Comparing foundation models and nnU-Net for segmentation of primary brain lymphoma on clinical routine post-contrast T1-weighted MRI

Guanghui Fu, Lucia Nichelli, Dario Herran, Romain Valabregue, Agusti Alentorn, Khê Hoang-Xuan, Caroline Houillier, Didier Dormont, Stéphane Lehéricy, Olivier Colliot

Comparing foundation models and nnU-Net for segmentation of primary brain lymphoma on clinical routine post-contrast T1-weighted MRI

Guanghui Fu[1] GUANGHUI.FU@ICM-INSTITUTE.ORG
Lucia Nichelli[1,3,4] LUCIA.NICHELLI@APHP.FR
Dario Herran[3] DARIO.HERRANDELAGALA@APHP.FR
Romain Valabregue[2,3] ROMAIN.VALABREGUE@ICM-INSTITUTE.ORG
Agusti Alentorn[2,5] AGUSTI.ALENTORN@APHP.FR
Khê Hoang-Xuan[2,5] KHE.HOANG-XUAN@APHP.FR
Caroline Houillier[2,5] CAROLINE.HOUILLIER@APHP.FR
Didier Dormont[1,4] DIDIER.DORMONT@ICM-INSTITUTE.ORG
Stéphane Lehéricy[2,3,4] STEPHANE.LEHERICY@ICM-INSTITUTE.ORG
Olivier Colliot[1] OLIVIER.COLLIOT@CNRS.FR

[1] *Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France.*
[2] *Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France.*
[3] *ICM, Centre de NeuroImagerie de Recherche-CENIR, Paris, France.*
[4] *AP-HP, Pitié Salpêtrière, DMU DIAMENT, Dep. of Neuroradiology, Paris, France.*
[5] *AP-HP, Pitié Salpêtrière, DMU Neurosciences, Dep. of Neurology 2, Paris, France.*

Abstract

Primary Central Nervous System (PCNS) lymphoma is an aggressive brain tumor with variable response rates to current therapeutic strategies. Segmentation of PCNS lymphoma on magnetic resonance imaging (MRI) is important to enhance diagnosis and follow-up. However, it is a difficult task due to the highly variable presentation of lymphoma lesions and the heterogeneity of acquisitions in clinical routine. The recent rise of foundation models (FM) for segmentation, such as Segment Anything Model (SAM), offers a potential alternative to classical supervised deep learning approaches. Nevertheless, their value on challenging tasks such as lymphoma segmentation on clinical routine data has not been thoroughly studied. In this paper, we assessed the performance of several FMs (SAM, MedSAM, UniverSeg) for lymphoma segmentation and compared them to a classical supervised learning with nnU-net. In addition, we performed experiments on the public dataset MSD-BraTS so that others can reproduce our findings. We found that nnU-net outperformed FMs by vast and statistically significant margins. For the lymphoma clinical routine dataset, the difference between nnU-Net and the best FM was about 19 percent points of Dice. For MSD-BraTS, it was about 10 percent points. Our findings suggest that current FMs fall short in handling complex segmentation tasks in particular on clinical routine data and that 3D supervised learning (e.g. with nnU-Net) is essential at present. Trained models for both tasks, code and box prompts (for MSD-BraTS) are available at https://github.com/GuanghuiFU/medical_cv_foundation_eval.

Keywords: Lymphoma, foundation model, SAM, Brain tumour, Segmentation, Deep learning

1. Introduction

Primary central nervous system (CNS) lymphoma (PCNSL) is a rare cancer, characterized by its rapid development and high mortality rate, which poses significant challenges in clinical diagnosis, treatment planning and follow-up (Schaff and Grommes, 2022). Automated segmentation tools are critical for clinical care as well as for research to deepen our understanding of the disease. However, few studies have been devoted to segmentation of PCNSL (Pennig et al., 2021; El Jurdi et al., 2024) unlike other types of brain tumours such as gliomas, e.g. (Ghaffari et al., 2019; Liu et al., 2021; van Kempen et al., 2021).

Recently, foundation models (FM) for image segmentation have been proposed, e.g. (Kirillov et al., 2023). They are highly effective in computer vision and have raised interest in medical image analysis (Huang et al., 2023; Mazurowski et al., 2023). They have several potential advantages over classical fully-supervised deep learning segmentation models: they require none or limited annotated training data, reduce the need for expert labelling as well as computational burden. Specifically, the application of FMs to new tasks requires only providing prompts (Kirillov et al., 2023) or limited additional data (Butoi et al., 2023). Segment Anything Model (SAM) (Kirillov et al., 2023) is a computer vision FM trained on 1 billion masks which allows interactive segmentation with prompts. MedSAM (Ma et al., 2024) is a version of SAM for medical data. UniverSeg (Butoi et al., 2023) is an FM for medical segmentation trained on extensive medical imaging data. It requires support data (i.e. a limited set of annotated data) to perform specific segmentation tasks.

One can thus legitimately asks whether current FMs can replace fully-supervised segmentation. Several works have assessed the value of FMs for medical image segmentation. Some of these works (He et al., 2023; Huang et al., 2023; Mazurowski et al., 2023) rely on prompts generated from ground truth. However, this is an ideal scenario which may not reflect real use. If we have the ground truth, we do not need an automatic method. Other studies (Cheng et al., 2023; Wald et al., 2023) introduced varying scales of jitter to simulate degrees of user inaccuracy and generate Oracle prompts. However, it is unclear if such simulated jitter reflects the users' habitual patterns. Regarding UniverSeg (Butoi et al., 2023), while it has undergone external validation (Kim et al., 2023), its effectiveness on clinical routine datasets remains unverified. (Kim et al., 2023) evaluated UniverSeg for prostate segmentation using an in-house dataset and found that the performance was comparable to that of nnU-Net (Isensee et al., 2021). However, it is important to note that UniverSeg was already pretrained on prostate segmentation tasks and that prostate has much less variability than brain tumours such as PCNSL making its segmentation easier. A notable limitation of most of these studies is that they lacked validation on clinical routine, heterogeneous, data.

