Fully-Automated FDG Lesion Segmentation in PET/CT Imaging via Deep Learning

Manuel Pires*[,1], Sebastian Gutschmayer*[,1], Daria Ferrara[1], Josef Yu[1, 2], Thomas Beyer[1], Zacharias Chalampalakis[1], Lalith Kumar Shiyam Sundar[1]

1 Quantitative Imaging and Medical Physics Team, Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, Vienna, Austria.

2 Department of Biomedical Imaging and Image-Guided Therapy, Medical University of Vienna, Vienna, Austria.

*First authors

Abstract:

Quantitative analysis of Positron Emission Tomography (PET) in cancer cases has a significant hurdle in the necessity for segmentation of malignant lesions. This segmentation is essential for extracting crucial tumor metrics, such as metabolic tumor volume and gross tumor volume, which play vital roles in patient prognosis and therapy planning. Currently, the segmentation process relies on expert knowledge and consumes a substantial amount of time.

In this paper, we introduce a fully automated lesion segmentation tool developed for the Autopet-II challenge. Our approach utilizes two U-Net models. The first model, using PET and Computed Tomography (CT) images, discriminates between cases without tumors and those with tumor presence. The second modelis activated for cancer cases and uses the PET scan to perform lesion segmentation. To enhance accuracy and reduce false positives from healthy regions with high uptake, organ context was incorporated in both models during training.

Our tool demonstrates strong performance in distinguishing negative controls, and the segmentation model achieves a commendable Dice score of 0.71 when segmenting tumor-positive cases in the test set. This innovative approach paves the way for automated tumor segmentation in PET/CT scans, contributing towards quantitative analysis of PET imaging.

Introduction

Positron Emission Tomography in conjunction with Computed Tomography (PET/CT) stands as an indispensable tool in the diagnostic arsenal for a multitude of malignant solid tumors. Predominantly, Fluorodeoxyglucose (FDG) serves as the PET tracer of choice within oncological domains, facilitating the quantification of tissue glucose metabolism — an attribute invariably amplified within tumor lesions. Notwithstanding its prominence, PET/CT analyses are traditionally approached qualitatively by adept nuclear medicine physicians. Delving into a quantitative analysis spectrum could usher in a new era of precise and individualized diagnostic paradigms, optimizing patient management.

Tumor lesion segmentation is the linchpin in the grand scheme of quantitative PET/CT evaluation. By ensuring meticulous segmentation, the pathway to detailed feature extraction is cleared, thereby enabling thorough tumor characterization, comprehensive oncologic staging, and an empirical assessment of therapeutic responses via imaging modalities. The conventional manual approach to lesion segmentation, while thorough, poses significant logistical challenges, both in terms of time commitment and associated costs. Automation emerges as the quintessential solution, making it imperative to bring in-depth PET image analysis within the ambit of regular clinical practice.

Advancements in the realm of deep learning present promising prospects for automated PET/CT lesion segmentation. While these methods illuminate the potential of automation, the segmentation landscape is riddled with intricate challenges. The crux of the matter lies in distinguishing the FDG uptakes. Tumor lesions, though characterized by high FDG uptake, are not alone in this regard. Physiological activities in organs, notably the brain, mirror similar uptakes, making the differentiation a nuanced endeavor. This complexity is accentuated by the paucity of robust training datasets essential for the refinement of segmentation algorithms.

The quantification of Metabolic Tumor Volume (MTV) and Gross Tumor Volume (GTV) holds paramount importance in the clinical management of patients. MTV and GTV serve as pivotal indicators of tumor aggressiveness, potential therapeutic strategies, and prognostic outcomes. Accurate lesion segmentation is at the heart of these quantifications, reaffirming the necessity for its precision. Inaccuracies or inconsistencies in segmentation could lead to suboptimal therapeutic strategies, underscoring the clinical imperative of accurate and consistent lesion delineation.

In this research, we employ a dual-model strategy to meticulously address the challenges of FDG PET/CT tumor segmentation for individual cases. Employing the strengths of nnU-Net, our approach is structured around two sequential models. The foremost model, utilizing both PET and CT imaging data, primarily functions as a discriminator for each specific case. It aims to discern whether the given case is devoid of a tumor. Notably, this model occasionally exhibits false negatives, inadvertently missing tumors in some cases. When a tumor presence is detected by the primary model, the case advances to the second model, which operates exclusively on PET imaging data for precise lesion delineation. This model, while adept at segmenting tumors, sometimes errs by interpreting non-tumorous regions as malignancies.

