Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Expert-centered Evaluation of Deep Learning Algorithms for Brain Tumor Segmentation

Katharina V. Hoebel, MD, PhD • Christopher P. Bridge, PhD • Sara Ahmed, MD • Oluwatosin Akintola, MD • Caroline Chung, MD • Raymond Y. Huang, MD, PhD • Jason M. Johnson, MD • Albert Kim, MD • K. Ina Ly, MD • Ken Chang, MD, PhD • Jay Patel, PhD • Marco Pinho, MD • Tracy T. Batchelor, MD, MPH • Bruce R. Rosen, MD, PhD • Elizabeth R. Gerstner, MD • Jayashree Kalpathy-Cramer, PhD

From the Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology (K.V.H., C.P.B., A.K., K.I.L., K.C., J.P., B.R.R., E.R.G., J.K.C.), and Stephen E. and Catherine Pappas Center for Neuro-Oncology (O.A., A.K., K.I.L., E.R.G.), Massachusetts General Hospital, 149 13th St, Charlestown, MA 02129; Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Mass (K.V.H., K.C., J.P.); MGH and BWH Center for Clinical Data Science, Boston, Mass (C.P.B., J.K.C.); Department of Radiation Oncology, Division of Radiation Oncology (S.A., C.C.); Department of Diagnostic Radiology, Division of Diagnostic Imaging (C.C.), and Department of Neuroradiology (J.M.J.), Division of Diagnostic Imaging, The University of Texas MD Anderson Cancer Center, Houston, Tex; Departments of Radiology (R.Y.H.) and Neurology (T.T.B.), Brigham and Women's Hospital, Boston, Mass; Department of Radiology and Advanced Imaging Research Center, University of Texas Southwestern Medical Center, Dallas, Tex (M.P.); and Department of Ophthalmology, University of Colorado Anschutz Medical Campus, Aurora, Colo (J.K.C.). Received October 31, 2022; revision requested January 17, 2023; revision received September 13; accepted November 1. **Address correspondence to** J.K.C (email: *jayashree.kalpathy-cramer@cuanschutz.edu*).

K.V.H. and J.K.C. supported by the National Institutes of Health (grant no. U01CA242879). E.R.G. supported by the National Institutes of Health (grant nos. R01CA129371 and K23CA169021). This research was carried out in whole or in part at the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital using resources provided by the Center for Functional Neuroimaging Technologies and a P41 Biotechnology Resource Grant (grant no. P41EB015896) supported by the National Institute of Biomedical Imaging and Bioengineering.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2024; 6(1):e220231 • https://doi.org/10.1148/ryai.220231 • Content codes: AI OI NR

Purpose: To present results from a literature survey on practices in deep learning segmentation algorithm evaluation and perform a study on expert quality perception of brain tumor segmentation.

Materials and Methods: A total of 180 articles reporting on brain tumor segmentation algorithms were surveyed for the reported quality evaluation. Additionally, ratings of segmentation quality on a four-point scale were collected from medical professionals for 60 brain tumor segmentation cases.

Results: Of the surveyed articles, Dice score, sensitivity, and Hausdorff distance were the most popular metrics to report segmentation performance. Notably, only 2.8% of the articles included clinical experts' evaluation of segmentation quality. The experimental results revealed a low interrater agreement (Krippendorff α, 0.34) in experts' segmentation quality perception. Furthermore, the correlations between the ratings and commonly used quantitative quality metrics were low (Kendall tau between Dice score and mean rating, 0.23; Kendall tau between Hausdorff distance and mean rating, 0.51), with large variability among the experts.

Conclusion: The results demonstrate that quality ratings are prone to variability due to the ambiguity of tumor boundaries and individual perceptual differences, and existing metrics do not capture the clinical perception of segmentation quality.

Clinical trial registration nos. NCT 00756106 and NCT 00662506

Supplemental material is available for this article.

The segmentation of organs and pathologic conditions in medical imaging is a required step in many clinical and research pipelines (eg, for radiation treatment planning and treatment response assessment). Over the last decade, advances in computer vision—enabled by deep learning (DL)—have revolutionized automated segmentation for medical use cases. Particularly for the segmentation of complex target structures such as tumors, DL promises to remedy two challenges. First, the manual delineation of such complicated and often ambiguous structures is time-consuming (1). Second, tumor segmentations suffer from a high interrater variability due to the absence of clearly defined boundaries (2–4). DL-assisted segmentation can substantially decrease the time requirement and variability among individual physicians (5,6).

Multiple factors, including technical difficulties in integrating DL models into clinical workflows, regulatory uncertainties, and ethical concerns such as a lack of trust, are inhibiting the widespread adoption of these algorithms in clinical settings (7,8). This distrust stems not only from the algorithms' black-box nature but also from their propensity to fail without notice due to often unknown reasons, such as imaging artifacts and domain shift (9–11). Therefore, DL algorithms are generally intended as a decision-support tool rather than autonomous models that make decisions of clinical importance without human oversight. Instead of spending long hours segmenting highly challenging target structures, clinical experts would in the future be tasked to determine whether the output of a segmentation model requires further manual editing or is of sufficient quality to be used (eg, for treatment planning).

Abbreviations

DL = deep learning, FLAIR = fluid-attenuated inversion recovery

Summary

Current practices for the evaluation of deep learning models for brain tumor segmentation did not reflect the segmentation quality perception of clinical experts.

Key Points

- Only 2.8% (five of 180 articles) of the evaluated literature on deep learning models for brain tumor segmentation included an evaluation of segmentation quality performed by clinical experts.
- The metric most commonly used to evaluate segmentation performance, the Dice score, showed a poor correlation with the segmentation quality perception of clinical experts (Kendall tau, 0.23) in the experimental study.
- Segmentation quality ratings performed by clinical experts were prone to high interrater variability (Krippendorff α, 0.34).

Keywords

Brain Tumor Segmentation, Deep Learning Algorithms, Glioblastoma, Cancer, Machine Learning

However, humans have a more contextual response to segmentation quality, leading to a disagreement between human perception and the metrics used to optimize and evaluate segmentation quality, particularly the Dice score (12,13). Additionally, similar to the substantial interrater variability in manual tumor delineations, the perception of what constitutes a good segmentation differs among human experts, which is reflected in high disagreements in peer review for radiation therapy planning (14). Consequently, using DL algorithms as an effective clinical decision support system will require a detailed understanding of the interaction between humans and algorithms.

