minimizing false negatives.

Intestinal metabolic tumor volume was calculated for each of the predicted tumor masks generated by both approaches and compared to the volumes in the ground truth masks. Figure 6 and 7 show comparison of automated total metabolic tumor volume with "ground truth" values in DLBCL and FL test patients respectively for whole body and organ-based approaches. Red lines in these figures are the linear fitted line and the dashed black lines are the line between minimum (bottom left) and maximum (top right) points. These figures show that both approaches are well correlated with the ground truth in estimation of metabolic tumor burden, although there is a greater spread in the data for the whole body approach. Whole-body and organ-based approaches generated Spearman's correlations of 0.86 and 0.93 respectively for the Goya test set and 0.77 and 0.88 respectively for the Gallium test set. In addition, Whole-body and organ-based approaches generated coefficients of determination ($R^2$) of 0.84 and 0.91 respectively for the Goya test set and 0.50 and 0.89 respectively for the Gallium test set.The results of the both computed Spearman's correlation and $R^2$ support that there is a more consistent relationship between the organ-focused approach and the ground truth in comparison to the whole body approach.



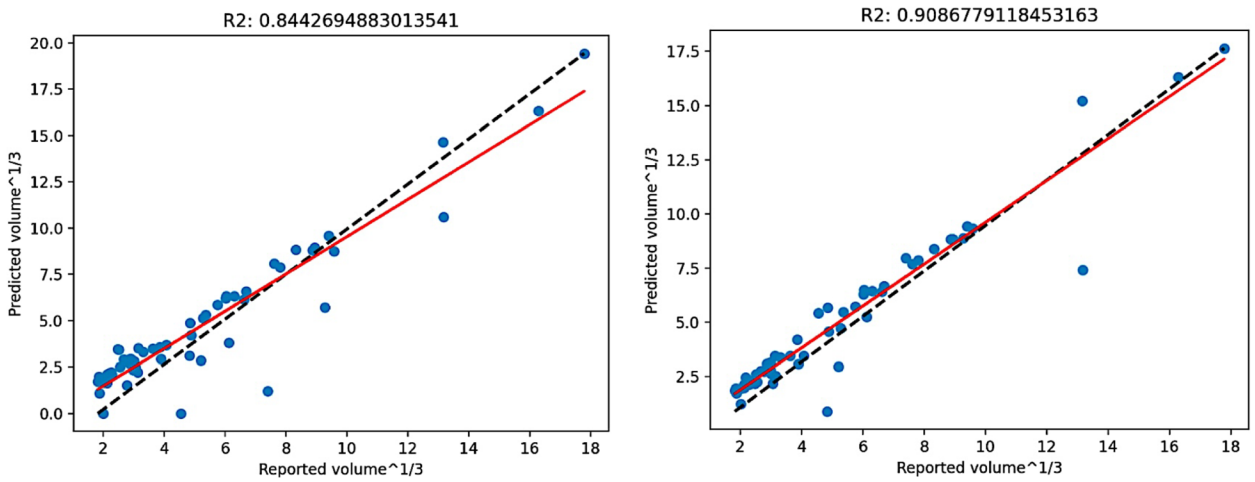

(a) error bars (b) 95% confidence intervals for Dice

**Fig. 4** Error bars and 95% confidence intervals for Dice in Goya tests comparing both methods




(a) error bars (b) 95% confidence intervals for Dice

**Fig. 5** Error bars and 95% confidence intervals for Dice in Gallium tests comparing both methods




(a) whole body (b) organ-based

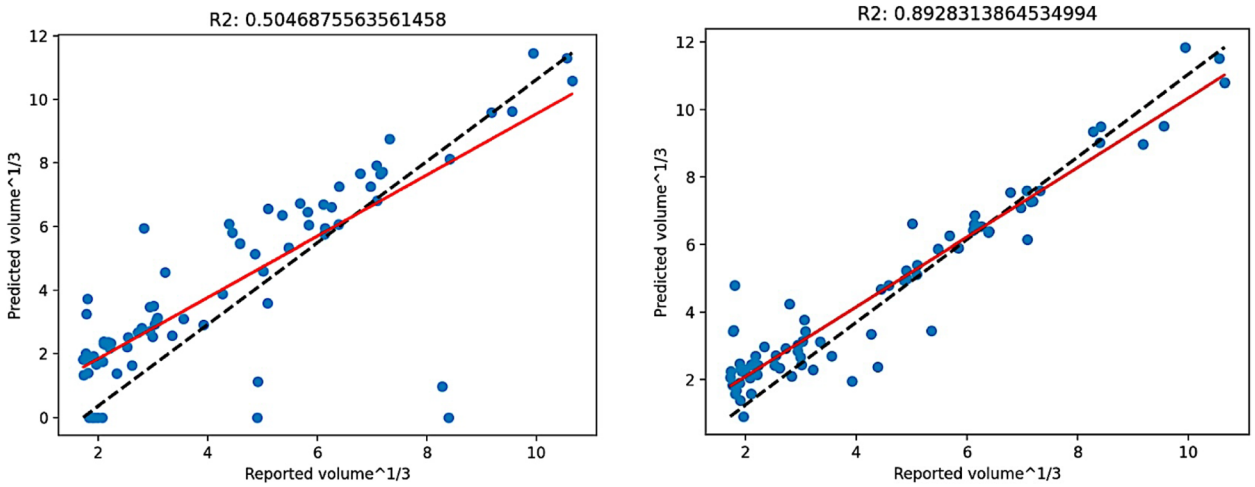
**Fig. 6** Comparison of automated total metabolic tumor volume with "ground truth" values in DLBCL test patients

Supplementary Fig. S1, Fig. S2, and Fig. S3 provide visualizations showcasing instances where organ-based approach outperform whole body approach in terms of Dice score, vice versa, or scenario where both methods exhibit comparable performance.

## Discussion

FDG-PET is a well-validated imaging technique which is widely used in oncology to quantify physiological and biochemical processes in a highly sensitive and non-invasive way. Combined PET/CT scanners are commonly used in clinics and additional anatomic information from CT in these scans can help identify and localize abnormal metabolic activities. Most cancer types show substantial FDG uptake by the cancer cells due to increased metabolism of glucose and this provides a means to visualize and quantify the active tumors from FDG-PET/CT scans. Among different malignancies, non-Hodgkin's lymphoma cancer cells have demonstrated high FDG uptake which makes them highly visible in the FDG-PET images.