An integral aspect of our approach was the inclusion of normal organ segmentations from MOOSE, which were meticulously cleaned and verified by medical professionals. In our analytical framework, tumors were conceptually treated akin to another 'organ'. By adopting this perspective, we combined the segmentations into multilabel masks, effectively streamlining the segmentation process.By methodically sequencing these models for each individual case, our intent is to harmonize

their distinct strengths and offset their respective limitations, thereby striving for an optimized segmentation outcome in clinical PET/CT analysis.

Subsequent sections will elucidate our methodologies in depth, showcase our empirical findings, and reflect on the broader implications of our dual-model strategy within the realm of medical image analysis.

Methods

1. Data Collection and Preprocessing

The dataset used in this investigation includes FDG-PET/CT scans from patients diagnosed with malignant melanoma, lymphoma, or lung cancer. Alongside these, negative control subjects were also included. These scans were conducted at two primary institutions: the University Hospital Tübingen and the University Hospital of the LMU in Munich, Germany. The FDG-PET/CT scans were acquired adhering to international standards, utilizing PET/CT scanners, namely the Siemens Biograph mCT series and the GE Discovery 690.

University Hospital Tübingen: Patients underwent a 6-hour fasting period before being administered with approximately 350 MBq 18F-FDG. 60 minutes post this intravenous tracer infusion, PET/CT images were obtained using the Biograph mCT. Diagnostic CT scans covering the neck, thorax, abdomen, and pelvis were initiated 90 seconds after the introduction of Ultravist 370 (Bayer AG) as a contrast agent. PET images underwent iterative reconstruction, with Gaussian post-reconstruction smoothing. Contrast-enhanced CT scans had a slice thickness ranging from 2 to 3 mm.

University Hospital of the LMU in Munich: Adhering to a similar protocol as Tübingen, the dosage of 18F-FDG was set at approximately 250 MBq post a 6-hour fasting phase. The contrast agents used were either Ultravist 300 (Bayer AG) or Imeron 350 (Bracco Imaging Deutschland GmbH). The resultant slice thickness for the contrast-enhanced CT was set at 3 mm.

Data Composition:

The training set comprised 1,014 studies from the University Hospital Tübingen, accounting for a total of 900 patients. Each study included one 3D whole body FDG-PET volume, an aligned 3D whole body CT volume, and a corresponding 3D binary mask of manually segmented tumor lesions on the FDG-PET. This alignment ensured congruence between the CT and PET volumes, with minor variations due to physiological motion.

2. Preliminary Organ Segmentation

Training datasets underwent an initial pass using MOOSE for organ segmentation. Recognizing that MOOSE isn't perfect in segmenting to the pixel level, we enlisted 13 medical students to refine and clean up the generated organ segmentations.

Hypothesis and Model Development

Operating under the hypothesis that the inclusion of organ information as multilabels would refine tumor segmentation accuracy, we directed our attention to organs exhibiting high FDG uptake. These organs were frequently misinterpreted as tumors (false positives), leading to our iterative model development from M0 to M5. Modeling began with M0, using only PET images as input without

organ information. Progressing from M1 to M5, we incrementally added organ data. All models utilized PET images as their input data. However, the target classes varied from M0 to M5. In the case of M0, only tumor segmentations were used as target classes. With M1, we expanded the target classes to include the segmentation of organs that usually have the highest uptake in FDG PET scans when they are healthy, specifically the brain, heart, liver, kidneys, and bladder. M2 extended this list by incorporating the pancreas and stomach, both of which exhibit some uptake even when healthy. Building on M2, M3 included all the organs present in M2 and introduced the bones as target classes, aiming to address false positives in this region. In contrast, M4 replaced the bones with the lungs as target classes. Finally, M5 encompassed all available organs in its target classes.

Through evaluation, we discerned that M5 outperformed the rest. However, a notable limitation surfaced across all models: poor performance in tumor-negative cases. Despite being tumor-free, healthy structures were often misclassified as tumors.