Our study undertakes a comparison of supervised learning and foundation models for segmentation of primary CNS lymphoma on clinical routine, heterogeneous, MRI. To that purpose, we used a dataset of 112 patients with PCNSL acquired with different sequence parameters, sequence types (spin echo vs gradient echo), 2D or 3D sequences, resolutions, and MRI scanners. Since this clinical routine dataset cannot be shared, we complement our work with experiments on a subset of the open glioma dataset MSD-BraTS (Antonelli et al., 2022; Menze et al., 2014) in order for others to reproduce our work. We compared FMs

to nnU-net (Isensee et al., 2021) which has been shown to be one of the most consistently effective segmentation algorithms across many medical applications (Antonelli et al., 2022).

2. Materials and Methods

2.1. Datasets

2.1.1. Lymphoma dataset

We studied 117 patients with primary CNS lymphoma. The study was approved by the Institutional Ethical Committee (Pitié Salpêtrière Hospital, Ile-de-France VI, n°DC-2009-957) and by the French Data Protection Authority (CNIL, Commission Nationale de l'Informatique et des Libertés, DR 2013-279). According to French regulation, consent was waived as these images were acquired as part of the routine clinical care of the patients. Each patient had a T1-weighted MRI after gadolinium injection. The images were acquired as part of clinical routine and were thus not harmonized (different MRI scanners, types of sequences, resolutions, contrast to noise, etc.).

The full dataset was annotated by a certified neuroradiologist (L.N.) with four years of experience in neuro-oncology imaging. A subset of 50 cases was independently annotated by a certified radiologist (D.H.) with one year of experience in neuro-oncology imaging to calculate inter-rater reliability. The computed inter-rater variability (mean and 95% confidence interval computed using bootstrap on the 50 samples) metrics were: Dice coefficient 81.17 [77.76, 84.50]; Hausdorff 95th percentile distance: 6.99 [3.44, 11.46]; topological 3D error: 2.12 [1.16, 3.38]. Please refer to section 2.3 for details about the metrics.

Each MRI volume in this dataset was resized and resampled to isotropic $1\times1\times1\text{mm}^3$ voxels and dimensions of $240\times240\times160$ using resample and resize operations from TorchIO (Pérez-García et al., 2021). We also rescale the intensity to a range of 0 and 1. We removed five subjects because the preprocessing failed (wrong dimensions). The final lymphoma MRI dataset comprises 112 patients (60 females, 51 males, one unspecified) aged 32-86 years (mean 66.13). It includes 92 isotropic (or quasi-isotropic) 3D acquisitions and 20 non-isotropic 2D acquisitions across different field strengths (3T: 53 subjects, 1.5T: 53, 1.0T: 6), manufacturers (GE Healthcare: 93, Philips: 11, Siemens: 8). In-plane resolution ranges from 0.39 mm to 1.2 mm and slice thickness ranges from 0.29 mm to 8.0 mm. The example images in Figure 1 illustrate the variability in lesion appearance, location and image quality. The dataset was randomly split into a training/validation set with 57 patients and a test set with 55 patients. The test set was only used to validate the models.

2.1.2. MSD-BraTs dataset

We also conducted experiments on the MSD-BRATS public dataset from the medical segmentation decathlon challenge (Antonelli et al., 2022). BraTS is a dataset of patients with glioblastoma or lower-grade glioma, manually segmented by expert raters (Simpson et al., 2019). The downloaded dataset has already undergone uniform pre-processing, including co-registration to a common template, standardization to 1 mm^3 resolution, and skull stripping. To be as similar as possible to lymphoma segmentation task, we selected the T1-weighted MRI after gadolinium injection and focused on the enhancing tumor as the target region. In order to have the approximately the same sample size for the two tasks,

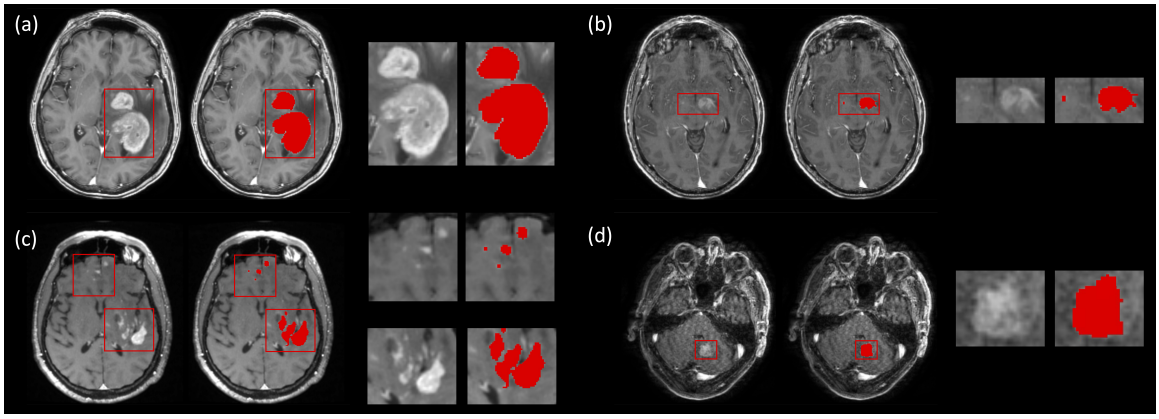

Figure 1: Examples of clinical T1-weighted MR images of patients with PCNS lymphoma. Each example shows the MRI image, the same image overlaid with the ground truth segmentation, and a zoom in of the area(s) marked with a box. Please note that the boxes shown here correspond the zoomed-in area and *not to the box prompts provided to the foundation models*.

we randomly selected 113 patients for our experiments[1]. They were subsequently randomly split into a training/validation set with 58 patients and a test set comprising 55 patients. The test set was only used to evaluate the models.

2.2. Segmentation methods

2.2.1. nnUNet

nnU-Net is a supervised deep-learning segmentation approach that dynamically adapts to specific datasets by analyzing provided training cases and automatically configuring a corresponding UNet-based segmentation pipeline (Isensee et al., 2021). Schematic architectures of the different tested approaches are presented in the appendix (Figure S1). During the training time, 5-fold cross-validation is executed on the training/validation data for 1000 epochs of training, to select the optimal configuration.

