Radiology: Imaging Cancer

ORIGINAL RESEARCH

Automatic Segmentation of Primary Central Nervous System Lymphoma at Clinical Routine Postcontrast T1-weighted MRI

Guanghui Fu, PhD[1] • *Lucia Nichelli, MD*[1,2] • *Darío Herrán de la Gala, MD*[2] • *Sophie Loizillon, PhD*[1] • *Camile Bousfiha*[1] • *Romain Valabregue, PhD*[3,4] • *Agusti Alentorn, MD*[3,5] • *Khê Hoang-Xuan, MD, PhD*[3,5] • *Bertrand Mathon*[3,6] • *Carole Soussain*[7] • *Jean Pierre Marolleau*[8] • *Jérôme Paillassa*[9] • *Luc Taillandier*[10] • *Philippe Agapé*[11] • *Anna Schmitt*[12] • *Olivier Chinot*[13] • *Guido Ahle*[14] • *Didier Dormont, MD, PhD*[1,2] • *Caroline Houillier*[3,5] • *Stéphane Lehéricy, PhD*[2,3,4] • *Olivier Colliot, PhD*[1] • *for the French Network for Oculocerebral Lymphoma (LOC Network)*[15]

Author affiliations, funding, and conflicts of interest are listed at the end of this article.

[15] Members of the French Network for Oculocerebral Lymphoma (LOC Network) are listed in the Acknowledgments.

Radiology: Imaging Cancer 2025; 7(5):e240446 • https://doi.org/10.1148/rycan.240446 • Content codes: NR OI MR

Purpose: To develop and validate a deep learning model for automatic segmentation of primary central nervous system lymphoma (PCNSL) at postcontrast T1-weighted MRI.

Materials and Methods: Data were retrospectively collected from patients with pathologically proven immunocompetent PCNSL between September 2010 and February 2022. Postcontrast T1-weighted MRI scans were used to train and validate a deep learning model based on the nnU-Net framework. Manual segmentation by neuroradiologists served as the reference standard. The model was trained using an internal dataset from a single center and tested on both internal and external test sets from seven additional centers. Performance was assessed using Dice score, mean average surface distance, and F1 score. Statistical comparisons were performed using Mann–Whitney *U* test and bootstrap resampling for CIs.

Results: The study included 135 patients (68 female, 66 male, and one of unspecified sex; internal dataset: mean age ±SD, 67.0 years ± 12.0; external dataset: mean age, 75.5 years ± 13.6). The model achieved a mean Dice score of 0.84 (95% CI: 0.79, 0.88) on the internal test set (n = 44) and 0.88 (95% CI: 0.84, 0.91) on the external test set (n = 48), with no evidence of a difference between test sets (P = .59). Performance varied by lesion type; accuracy was highest in homogeneous discrete lesions, and performance was slightly decreased when numerous poorly defined infracentimetric lesions occurred. Strong volumetric correlation was observed between automatic and manual segmentations (internal: r = 0.99, P < .001; external: r = 0.98, P < .001).

Conclusion: A deep learning model achieved accurate and robust automatic segmentation of PCNSL across multiple clinical centers with different MRI acquisition parameters.

Supplemental material is available for this article.

Primary central nervous system lymphoma (PCNSL) is a rare, highly malignant brain tumor within diffuse large B-cell lymphoma whose incidence is continuously rising, especially in older persons (1,2). Several mutations, molecular alterations, and tumor microenvironment signatures have been linked to PCNSL in the past few decades, increasing the stratified complexity behind an apparently homogeneous abnormality (3). Recently, multiomics analysis has identified four different molecular PCNSL subtypes with distinct clinical behaviors and survival characteristics, laying the foundation for a deeper and more integrated clinicopathologic understanding (4,5). The oncologic management of PCNSL has also progressed substantially and is still rapidly evolving with the advent of targeted therapies and CD19-directed chimeric antigen receptor T cells (6–8).

In this flourishing, personalized, therapeutic scenario, accurate and objective evaluation of tumor burden is pivotal. To date, according to the International Primary CNS Lymphoma Collaborative Group (IPCG) (9), lymphoma tumor burden is defined by blood–brain barrier leakiness visible as abnormal contrast enhancement on gadolinium-enhanced T1-weighted MRI sequences. Brain lymphoma pathologic abnormality is known to be infiltrative and is not limited to the rupture in blood–brain barrier visible on a gadolinium-enhanced T1-weighted sequence; it is also visible as hyperintense on T2-weighted fluid-attenuated inversion recovery (FLAIR) images. However, T2-weighted FLAIR hyperintensities have lower specificity than the blood–brain barrier disruption visible at gadolinium-enhanced T1-weighted MRI, and evidence on the role of T2-weighted FLAIR sequences in brain lymphoma is scarce. Thus, nonenhancing tumor burden evaluation has not been included in IPCG criteria.

Although deep learning segmentation models have already been proposed for diffuse gliomas segmentation (10,11), no publicly available segmentation tool exists for brain lymphoma, probably given the rarity of the disease. The only automated volumetric models lack independent clinical validation (12,13). Pennig et al (13) trained a deep learning segmentation model (DeepMedic [14]) using MRI data from 220 patients with glioma and applied the model to 69 scans from 43 patients with lymphoma, achieving a median Dice score of 0.76. However, because the model was trained with gliomas, it is unclear whether this approach is optimal for lymphomas. Lymphomas often include a varying number of lesions, which is rare for gliomas (15,16). Multiple periventricular lesions are also common in lymphomas and less common in gliomas. In addition, in immunocompetent individuals, lymphomas usually display intense homogeneous contrast enhancement; in contrast, gliomas are more commonly heterogeneous, with several necrotic and hemorrhagic components. Finally, there are no reports on external generalization of models

Abbreviations

FLAIR = fluid-attenuated inversion recovery, IPCG = International Primary CNS Lymphoma Collaborative Group, MASD = mean average surface distance, PCNSL = primary central nervous system lymphoma, 3D = three-dimensional

Summary

The trained deep learning model accurately segmented primary central nervous system lymphoma with high generalizability across eight centers, offering an open-source tool that may enhance quantitative imaging evaluation in brain lymphoma.

Key Points

- This retrospective study, based on a cohort of 135 patients and using postcontrast T1-weighted MRI performed in a clinical context, led to the development of the first publicly available automatic segmentation model for primary central nervous system lymphoma; the model achieved high mean Dice scores of 0.84 on internal test data (n = 44) and 0.88 on external test data (n = 48).
- The model demonstrated strong volumetric correlation with manual annotations (internal: r = 0.99, P < .001; external: r = 0.98, P < .001) and achieved F1 scores of 76% and 67% on internal and external test sets, respectively.
- Performance varied by lesion type, with the best results in homogeneous, discrete lesions and decreased performance when multiple ill-defined infracentimetric lesions occurred.

Keywords

Brain Lymphoma, Brain Tumor, Automatic Segmentation, Artificial Intelligence, Deep Learning

using the current segmentation algorithms, which is crucial for clinical applications.

Zhang et al (12) built deep learning methods for automatic classification of brain tumors, including in 92 patients with lymphoma. Their approach involved a segmentation network, the results of which were used for classification, but did not evaluate the segmentation accuracy. Furthermore, none of these models can be downloaded and tested. Hence, a publicly available, automatic, validated segmentation model for brain lymphomas is highly desirable. Brain tumor segmentation is a complex task because of the high variability of lesion location, morphologic features, and dimension. The task is even more complex for real-life data because in daily clinical settings, MRI examinations are heterogeneous in terms of scanners, magnetic fields, acquisition parameters, two-dimensional or three-dimensional (3D) sequences, image resolutions, and sequence types (spin echo vs gradient echo).

In this study, we trained and evaluated, with both internal and external testing, a nnU-Net model (17) for automatic segmentation of histologically proven primary brain lymphomas at clinical routine gadolinium-enhanced T1-weighted MRI.

Materials and Methods

Study Sample

A multicenter retrospective study was conducted on 135 immunocompetent patients with pathologically confirmed Epstein–Barr virus–negative PCNSL from the French Network for Oculocerebral Lymphoma network database. The study was approved by the Institutional Ethical Committee (Pitié-Salpêtrière Hospital, Île-de-France VI, no. DC-2009-95) and by the French Data Protection Authority (Commission Nationale de l'Informatique et des Libertés, DR-2013-279). According to French regulation, consent was waived because these images were acquired as part of the routine clinical care of the patients.