Sequential Dual-Model Approach

To mitigate the aforementioned false-positive issue, a two-fold strategy was adopted (Figure 1):

Primary Model (PET/CT Discriminator): A primary model utilizing both PET and CT images was designed to discern if a tumor is present or absent. If the primary model identified a tumor, the case was forwarded to the secondary model.

Secondary Model (PET-based Segmenter): This model employed only PET images and optimized organ information from M5 to execute precise tumor segmentation.

The choice to sequence these models was informed by their respective strengths. The primary model exhibited proficiency in discerning tumor absence, whereas the secondary model excelled at tumor segmentation.

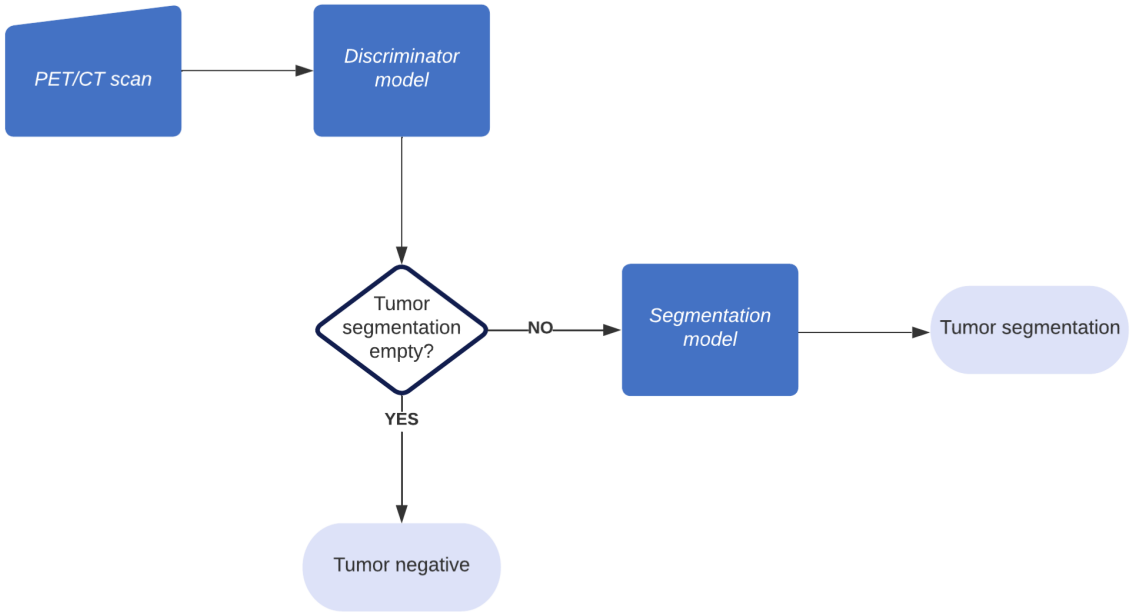

Figure 1- Schematic representation of the approach proposed.

5. Implementation and Availability

Our methodology harnesses the capabilities of nnU-Net V2, an automatic segmentation tool that serves as the backbone of our solution. For accessibility to the broader research community, our developed tool has been made freely available. The package, named "LION" (Lesion segmentatION), can be accessed on GitHub (https://github.com/lalithShiyam/LION) and easily integrated into one's workflow using the command pip install lionz.

6. Post-Processing for Enhanced Precision

To further refine our segmentation results, a post-processing stage was introduced, targeting two primary artifacts:

Edge Artifacts Removal: A recurring issue observed in segmentations was the presence of edge artifacts, which tend to hamper the accuracy of the segmentation. To mitigate this, we systematically pruned 10 voxels from the borders of all segmentations. This margin removal effectively eliminated the extraneous edge artifacts without affecting the core segmented region.

Small Component Removal: During our analysis, we identified sporadic presence of isolated small components within the segmentation. Given the improbability of such tiny structures representing valid tumor regions, we implemented a size-based filtering technique. Specifically, any connected component spanning less than 10 voxels was discarded, ensuring that only significant and relevant tumor regions were retained in the final segmentation.

These post-processing measures were crucial not only in eliminating systematic and random artifacts from our segmentations but also in enhancing the overall reliability and clinical relevance of our findings.