We hypothesized that the quality perception of experts would not correlate well with the popular segmentation quality metrics from the literature and that the agreement between the quality ratings of individual experts would be low. In this study on expert-centered evaluation of DL segmentation models, we first studied how DL segmentation models are currently being evaluated. We reviewed studies on DL segmentation models for brain tumors and assessed which quantitative metrics are used and whether clinical experts were involved in evaluating the algorithms' performance. Second, we performed an experimental study to assess *(a)* interrater variability in segmentation quality perception and *(b)* variability in the agreement between quantitative metrics and experts' quality perception for segmentations of postoperative glioblastomas.

Materials and Methods

Literature Review

We searched PubMed for English-language original research articles published between August 2017 and September 2022 reporting on DL segmentation models for macroscopic brain tumors using the following search terms: "('deep learning' OR 'neural network') AND 'segmentation' AND 'brain tumor,'" explicitly excluding articles categorized as reviews or surveys. Only articles that described a segmentation algorithm for human brain tumors on macroscopic images and contained a performance evaluation of the segmentations were included in the analysis. We recorded the metrics used in the performance evaluation of the segmentation models. We also categorized the quantitative metrics into seven groups such that metrics that measure similar concepts were grouped together. For example, both the Dice score and Jaccard index are computed based on overlap between the ground truth and the predicted segmentation and were therefore categorized as overlap-based metrics. The seven groups were *(a)* overlap-based metrics, such as Dice score; *(b)* volume-based metrics, such as relative volume error; *(c)* voxel-level confusion matrix–derived metrics, such as sensitivity; *(d)* distance-based metrics, such as Hausdorff distance; *(e)* threshold metrics, such as area under the receiver operating characteristic curve; *(f)* information-based metrics, such as variation of information; and *(g)* boundary-based metrics, such as the boundary F1 score. Appendix S1 contains a complete list of all metrics in each group.

Experimental Study

Dataset.— We performed a secondary analysis of imaging data from two clinical trials (ClinicalTrials.gov identifiers, NCT 00756106 and NCT 00662506). The analysis was approved by the institutional review board with a waiver for written consent. The dataset contains a total of 713 postoperative MRI visits from 54 individuals newly diagnosed with glioblastoma (33 male participants, 21 female participants; mean age, 57 years) (15). All individuals had undergone tumor biopsy or partial tumor resection with a remaining contrast-enhancing tumor of at least 1 cm in diameter at the time of enrollment. Two experts, one neuro-oncologist with 12 years of experience (E.R.G.) and one neuroradiologist with 11 years of experience (M.P.), performed manual ground truth segmentations of areas of T2-weighted fluid-attenuated inversion recovery (FLAIR) abnormality corresponding to the total tumor burden. Each case was annotated by one expert, resulting in a single set of manual segmentations for the full dataset. These segmentations were used for the development and evaluation of the segmentation algorithm. A third expert from a different institution trained in neuroradiology with 5 years of experience provided manual annotations for a subset of the dataset. These segmentations were used only to provide an estimate of interrater variability. All experts were blinded to participant identity, the order of scans, and participant treatment status. See Appendix S2 for more information.

Data preprocessing.— Since the dataset originates from longitudinal clinical studies, it contained images from multiple study visits for each participant. We split the dataset into training, validation, and test subsets at the participant level such that all available images of a respective participant were part of only one subset. The training, validation, and test datasets consisted of images from 34 (study visits, 464), nine (study visits, 128), and 11 (study visits, 119) participants, respectively. We used T1-weighted pre- and postcontrast sequences and T2-weighted

FLAIR sequences as the three input channels for model development. All images were registered to the T2-weighted FLAIR image of the respective study visit. Preprocessing consisted of brain extraction, N4 bias correction, and *z* score normalization of the brain region of each scan (16,17).

Segmentation model.— Monte Carlo dropout networks approximate Bayesian neural networks by dropping out full activation maps after each convolutional layer at training and inference time (18). At inference time, slightly varying segmentations can be sampled from an approximate posterior distribution and used to quantify model uncertainty.

We trained a Monte Carlo dropout three-dimensional U-Net, with a dropout probability of 0.2 for each activation map, on patches of size 64 × 64 × 16 voxels to segment areas of T2-weighted FLAIR abnormality using weighted cross-entropy loss. The simultaneous truth and performance level estimation, or STAPLE, algorithm created the final segmentation from 10 Monte Carlo dropout samples (19). The segmentation model was implemented in DeepNeuro (20). Details on model development, hyperparameters, and the calculation of model uncertainty are provided in Appendix S3. Quantitative segmentation quality metrics were computed using the Python (Python Software Foundation) package, pymia (21).

Expert ratings of segmentation quality.— Eight experts (fellows and attending physicians from the neuro-oncology, neuroradiology, and radiation oncology departments at three academic centers in the United States) graded the quality of segmentations that the DL model generated. We split 60 randomly selected cases from the test set into two datasets with 30 cases each and assigned four experts to each set. Experts were stratified by specialty and experience (see Appendix S4 for details on each experts' specialty and level of experience).

The experts viewed the automatically generated tumor outline overlayed on the T2-weighted FLAIR image as a stack of axial two-dimensional sections. They then graded the quality of each segmentation with a score between 1 and 4. The rating categories were as follows: 1, not acceptable; 2, acceptable with moderate changes; 3, acceptable with minor changes; and 4, acceptable without changes. All experts were blinded toward participant identity and treatment status and viewed the cases in a randomized order. Each expert was provided 15 cases to practice, and six experts rated four cases twice during different sessions to assess intrarater variability. Appendix S4 contains a detailed description of the study setup and example cases.

Statistical analysis.— Agreement between ratings was assessed using Krippendorff α (type ordinal) for more than two ratings per case and Gwet AC2 with linear weights for two ratings per case for pairwise expert comparisons. We chose Gwet AC2 instead of Cohen κ for pairwise comparisons and the assessment of intrarater reliability because Gwet AC is not affected by the frequency distribution of the ratings (22). Kendall rank correlation coefficient (Kendall tau) was used to measure the correlations between the continuous quantitative segmentation quality metrics and ordinal quality ratings (23). Since the number of ordinal categories was relatively low, Kendall tau is preferable to similar measures like the Spearman rank correlation coefficient (24).

To determine whether specific factors were associated with high or low agreement between the experts, we separated all cases into two groups according to the difference between their lowest and highest rating: a high agreement group (rating difference, ≤1) consisting of 40 cases and a low agreement group (rating difference, ≥2) consisting of 20 cases. Comparison between the distributions of continuously valued features was determined using the Kruskal-Wallis test.

Statistical significance was defined as a *P* value less than .05. Data analysis was performed in Python (version 3.6) and R (version 4.0.2; R Foundation for Statistical Computing).

Results

Literature Review

We searched PubMed full-text articles that covered DL segmentation algorithms of brain tumors. The search identified 248 articles. We excluded 53 articles after screening titles and abstracts, and 15 more were excluded after a review of the full-text articles. Thus, 180 studies were included in the final analysis. Figure 1 illustrates the literature selection process.