The aim of this study is to investigate the impact of training data homogeneity by contrasting an organ-focused model with a whole-body training approach in segmenting intestinal tumors from FDG PET/CT scans. Specifically, we conduct a comparative analysis between a whole-body fully automated metabolic tumor segmentation and an organ-specific model for the segmentation of metabolically active lesions in the gastro-intestinal tract.




(a) whole body (b) organ-based

**Fig. 7** Comparison of automated total metabolic tumor volume with "ground truth" values in FL test patients

The results on Dice score has p-value of less than 0.005 which shows the statistical significance of the organ-focused approach. Both F1 metric in lesions-level analysis and metabolic burden computation support the superiority of the organ-focused approach in intestinal tumor segmentation. These findings indicate that taking into account the complex anatomical structures and non-malignant metabolic activity within the intestinal region, intestinal lesion segmentation benefits from homogeneity of training data in the organ-based approach. This targeted training approach by focusing on the intestine and minimizing the influence of neighboring tissues generates more accurate outcomes with much less training data compared to the whole-body approach (215 vs. 933 visits), which is characterized by a higher heterogeneous anatomical background.

An interesting discovery is the generalizability of the results. By assessing the models on both the Goya and Gallium hold-out datasets, we illustrated their ability to generalize results across diverse NHL subtypes, specifically DLBCL and FL, and across data sourced from multiple countries (supplementary Tab. 1). This aspect is crucial for mitigating bias and ensuring consistent and applicable performance across the target patient demographic.

We believe that extending the organ-focused methodology to encompass all major organs could yield a more precise evaluation of total metabolic tumor burden. This extension is grounded in the observation that regional homogeneity within most organs surpasses the homogeneity observed across the entire body. Total metabolic tumor burden stands as a prognostic indicator of clinical outcomes [14].

Our comprehensive preprocessing steps, guided by a data-centric model development approach, support the importance of meticulous data curation in the model development process. This approach proves particularly advantageous when dealing with relatively small datasets [43].

It is noteworthy that our organ-focused training set comprises fewer data points compared to the whole-body approach, yet it yields superior quantitative results. This indicates that the regional specificity inherent in the organ-focused approach, along with the thorough preprocessing guided by a data-centric model development, plays a crucial role in its effectiveness.