3 Results

As anticipated, our baseline model, M0, exhibited poor performance, achieving only a dice score of 0.46 on the tumor-positive cases in our test set. However, with the introduction of organ context, the models showed a significant improvement in performance. M2, M3, and M4 were the worst performing models, achieving a dice score of 0.69 on cancer cases. This marked a substantial enhancement compared to the baseline results. M1 emerged as the second-best performing model, boasting a dice score of 0.70 on cancer cases. Ultimately, the top-performing model was M5, which achieved a dice score of 0.71 on cancer cases. This represented a remarkable 54% increase in performance when compared to the baseline model, M0.

The performance of our tool was evaluated using a test set of 100 held-out cases. When distinguishing negative controls from tumor-positive cases, our tool achieved a satisfactory performance, correctly identifying 67% of the negative controls. While some false positives persisted, this discrimination step effectively reduced the number of negative control cases considered for segmentation.

To select the segmentation model, all models were assessed on the positive cases from the held-out data, and the Dice scores for each model are presented in Figure 2. Notably, all new models outperformed the baseline model. As anticipated, the inclusion of healthy organ information played

a pivotal role in mitigating false positive regions. It's also worthy to note that when checking the segmentations of the test set the PET only model identified some tumors that the PET/CT model missed, as can be seen in Figure 3.

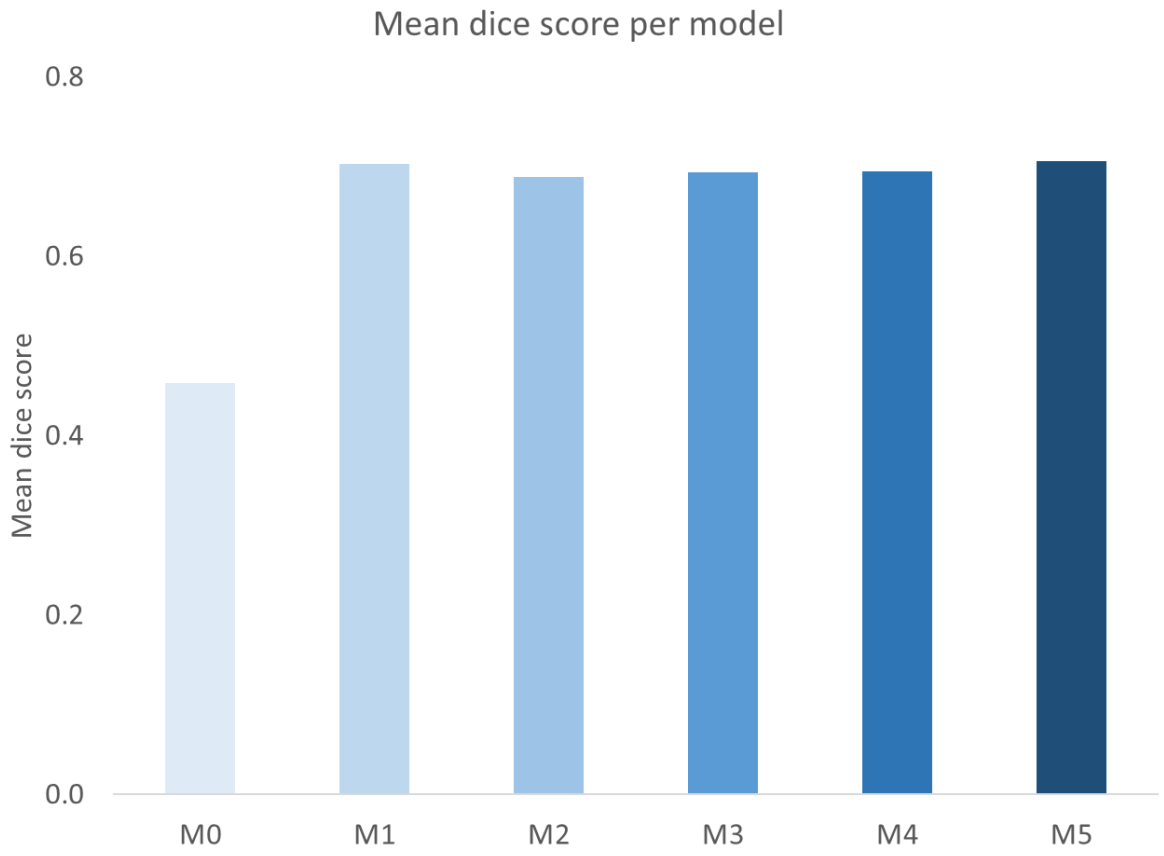

Figure 2- Dice scores obtained in tumor cases from the test for the different models trained.