2.2.2. SAM and MedSAM

The Segment Anything Model (SAM) is an FM which has undergone training on a dataset comprising 11 million images and 1.1 billion masks, demonstrating robust zero-shot performance across a diverse range of segmentation tasks (Kirillov et al., 2023). Our experiments included various versions of the SAM model, including SAM-vit-b, SAM-vit-l, and SAM-vit-h. These models, trained with progressively larger numbers of masks, vary in their parameter sizes. MedSAM (Ma et al., 2024) is a medically fine-tuned version of the smallest SAM model variant (vit-b) which has been trained with over one million medical image-mask pairs and was reported to consistently surpass the original SAM model in performance.

1. Experiments on lymphoma include 112 patients because one preprocessing defect was only identified afterwards.

Both SAM and MedSAM require a prompt from the user for each image to be segmented. The prompt can take different forms including points and boxes. We chose box prompts which have been suggested to lead to better performance in medical segmentation tasks (Wald et al., 2023). One author (G.F.) manually annotated 3D box prompts in each test volume. The number of boxes is variable, to ensure complete coverage when multiple lesions are present. These 3D boxes are then transformed into 2D slice-level prompts. Utilizing these prompts, SAM (or MedSAM) infers the segmentation in each 2D slice. Subsequently, these predictions are stacked into 3D volume. The time taken to draw the box prompt for one image was 76.18±26.7 (in seconds; mean± standard-deviation) for lymphoma and 72.15±16.67 for MSD-BraTs. An example of 3D box prompt is depicted in Figure 2. In several other papers (He et al., 2023; Huang et al., 2023; Mazurowski et al., 2023; Ma et al., 2024), box prompts were generated from ground truth segmentation (automatically computing a bounding box from the ground truth, possibly with perturbation). We believe this can only be an ideal scenario and not reflect realistic usage. However, as a comparison, we also provide results obtained by computing box prompts from ground truth in appendix C.

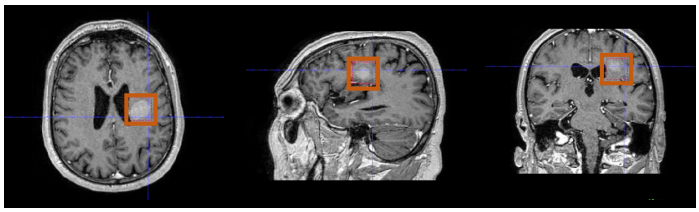

Figure 2: Example of a manually drawn 3D box prompt in a patient with lymphoma.

2.2.3. UniverSeg

UniverSeg is designed to solve new medical segmentation tasks without the need for retraining, leveraging the newly proposed CrossBlock mechanism for information transfer from the support set to the query image (Butoi et al., 2023). The support set is a small annotated dataset corresponding to the target task. UniverSeg's training encompasses 53 datasets across 26 medical domains and 16 imaging modalities. The size of the support set can range from 1 to 64.

In the UniverSeg experiments, we initially converted the 3D volumes into 2D slices, resizing them from their original dimensions of 240×240 to 128×128 to meet the model's architectural requirements. We then devised various support set configurations. UniverSeg then processes each MRI slice in sequence. Upon completing all 2D slice predictions, these are resized to their original resolution (from 128×128 back to 240×240) and reconstituted into a 3D volume.

The support set can include up to 64 images. We employed different strategies to examine the performance impact of different support sets. Moreover, UniverSeg generates soft predictions and we experimented with different thresholds. Results with different support sets and different thresholds are presented in Appendix D. In the rest of the paper, we present results obtained with the best performing support sets (mid for lymphoma, large for MSD-BraTs) and thresholds (0.8 for lymphoma and 0.7 for BraTS).

2.3. Performance metrics and statistical analysis

For selecting performance metrics, we followed the recommendations of Metrics Reloaded (Maier-Hein et al., 2023), specifically employing the 3D Dice coefficient and the 95% 3D Hausdorff distance (HD95). We also computed the 3D connected component error (Topo 3D) which is the absolute difference in number of 3D connected components between the ground truth and prediction. For each metric, we report its mean value as well as the corresponding 95% confidence interval (CI) computed using bootstrap over the independent test set. For a subset of results (see Results and Discussion sections), we assessed whether the observed differences in Dice score between models were statistically significant using paired T-tests on the independent test set.

2.4. Code, data and trained models availability

We made the repository[2] publicly available, it contains: code, list of patients, box prompts, support sets for MSD-BraTS and trained models for both lymphoma and MSD-BraTS. MSD-BraTS data can be downloaded from the MSD website[3]. We cannot provide data and prompts for the lymphoma dataset due to regulatory constraints.

3. Results

Results are presented in Table 1. Our experiments, conducted on both a private PCNS lymphoma and the public MSD-BraTS glioma dataset, yielded consistent findings. Supervised training with a 3D nnU-Net outperformed the best FM (SAM, vit-b version) by a vast margin: there was a statistically significant improvement of about 19 percent points of Dice ($p < 10^{-9}$) for the lymphoma dataset and of about 10 percent points for the MSD-BraTS ($p < 10^{-5}$). Similar results were observed for the HD95 metric: improvement of about 8m and 14mm with the 3D nnU-net compared to the best FM, for the lymphoma and the MSD-BraTS datasets respectively. For the 3D topological errors, nnU-net also performed in general better compared to FMs, to the exception of MedSAM-vit-b which had similar results for this specific metric. Overall, findings from both tasks underscore that nnUNet-3D consistently outperformed FMs. Comparative examples of lymphoma segmentation by different models are illustrated in Figure 3. In addition, distribution of Dice scores on the test set are presented in Figures S2 and S3 of Appendix B.

4. Discussion

In this paper, we compared the segmentation performance of supervised learning with nnU-Net and various FMs on two datasets: a local clinical routine lymphoma dataset and the public MSD-BraTS glioma dataset. Results were consistent across datasets and demonstrated that 3D nnU-Net outperformed FMs by a vast margin.

Even though the performance gap was already large (10 percent points of Dice) and statistically significant for the MSD-BraTS dataset, it was huge for the lymphoma dataset (19 percent points). This difference comes mainly from a lower performance of FMs while that

2. https://github.com/GuanghuiFU/medical_cv_foundation_eval
3. http://medicaldecathlon.com/

Table 1: The performance of lymphoma and brain glioma segmentation on T1-weighted MRI from clinical dataset and MSD-BraTS dataset. Results presented as mean with 95% bootstrap confidence interval computed on the independent test set.