Each patient underwent T1-weighted MRI after injection of gadolinium-based contrast agent. Patient selection process is described in Figure 1, and additional details are provided in Figure S1. First, we included patients from a single tertiary center (Pitié Salpêtrière University Hospital) based on the following inclusion criteria: *(a)* pathologic diagnosis of diffuse large B-cell lymphoma of the central nervous, *(b)* MRI availability, *(c)* absence of systemic involvement assessed with full-body CT or PET, *(d)* immunocompetent status as well as negative HIV serology results, and *(e)* age older than 18 years.

All images were converted from Digital Imaging and Communications in Medicine (DICOM) to NIfTI using dicom2niix (18) and organized according to the Brain Imaging Data Structure standard. Images were then selected by a certified neuroradiologist (L.N.) according to the following exclusion criteria: *(a)* absence of a T1-weighted sequence after injection of gadolinium-based contrast agent, *(b)* severe image artifacts at gadolinium-enhanced T1-weighted MRI that affected the diagnostic interpretation of the images, and *(c)* errors in DICOM–NIFTI conversion that affected the diagnostic interpretation of the images. Previous

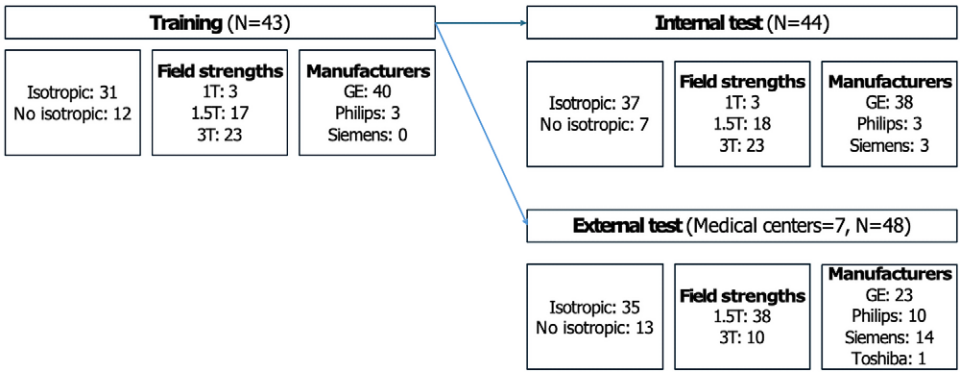

Figure 1: Data distribution of training and test sets. "External" refers to the sets with patients scanned from September 2010 to February 2022 at institutions different from the one where the training patients were scanned. N denotes the number of patients.

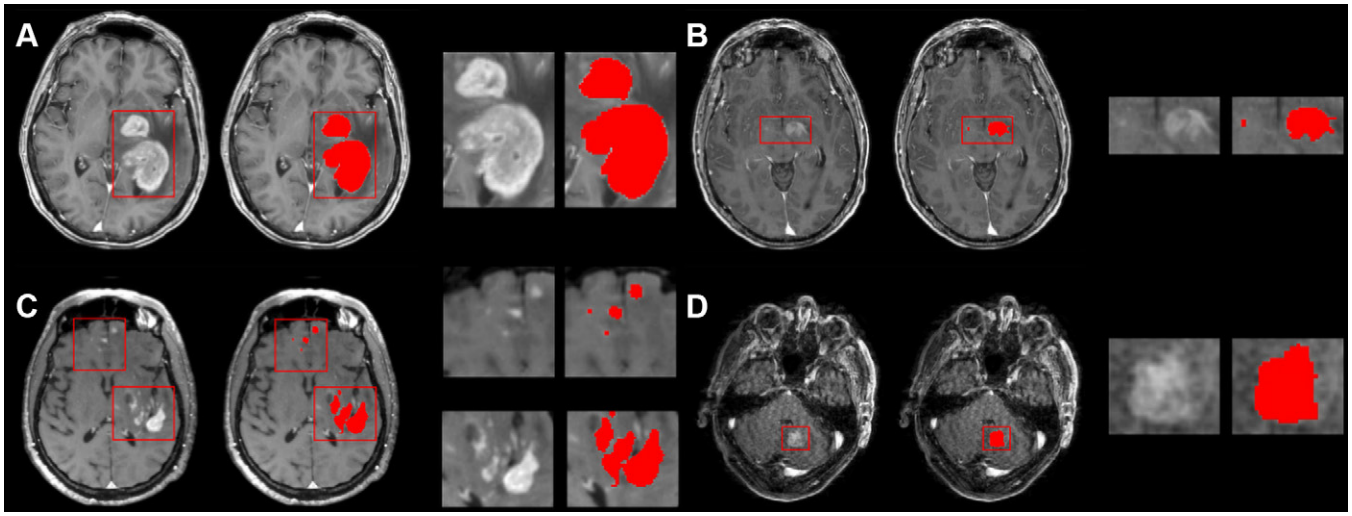

Figure 2: Gadolinium T1-weighted contrast-enhanced MRI scans acquired in clinical practice for patients with primary central nervous system lymphoma. For each of four patients **(A–D)**, an image on an axial plane is shown, followed by the same image overlaid with the annotated segmentation, and a zoom-in of the area(s) marked with a box.

treatment status (corticotherapy as well as different antitumor treatment strategies) was not an exclusion criterion, but in cases with multiple MRI examination for a single individual, baseline MRI (eg, MRI before therapy) was chosen. Artifacts of various origins (eg, motion, blooming, aliasing) that did not interfere substantially with image interpretation were not excluded. In case of multiple gadolinium-enhanced T1-weighted MRI sequences, high-resolution thin-section volumetric sequences were preferred. If more than one high-resolution sequence was present, a 3D spin-echo sequence was preferred to a 3D gradient-echo sequence because it provides an improved signal-to-noise ratio for the detection of T1 abnormal contrast enhancement (9).

Eighty-seven patients from a single center (Pitié Salpêtrière University Hospital) met the inclusion criteria. This cohort was used for training/validation and for internal testing. Subsequently, we included patients from other hospitals using the same selection process as for the internal cohort. Forty-eight patients met the criteria and are referred to as the external dataset. They came from seven centers (center I: Clermont-Ferrand University Hospital, seven patients; center II: Amiens University Hospital, seven patients; center III: Nancy University Hospital, seven patients; center IV: Timone University Hospital, seven patients; center V: Pasteur University Hospital, seven patients; center VI: Bergonié Institute, six patients; center VII: Oncologic Institute of the West, seven patients.). Forty-eight patients met the inclusion criteria. The study flowchart is shown in Figure S1.

MRI Characteristics

The images were acquired as part of clinical routine and were thus not harmonized (eg, different MRI scanners, sequence types, resolutions, contrast-to-noise ratio).

Internal data included 68 isotropic (or quasi-isotropic) 3D acquisitions and 19 nonisotropic two-dimensional acquisitions across different field strengths (3-T: 46; 1.5-T: 35; 1.0-T: six) and manufacturers (GE HealthCare: 78; Philips: six; Siemens: three). The dataset contained 53 gradient-echo and 34 spin-echo images. In-plane resolution ranged from 0.47 mm to 2 mm and section thickness ranged from 0.47 mm to 6.5 mm. The mean ± SD number of lesions per scan was 3.68 ± 3.26. Seventy-six MRI scans were acquired at baseline, 10 were acquired after induction therapy, and one patient had an unspecified treatment status.

External data included 35 isotropic (or quasi-isotropic) 3D acquisitions and 13 nonisotropic two-dimensional acquisitions. Field strengths and manufacturers also varied (field strengths: 3-T: 10 patients; 1.5-T: 38 patients; manufacturers: GE HealthCare: 23 patients; Philips: 10 patients; Siemens: 14 patients; Toshiba: one patient), as did sequence types (37 gradient echo, 11 spin echo). In-plane resolution ranged from 0.38 mm to 1.4 mm and section thickness ranged from 0.47 mm to 6.3 mm. The mean number of lesions per scan was 5.42 ± 5.54. The dataset included 34 pretreatment patients, 13 posttreatment patients, and one patient with an unspecified treatment status.

A representation of the cohorts' division used for the development of model is shown in Figure 1. Figure 2 illustrates the variability in lesion appearance, location, and image quality.