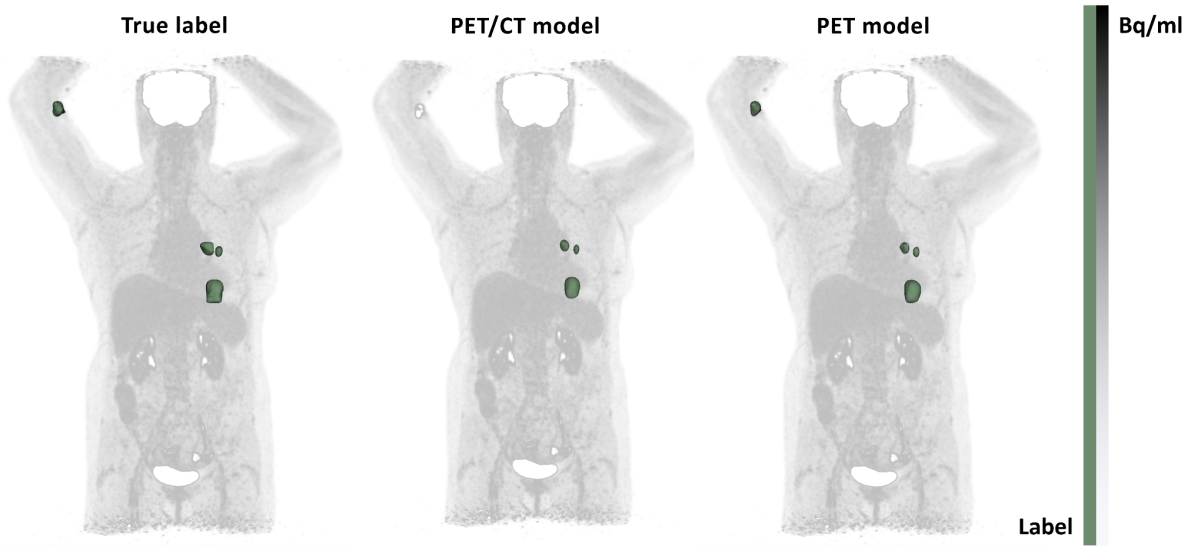

Figure 3- Example of a case where the PET only model segments a tumor missed by the PET/CT model.

This context proved instrumental in avoiding regions with heightened uptake caused by other pathologies (Figure 4). Importantly, the model exhibited consistently robust performance, resulting in a relatively low standard deviation for the Dice score when compared to the preliminary leaderboard. Models M1 and M5 emerged as the top performers, achieving Dice scores of 0.703 ± 0.253 and 0.706 ± 0.236, respectively. We opted for M5 due to its slightly superior performance and incorporated it into the pipeline. The dual model approach also contributed to delineate some tumors that the PET/CT model would have missed.