Task	Model type	Model	Dice (in %)	HD95 (in mm)	Topo 3D
Lymphoma	2D	UniverSeg	35.91 [29.33, 42.70]	73.89 [69.47, 78.41]	56.06 [51.22, 60.91]
		SAM-vit-b	63.21 [57.96, 68.15]	21.98 [16.45, 28.11]	10.72 [5.54, 17.52]
		SAM-vit-l	60.80 [55.66, 65.77]	21.12 [15.87, 27.02]	8.94 [4.93, 13.98]
		SAM-vit-h	60.13 [54.98, 65.14]	22.46 [17.04, 28.40]	9.63 [5.65, 14.52]
		MedSAM-vit-b	55.44 [49.78, 60.96]	22.83 [16.94, 29.35]	2.69 [1.80, 3.69]
		nnUNet-2d	73.99 [67.49, 80.06]	9.78 [6.03, 14.19]	1.41 [0.93, 1.94]
	3D	nnUNet-3d	82.08 [77.05, 86.44]	7.96 [4.63, 11.92]	1.19 [0.81, 1.59]
MSD-BraTS	2D	UniverSeg	51.42 [45.03, 57.89]	68.73 [63.81, 73.79]	98.33 [98.85, 104.51]
		SAM-vit-b	69.96 [65.78, 73.80]	11.63 [9.28, 15.07]	17.58 [12.44, 23.56]
		SAM-vit-l	69.77 [65.60, 73.60]	12.00 [9.63, 15.36]	14.62 [9.96, 19.95]
		SAM-vit-h	67.76 [63.77, 71.43]	12.55 [10.23, 15.81]	14.80 [10.71, 19.58]
		MedSAM-vit-b	46.11 [41.39, 50.78]	12.28 [9.95, 15.80]	8.31 [5.67, 11.69]
		nnUNet-2d	73.46 [66.76, 79.30]	11.90 [7.76, 17.19]	7.67 [5.07, 11.02]
	3D	nnUNet-3d	79.98 [74.42, 85.12]	6.47 [4.93, 8.31]	8.33 [5.61, 11.78]

of nnU-Net was quite stable. There may be different explanations to this: the lymphoma dataset has a wide heterogeneity which reflects clinical routine and lymphoma characteristics are highly variable. This highlights the interest of performing experiments on clinical routine datasets. The lower performance of SAM could in principle be expected as it was not trained medical data and this limitation is acknowledged by the authors (Kirillov et al., 2023). (He et al., 2023) also found that SAM was outperformed by U-nets by vast margins over different medical tasks. However, the massive superiority of nnU-net is at odds with the results reported in the original papers of MedSAM (Ma et al., 2024) and Universeg (Butoi et al., 2023). The MedSAM paper reports excellent Dice scores which are comparable to that of a U-Net (Butoi et al., 2023). UniverSeg was reported to have lower performance than nnU-net (Butoi et al., 2023) but the difference was much lower than in our experiments (13 Dice points in original paper versus 46 points for lymphoma and 28 points for MSD-BraTS).

Among FMs, SAM-based models gave in general better performance. The smaller SAM-vit-b model tended to work slightly better than the larger version, but confidence intervals are overlapping. Surprisingly, MedSAM performed worse than the corresponding non-medical SAM-bit-b, by vast and statistically significant margins (8 points of Dice for lymphoma, $p < 10^{-6}$; 23 points for MSD-BraTS, $p < 10^{-13}$). This is in contradiction with what was reported in the original paper (Ma et al., 2024) where superior performance compared to SAM was observed. UniverSeg resulted in very poor performances, as it was not only vastly outperformed by nnU-Net but also by SAM (28 Dice points for lymphoma, $p < 10^{-8}$; 18 points for MSD-BraTS, $p < 10^{-7}$). Lower performance on MSD-BraTS is particularly surprising as UniverSeg was trained on BraTS (even though not the same task and with more modalities than in our experiments). The results we report are the best that we obtained across different thresholds and support sets (see Appendix D). Qualitatively, we could notice that UniverSeg tended to focus only on segments with similar contrast,

neglecting the anatomical location inside the data. This limitation becomes particularly evident given the variable position and contrast of lymphoma (Barajas Jr et al., 2021).

The 3D nnUNet demonstrated superior performance over its 2D version. It is thus possible that the poor performance of FMs is in part due to their 2D nature which prevents them from using the complex 3D information. Nevertheless, the performance of FMs was still inferior to that of the 2D nnUNet, indicating that this is likely not the only limitation.

Some previous works have generated prompts from ground truth. For comparison, we also provided results obtained when generating prompts from ground truth (Appendix C). The only setting that led to comparable results between SAM and nnU-Net was when 2D prompts generated from ground truth were provided. We believe this is clearly not a realistic scenario as the box perfectly fits the ground truth on every slice.

Our work has the following limitations. First, our results only concern two segmentation tasks, which are particularly difficult, and the situation could be different for other tasks. Moreover, the field of FMs is advancing fastly. It is very well possible that better results could be obtained with newer FMs. Furthermore, our provided prompts may have imperfections. It is also possible that we have used UniverSeg in a suboptimal way, even though we experimented with different support sets and thresholds. We have made prompts and support sets openly available for MSD-BraTS for other researchers to experiment with them and would be happy to be shown that there was a more optimal way to use FMs.

5. Acknowledgments

We acknowledge funding from the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6). Guanghui Fu is supported by the Chinese Government Scholarship provided by China Scholarship Council (CSC). Lucia Nichelli is supported by an Inria/AP-HP “Poste d’accueil.

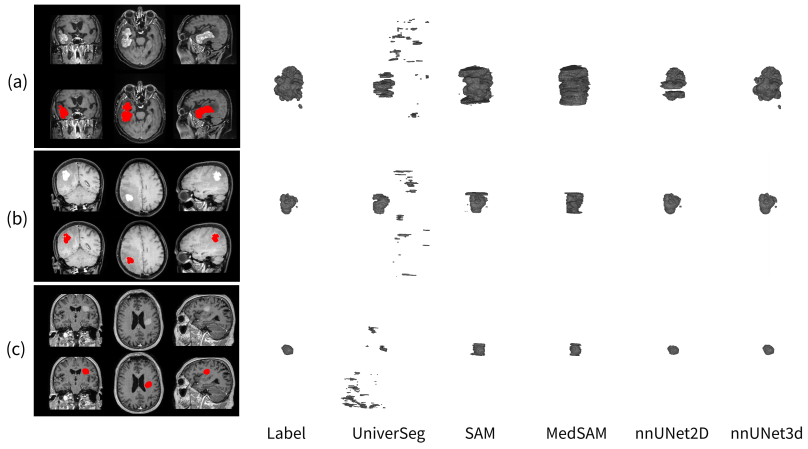

Figure 3: Ground truth vs predictions on lymphoma data. Left: ground-truth superimposed on MRI. Right: 3D renderings of ground-truth (label) and of predictions with different methods.