Quantitative evaluation of segmentation model performance.— Among the 180 articles reviewed, the three most popular strategies to evaluate segmentation quality were the Dice score as the only evaluation metric (42 of 180 articles; 23.3%); a combination of Dice score, Hausdorff distance, sensitivity, and specificity (26 of 180 articles; 14.4%); or a combination of Dice score and Hausdorff distance (21 of 180 articles; 11.7%). Overall, the Dice score was used in 170 of the 180 articles (94.4%) either exclusively or in combination with other metrics. Sensitivity and Hausdorff distance were used in combination with other metrics in 86 (47.8%) and 69 (38.3%) of the 180 articles, respectively (Fig 2A).

The most widely used metric groups were overlap-, confusion matrix–, and distance-based metrics (Fig 2B). From the 180 articles, 28.3% (51 of 180) used metrics from only one metric group, 38.3% (69 of 180) used two metric groups, and 31.1% (56 of 180) used three metric groups (Fig 2C). We found that overlap-based metrics were most frequently combined with confusion matrix–based metrics. Distance-based metrics were always associated with overlap-based metrics, as illustrated in Figure 2D. Forty-two of 180 studies (23.3%) combined metrics from all three most common metric groups.

Segmentation model performance evaluation by clinical experts.— In addition to the purely quantitative evaluation of segmentation performance described above, five of the 180 studies (2.8%) included an assessment by clinical experts (Table). The diverse evaluation approaches involved measuring the time it took clinical experts to manually correct an automatically generated segmentation, assessing the consensus between automatic segmentations edited by experts, and rating the segmentation quality. While a few articles included quali-

tative evaluation performed by clinical experts, none linked quantitative and qualitative segmentation quality measurements.

User Study on Segmentation Quality Perception of Clinical Experts

Our brain tumor segmentation model achieved a mean Dice score of 0.72 on the held-out test dataset. The previously reported Dice score for this dataset was 0.70 (25), and the average Dice score for the interreader agreement was 0.64 (Fig S2). The sensitivity, relative volume error, and 95th percentile Hausdorff distance were 77%, 0.33, and 6.5 mm, respectively. We obtained 264 segmentation quality ratings for 60 cases, including 24 double reads from six domain experts, to determine intrarater variability. Given the longitudinal nature of the imaging, segmentation quality measurements (ie, quantitative segmentation metrics or ratings) from repeated imaging of an individual cannot be assumed to be independent. While the intraclass correlation coefficient of the Dice score of segmentations for repeated MRI from the same individual was high (0.82), segmentation quality ratings were not influenced by this relationship (intraclass correlation coefficient, 0.29).

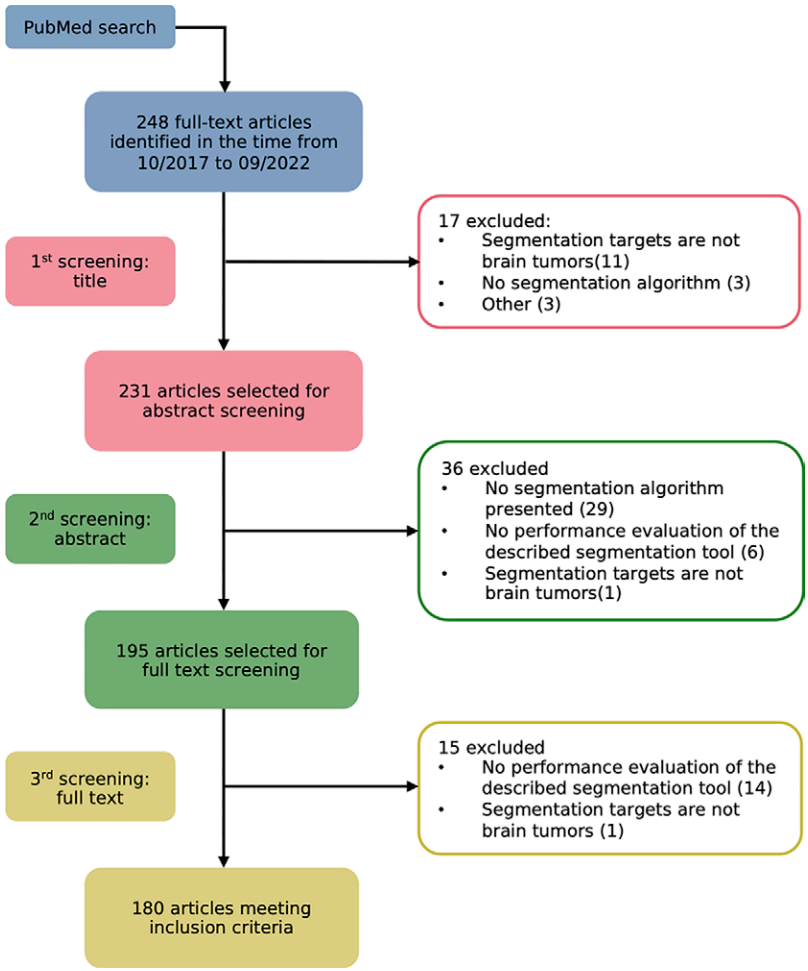

Figure 1: Consort diagram of the literature review process. The process to identify suitable articles for review consisted of an initial screening of PubMed (blue), a first screening of the titles (red), a second screening of the abstracts (green), and final screening of the full texts (yellow).

Differences among experts.— The intrarater agreement based on double reads ranged from 0.24 to 1 with a median of 0.88 (Gwet AC2). The interrater agreement was low among all quality ratings with a Krippendorff α value of .34. For pairwise comparisons (panels A1 and A2 of Fig 3), we found that the agreement between the ratings of individual experts showed wide variability, ranging between 0.37 and 0.79 (median, 0.59; Gwet AC2).

After sorting the cases in each subset according to their mean rating, it became apparent that the low agreement between raters was possibly caused by different internal quality class thresholds (panels A1 and A2 in Fig 4). The threshold between what constitutes an acceptable segmentation and one that requires minor changes varied between the experts.

However, as indicated by the variability in the pairwise correlations between experts' ratings (panels B1 and B2 in Fig 3), these individual thresholds did not account for the total variability we observed in the quality ratings. Therefore, we assessed whether there were additional factors that influenced these differences. A qualitative analysis comparing cases with a low and high disagreement between experts revealed that the ratings disagreed for cases with diffuse tumor boundaries, heterogeneous tumor intensity, and multiple lesions. We found high agreement in the quality ratings for cases with clear tumor boundaries and false-positive segmented areas that were clearly not connected to the primary lesions (see Figs S3 and S4).