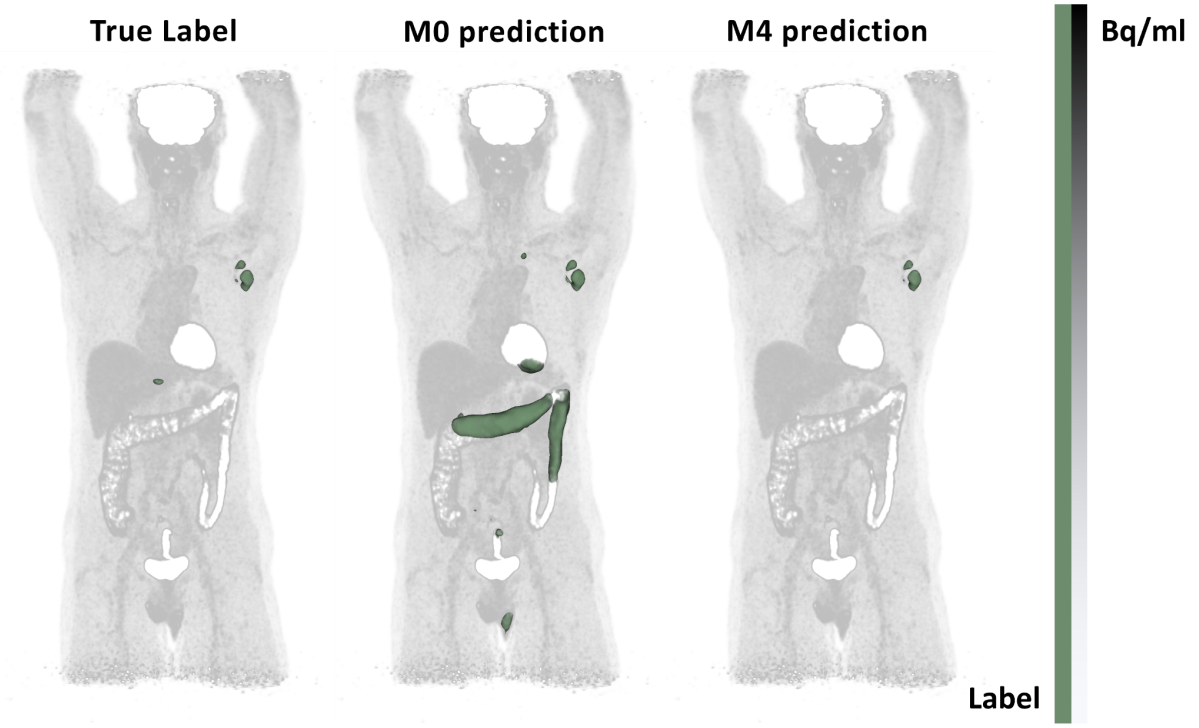

Figure 4- Maximum intensity projection for a case with non-cancer related higher FDG uptake in the bowel region and a) true segmentation; b) M0 predicted segmentation; c) M4 predicted segmentation.

With the integration of the M5 model into the pipeline, the entire testing process was repeated. The results for tumor cases exhibited a slight improvement, with an average Dice score of 0.711 ± 0.201. This underscores the importance of post-processing steps in mitigating border regions and identifying true malignant lesions.

4 Conclusion

The approach presented in this paper, submitted to the Autopet-II challenge, has demonstrated promising results by adopting a dual-model approach to first identify negative control cases and then in the presence of tumour segment it. Furthermore, the incorporation of organ segmentations has significantly enhanced tumor segmentation by providing valuable context to the model, particularly concerning healthy high uptake regions.

References

1. Salaün P-Y, Abgral R, Malard O, Querellou-Lefranc S, Quere G, Wartski M, et al. Good clinical practice recommendations for the use of PET/CT in oncology. Eur J Nucl Med Mol Imaging. 2020;47: 28–50.

2. Hofman MS, Hicks RJ. Moving Beyond "Lumpology": PET/CT Imaging of Pheochromocytoma and Paraganglioma. Clinical cancer research: an official journal of the American Association for Cancer Research. 2015. pp. 3815–3817.

3. Wen W, Xuan D, Hu Y, Li X, Liu L, Xu D. Prognostic value of maximum standard uptake value, metabolic tumor volume, and total lesion glycolysis of positron emission tomography/computed tomography in patients with breast cancer: A systematic review and meta-analysis. PLoS One. 2019;14: e0225959.

4. Ito K, Schöder H, Teng R, Humm JL, Ni A, Wolchok JD, et al. Prognostic value of baseline metabolic tumor volume measured on 18F-fluorodeoxyglucose positron emission tomography/computed tomography in melanoma patients treated with ipilimumab therapy. Eur J Nucl Med Mol Imaging. 2019;46: 930–939.

5. Guo B, Tan X, Ke Q, Cen H. Prognostic value of baseline metabolic tumor volume and total lesion glycolysis in patients with lymphoma: A meta-analysis. PLoS One. 2019;14: e0210224.

6. Burnet NG, Thomas SJ, Burton KE, Jefferies SJ. Defining the tumour and target volumes for radiotherapy. Cancer Imaging. 2004;4: 153–161.

7. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18: 203–211.

8. Shiyam Sundar LK, Yu J, Muzik O, Kulterer OC, Fueger B, Kifjak D, et al. Fully Automated, Semantic Segmentation of Whole-Body 18F-FDG PET/CT Images Based on Data-Centric Artificial Intelligence. J Nucl Med. 2022;63: 1941–1948.

9. Wasserthal J, Breit H-C, Meyer MT, Pradella M, Hinck D, Sauter AW, et al. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiology: Artificial Intelligence. 2023;5: e230024.