References

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. Nature communications, 13(1):4128, 2022.

Ramon F Barajas Jr, Letterio S Politi, Nicoletta Anzalone, Heiko Schöder, Christopher P Fox, Jerrold L Boxerman, Timothy J Kaufmann, C Chad Quarles, Benjamin M Ellingson, Dorothee Auer, et al. Consensus recommendations for MRI and PET imaging of primary central nervous system lymphoma: guideline statement from the international primary CNS lymphoma collaborative group (IPCG). Neuro-oncology, 23(7):1056–1071, 2021.

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. UniverSeg: Universal medical image segmentation. In Proc. ICCV 2023, pages 21438–21451, 2023.

Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. SAM on medical images: A comprehensive study on three prompt modes. arXiv preprint arXiv:2305.00035, 2023.

Rosana El Jurdi, Lucia Nichelli, Agusti Alentorn, Ghislain Vaillant, Guanghui Fu, Khe Hoang-Xuan, Caroline Houillier, Stéphane Lehéricy, and Olivier Colliot. Border irregularity loss for automated segmentation of primary brain lymphomas on post-contrast MRI. In Medical Imaging 2024: Image Processing. SPIE, 2024.

Mina Ghaffari, Arcot Sowmya, and Ruth Oliver. Automated brain tumor segmentation using multimodal brain scans: a survey based on models submitted to the BraTS 2012–2018 challenges. IEEE reviews in biomedical engineering, 13:156–168, 2019.

Sheng He, Rina Bao, Jingpeng Li, Jeffrey Stout, Atle Bjornerud, P Ellen Grant, and Yangming Ou. Computer-vision benchmark segment-anything model (SAM) in medical images: Accuracy in 12 datasets. arXiv preprint arXiv:2304.09324, 3, 2023.

Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? Medical Image Analysis, page 103061, 2023.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2):203–211, 2021.

Heejong Kim, Victor Ion Butoi, Adrian V Dalca, and Mert R Sabuncu. Empirical analysis of a segmentation foundation model in prostate imaging. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 140–150. Springer, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In Proc. ICCV 2023, pages 4015–4026, 2023.

Zhihua Liu, Lei Tong, Long Chen, Feixiang Zhou, Zheheng Jiang, Qianni Zhang, Yinhai Wang, Caifeng Shan, Ling Li, and Huiyu Zhou. Canet: Context aware network for brain glioma segmentation. IEEE Transactions on Medical Imaging, 40(7):1763–1777, 2021.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. Nature Communications, 15(1):654, 2024.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A. Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Hannes Kenngott, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Maarten van Smeden, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F. Jäger. Metrics reloaded: Recommendations for image analysis validation. arXiv preprint arXiv:2206.01653, 2023.

Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. Medical Image Analysis, 89:102918, 2023.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BraTS). IEEE transactions on medical imaging, 34(10):1993–2024, 2014.

Lenhard Pennig, Ulrike Cornelia Isabel Hoyer, Lukas Goertz, Rahil Shahzad, Thorsten Persigehl, Frank Thiele, Michael Perkuhn, Maximilian I Ruge, Christoph Kabbasch, Jan Borggrefe, et al. Primary central nervous system lymphoma: clinical evaluation of automated segmentation on multiparametric MRI using deep learning. Journal of Magnetic Resonance Imaging, 53(1):259–268, 2021.

Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. TorchIO: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Computer Methods and Programs in Biomedicine, 208:106236, 2021.

Lauren R Schaff and Christian Grommes. Primary central nervous system lymphoma. Blood, The Journal of the American Society of Hematology, 140(9):971–979, 2022.

Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063, 2019.

Evi J van Kempen, Max Post, Manoj Mannil, Richard L Witkam, Mark Ter Laan, Ajay Patel, Frederick JA Meijer, and Dylan Henssen. Performance of machine learning algorithms for glioma segmentation of brain MRI: a systematic literature review and meta-analysis. European Radiology, 31(12):9638–9653, 2021.

Tassilo Wald, Saikat Roy, Gregor Koehler, Nico Disch, Maximilian Rouven Rokuss, Julius Holzschuh, David Zimmerer, and Klaus Maier-Hein. SAM. MD: Zero-shot medical image segmentation capabilities of the segment anything model. In Medical Imaging with Deep Learning, short paper track, 2023.

Appendix A. Schematic overview of the tested segmentation methods

Figure S1 displays schematic overviews of the different tested segmentation methods.

Appendix B. Distribution of Dice score on the test set for different models

Figure S2 and S3 display the Dice score distribution across the test set for different models for lymphoma and MSD-BraTs tasks. For the lymphoma dataset, the variability (represented by the interquartile range - IQR) is lower for the best performing model (3D nnU-net) compared to other models. For the BraTS dataset, the IQR is similar for 3D nnU-Net and SAM, despite a much higher median value for the former. MedSAM, UniverSeg and nnU-Net display a larger variability. In general, the variability tends to be higher for lower

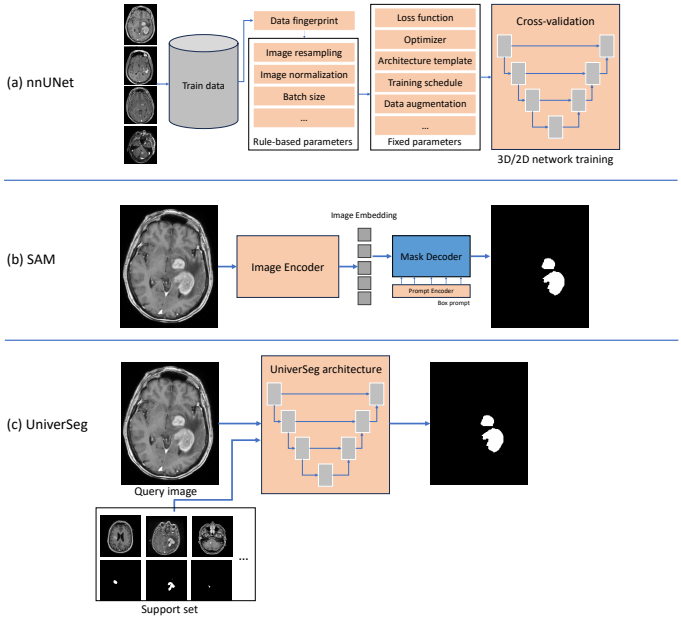

Figure S1: The model architectures of (a) nnUNet, (b) SAM and (c) UniverSeg.

performing models. Nevertheless, note that for all models there exist a few patients for which the performance is very poor.