Upon quantitative comparison between cases with a low and high disagreement between the experts, the following factors showed statistically significant associations with a lower agreement between experts: higher segmentation volume of the automatic segmentation (P = .04), lower Dice score (P < .001), higher 95th percentile Hausdorff distance (P = .046) between the automatic and the manual ground truth segmentation, and a higher segmentation uncertainty of the segmentation model (P < .001). The different distributions of these metrics in the low and high agreement groups are illustrated in Figure 4B. We found no evidence of differences between the low and high agreement groups' surface area (P = .12), sphericity (P = .62), and volume similarity (P = .51).

Differences between commonly used metrics and expert quality perception.— Last, we compared the correlation between the ratings of each expert and the most used quantitative segmentation quality metrics. Figure 5 presents the Kendall tau correlation coefficients between seven selected segmentation metrics and the ratings of all eight raters. Overall, we observed a high variability between the metrics and among the experts.

The highest correlations and lowest variability among raters were found with the 95th percentile Hausdorff distance.

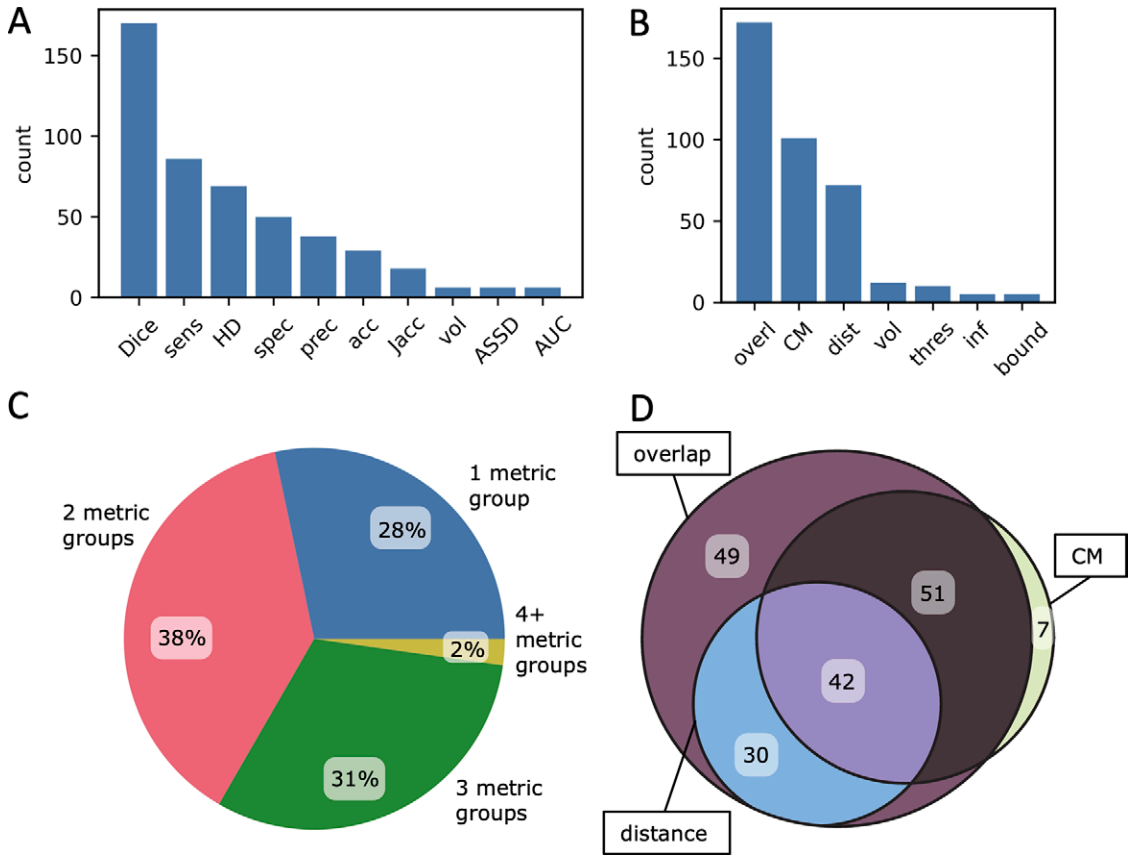

Figure 2: Use of quantitative segmentation quality metrics in the reviewed literature. **(A)** Graph shows the count of how often the 10 most popular segmentation quality metrics were used to evaluate the performance of the segmentation models. **(B)** Graph shows the count of how often metrics belonging to one of the seven defined metric groups were used to evaluate the performance of the segmentation models. **(C)** Pie chart shows the percentage of studies that used metrics from one, two, three, or four or more metric groups. **(D)** Venn diagram illustrates the frequency of metric group combinations for segmentation model evaluation between the three most popular groups of segmentation quality metrics. acc = accuracy, ASSD = average symmetric surface distance, AUC = area under the receiver operating characteristic curve, bound = boundary-based metrics, CM = confusion matrix–based metrics, dist = distance-based metrics, inf = information-based metrics, HD = Hausdorff distance, Jacc = Jaccard index, overl = overlap-based metrics, prec = precision, sens = sensitivity, spec = specificity, thres = threshold-based metrics, vol = volume-based metrics.

Studies including Segmentation Model Performance Evaluation by Quantitative Metrics and Clinical Experts

Study	Segmentation Target	Expert Evaluation Metric	No. of Experts	Expert Background (No. of Experts)
Lu et al, 2021 (34)	Brain metastases, meningioma, vestibular schwannoma	Time to correct Agreement between corrected segmentations	8	Neuroradiology (n = 1) Radiation oncology (n = 5) Neurosurgery (n = 2)
Conte et al, 2021 (35)	Preoperative glioma	Time to correct	2	Neuroradiology (n = 2)
Di Ieva et al, 2021 (36)	Preoperative glioma	Binary quality classification (acceptable or not acceptable)	4	Neuroradiology (n = 1) Radiation oncology (n = 1) Neurosurgery (n = 2)
Mitchell et al, 2020 (37)	Preoperative glioma	Comparison between manual ground truth and automatic segmentation Segmentation quality on a scale from 0 (poor) to 10 (perfect)	20	Neuroradiology (n = 20)
Wang et al, 2018 (38)	Preoperative glioma, fetal organs	User interaction time to refine the segmentation	2	Radiology (n = 1) Obstetrics (n = 1)