Manual Annotation of Tumors

The full dataset was annotated by a certified neuroradiologist (L.N.) with 5 years of experience in neuro-oncology imaging. According to IPCG consensus recommendations, lymphoma tumor burden is defined by the measure of blood–brain barrier leakiness visible as abnormal contrast enhancement at gadolinium-enhanced T1-weighted MRI. Therefore, in this study we focused on enhancing tumor burden and did not evaluate the nonenhancing tumor burden visible on T2-weighted FLAIR sequences.

All contrast-enhancing lesions were segmented using ITK-SNAP software, version 3.6 *(http://www.itksnap.org)*, with a semiautomatic active contour segmentation method and subsequent manual correction. To evaluate the consistency of manual segmentations, which we considered the reference standard, we assessed both intrareader and interreader variability using key segmentation metrics. Intrareader variability was quantified after the primary reader (L.N.) reannotated a random subset of 50 cases with a 1-month interval. Interreader variability was assessed after a second board-certified neuroradiologist (D.H.G.), with 1 year

of experience in neuro-oncology imaging, independently annotated the same 50 cases.

The primary reader (L.N.) also assessed the type of MRI sequence (spin echo or gradient echo) and classified the PCNSL imaging patterns. Because there is no reference standard for MRI characteristics of PCNSL, based on figure 2 of the IPCG consensus recommendations (9), the radiologic presentation of PCNSL was divided as follows: unique, homogeneous lesion (group A); multiple discrete homogeneous lesions (group B); single heterogeneous lesions (group C); multiple heterogeneous lesions, including linear ependymal enhancement (group D); and linear perivascular enhancement (group E).

Development of Automatic Segmentation Model

Each MRI volume in this dataset was resized and resampled to isotropic 1 × 1× 1–mm^3 voxels and dimensions of 240 × 240 × 160 using resample and resize operations from TorchIO (19) for model training purposes. We also rescaled the intensity to a range of 0 and 1. At the evaluation step, we resample the images back to the original resolution.

The segmentation approach was based on nnU-Net V2, a supervised deep learning segmentation method (17) that has been shown to be one of the most consistently effective algorithms across many medical applications (20). Initially, the pipeline autonomously extracted features from the dataset, including intensity distribution, image size, and modalities, to create data fingerprints. Subsequently, rule-based parameters used these dataset fingerprints to modify specific properties of the segmentation pipeline, following established heuristic rules. This included the automatic selection of appropriate image resampling methods, batch sizes, and other features. Certain robust configurations were predetermined, including the automatic determination of model architecture, training schedules, and implementation of data augmentation strategies. During the training time, fivefold cross-validation was executed on the training/validation data for 1000 epochs of training, to select the optimal configuration. No modifications were made to the nnU-Net algorithm, ensuring full compatibility, reproducibility, and easy deployment for researchers and clinicians. The model was trained based on a single Tesla V100SXM2-32GB GPU. This may vary depending on computational hardware in other environments.

We first tested the model trained on contrast-enhancing glioma data using the same nnU-Net framework to assess cross-disease transferability (Appendix S1, part H). As shown in Table S3, this model performed poorly on PCNSL, with Dice scores of 0.14 (internal) and 0.07 (external), and mean average surface distance (MASD) values of 44 mm (internal) and 69 mm (external). Given these low performances, we therefore used the same nnU-Net framework, trained on our study sample, and tested it.

Evaluation Metrics

To evaluate the performance of the nnU-Net model we used the following metrics, based on the consensus recommendations of Metrics Reloaded (21). Segmentation accuracy was assessed using the 3D Dice score and the MASD, which are computed at the voxel level and result in one measure per patient. The ability to detect the multiple lesion instances was assessed using the F1 score, which is computed at the instance level and results in one measure per patient. The F1 score evaluates the detection performance by balancing precision and recall for the different lesion instances of a given patient.

The segmentation map of a given patient was broken down into instances using connected-components analysis. The localization criterion was the mask intersection over union. We chose a small overlap threshold (intersection over union > 0.1 in our experiment) because of a high variability of structure sizes. The localization criterion allows computing the precision (proportion of predicted instances, which are present in the manual segmentation) and the recall (proportion of manual segmentation instances, which are retrieved by the automatic segmentation). The F1 score is then the harmonic mean of precision and recall. It is computed for each patient and quantifies how well the different lesions of a given patient are detected. In addition, even though this metric is not present in Metrics Reloaded recommendations, we assessed the concordance of manual and automatic volumetric measurements using Pearson correlation because it is clinically useful information.

Statistical Analysis

For each metric reported, we provide the mean value across patients as well as the corresponding 95% CI computed using bootstrap resampling (5000 samples) over the independent test set. A *P* value less than .05 was considered to indicate a statistically significant difference. To compare model performance across subgroups, we assessed the statistical significance of Dice score differences between the internal and external test sets using the Mann–Whitney *U* test. This subgroup analysis was prespecified as part of our evaluation protocol. Statistical analyses were performed using Python, version 3.11.4.

Data Availability

The repository was made publicly available, containing the trained model to automatically segment brain lymphoma. We could not make the images of individual patients available because of regulatory constraints. The trained model is available at: *https://github.com/GuanghuiFU/nnunet_lymphoma_segment/tree/main*.

Results

Patient Characteristics

The internal dataset was randomly split into a training/validation set with 43 patients (25 female and 18 male) aged 32–86 years (mean age, 67.07 years ± 11.79) and an internal test set with 44 patients (23 female, 20 male, and one patient of unspecified sex) aged 32–86 years (mean age, 67.03 years ± 12.23). Forty-eight patients came from seven other centers and formed the external test set (20 female and 28 male) aged 33–95 years (mean age, 75.52 years ± 13.60). The test sets were used to evaluate the performance of the models. MRI examinations were performed between 2010 and 2021. Data distribution is reported in Table 1. The distribution of Karnofsky scores for the internal and external test sets is shown in Figure S2.

Manual Segmentation Reliability

For intrareader reliability, where the primary annotator resegmented cases after a 1-month interval, the variability metrics

Table 1: Distribution of Datasets for Model Development

Medical Center	Data Split	Patients	No. of Lesions per Scan	Age (y)	Sex	Location	Categories
Internal	Train/validation	43	2.76 ± 2.64	67.07 ± 11.79 (32–86)	M: 18; F: 25	S: 41; I: 5	A: 9; B: 11; C: 8; D: 12; E: 3
Internal	Test	44	3.68 ± 3.22	67.03 ± 12.23 (32–86)	M: 20; F: 23, U: 1	S: 42; I: 13	A: 14; B: 12; C: 9; D: 5; E: 4
External							
I	Test	7	7.29 ± 8.92	72.43 ± 7.78 (33–95)	M: 2; F: 5	S: 5; I: 3	A: 2; B: 1; C: 2; D: 2; E: 0
II	Test	7	9 ± 6.93	80.29 ± 2.83 (72–89)	M: 6; F: 1	S: 6; I: 2	A: 2; B: 1; C: 1; D: 2; E: 1
III	Test	7	4.57 ± 3.29	76.14 ± 16.97 (59–91)	M: 4; F: 3	S: 5; I: 3	A: 0; B: 4; C: 1; D: 2; E: 0
IV	Test	7	4.43 ± 4.07	71.71 ± 12.02 (48–88)	M: 3; F: 4	S: 7; I: 0	A: 2; B: 1; C: 3; D: 1; E: 0
V	Test	7	4.29 ± 3.49	72.29 ± 15.56 (60–91)	M: 4; F: 3	S: 7; I: 1	A: 2; B: 3; C: 2; D: 0; E: 0
VI	Test	6	3.83 ± 2.61	83.5 ± 1.41 (75–89)	M: 5; F: 1	S: 5; I: 1	A: 3; B: 1; C: 1; D: 1; E: 0
VII	Test	7	3.28 ± 2.60	72.43 ± 7.78 (45–90)	M: 4; F: 3	S: 6; I: 3	A: 1; B: 3; C: 2; D: 1; E: 0
External total		48	5.47 ± 5.48	75.52 ± 13.60 (33–95)	M: 28; F: 20	S: 41; I: 13	A: 12; B: 14; C: 12; D: 9; E: 1

Note.—Unless otherwise noted, values are expressed as numbers of patients (corresponding to the number of scans). Values are the means ± SDs, with ranges in parentheses. The model was trained using an internal dataset and validated both internally (internal) and across seven external centers (external I–VII). In the location column, "S" denotes supratentorial and "I" denotes infratentorial. In the categories column, the classifications are as follows: unique, homogeneous lesion (group A); multiple discrete homogeneous lesions (group B); single discrete heterogeneous lesions (group C); multiple heterogeneous lesions, including linear ependymal enhancement (group D); and linear perivascular enhancement (group E). A tumor may be located in both supratentorial and infratentorial regions simultaneously and may belong to multiple categories.