Appendix C. Generating box prompts from ground truth segmentation

In this study, we manually constructed prompts for SAM and MedSAM experiments. Some other works have generated prompts from ground truth. We do not advocate for such approach as we believe it is a circular reasoning: if real labels are already available, automated methods would be redundant. Nevertheless, we conducted experiments to compare the performance when generating prompts from ground-truth to our manual box prompts. The details of these experiments are as follows:

- 3D-manual: our manually labeled 3D box prompts.
- 3D-gt-single: a single 3D prompt box generated from the ground-truth defined as the smallest 3D area which encompasses all lesions.
- 3D-gt-multi: multiple 3D boxes are generated: one for each connected component of the ground-truth (only relevant for lymphomas).
- 2D-gt-single: a different 2D box prompt for each slice in the axial plane. Each box is the smallest 2D area encompassing all lesions in a given slice.
- 2D-gt-multi: same as 2D-gt-single but with possibly multiple prompts per slice, corresponding to the different lesions (only relevant for lymphomas) .

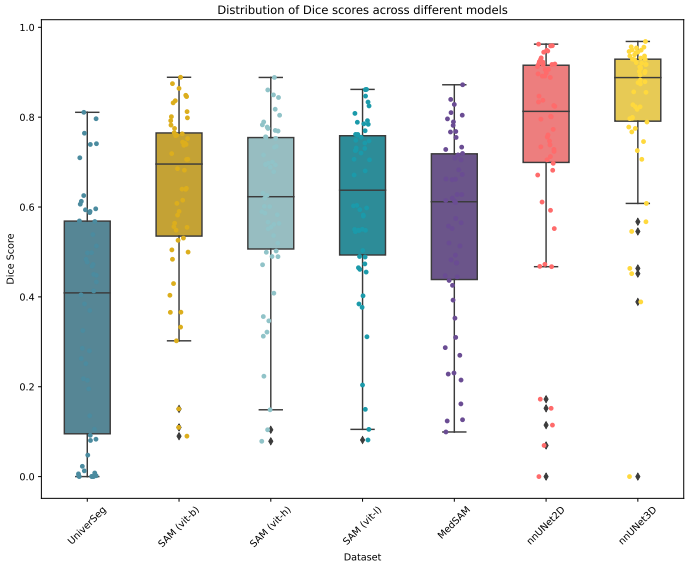

Figure S2: Boxplot illustrating the distribution of Dice score in different models for lymphoma data. Every box is bounded by the lower quartile (Q1) and the upper quartile (Q3), with the centre line representing the median. Whiskers extend from the box to the furthest data point within 1.5x the interquartile range of the box.

Based on the results presented in Table 1, the SAM-vit-b model achieved the best performance among SAM models. Consequently, we selected this version for our experiments. Results with different box prompts are provided in Table S1.

The experimental results indicate that SAM-b with manual annotation (3D-manual) outperforms a single 3D box prompt box generated from real labels (3D-gt-single). The advantage of 3D-manual over 3D-gt-single is attributed to our manual annotations focusing on the main lesion area, whereas 3D-gt-single may include unnecessarily large borders due to pixels distant from the lesion center, adversely affecting performance. This effect is particularly noticeable in lymphoma data. On the contrary, the 3D-gt-multi approach, where a bounding box is drawn for each connected region, resulted in a substantial improvement but the Dice was still about 9 percent points lower than that of the 3D nnU-Net. Nevertheless, it resulted in a small improvement in HD95. The generation of 2D box prompts from ground truth resulted in massive performance enhancement providing results which are better than those of nnU-Net (the difference is moderate in terms of Dice score but substantial in terms of HD95). This demonstrates that SAM-based segmentation models can well capture lesion borders when given optimal prompts. However, this correspond to a completly unrealistic scenario as the ground-truth was used to generate 2D boxes that perfectly match the lesions on every slice.

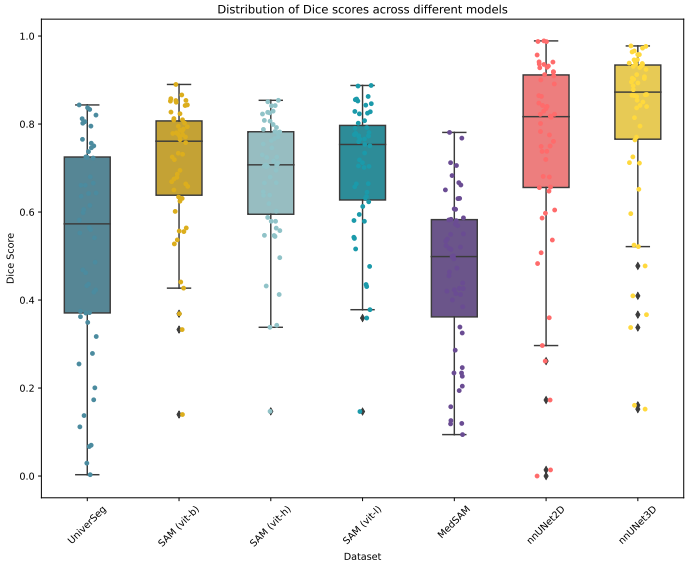

Figure S3: Boxplot illustrating the distribution of Dice score in different models for MSD-BraTs data. Every box is bounded by the lower quartile (Q1) and the upper quartile (Q3), with the centre line representing the median. Whiskers extend from the box to the furthest data point within 1.5x the interquartile range of the box.