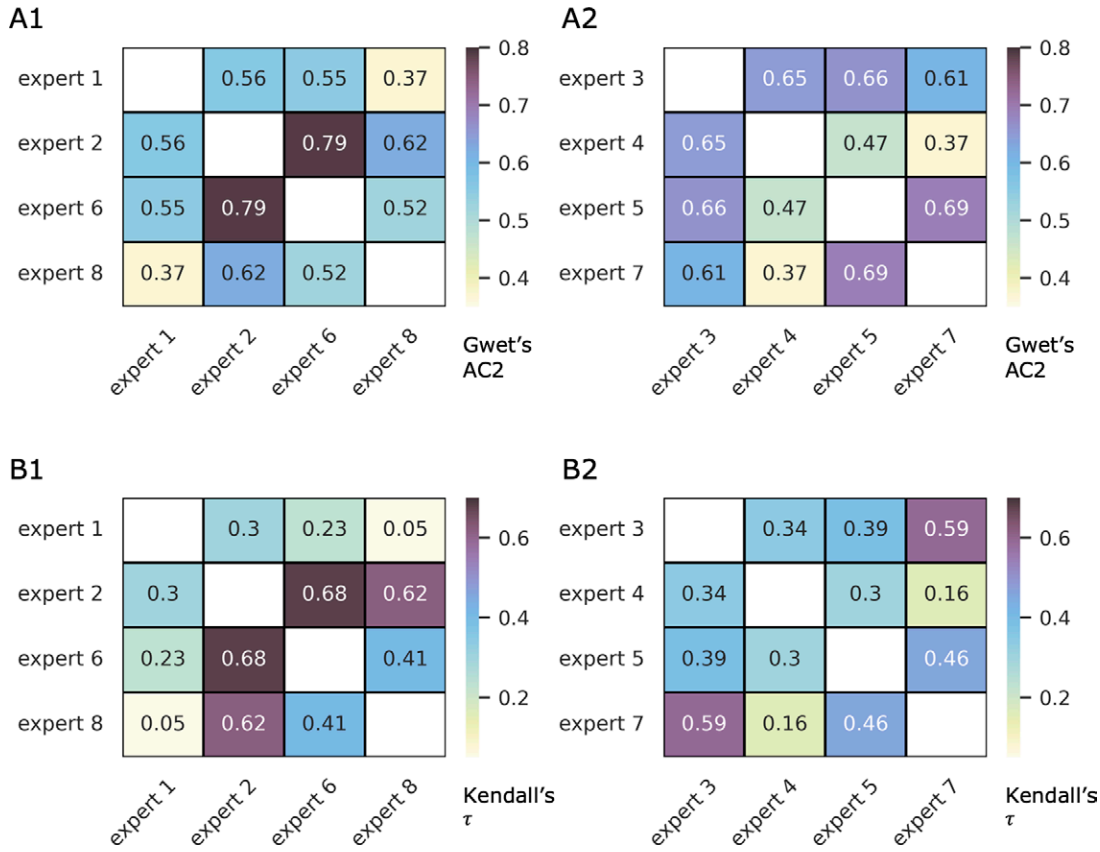

Figure 3: Pairwise agreement and correlation between experts. Heat maps show pairwise agreement and correlation that were computed between all pairs of experts on the two sets of cases. Each panel (A1 and A2 and B1 and B2) represents one set of 30 cases that were rated by the same group of experts without overlap between the sets. A1 and A2 maps show pairwise agreement between experts' ratings using Gwet AC2, and B1 and B2 maps show pairwise correlation between experts' ratings using Kendall tau.

The sensitivity and surface Dice score (26) showed acceptable correlation values with limited variability among the raters. However, similarly to most other quantitative metrics, we found one outlier in the correlations between experts' ratings and the surface Dice score; the ratings provided by expert 4 showed no correlation (Kendall tau, -0.01). The most popular overlap-based metric, the Dice score, showed a surprisingly low correlation with experts' ratings with a median Kendall tau of 0.185. We found no meaningful relationship between volume similarity and relative volume error with median Kendall tau values of 0.11 (range, -0.22–0.3) and -0.1 (range, -0.22–0.24), respectively. Similarly, low correlations were found for specificity with a median Kendall tau value of 0.11 (range, -0.19–0.21). However, due to the high background-to-foreground ratio, specificity values, which convey little information about segmentation quality, were very high.

Discussion

We present a study on expert-centered segmentation quality evaluation. First, we identified how articles reporting on DL brain tumor segmentation evaluate those models. Second, we performed an experimental study on clinical experts' perception of brain tumor segmentation and showed how current evaluation practices identified in the literature review relate to clinical experts' segmentation quality assessment.

The Dice score was the most popular metric, and 23.3% of the analyzed articles relied on the Dice score as the only metric for segmentation performance evaluation. However, for segmentation of preoperative gliomas, overlap metrics correlate poorly with the human quality perception (12). The experimental data on the quality perception of postoperative brain tumor segmentation confirm this finding, indicating that certain metrics exhibit better agreement with expert quality perception while also revealing differences in the agreement between experts.

Distance-based metrics that showed the highest agreement and lowest variability among experts were never used as the primary performance metric for segmentation quality. Unlike overlap- and confusion matrix–based metrics, most distance metrics, such as the Hausdorff distance, are not bounded between 0 and 1. Therefore, they are harder to interpret and compare between studies that are using different datasets. Our findings suggest that the surface Dice score (26) can be a promising alternative, as it had a higher agreement with experts' ratings than the Dice score and is constrained to values between 0 and 1.

Even though DL-assisted segmentation algorithms can increase interrater agreement of segmentations and other downstream measurements, like time to progression (5,27), the interrater reliability in segmentation quality perception has not been studied to date. Given the known low interrater agreement in the manual segmentations for challenging targets such as tumors