Despite the comprehensive analysis, this study has several limitations. We employed an in-house algorithm to register PET and corresponding CT data, ensuring precise alignment and minimizing mismatches between the two modalities. However, in rare cases, mismatches between the two modalities are unavoidable, particularly when working with international multicenter data, as in our study. The performance of the organ-focused approach relies on the accuracy of the organ segmentation model. If the model does not segment the intestine accurately, the error inevitably propagates downstream to the organ-based approach. We used low-dose CT data, typically acquired alongside PET images, to train our organ segmentation algorithm. Using a CT-free, PET organ segmentation approach could address the limitations that may happen due to mismatches between the

PET and CT information by only utilizing PET images for organ segmentation [44]. In addition, in instances where the organ-focused approach exhibits suboptimal performance, our preliminary assessment suggests that the whole-body training, leveraging a larger and more diverse dataset, equips the model with enhanced capability to accurately segment tumors with lower PET uptake. This is especially important for signals that may not dominate in a specific organ, helping to mitigate challenges in detection signals. The organ-focused approach requires preprocessing steps to prepare the data for training and testing. The average preprocessing time to prepare a PET/CT scan for applying the organ-focused model is 25 seconds for the Goya trial and 33.58 seconds for the Gallium trial. The inference time for the organ-focused approach is 1.38 seconds on average for each test case in Goya and 1.37 seconds on average for each test case in Gallium, whereas for the whole-body approach, it is 54.9 seconds for Goya and 54.7 seconds for Gallium on the same test sets. The whole body has a higher inference time due to memory limitations associated with large 3D volumes, necessitating both 2D and 3D segmentation [8]. In contrast, organ-focused segmentation has a shorter inference time because it deals with smaller regions and does not require the same techniques needed for large whole-body volumes. The experiments for this task were conducted on a CPU node, featuring 36 CPU cores and 278 GB of RAM. As a future work, software optimization is needed to reduce the execution time of the preprocessing step in the organ-focused approach. We also aim to apply this model to real-world data to assess its generalizability. Additionally, we plan to explore the advantages of using a similar organ-specific approach for lesion segmentation in other organs.

In summary, this manuscript demonstrates that an organ-focused approach, leveraging a hierarchical approach to lesion segmentation achieves superior results in complex organs like the intestine for lymphoma patients. Organ-focused approaches also use significantly less data and memory compared to a whole-body approach. By emphasizing homogeneity in training and leveraging robust organ segmentation, our method highlights the benefits of a data-centric approach. Our findings hold promise for advancing automation, improving efficiency, and facilitating personalized approaches in oncology. This progress is anticipated to translate into enhanced patient outcomes.

## Conclusions and future perspective

We propose a novel 3D deep learning, organ-focused approach for intestinal tumor segmentation using FDG-PET/CT data from lymphoma patients, enabling robust and automated segmentation of intestinal tumors. Segmentation of FDG-avid tumors and the computation of total metabolic tumor volume on FDG-PET have demonstrated significant prognostic value in NHL patients. Our method outperformed the whole-body segmentation approach, achieving superior results with four times less data. Given the challenges of tumor annotation in medical imaging, such as time-consuming processes, potential errors, and inter- and intra-observer variability, our approach demonstrates the value of reducing data requirements without compromising accuracy. The model was trained on DLBCL lymphoma patient data and tested on both DLBCL and FL subtypes, showcasing its generalizability across different lymphoma types. Quantitative metrics at both the lesion and image levels, along with statistical analysis, demonstrated a strong agreement between the proposed method and the radiologist-generated ground truth segmentations.

Future work will extend this approach to other organs and assess its performance on real-world data beyond clinical trial settings. Additionally, if PET images with tracers other than FDG become available, we plan to explore PET-only training methods, as outlined in [44], to overcome the limitations of low-dose CT images used for the training of organ segmentation. In addition, we plan to explore extracting radiomic features from the segmented tumors, as these can enhance prognostic value and support personalized treatment planning. Integrating radiomics into our pipeline will expand the applicability of our method and strengthen AI-driven medical decision-making in oncology. Given the availability of data from gastric cancer patients, we plan to adapt and refine our current model for application to this patient population. This effort is particularly significant as gastric cancer remains the second leading cause of cancer-related deaths worldwide [28].