(expressed as mean with 95% bootstrap CIs) were as follows: Dice score, 0.85 (0.82, 0.88); F1 score, 75% (68, 82), and MASD, 1.05 (0.77, 1.39). For interreader reliability, where segmentations were independently performed by a second certified radiologist, the variability metrics were as follows: Dice score, 0.81 (0.78, 0.85); MASD, 1.64 (1.20, 2.17); F1 score, 74% (68, 80).

Overall Segmentation Performance

The model achieved a mean (95% CI) Dice score of 0.84 (0.79, 0.88) on the internal test set and 0.88 (0.84, 0.91) on the external test set. It also achieved F1 scores of 76% on the internal test set and 67% on the external test set. Table 2 details the performance of PCNSL segmentation in internal and external datasets using the three chosen evaluation metrics (Dice score, MASD, F1 score). Figure 3 displays the Dice score distribution of our model across the internal and external test sets; Figure S4 and Figure S5 show the distributions of the MASD and F1 score. The Mann–Whitney *U* test showed that the performance between the internal and external datasets were not statistically significant ($P = .59$).

Volumetric Assessment

Volumetric correlation between the automatic segmentation and the manual annotation had a correlation coefficient of 0.99 ($P < .001$) for the internal test set and 0.98 ($P < .001$) for the external test set (Fig 4). After exclusion of patients with very large reference standard label tumor volumes (≥30 cm^3), the correlation coefficients were 0.98 ($P < .001$) for the internal test set and 0.94 ($P < .001$) for the external test set (Fig S6). Bland–Altman analysis (Fig S3) further confirmed good agreement, with the model showing a slight systematic underestimation compared with manual segmentation.

Subgroup Analysis

Figure 6 illustrates the mean performance metrics of the model across different PCNSL tumor subtypes, as previously described. Table 3 details this performance in each center. The model performed best on homogeneous and discrete lesions (groups A and B), whereas performance slightly decreased for single heterogeneous lesions and periventricular pattern (group D) and markedly decreased in case of perivascular enhancement (group E).

We also reevaluated model performance after excluding post-treatment or unknown treatment status cases from both the internal ($n = 5$) and external ($n = 14$) test sets (Appendix S1, part F). As shown in Table S2, Dice scores remained consistent (0.84 for internal data and 0.87 for external), with no statistically significant difference compared with the full dataset ($P = .81$ and 0.82, respectively).

Table 2: Performance of Primary Central Nervous System Lymphoma Model Segmentation in Internal and External Test Set and Centers I–VII

Hospitals (Patients)	Dice Score	MASD (mm)	F1 Score (%)
Internal (*n* = 44)	0.84 (0.79, 0.88)	1.69 (0.94, 2.67)	76 (69, 83)
External (*n* = 48)	0.87 (0.84, 0.90)	1.44 (0.90, 2.21)	67 (59, 74)
External: I (*n* = 7)	0.85 (0.76, 0.92)	1.37 (0.76, 2.07)	62 (40, 80)
External: II (*n* = 7)	0.88 (0.82, 0.92)	1.42 (0.88, 1.94)	53 (37, 72)
External: III (*n* = 7)	0.75 (0.55, 0.89)	3.13 (0.40, 7.31)	71 (50, 90)
External: IV (*n* = 7)	0.91 (0.89, 0.93)	1.33 (0.70, 2.17)	69 (54, 84)
External: V (*n* = 7)	0.91 (0.86, 0.95)	0.48 (0.24, 0.72)	73 (57, 90)
External: VI (*n* = 6)	0.91 (0.88, 0.93)	0.53 (0.42, 0.68)	66 (42, 90)
External: VII (*n* = 7)	0.92 (0.89, 0.94)	1.69 (0.52, 3.44)	71 (56, 87)

Note.—The results are presented as means with 95% bootstrap CIs computed on the independent test set. MASD = mean average surface distance.

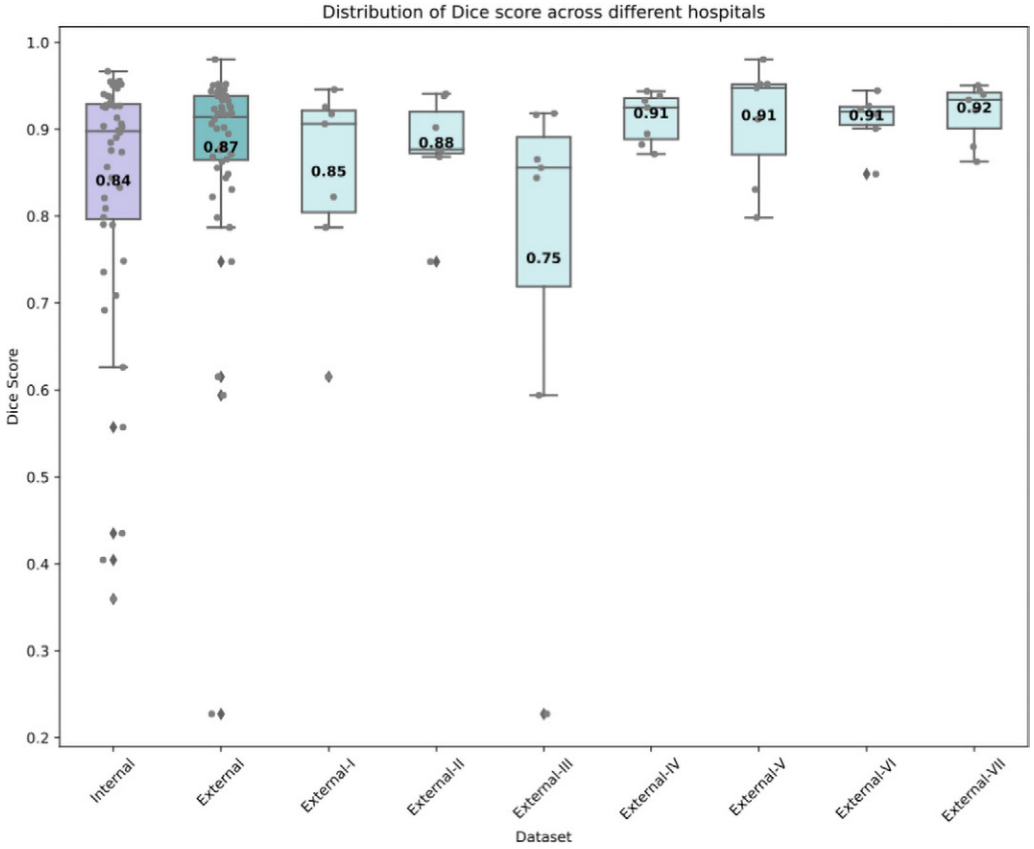

Figure 3: Distribution of Dice score for the internal test set, the external test set, and the seven medical centers of the external test set. Each box is bounded by the lower quartile and the upper quartile, with the center line representing the median. Whiskers extend from the box to the farthest data point within 1.5 times the IQR of the box.

To further understand failure modes, we analyzed test cases with Dice scores below 0.60, as summarized in Table S4 (Appendix S1, part I). Poor image quality (eg, strong motion artifacts), low lesion contrast, and undersegmentation of small (<1 cm) lesions, especially in the periventricular location were common error sources.

Figure 5 shows an example of expert annotation and automatic segmentation in a patients with movement artifacts.

Discussion

We trained and extensively evaluated an automated segmentation model for PCNSL at clinical routine gadolinium-enhanced T1-weighted MRI, demonstrating its accuracy and robustness. Our model, developed at a single medical center and internally and externally tested, was highly accurate (Dice scores of 0.84 and 0.88 on internal and external test sets, respectively).

Of note, our model is publicly available to encourage its widespread use. To the best of our knowledge, no other open-source model with external testing exists for this primary brain tumor. There is only one published article on automated segmentation in PCNSL (13). In contrast to our model, which was trained on pathologically proven PCNSL, the deep learning model was trained on MRI examinations from patients with glioblastoma.