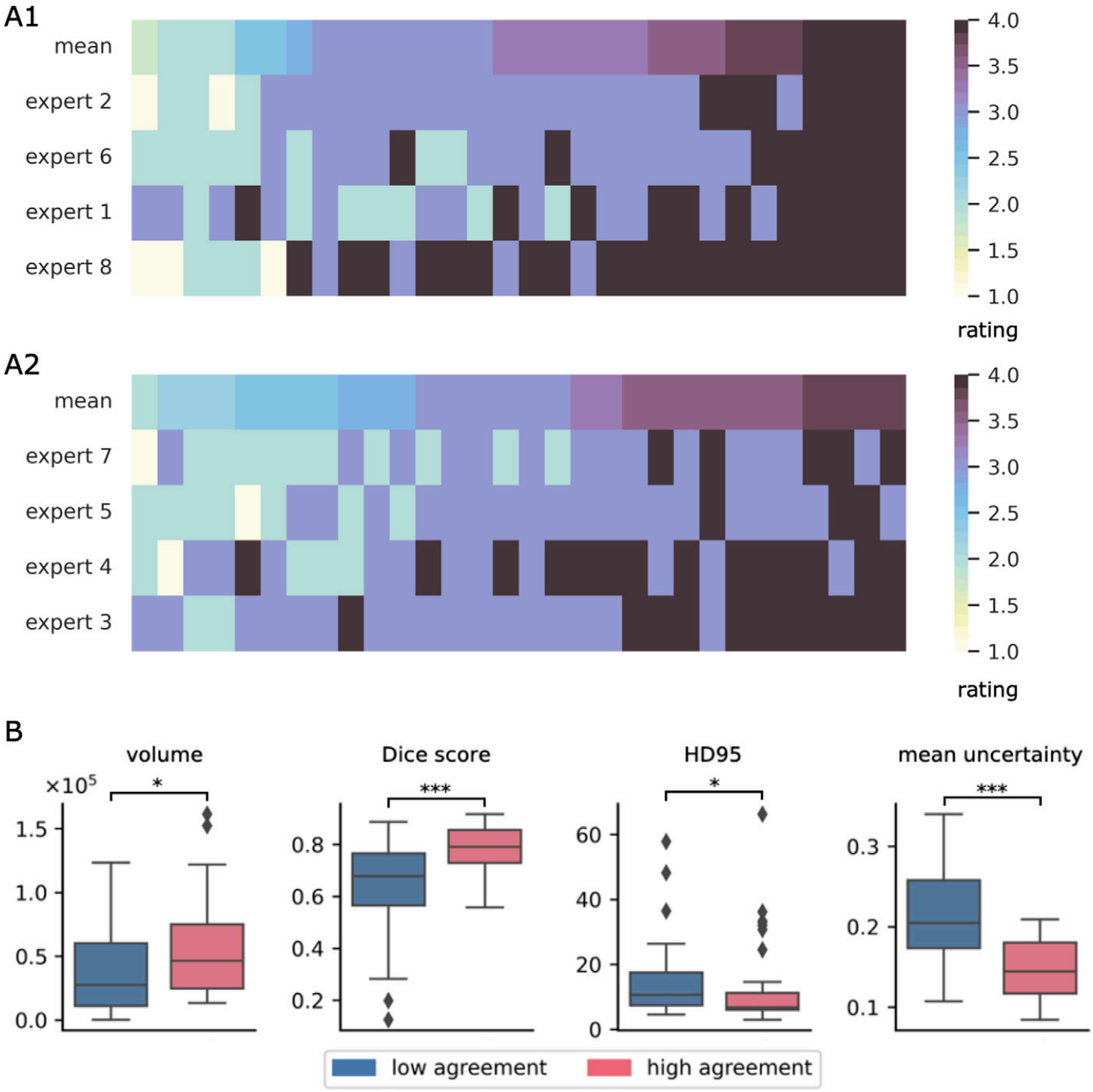

Figure 4: Factors influencing disagreement between experts. **(A)** Heat maps show agreement between raters for single cases. Each column represents the ratings for one case, and each row represents one expert. The cases within both subsets of data were ranked based on their average rating. Experts were ordered based on the average rating assigned to all cases from lowest (top) to highest (bottom) average rating. Each panel (A1 and A2) represents one set of 30 cases that were rated by the same group of experts without overlap. **(B)** Box and whisker plots show distributions of rater-independent metrics for the group with low (blue) and high agreement between experts (red). Statistically significant differences between the distributions are indicated by brackets above the box plots. Boxes represent the IQR (25th–75th percentile), and the horizontal line inside the boxes represents the median value of each parameter. Whiskers represent the minimum and maximum values. ♦ = outliers. * = $P \leq .05$, ** = $P \leq .01$, *** = $P \leq .001$, HD95 = 95th percentile Hausdorff distance.

(2,3,4), we expected variability in the segmentation quality perception of the participating experts.

The intrarater reliability in our experiments was high, with a median of 0.88 (Gwet AC2) for the six experts. This finding contrasts the low interrater agreement between all raters (Krippendorff α, .39). In part, individually varying thresholds between adjacent segmentation quality grades can account for the observed variability. Similar variability in individual thresholds between ordinal rating categories has been observed in disease severity classification (28,29). In contrast, experts are more consistent in their assessment of disease severity when comparing images rather than assigning absolute ratings (30). Therefore, alternative evaluation techniques based on comparisons between segmentations and a defined segmentation quality standard may warrant higher agreement between raters.

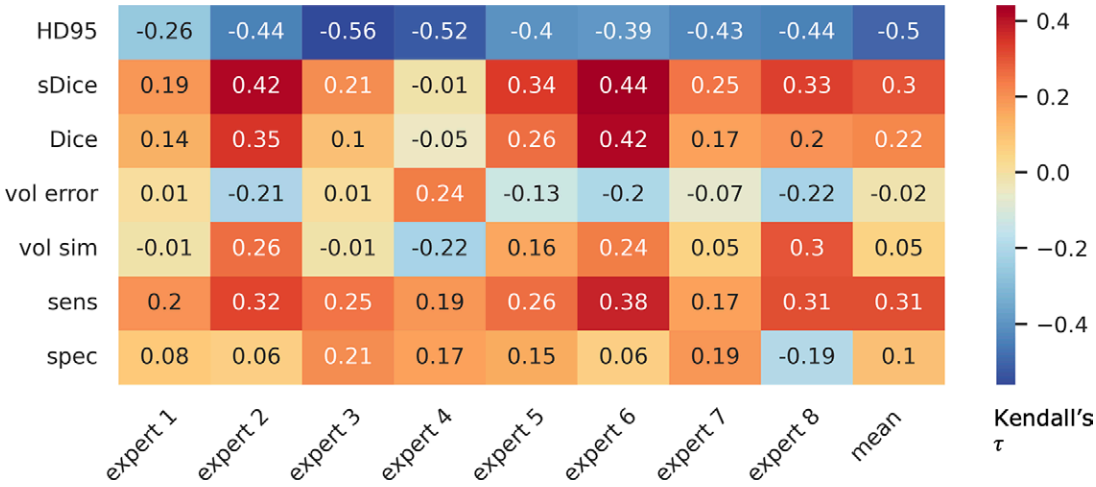

Figure 5: Heat map shows correlation between segmentation metrics and ratings. Kendall tau correlation coefficient between the ratings provided by each expert and selected segmentation quality metrics: 95th percentile Hausdorff distance (HD95th), surface Dice score (sDice), Dice score, relative volume error (vol error), volume similarity (vol sim), sensitivity (sens), and specificity (spec).

Furthermore, we identified several factors significantly associated with a lower agreement between raters. Among these factors were smaller volumes of the automated segmentation, indicating that the experts saw aberrations in these cases as of different importance. Additionally, cases with a high disagreement in the ratings were associated with higher model uncertainty. Epistemic uncertainty is expected to be higher for cases dissimilar to those seen during model training (31). Model uncertainty, which can be computed independently of any manual ground truth, is highly correlated with the segmentation Dice score (25,32). Our finding that epistemic uncertainty is higher for samples with a lower agreement between raters indicates that some cases may be more challenging for experts and DL algorithms alike. Consequently, cases with a high uncertainty could be automatically routed for review by multiple experts (eg, during a peer review session).