Table S1: Comparison between our 3D manual box prompts and different box prompts generated from ground-truth (denoted as gt). Please refer to the text for the explanation of the different strategies used to generate prompts from ground truth. Results presented as mean with 95% bootstrap confidence interval computed on the independent test set.

Data	Prompt	Model	Dice	HD95	Topo3D
Lymphoma	3D-manual	SAM-vit-b	63.21 [57.96, 68.15]	21.98 [16.45, 28.11]	10.72 [5.54, 17.52]
	3D-gt-single		45.82 [37.84, 53.80]	19.61 [15.89, 23.60]	58.88 [38.80, 82.54]
	3D-gt-multi		73.12 [70.03, 76.28]	6.17 [5.27, 7.12]	10.19 [6.24, 14.74]
	2D-gt-single		75.17 [69.73, 80.16]	7.50 [5.32, 9.90]	9.09 [4.57, 14.22]
	2D-gt-multi		86.33 [84.35, 88.24]	1.91 [1.68, 2.19]	3.96 [2.20, 5.91]
	None	nnUNet-2d	73.99 [67.49, 80.06]	9.78 [6.03, 14.19]	1.41 [0.93, 1.94]
	None	nnUNet-3d	82.08 [77.05, 86.44]	7.96 [4.63, 11.92]	1.19 [0.81, 1.59]
MSD-BraTs	3D-manual	SAM-vit-b	69.96 [65.78, 73.80]	11.63 [9.28, 15.07]	17.58 [12.44, 23.56]
	3D-gt-single		64.68 [60.55,68.95]	14.23 [12.46, 16.46]	25.62 [18.89, 33.78]
	2D-gt-single		82.22 [79.39, 84.90]	4.46 [3.78, 5.23]	7.45 [5.04, 10.35]
	None	nnUNet-2d	73.46 [66.76, 79.30]	11.9 [7.76, 17.19]	7.67 [5.07, 11.02]
	None	nnUNet-3d	79.98 [74.42, 85.12]	6.47 [4.93, 8.31]	8.33 [5.61, 11.78]

Appendix D. UniverSeg under different thresholds

As mentioned above, we test different strategies to build the support set. Specifically, we chose one slice from each training scan, selecting those with the largest, smallest, and medium lesion proportions (denoted as large, small, and mid). We conducted an additional experiment where we combined slices with the largest and smallest lesion proportions to investigate performance at the maximum setting of 64 images (large+small experiment). Also, we experimented with different thresholdings of the soft predictions generated by UniverSeg, from 0.1 to 0.9. Results are presented in Table S2 and Figures S4 and S5. The small support set setting produces almost no predictions. For lymphoma, large and mid setting provided comparable results. For MSD-BraTS, the large setting proved more efficient. Predictions with thresholds below 0.7 tended to segment all high-signal regions, resulting in many meaningless regions.

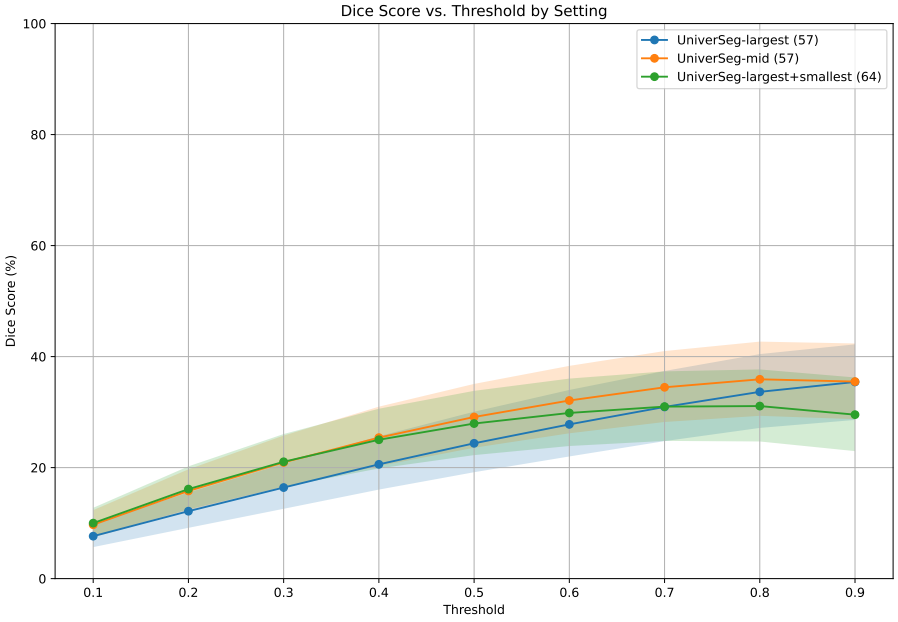

Figure S4: The Dice score of UniverSeg across varying thresholds for the lymphoma dataset.

Table S2: Influence of different strategies to define the support set and of different threshold on UniverSeg results. Results presented as mean with 95% bootstrap confidence interval computed on the independent test set.