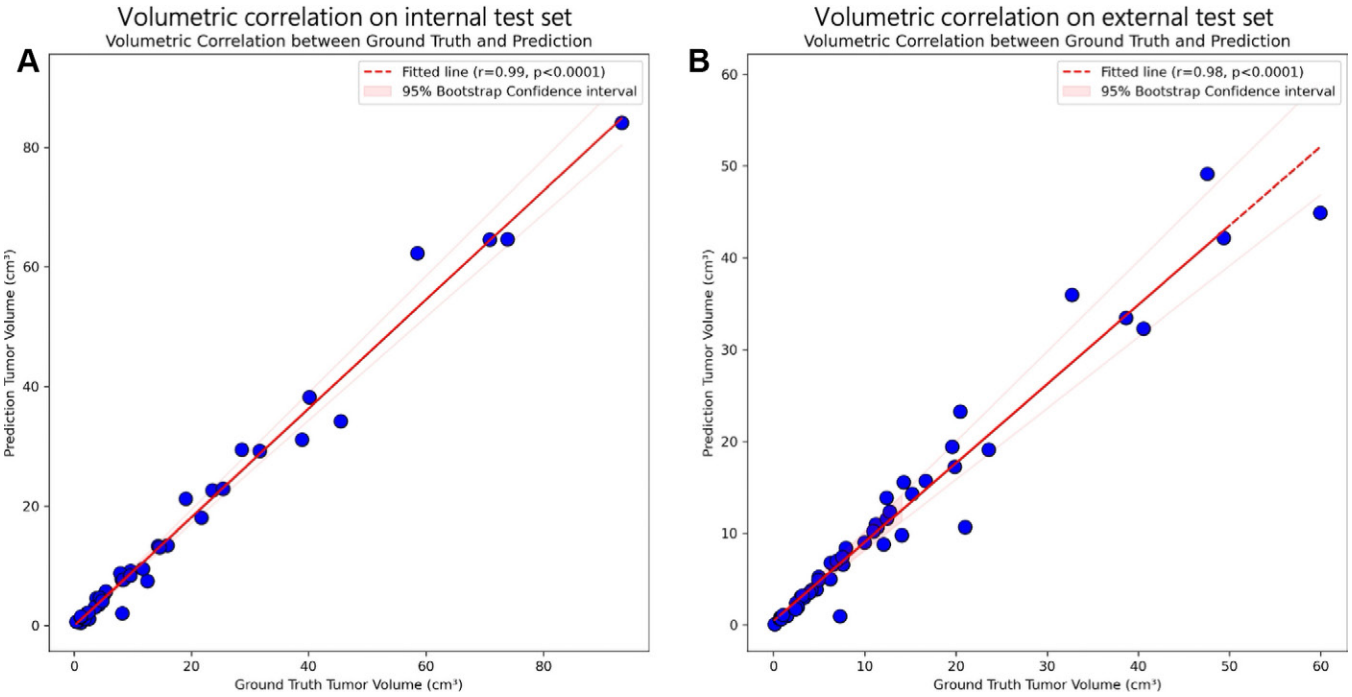

Figure 4: Volumetric correlation between the manually annotated reference standard label and the automatic segmentation for total tumor volume per MRI scan was assessed using Pearson correlation coefficient (*r*) for both the internal and external test sets. The shaded area represents the 95% bootstrap CI of the fitted line.

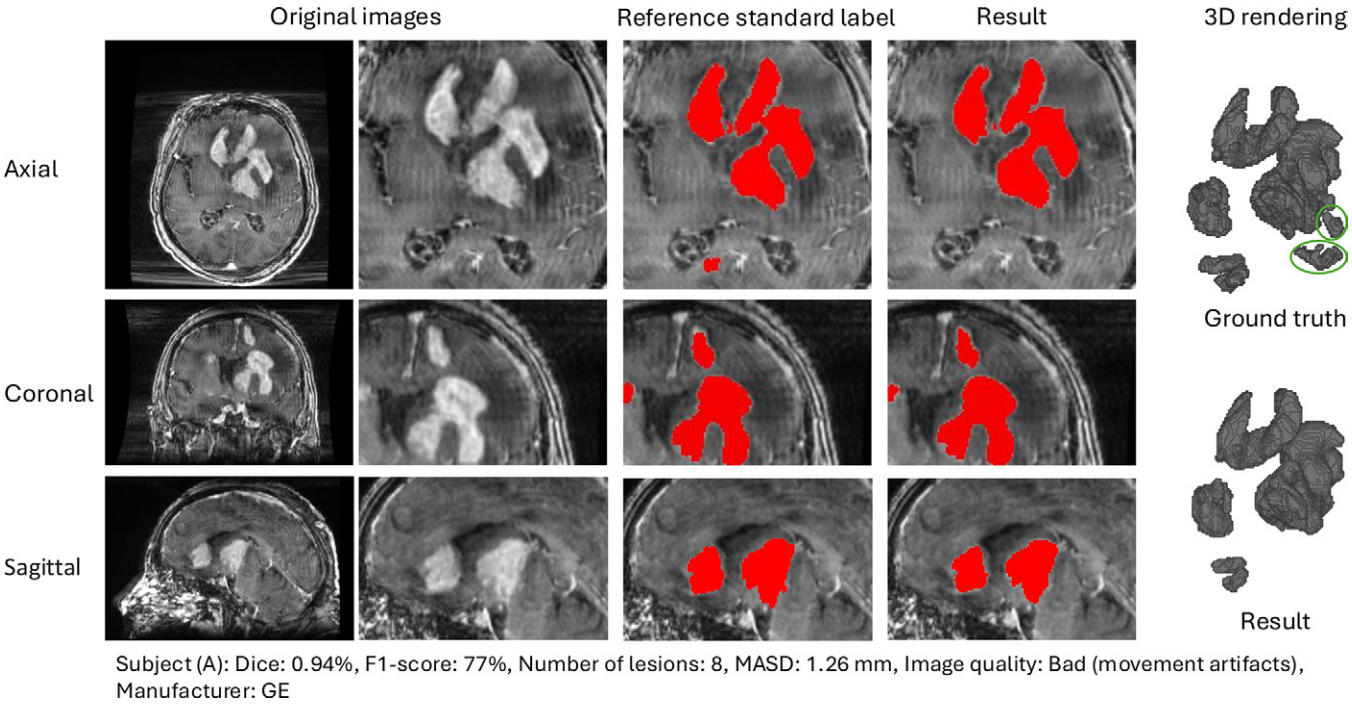

Figure 5: Reference standard label versus automatic segmentation results for a primary central nervous system lymphoma visible on a gadolinium T1-weighted contrast-enhanced MRI sequence with movement artifacts in a 67-year-old female patient. The main prediction error regions are marked with green circles. MASD = mean average surface distance. 3D = three-dimensional.

The experiments conducted on our population do not support the translation of nnU-Net models trained on glioma data to PCNSL (Dice scores of 0.14 and 0.07, in the internal and external population, respectively, and MASD values > 40 mm). Instead, they advocate for disease-specific models to achieve accurate tumor segmentation. High-grade gliomas and PCNSL can have similar MRI findings, but they usually differ. Grade 4 gliomas are typically necrotic and highly heterogeneous, whereas lymphomas display more uniform, homogeneous contrast enhancement. In addition, brain lymphomas often have multiple distant locations, unlike gliomas, and frequently display periventricular and ependymal enhancement, which is rarer at the time of glioma diagnosis. Furthermore, the performance of the proposed deep learning model for PCNSL segmentation is lower than that of our model, with a mean Dice score of 0.76, and it did not include external validation (13). In contrast, one of the main