In the surveyed articles, expert-centered evaluation, if performed, played a supplementary role to a primary quantitative assessment of the test data using established quantitative metrics. Furthermore, we observed a high variability in the evaluation processes that may reflect the diversity of use cases for segmentation (eg, treatment monitoring or planning). Solely relying on quantitative evaluation is insufficient to assess the readiness of DL segmentation models for downstream clinical tasks (33). As a solution, Jha et al (32) advocated that artificial intelligence algorithms should be evaluated with the involvement of clinical experts and their clinical context in mind.

Most research on brain tumor segmentation is performed using the imaging of patients who have not undergone surgery. In the representative sample of the literature on DL-based brain tumor segmentation, only 1.7% (three of 180) of articles explicitly state the inclusion of postoperative imaging.

While segmentation of the preoperative tumor has value for neurosurgical purposes, neuro-oncologic response assessment and the planning of radiation therapy are based on imaging following surgery or biopsy. However, due to the presence of treatment-related changes in addition to the naturally blurry tumor outlines, the manual and automatic segmentation of postoperative gliomas is more challenging than the segmentation on preoperative images. For this reason, we have chosen to focus on assessing the quality of postoperative tumor segmentations. However, our findings may not generalize to preoperative segmentations. Furthermore, it would be valuable to study whether experts' perception of segmentation quality for the segmentation of other structures, such as stroke lesions or lung tumors, shows similar variability as described here.

There were some limitations to this study. The assessment of T2-weighted FLAIR abnormality segmentations was based only on T2-weighted FLAIR images and no additional sequences. In clinical settings, radiologists may use other sequences such as T2-weighted and T1-weighted sequences with and without contrast agent administration for additional information. Therefore, the simplified study setup did not fully mimic segmentation quality assessment as it would be performed in a clinical workflow. Furthermore, the experiments were limited to the segmentation quality perception of postoperative brain tumors on T2-weighted FLAIR images, a highly complex and ambiguous segmentation target. Our findings may not generalize to other segmentation targets. Future studies should evaluate whether the observed disagreement between segmentation quality metrics and the quality perception of experts can be observed for other segmentation targets as well. Based on our findings, we suggest that the performance of segmentation models should include a use-case–focused assessment performed by clinical experts. If this is not feasible, a purely quantitative analysis should use selected segmentation quality metrics that correlate with their usefulness for the desired clinical application.

In conclusion, a better understanding of the requirement for high-quality segmentation and the quality perception of clinical experts is required. This knowledge will catalyze the development of tailored quantitative metrics to develop clinically helpful segmentation models.

Author contributions: Guarantors of integrity of entire study, **K.V.H.**, **J.K.C.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual con-

tent, 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, **K.V.H.**, **C.C.**, **R.Y.H.**, **K.C.**; clinical studies, **S.A.**, **O.A.**, **R.Y.H.**, **A.K.**, **K.I.L.**, **M.P.**, **T.T.B.**, **E.R.G.**; experimental studies, **K.V.H.**, **O.A.**, **C.C.**, **R.Y.H.**, **J.M.J.**, **K.I.L.**, **K.C.**, **M.P.**, **J.K.C.**; statistical analysis, **K.V.H.**, **J.K.C.**; and manuscript editing, **K.V.H.**, **C.P.B.**, **S.A.**, **C.C.**, **R.Y.H.**, **J.M.J.**, **K.C.**, **J.P.**, **M.P.**, **B.R.R.**, **E.R.G.**, **J.K.C.**

Data sharing: Data generated or analyzed during the study are available from the corresponding author by request.

Disclosures of conflicts of interest: **K.V.H.** Member of the trainee editorial board for *Radiology: Artificial Intelligence*. **C.P.B.** Support from the Rappaport Research Fellowship in radiology; grants from the National Cancer Institute (HHSN261200800001E; Integrative Cancer Imaging Data Commons), the National Institutes of Health (5R01EB023281-07; Freesurfer Maintenance, Development, and Hardening), the European Union (H2020-SC1- FA-DTS-2019-1; Procancer-I Consortium), and the Nvidia Corporation through the Mass General Brigham Data Science Office; patent application for CTA Large Vessel Occlusion Model (17/083,761), Computed Tomography Medical Imaging Intracranial Hemorrhage Mode (16/587,828), and Medical Imaging Stroke Model (16/588,080). **S.A.** No relevant relationships. **O.A.** No relevant relationships. **C.C.** No relevant relationships. **R.Y.H.** Consulting fees from Nuvation Bio; advisor for Vysioneer. **J.M.J.** Grant from Blue Earth Diagnostics to institution; consulting fees from Bioclinica, Kura Oncology, and InformAI. **A.K.** No relevant relationships. **K.I.L.** Grant support to author from Neurofibromatosis Therapeutic Acceleration Program and the Department of Defense; unpaid consultant to SpringWorks Therapeutics. **K.C.** Member of the trainee editorial board for *Radiology: Artificial Intelligence*. **J.P.** No relevant relationships. **M.P.** No relevant relationships. **T.T.B.** Grants from the National Cancer Institute and ONO Pharmaceutical; royalties from UpToDate, Oxford University Press, the American Cancer Society, and Springer Publishing; payment from Oakstone Publishing, the Chinese Association of Hematology and Oncology of North America; expert consultant in glioblastoma case: travel support for the Japan Society for Neuro-Oncology meeting, the World Congress of Neurology meeting, Society for Neuro-Oncology meetings, International Brain Tumor Alliance meeting, Society for Neuro-Oncology for Sub-Saharan Africa, and the European Association for Neuro-Oncology; advisory board member for DSMB for CODEL Study (Europe); president of the Society for Neuro-Oncology; chair of the Neuro-Oncology Specialty Group, World Federation of Neurology; co-chair of the NCI Brain Malignancies Steering Committee (until 2022). **B.R.R.** No relevant relationships. **E.R.G.** No relevant relationships. **J.K.C.** Deputy editor of *Radiology: Artificial Intelligence*; grants from GE and Genetech; consulting fees from Siloam Vision; technology licensed to BostonAI.