**Abbreviations**

| | |
|---|---|
| FDG | 18-Fluoro-deoxyglucose |
| PET | positron emission tomography |
| CT | computed tomography |
| NHL | Non-Hodgkin's lymphoma |
| DLBCL | diffuse large B cell |
| FL | Follicular lymphoma |
| SUV | standardized uptake valu |
| ROI | region of interest |
| DL | deep learning |
| CNN | convolutional neural networks |
| FN, FP, TP, TN | false negative, false positive, true positive, true negative |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12880-025-01587-3.

Supplementary Material 1.

**Acknowledgements**
Not applicable.

**Clinical trial numbers**
Goya ClinicalTrials.gov identifier: NCT01287741.
Gallium ClinicalTrials.gov identifier: NCT01332968.

**Authors' contributions**
MT, SJ, and RC worked on the methodology, study design, and prepared the main manuscript and all the figures. JF, ADC, and AFC worked on clinical trial design, imaging protocol, data collection and preparation. All authors read, contributed to the editing of the manuscript, and approved the final manuscript.

**Funding**
The authors state that this work has not received any external funding.

**Data availability**
The datasets generated and/or analyzed during the current study are not publicly available due to regulatory restrictions pertaining to patients' privacy but are available from the corresponding author on reasonable request. Data sharing requests for patient-level Roche clinical trials data (excluding imaging) through a data request platform: At the time of writing this request platform is Vivli. https://vivli.org/ourmember/roche/. For up to date details on Roche's global policy on the sharing of clinical information and how to request access to related clinical study documents, see here: https://go.roche.com/data_sharing. Requests for access to the imaging data should be directed to imaging_data-sharing-d@gene.com.

## Declarations

**Ethics approval and consent to participate**
This is a retrospective analysis of two multi-institutional clinical trials, NCT01287741 [33] and NCT01332968 [34]. The trials were conducted in accordance with the Declaration of Helsinki and the International Conference on Harmonization of Good Clinical Practice guidelines. The retrospective analysis of this data presented in this manuscript was approved by Genentech Data Governance Committee (Genentech Inc) and is consistent with the intended use of this data defined by the original ethics committee/institutional review board at each participating institution. This study involves data from over 400 sites, with written informed consent obtained from all participants at each site.

**Consent for publication**
Not applicable.

**Competing interests**
The only potential conflicts of interest that may be relevant is that all authors are employees and stockholders in Roche/Genentech.

Received: 15 October 2024 Accepted: 10 February 2025
Published online: 17 February 2025