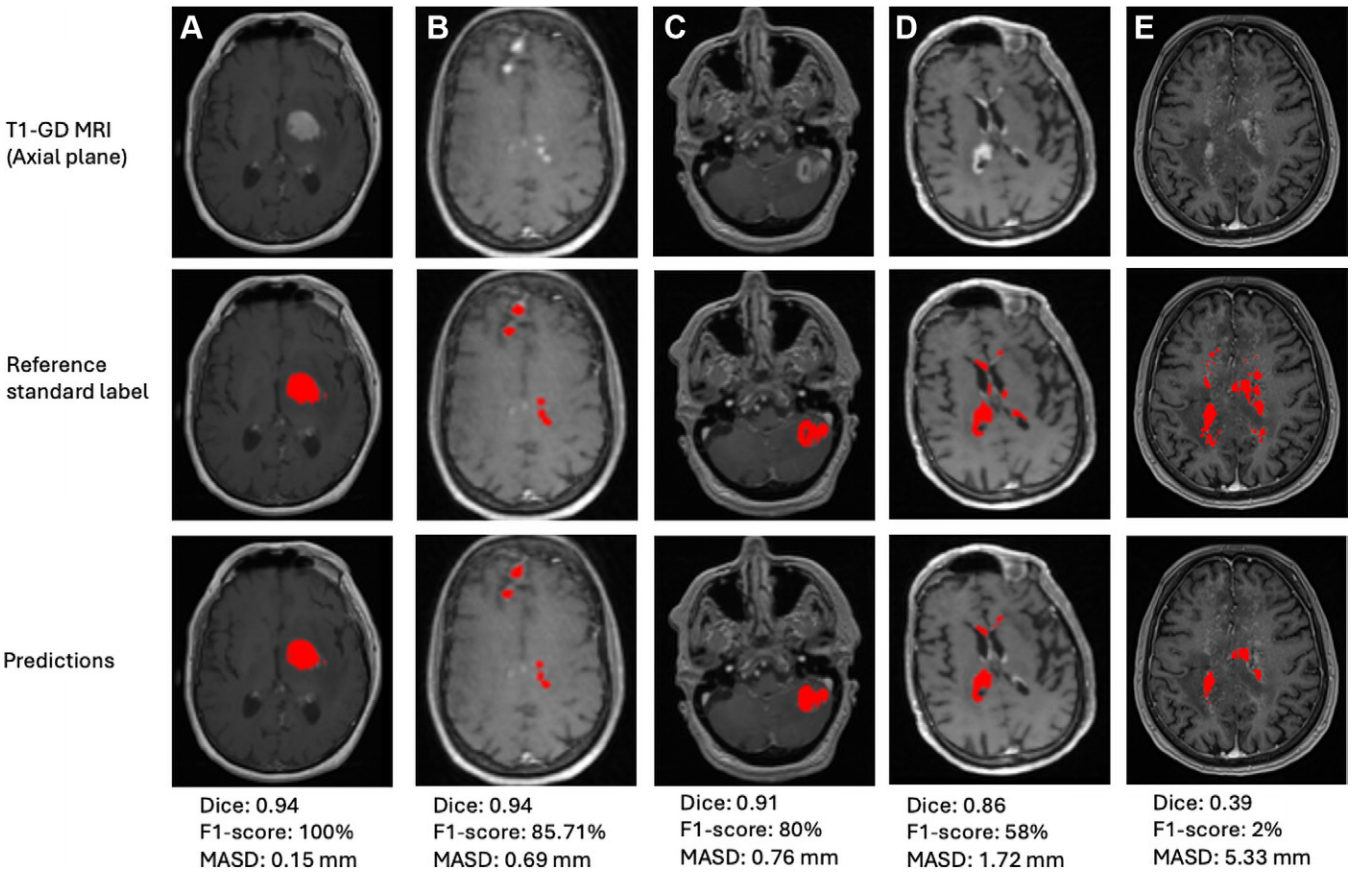

Figure 6: Automatic segmentation results in different primary central nervous system lymphoma (PCNSL) cases. **(A)** Single, homogeneous lesion in a 72-year-old male patient. **(B)** Multiple discrete homogeneous PCNSL lesions in a 67-year-old female patient. **(C)** Single heterogeneous lesion in a 73-year-old female patient. **(D)** Multiple ependymal enhancements in a 75-year-old male patient. **(E)** Multiple perivascular enhancements in a 64-year-old male patient. Red indicates annotated lesions. MASD = mean average surface distance, T1-GD = T1-weighted gadolinium-enhanced.

Table 3: Performance of Primary Central Nervous System Lymphoma Model Segmentation in Internal and External Test Set Group by Tumor Type

Tumor (Patients)	Dice Score	MASD (mm)	F1 Score (%)
Internal			
A (*n* = 11)	0.84 (0.75, 0.91)	0.93 (0.63, 1.36)	93 (86, 100)
B (*n* = 11)	0.83 (0.71, 0.92)	2.95 (0.67, 6.63)	77 (64, 88)
C (*n* = 14)	0.87 (0.83, 0.91)	0.90 (0.66, 1.14)	68 (56, 81)
D (*n* = 4)	0.78 (0.63, 0.88)	3.01 (1.00, 6.35)	63 (48, 76)
A + D (*n* = 3)	0.74 (0.36, 0.93)	2.02 (0.39, 5.08)	79 (36, 100)
B + E (*n* = 1)	0.93 (0.93, 0.93)	0.89 (0.89, 0.89)	40 (40, 40)
External			
A (*n* = 12)	0.93 (0.91, 0.94)	0.92 (0.44, 1.59)	71 (54, 86)
B (*n* = 12)	0.87 (0.81, 0.92)	1.39 (0.48, 2.62)	78 (69, 86)
C (*n* = 21)	0.85 (0.78, 0.91)	1.78 (0.89, 3.26)	60 (49, 71)
B + D (*n* = 3)	0.82 (0.75, 0.87)	1.36 (0.92, 2.16)	53 (44, 67)

Note.—Tumor types were as follows: unique, homogeneous lesion (group A); multiple discrete homogeneous lesions (group B); single heterogeneous lesions (group C); multiple heterogeneous lesions, including linear ependymal enhancement (group D); and linear perivascular enhancement (group E). The results are presented as means with 95% bootstrap CI computed on the independent test set.

strengths of our model is its ability to generalize well in real-world clinical settings, achieving similar Dice scores between the internal and external test sets. Among the seven external centers, the performance was generally consistent, except for center III, which showed a lower but acceptable Dice score (0.75).

Another strength of our study is that we detailed the F1 score, according to Metrics Reloaded guidelines (21). The model achieved F1 scores of 76% on the internal test set and 67% on the external test set, indicating good lesion detection performance. Because previous studies did not document these detection measures, a direct comparison with our results is not possible. The F1 score is an infrequently reported accuracy metric in machine learning models. However, this score is important, especially in a real-world dataset, because it comprehensively computes the number of times a model makes a correct prediction on the dataset by balancing precision and recall.

In terms of reliability of manual annotations, both intrareader and interreader Dice scores exceeded 0.80, indicating a high level of segmentation consistency. The lower MASD in intrareader comparisons (1.05 mm) suggests that the same annotator achieved greater spatial precision, whereas the slight decrease in interreader agreement (1.64 mm) reflects expected variability between different radiologists.

In terms of the research dataset, although we did not conduct a formal power analysis, our cohort of 135 patients represents the largest PCNSL segmentation dataset to date, surpassing that of previous studies, such as those conducted by Pennig et al (13) (43 patients) and Zhang et al (22) (92 patients). Given the rarity of PCNSL, this dataset constitutes a substantial, clinically representative multicentric collection. The strong and consistent performance of our model on both internal and external test sets supports the adequacy of this sample size for training and evaluation.

A strong volumetric correlation between automatic segmentation and manual annotation was also observed (Fig 4). These findings indicate that our model offers a consistent approach to volumetric assessment, potentially reducing current interobserver variability in clinical settings. Although PCNSL response assessment is often straightforward, interobserver variability in central review remains a challenge (24). A recent study conducted with PCNSL MRI also suggests that standardized volumetric measurements could improve tumor monitoring and enable earlier identification of responders versus nonresponders (25). The integration of automated segmentation tools with longitudinal imaging aims to refine response assessment protocols beyond conventional sum-product methods, especially for cases with multiple, irregularly shaped lesions. The robustness and reliability of our model in clinical practice provide a strong foundation for fully automated PCNSL MRI evaluation.

To better understand our model's performance, we visually categorized our dataset into five representative PCNSL patterns at gadolinium-enhanced T1-weighted MRI (Fig 6), and we analyzed test cases with low Dice scores (Appendix S1, part I). The model performed best on homogeneous and discrete lesions (eg, subtypes A and B). Its performance decreased for periventricular and perivascular patterns, which often involve multiple, poorly defined, and highly heterogeneous infracentimetric lesions. These findings indicate that although the model is reliable, its performance may be limited under these specific conditions. In addition, because our dataset lacked cases with purely meningeal spread, our model was neither trained nor validated for this pattern, limiting its applicability in such cases. Future work should address this limitation by incorporating a broader dataset with more representation of these rare patterns.