References

1. Egger J, Kapur T, Fedorov A, et al. GBM volumetry using the 3D Slicer medical image computing platform. Sci Rep 2013;3(1):1364.
2. Moltz JH, Braunewell S, Rühaak J, et al. Analysis of variability in manual liver tumor delineation in CT scans. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. 2011; 1974–197.
3. van der Veen J, Gulyban A, Nuyts S. Interobserver variability in delineation of target volumes in head and neck cancer. Radiother Oncol 2019;137:9–15.
4. Chisholm RA, Stenning S, Hawkins TD. The accuracy of volumetric measurement of high-grade gliomas. Clin Radiol 1989;40(1):17–21.
5. Bi N, Wang J, Zhang T, et al. Deep Learning Improved Clinical Target Volume Contouring Quality and Efficiency for Postoperative Radiation Therapy in Non-small Cell Lung Cancer. Front Oncol 2019;9:1192.
6. Ma CY, Zhou JY, Xu XT, et al. Deep learning-based auto-segmentation of clinical target volumes for radiotherapy treatment of cervical cancer. J Appl Clin Med Phys 2022;23(2):e13470.
7. Esmaeilzadeh P. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. BMC Med Inform Decis Mak 2020;20(1):170.
8. Lambert SI, Madi M, Sopka S, et al. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. NPJ Digit Med 2023;6(1):111.
9. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete Problems in AI Safety. arXiv 1606.06565 [preprint] https://arxiv.org/abs/1606.06565. Published June 21, 2016. Accessed September 2021.
10. Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. J Med Internet Res 2020;22(6):e15154.
11. Cadario R, Longoni C, Morewedge CK. Understanding, explaining, and utilizing medical artificial intelligence. Nat Hum Behav 2021;5(12):1636–1642.
12. Kofler F, Ezhov I, Isensee F, et al. Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient. arXiv 2103.06205 [preprint] https://arxiv.org/abs/2103.06205. Published March 10, 2021. Accessed April 2021.
13. Shi R, Wang W, Li Z, et al. Human Perception-based Evaluation Criterion for Ultra-high Resolution Cell Membrane Segmentation. arXiv 2010.08209 [preprint] https://arxiv.org/abs/2010.08209. Published October 16, 2020. Accessed October 2020.
14. Lo AC, Liu M, Chan E, et al. The impact of peer review of volume delineation in stereotactic body radiation therapy planning for primary lung cancer: a multicenter quality assurance study. J Thorac Oncol 2014;9(4):527–533.
15. Batchelor TT, Gerstner ER, Emblem KE, et al. Improved tumor oxygenation and survival in glioblastoma patients who show increased blood perfusion after cediranib and chemoradiation. Proc Natl Acad Sci USA 2013;110(47):19059–19064.
16. Iglesias JE, Liu CY, Thompson PM, Tu Z. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans Med Imaging 2011;30(9):1617–1634.
17. Gorgolewski K, Burns CD, Madison C, et al. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Front Neuroinform 2011;5:13.
18. Kendall A, Badrinarayanan V, Cipolla R. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. arXiv 1511.02680 [preprint] https://arxiv.org/abs/1511.02680. Published November 9, 2015. Accessed June 2019.
19. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004;23(7):903–921.
20. Beers A, Brown J, Chang K, et al. DeepNeuro: an open-source deep learning toolbox for neuroimaging. Neuroinformatics 2021;19(1):127–140.
21. Jungo A, Scheidegger O, Reyes M, Balsiger F. pymia: A Python package for data handling and evaluation in deep learning-based medical image analysis. Comput Methods Programs Biomed 2021;198:105796.
22. Quarfoot D, Levine RA. How Robust Are Multirater Interrater Reliability Indices to Changes in Frequency Distribution? Am Stat 2016;70(4):373–384.
23. Kendall MG. A new measure of rank correlation. Biometrika 1938;30(1–2):81–93.
24. Khamis H. Measures of association: How to choose? Vol. 24. J Diagn Med Sonogr 2008;24(3):155–162.
25. Chang K, Beers AL, Bai HX, et al. Automatic assessment of glioma burden: a deep learning algorithm for fully automated volumetric and bidimensional measurement. Neuro Oncol 2019;21(11):1412–1422.
26. Nikolov S, Blackwell S, Zverovitch A, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv 1809.04430 [preprint] https://arxiv.org/abs/1809.04430. Published September 12, 2018. Accessed May 2021.
27. Vollmuth P, Foltyn M, Huang RY, et al. Artificial intelligence (AI)-based decision support improves reproducibility of tumor response assessment in neuro-oncology: An international multi-reader study. Neuro Oncol 2023;25(3):533–543.
28. Campbell JP, Kalpathy-Cramer J, Erdogmus D, et al. Plus Disease in Retinopathy of Prematurity: A Continuous Spectrum of Vascular Abnormality as a Basis of Diagnostic Variability. Ophthalmology 2016;123(11):2338–2344.
29. Li MD, Little BP, Alkasab TK, et al. Multi-Radiologist User Study for Artificial Intelligence-Guided Grading of COVID-19 Lung Disease Severity on Chest Radiographs. Acad Radiol 2021;28(4):572–576.
30. Kalpathy-Cramer J, Campbell JP, Erdogmus D, et al. Plus Disease in Retinopathy of Prematurity: Improving Diagnosis by Ranking Disease Severity and Using Quantitative Image Analysis. Ophthalmology 2016;123(11):2345–2351.
31. Der Kiureghian A, Ditlevsen O. Aleatory or epistemic? Does it matter? Struct Saf 2009;31(2):105–112.
32. Jha AK, Myers KJ, Obuchowski NA, et al. Objective Task-Based Evaluation of Artificial Intelligence-Based Medical Imaging Methods: Framework, Strategies, and Role of the Physician. PET Clin 2021;16(4):493–511.
33. Cha E, Elguindi S, Onochie I, et al. Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy. Radiother Oncol 2021;159:1–7.
34. Lu SL, Xiao FR, Cheng JCH, et al. Randomized multi-reader evaluation of automated detection and segmentation of brain tumors in stereotactic radiosurgery with deep neural networks. Neuro Oncol 2021;23(9):1560–1568.
35. Conte GM, Weston AD, Vogelsang DC, et al. Generative Adversarial Networks to Synthesize Missing T1 and FLAIR MRI Sequences for Use in a Multisequence Brain Tumor Segmentation Model. Radiology 2021;299(2):313–323.

36. Di Ieva A, Russo C, Liu S, et al. Application of deep learning for automatic segmentation of brain tumors on magnetic resonance imaging: a heuristic approach in the clinical scenario. Neuroradiology 2021;63(8):1253–1262.
37. Mitchell JR, Kamnitsas K, Singleton KW, et al. Deep neural network to locate and segment brain tumors outperformed the expert technicians who created the training data. J Med Imaging (Bellingham) 2020;7(5):055501.
38. Wang G, Li W, Zuluaga MA, et al. Interactive Medical Image Segmentation Using Deep Learning With Image-Specific Fine Tuning. IEEE Trans Med Imaging 2018;37(7):1562–1573.