**References**

1. Chen L, Wu Y, DSouza AM, Abidin AZ, Wismüller A, Xu C. MRI tumor segmentation with densely connected 3D CNN. In: Medical Imaging 2018: Image Processing. vol. 10574. Bellingham: SPIE – The International Society for Optics and Photonics; 2018. pp. 357–64.
2. Petrov Y, Malik B, Fredrickson J, Jemaa S, Carano RA. Deep Ensembles Are Robust to Occasional Catastrophic Failures of Individual DNNs for Organs Segmentations in CT Images. J Digit Imaging. 2023;36:1–15.
3. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S. and Bach, M.. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence. 2023;5(5).
4. Dabeer S, Khan MM, Islam S. Cancer diagnosis in histopathological image: CNN based approach. Inform Med Unlocked. 2019;16:100231.
5. Nielsen A, Hansen MB, Tietze A, Mouridsen K. Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning. Stroke. 2018;49(6):1394–401.
6. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE; 2015. pp. 3431–40.
7. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. Piscataway: IEEE; 2017. pp. 2961–9.
8. Jemaa S, Fredrickson J, Carano RA, Nielsen T, de Crespigny A, Bengtsson T. Tumor segmentation and feature extraction from whole-body FDG-PET/CT using cascaded 2D and 3D convolutional neural networks. J Digit Imaging. 2020;33:888–94.
9. Capobianco N, Meignan M, Cottereau AS, Vercellino L, Sibille L, Spottiswoode B, et al. Deep-learning 18F-FDG uptake classification enables total metabolic tumor volume estimation in diffuse large B-cell lymphoma. J Nucl Med. 2021;62(1):30–6.
10. Weisman AJ, Kieler MW, Perlman SB, Hutchings M, Jeraj R, Kostakoglu L, et al. Convolutional neural networks for automated PET/CT detection of diseased lymph node burden in patients with lymphoma. Radiol Artif Intell. 2020;2(5):e200016.
11. Sibille L, Seifert R, Avramovic N, Vehren T, Spottiswoode B, Zuehlsdorff S, et al. 18F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks. Radiology. 2020;294(2):445–52.
12. Jemaa S, Ounadjela S, Wang X, El-Galaly TC, Kostakoglu L, Knapp A, Ku G, Musick L, Sahin D, Wei MC, Yin S, Bengtsson T, De Crespigny A, Carano RAD. Automated Lugano Metabolic Response Assessment in 18F-Fluorodeoxyglucose-Avid Non-Hodgkin Lymphoma With Deep Learning on 18F-Fluorodeoxyglucose-Positron Emission Tomography. J Clin Oncol. 2024;42:2966–77.
13. Barrington SF, Mikhaeel NG, Kostakoglu L, Meignan M, Hutchings M, Müeller SP, et al. Role of imaging in the staging and response assessment of lymphoma: consensus of the International Conference on Malignant Lymphomas Imaging Working Group. J Clin Oncol. 2014;32(27):3048–58.
14. Cheson BD, Fisher RI, Barrington SF, Cavalli F, Schwartz LH, Zucca E, et al. Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. J Clin Oncol. 2014;32(27):3059.
15. Naqa IE. The role of quantitative PET in predicting cancer treatment outcomes. Clin Transl Imaging. 2014;2:305–20.
16. James AP, Dasarathy BV. Medical image fusion: A survey of the state of the art. Inf Fusion. 2014;19:4–19.
17. Zhang H, Tan S, Chen W, Kligerman S, Kim G, D'Souza WD, et al. Modeling pathologic response of esophageal cancer to chemoradiation therapy using spatial-temporal 18F-FDG PET features, clinical parameters, and demographics. Int J Radiat Oncol Biol Phys. 2014;88(1):195–203.
18. Ju W, Xiang D, Zhang B, Wang L, Kopriva I, Chen X. Random walk and graph cut for co-segmentation of lung tumor on PET-CT images. IEEE Trans Image Process. 2015;24(12):5854–67.
19. Kelloff GJ, Hoffman JM, Johnson B, Scher HI, Siegel BA, Cheng EY, et al. Progress and promise of FDG-PET imaging for cancer patient management and oncologic drug development. Clin Cancer Res. 2005;11(8):2785–808.
20. Weiler-Sagie M, Bushelev O, Epelbaum R, Dann EJ, Haim N, Avivi I, et al. 18F-FDG avidity in lymphoma readdressed: a study of 766 patients. J Nucl Med. 2010;51(1):25–30.
21. Young H, Baum R, Cremerius U, Herholz K, Hoekstra O, Lammertsma A, et al. Measurement of clinical and subclinical tumour response using [18F]-fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. Eur J Cancer. 1999;35(13):1773–82.
22. Mikhaeel NG, Heymans MW, Eertink JJ, de Vet HC, Boellaard R, Dührsen U, et al. Proposed new dynamic prognostic index for diffuse large B-cell lymphoma: international metabolic prognostic index. J Clin Oncol. 2022;40(21):2352.
23. Trotman J, Barrington SF, Belada D, Meignan M, MacEwan R, Owen C, et al. Prognostic value of end-of-induction PET response after first-line immunochemotherapy for follicular lymphoma (GALLIUM): secondary analysis of a randomised, phase 3 trial. Lancet Oncol. 2018;20(6):1530.
24. Kostakoglu L, Mattiello F, Martelli M, Sehn LH, Belada D, Ghiggi C, et al. Total metabolic tumor volume as a survival predictor for patients with