It is worth nothing that, in daily practice, leptomeningeal and perivascular enhancement are also challenging for clinicians because defining their boundaries is difficult and dependent on acquisition parameters. Interleukin-10 quantification in cerebrospinal fluid is useful in these complex imaging scenarios (22,23).

Our study had several limitations. First, despite being the largest PCNSL segmentation dataset to date, our sample size was limited by the rarity of the disease, potentially limiting the model's ability to learn uncommon imaging patterns, underscoring the need for larger studies to validate our model. The hope is that our open-source link will encourage further research and refinement of the model's performance. Second, the absence of T2-weighted FLAIR sequences in the segmentation framework limits our model. T2-weighted FLAIR hyperintense lesions are not specific, especially in older patients with concomitant vascular leukopathy or methotrexate-related leukoencephalopathy, and evidence on its evaluation in brain lymphoma is scarce. However, we recognize that accurate quantification of nonenhancing tumor burden is advisable, and we plan to expand our dataset with this sequence. Third, our model is intended for PCNSL segmentation at different stages of diagnosis and treatment, it was primarily trained and evaluated on pretreatment MRI scans. The relatively small number of posttreatment cases limits our ability to evaluate its performance for follow-up evaluations. Expanding the dataset with more posttreatment MRI examinations will help assess its robustness in long-term monitoring.

In summary, we constructed an automatic PCNSL segmentation model using postcontrast T1-weighted MRI data from clinical routine and tested the model on data from seven other medical centers. Our results concurrently agreed that the proposed deep learning method was accurate and generalizable. We achieved good performance, similar to that of manual annotation, in both the internal and external test sets. This publicly available segmentation model can serve as a valuable analytical tool for clinical research on PCNSLs. Beyond tumor burden quantification, this tool can also support clinical decision-making by integrating radiomic features with survival outcomes, thereby aiding in the prediction of therapeutic resistance and disease progression. The model's ability to evaluate volumetric changes over time could enhance response assessment in clinical trials and improve patient stratification for treatment. Further research will explore how integrating molecular and genetic tumor profiles with automated MRI analysis can refine prognostic and predictive models.

Author affiliations:

[1] Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, 47 Boulevard de l'Hôpital, 75013 Paris, France

[2] Department of Neuroradiology, DMU DIAMENT, AP-HP, Pitié Salpêtrière, Paris, France

[3] ICM, Centre de NeuroImagerie de Recherche-CENIR, Paris, France

[4] Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France

[5] Department of Neurology 2, DMU Neurosciences, AP-HP, Pitié Salpêtrière, Paris, France

[6] Department of Neurosurgery, AP-HP Pitié-Salpêtrière Hospital, Sorbonne University, Paris, France

[7] Department of Clinical Hematology, Institut Curie, Saint Cloud, Paris, France

[8] Department of Hematology, CHU Amiens-Picardie, Amiens, France

[9] Department of Hematology, CHU d'Angers, Angers, France

[10] Department of Neuro-Oncology, Université de Lorraine, CHRU Nancy, Nancy, France

[11] Department of Hematology, Centre Hospitalier de Laval, Laval, France

[12] Department of Hematology, Institut Bergonié, Bordeaux, France

[13] Department of Neuro-Oncology, AP-HM, University Hospital Timone, Marseille, France

[14] Department of Neurology, Hôpital Pasteur - Hôpitaux Civils de Colmar, Colmar, France

Received November 18, 2024; revision requested January 22; revision received June 11; accepted August 4.

Address correspondence to: O. Colliot.

Funding: 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), from the European Union's Horizon Europe Framework Programme (grant number 101136607, project CLARA) and from the Fonds Recherche Neurosciences under the project "NEIMO - Monitoring Neuroinflammation: a Paris Brain Institute-Yale Neurology Collaborative program". G.F. is supported by the Chinese Government Scholarship provided by China Scholarship Council (CSC). L.N. is supported by an Inria/AP-HP "Poste d'accueil.

Acknowledgments: On behalf of the French Network for Oculocerebral Lymphoma (LOC Network): Caroline Houillier (CHU Pitié Salpêtrière, Paris, France), Sylvain Choquet (CHU Pitié Salpêtrière, Paris, France), Khê Hoang-Xuan (CHU Pitié Salpêtrière, Paris, France), Valérie Touitou (CHU Pitié Salpêtrière, Paris, France), Carole Soussain (Institut Curie, Saint Cloud, France), Nathalie Cassoux (Institut Curie, Saint Cloud, France), Denis Malaise (Institut Curie, Saint Cloud, France), Renata Ursu (Hôpital Saint-Louis, Paris, France), Lise Willems (CHU Paris Centre Site Cochin, Paris, France), Hervé Ghesquières (CHU Lyon Sud, Lyon, France), Anna Schmitt (Institut Bergonié, Bordeaux, France), Olivier Chinot (CHU de la Timone, Marseille, France), Emeline Tabouret (CHU de la Timone, Marseille, France), Luc Taillandier (CHU de Nancy, Nancy, France), Marie Blonski (CHU de Nancy, Nancy, France), Roch Houot (CHU de Rennes, Rennes, France), Guido Ahle (Hôpitaux Civils, Colmar, France), Gandhi Damaj (CHU de Caen, Caen, France), Cécile Molucon-Chabrot (CHU de ClermontFerrand, Clermont-Ferrand, France), Vincent Delwail (CHU de ClermontFerrand, Clermont-Ferrand, France), Michel Fabbro (Institut du Cancer de Montpellier Val d'Aurelle, Montpellier, France), Fabrice Jardin (Centre Henri Becquerel, Rouen, France), Adrien Chauchet (CHU de Besançon, Besançon, France), Franck Morschhauser (CHRU de Lille, Lille, France), Olivier Casasnovas (CHU de Dijon, Dijon, France), Rémy Gressin (CHU de Grenoble, Grenoble, France), Luc-Matthieu Fornecker (CHU de Strasbourg, Strasbourg, France), Julie Abraham (CHU de Limoges, Limoges, France), Jean-Pierre Marolleau (CHU d'Amiens, Amiens, France), Adrian Tempescul (CHU de Brest, Brest, France), Philippe Agapé (Institut de Cancérologie de l'Ouest, Saint Herblain, France), Lucie Oberic (Institut Universitaire du Cancer de Toulouse, Toulouse, France), Emmanuel Gyan (CHU Tours, Tours, France), Frédéric Peyrade (Centre Antoine Lacassagne, Nice, France), Thomas Gastinne (CHU de Nantes, Nantes, France).

Author contributions: Guarantors of integrity of entire study, **G.F., L.N., R.V., K.H.X., J.P., O. Chinot**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **G.F., L.N., D.H.d.l.G., C.S., D.D., O. Colliot**; clinical studies, **L.N., S.L., A.A., B.M., C.S., J.P.M., P.A., A.S., G.A., D.D., C.H.**; experimental studies, **G.F., L.N., D.H.d.l.G., S.L.**; statistical analysis, **G.F., L.N., S.L., A.A.**; and manuscript editing, **G.F., L.N., S.L., C.B., R.V., A.A., K.H.X., B.M., J.P., O. Chinot, G.A., D.D., C.H., S.L., O. Colliot**