Data	Support set	Threshold	Dice	HD95	Topo 3D
Lymphoma	large (57)	0.1	7.65 [5.68, 9.89]	92.82 [88.81, 96.89]	168.81 [154.48, 182.70]
		0.2	12.15 [9.15, 15.51]	89.54 [85.32, 93.75]	170.02 [158.37, 181.76]
		0.3	16.41 [12.58, 20.67]	87.55 [83.31, 91.87]	165.20 [153.91, 176.44]
		0.4	20.59 [16.04, 25.65]	85.67 [81.41, 90.02]	158.72 [148.74, 168.59]
		0.5	24.38 [19.17, 30.07]	84.12 [79.81, 88.56]	136.41 [128.74, 143.98]
		0.6	27.79 [22.01, 33.99]	82.60 [78.08, 87.22]	115.89 [109.20, 122.65]
		0.7	30.93 [24.77, 37.41]	80.46 [75.77, 85.25]	91.35 [85.39, 97.33]
		0.8	33.65 [27.15, 40.45]	78.01 [73.32, 82.91]	68.76 [63.94, 73.54]
		0.9	35.43 [28.59, 42.21]	73.96 [69.13, 78.81]	41.33 [37.96, 44.87]
	mid (57)	0.1	9.71 [7.43, 12.34]	91.41 [87.30, 95.47]	223.13 [206.89, 239.02]
		0.2	15.80 [12.31, 19.69]	86.60 [82.44, 90.83]	202.15 [189.06, 214.89]
		0.3	20.93 [16.49, 25.77]	84.05 [79.90, 88.36]	176.35 [164.52, 187.57]
		0.4	25.40 [20.29, 30.98]	82.12 [77.89, 86.51]	152.70 [142.44, 162.69]
		0.5	29.12 [23.58, 35.08]	80.22 [75.99, 84.57]	131.52 [122.41, 140.28]
		0.6	32.09 [26.18, 38.33]	78.38 [74.16, 82.70]	104.54 [97.15, 111.57]
		0.7	34.47 [28.22, 41.00]	76.53 [72.18, 80.93]	79.13 [73.24, 85.09]
		0.8	35.91 [29.33, 42.70]	73.89 [69.47, 78.41]	56.06 [51.22, 60.91]
		0.9	35.50 [28.73, 42.37]	67.51 [62.50, 72.63]	31.07 [27.87, 34.41]
	large+small (64)	0.1	10.00 [7.60, 12.76]	91.73 [87.66, 95.73]	239.76 [224.17, 254.91]
		0.2	16.13 [12.46, 20.23]	88.53 [84.31, 92.70]	224.91 [210.74, 238.96]
		0.3	21.05 [16.55, 26.05]	86.26 [81.93, 90.61]	197.76 [184.67, 210.31]
		0.4	25.02 [19.85, 30.61]	84.02 [79.70, 88.47]	169.83 [158.31, 181.57]
		0.5	27.94 [22.25, 33.83]	81.82 [77.27, 86.44]	141.13 [130.81, 150.98]
		0.6	29.85 [23.89, 36.04]	80.09 [75.37, 84.81]	111.61 [103.20, 119.65]
		0.7	30.99 [24.81, 37.34]	78.17 [73.38, 82.90]	83.39 [76.89, 89.70]
		0.8	31.09 [24.71, 37.69]	75.81 [70.96, 80.62]	57.56 [52.69, 62.33]
		0.9	29.54 [22.97, 36.24]	70.49 [65.48, 75.59]	32.50 [29.39, 35.72]
MSD-BraTs	large (57)	0.1	14.85 [12.56, 17.30]	83.10 [80.14, 86.26]	120.85 [113.65, 128.31]
		0.2	22.02 [18.84, 25.34]	81.30 [78.27, 84.50]	121.87 [113.80, 129.69]
		0.3	29.46 [25.55, 33.57]	80.68 [77.59, 83.94]	139.18 [129.78, 148.73]
		0.4	37.50 [32.94, 42.24]	79.99 [76.72, 83.39]	153.24 [143.96, 162.05]
		0.5	44.65 [39.34, 50.04]	78.44 [74.63, 82.05]	154.25 [146.71, 161.65]
		0.6	49.64 [43.83, 55.53]	74.96 [70.84, 79.23]	141.45 [133.64, 148.80]
		0.7	51.42 [45.03, 57.89]	68.73 [63.81, 73.79]	98.33 [91.85, 104.51]
		0.8	49.16 [42.32, 56.08]	53.89 [47.27, 60.63]	53.24 [48.84, 57.62]
		0.9	41.43 [34.60, 48.43]	30.75 [24.14, 37.87]	18.76 [15.84, 21.91]
	mid (57)	0.1	11.14 [8.94, 13.45]	87.80 [84.56, 91.10]	208.62 [197.42, 220.80]
		0.2	18.06 [14.42, 21.83]	82.11 [77.84, 86.44]	215.65 [201.93, 229.13]
		0.3	18.66 [14.44, 23.15]	64.27 [59.86, 68.72]	123.73 [113.20, 134.11]
		0.4	12.86 [9.19, 16.88]	52.37 [46.91, 57.77]	63.98 [56.91, 70.96]
		0.5	7.02 [4.47, 9.94]	39.89 [35.11, 44.74]	29.38 [24.95, 33.67]
		0.6	3.24 [1.83, 4.96]	34.44 [30.73, 38.43]	13.96 [11.13, 17.05]
		0.7	1.20 [0.59, 1.92]	∞	8.60 [6.20, 11.76]
		0.8	0.38 [0.11, 0.75]	∞	8.55 [5.96, 11.71]
		0.9	0.11 [0.01, 0.27]	∞	9.22 [6.42, 12.60]

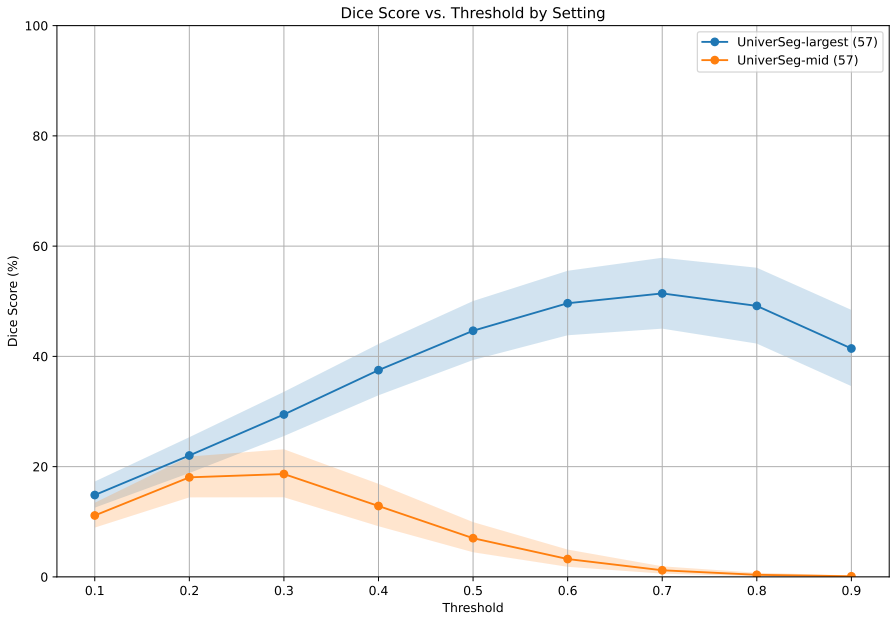

Figure S5: The Dice score of UniverSeg across varying thresholds for the MSD-BraTS dataset.