diffuse large B-cell lymphoma in the GOYA study. Haematologica. 2022;107(7):1633.
25. Eertink JJ, Zwezerijnen GJC, Heymans MW, Pieplenbosch S, Wiegers SE, Dührsen U, et al. Baseline PET radiomics outperforms the IPI risk score for prediction of outcome in diffuse large B-cell lymphoma. Blood. 2023;141(25):3055.
26. Krak NC, Boellaard R, Hoekstra OS, Twisk JW, Hoekstra CJ, Lammertsma AA. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. Eur J Nucl Med Mol Imaging. 2005;32:294–301.
27. Burggraaff CN, Rahman F, Kaßner I, Pieplenbosch S, Barrington SF, Jauw YW, et al. Optimizing workflows for fast and reliable metabolic tumor volume measurements in diffuse large B cell lymphoma. Mol Imaging Biol. 2020;22:1102–10.
28. Basirinia G, Ali M, Comelli A, Sperandeo A, Piana S, Alongi P, et al. Theranostic Approaches for Gastric Cancer: An Overview of In Vitro and In Vivo Investigations. Cancers. 2024;16(19):3323.
29. Shin SY, Shen TC, Wank SA, Summers RM. Fully-automated detection of small bowel carcinoid tumors in CT scans using deep learning. Med Phys. 2023;50(12):7865–78.
30. Shin SY, Lee S, Summers RM. A graph-theoretic algorithm for small bowel path tracking in CT scans. In: Medical Imaging 2022: Computer-Aided Diagnosis. vol. 12033. Bellingham: SPIE – The International Society for Optics and Photonics; 2022. pp. 849–54.
31. Bakker GJ, Vanbellinghen MC, Scheithauer TP, Verchere CB, Stroes ES, Timmers NK, et al. Pancreatic 18F-FDG uptake is increased in type 2 diabetes patients compared to non-diabetic controls. PLoS ONE. 2019;14(3):e0213202.
32. Zhang X, Ogihara T, Zhu M, Gantumur D, Li Y, Mizoi K, et al. Effect of metformin on 18F-fluorodeoxyglucose uptake and positron emission tomographic imaging. Br J Radiol. 2022;95(1130):20200810.
33. Vitolo U, Trněnỳ M, Belada D, Burke JM, Carella AM, Chua N, et al. Obinutuzumab or rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone in previously untreated diffuse large B-cell lymphoma. J Clin Oncol. 2017;35(31):3529–37.
34. Marcus R, Davies A, Ando K, Klapper W, Opat S, Owen C, et al. Obinutuzumab for the first-line treatment of follicular lymphoma. N Engl J Med. 2017;377(14):1331–44.
35. Zhang J, Huang Y, Zhang Z, Shi Y. Whole-body lesion segmentation in 18f-fdg pet/ct. 2022. arXiv preprint arXiv:220907851.
36. Jemaa S, Paulson J, Hutchings M, Kostakoglu L, Trotman J, Tracy S, et al. Full automation of total metabolic tumor volume from FDG-PET/CT in DLBCL for baseline risk assessments. Cancer Imaging. 2022;22(1):1–14.
37. Barrington SF, Zwezerijnen BG, de Vet HC, Heymans MW, Mikhaeel NG, Burggraaff CN, et al. Automated segmentation of baseline metabolic total tumor burden in diffuse large B-cell lymphoma: which method is most successful? A study on behalf of the PETRA consortium. J Nucl Med. 2021;62(3):332–7.
38. Driessen J, Zwezerijnen GJ, Schöder H, Drees EE, Kersten MJ, Moskowitz AJ, et al. The impact of semiautomatic segmentation methods on metabolic tumor volume, intensity, and dissemination radiomics in 18F-FDG PET scans of patients with classical Hodgkin lymphoma. J Nucl Med. 2022;63(9):1424–30.
39. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Cham: Springer; 2015. pp. 234–41.
40. Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014. arXiv preprint arXiv:14126980.
41. Galdran A, Carneiro G, Ballester MAG. On the Optimal Combination of Cross-Entropy and Soft Dice Losses for Lesion Segmentation with Out-of-Distribution Robustness. In: Diabetic Foot Ulcers Grand Challenge. Cham: Springer; 2022. pp. 40–51.
42. Mooney CZ, Duval RD, Duvall R. Bootstrapping: A nonparametric approach to statistical inference. 95. Thousand Oaks: SAGE Publications; 1993.
43. Torkaman M, Yang J, Shi L, Wang R, Miller EJ, Sinusas AJ, et al. Data Management and Network Architecture Effect on Performance Variability in Direct Attenuation Correction via Deep Learning for Cardiac SPECT: A Feasibility Study. IEEE Trans Radiat Plasma Med Sci. 2021;6(7):755–65.
44. Salimi Y, Mansouri Z, Shiri I, Mainta I, Zaidi H. Deep Learning-Powered CT-Less Multitracer Organ Segmentation From PET Images: A Solution for Unreliable CT Segmentation in PET/CT Imaging. Clin Nucl Med. 2025.

## Publisher's Note