Disclosures of conflicts of interest: G.F. Funding from the French government under management of Agence Nationale de la Recherche as part of the "France 2030" program (reference ANR-23-IACL-0008, project PRAIRIE-PSAI), as part of the "Investissements d'avenir" program (reference ANR-19- P3IA-0001, project PRAIRIE 3IA Institute and reference ANR-10- IAIHU-06, project Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6), from Agence Nationale de la Recherche (reference ANR-23- CE17-0054, project LYMP-PD), from the European Union's Horizon Europe Framework Programme (grant number 101136607, project CLARA), grant from Inserm and the French Ministry of Health in the context of the MESSIDORE 2023 call operated by IReSP (reference AAP-2023-MSD-341011), and by the Fonds Recherche Neurosciences under the project "NEIMO - Monitoring Neuroinflammation: a Paris Brain Institute-Yale Neurology Collaborative program"; supported by the Chinese Government Scholarship provided by China Scholarship Council (CSC). **L.N.** Payment or honoraria from Servier; support for attending meetings from Olea Medical. **D.H.d.l.G.** No relevant relationships. **S.L.** No relevant relationships. **C.B.** No relevant relationships. **R.V.** No relevant relationships. **A.A.** ANR-JCJC (ANR-23-CE17-0027), ANR-18-RHUS-0012, PRT-K/INCa grant LOC-model reference 2017-1-RT-04, INCa-DGOS-Inserm_12560 of the SiRIC CURAMUS [grant number INCa-DGOS-Inserm_12560], Cancéropôle Ile-de-France 2024 (Emergence). **K.H.X.** Member of a safety monitoring board of the Sonokid phase I trial (unpaid). **B.M.** No relevant relationships. **C.S.** No relevant relationships. **J.P.M.** No relevant relationships. **J.P.** No relevant relationships. **L.T.** No relevant relationships. **P.A.** No relevant relationships. **A.S.** No relevant relationships. **O. Chinot** Honoraria from Servier; travel support from Servier and Carthera. **G.A.** Payment or honoraria from Pfizer (outside this work); grants from Novartis, Roche, and CSL Behring, (outside this work). **D.D.** No relevant relationships. **C.H.** No relevant relationships. **S.L.** 2024-2026 EBRAINS 2.0: A Research Infrastructure to Advance Neuroscience and Brain Health) project number 1011473191, PI Maurizio Corbetta partner; 2023-2027 ANRT (CE17) – LYMP-PD: Lymphatic system dysfunction in Parkinson's disease, Principal investigator; 2023-2025 SESAME NEURO@7T: Development and validation of innovative 7T biomarkers for neurological and psychiatric pathologies, principal investigator; 2023-2026 ANR - JPND REACT: Rebalancing sleep-wake disturbances in Parkinson's Disease with deep brain stimulation , principal Investigator France for France. **O. Colliot** French government under management of Agence Nationale de la Recherche as part of the "France 2030" program (reference ANR-23-IACL-0008, project PRAIRIE-PSAI), as part of the "Investissements d'avenir" program (reference ANR-19- P3IA-0001, project PRAIRIE 3IA Institute and reference ANR-10- IAIHU-06, project Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6), from Agence Nationale de la Recherche (reference ANR-23- CE17-0054, project LYMP-PD), from the European Union's Horizon Europe Framework Programme (grant number 101136607, project CLARA), was supported by a grant from Inserm and the French Ministry of Health in the context of the MESSIDORE 2023 call operated by IReSP (reference AAP-2023-MSDR-341011), and by the Fonds Recherche Neurosciences under the project "NEIMO - Monitoring Neuroinflammation: a Paris Brain Institute-Yale Neurology Collaborative program" (payments made to the institution); other principal investigators affiliated to the team which I co-leads have received grants (paid to the institution) from Sanofi and Biogen (payments made to the institution. No relationship with this work); consulting fees from Therapanacea (no relationship to this work); patent registered at the International Bureau of the World Intellectual Property Organization (PCT/IB2016/0526993, Allassonniere S, Colliot O, Durrleman S, Schiratti J-B, A method for determining the temporal progression of a biological phenomenon and associated methods and devices) (2017). I have submitted a declaration of invention to the European Patent Office (File Number: EP23306632, Ayache N, Colliot O, Hamzaoui M, Soulier T, Stankoff B, Devices for generation of synthetic 3D representation representative of myelin content, submitted to the European Patent Office on September 28th, 2023); spouse was employee of myBrain Technologies and is an employee of DiamPark.

References

1. Shiels MS, Pfeiffer RM, Besson C, et al. Trends in primary central nervous system lymphoma incidence and survival in the US. Br J Haematol 2016;174(3):417–424.
2. Houillier C, Soussain C, Ghesquières H, et al. Management and outcome of primary CNS lymphoma in the modern era: an LOC network study. Neurology 2020;94(10):e1027–e1039.
3. Ferreri AJM, Calimeri T, Cwynarski K, et al. Primary central nervous system lymphoma. Nat Rev Dis Primers 2023;9(1):29.
4. Hernández-Verdin I, Kirasic E, Wienand K, et al. Molecular and clinical diversity in primary central nervous system lymphoma. Ann Oncol 2023;34(2):186–199.
5. Hernández-Verdin I, Morales-Martínez A, Hoang-Xuan K, Alentorn A. Primary central nervous system lymphoma: advances in its pathogenesis, molecular markers and targeted therapies. Curr Opin Neurol 2022;35(6):779–786.
6. Choquet S, Soussain C, Azar N, et al. CAR T-cell therapy induces a high rate of prolonged remission in relapsed primary CNS lymphoma: real-life results of the LOC network. Am J Hematol 2024;99(7):1240–1249.
7. von Roemeling C, Ferreri AJM, Soussain C, Tun HW, Grommes C. Targets and treatments in primary CNS lymphoma. Leuk Lymphoma 2024;65(8):1055–1067.
8. Hoang-Xuan K, Deckert M, Ferreri AJM, et al. European Association of Neuro-oncology (EANO) guidelines for treatment of primary central nervous system lymphoma (PCNSL). Neuro Oncol 2023;25(1):37–53.
9. Barajas RF, Politi LS, Anzalone N, 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 Oncol 2021;23(7):1056–1071.
10. Ghaffari M, Sowmya A, Oliver R. Automated brain tumor segmentation using multimodal brain scans: a survey based on models submitted to the BraTS 2012–2018 challenges. IEEE Rev Biomed Eng 2020;13:156–168.
11. van Kempen EJ, Post M, Mannil M, et al. Performance of machine learning algorithms for glioma segmentation of brain MRI: a systematic literature review and meta-analysis. Eur Radiol 2021;31(12):9638–9653.

12. Zhang Y, Liang K, He J, et al. Deep learning with data enhancement for the differentiation of solitary and multiple cerebral glioblastoma, lymphoma, and tumefactive demyelinating lesion. Front Oncol 2021;11:665891.
13. Pennig L, Hoyer UCI, Goertz L, et al. Primary central nervous system lymphoma: clinical evaluation of automated segmentation on multiparametric MRI using deep learning. J Magn Reson Imaging 2021;53(1):259–268.
14. Kamnitsas K, Ferrante E, Parisot S, et al. DeepMedic for brain tumor segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Second International Workshop, BrainLes 2016, with the Challenges on BRATS, ISLES and mTOP 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 2. Springer, 2016; 138–149. doi:10.1007/978-3-319-55524-9_14.
15. Haldorsen IS, Kråkenes J, Krossnes BK, Mella O, Espeland A. CT and MR imaging features of primary central nervous system lymphoma in norway. AJNR Am J Neuroradiol 2009;30(4):744–751.
16. Xing Z, You RX, Li J, Liu Y, Cao DR. Differentiation of primary central nervous system lymphomas from high-grade gliomas by rCBV and percentage of signal intensity recovery derived from dynamic susceptibility-weighted contrast-enhanced perfusion MR imaging. Clin Neuroradiol 2014;24(4):329–336.
17. 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(2):203–211.
18. Li X, Morgan PS, Ashburner J, Smith J, Rorden C. The first step for neuroimaging data analysis: Dicom to nifti conversion. J Neurosci Methods 2016;264:47–56.
19. Pérez-García F, Sparks R, Ourselin S. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput Methods Programs Biomed 2021;208:106236.
20. Antonelli M, Reinke A, Bakas S, et al. The medical segmentation decathlon. Nat Commun 2022;13(1):4128.
21. Maier-Hein L, Reinke A, Godau P, et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods 2024;21(2):195–212.
22. Zhang Y, Zou D, Yin J, et al. Changes in cerebrospinal fluid interleukin-10 levels display better performance in predicting disease relapse than conventional magnetic resonance imaging in primary central nervous system lymphoma. BMC Cancer 2021;21(1):183.
23. Herzi D, Le Garff-Tavernier M, Sourdeau E, et al. Prognostic value of CSF IL-10 at early assessment of induction chemotherapy in primary CNS lymphomas: A LOC network study. Neurology 2024;102(12):e209527.
24. van der Meulen M, Postma AA, Smits M, et al. Extent of radiological response does not reflect survival in primary central nervous system lymphoma. Neurooncol Adv 2021;3(1):vdab007.
25. Lauer EM, Riegler E, Mutter JA, et al. Improved early outcome prediction by MRI-based 3d tumor volume assessment in patients with cns lymphomas. Neuro Oncol 2024;26(2):374–386.
26. Simpson AL, Antonelli M, Bakas S, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv 2019. Preprint posted online February 25, 2019; doi:10.48550/arxiv.1902.09063.