Neuro-Oncology Advances

4(1), 1–10, 2022 | https://doi.org/10.1093/noajnl/vdac093 | Advance Access date 14 June 2022

Identifying clinically applicable machine learning algorithms for glioma segmentation: recent advances and discoveries

Niklas Tillmanns, Avery E. Lum, Gabriel Cassinelli, Sara Merkaj, Tej Verma, Tal Zeevi, Lawrence Staib, Harry Subramanian, Ryan C. Bahar, Waverly Brim, Jan Lost, Leon Jekel, Alexandria Brackett, Sam Payabvash, Ichiro Ikuta, MingDe Lin, Khaled Bousabarah, Michele H. Johnson, Jin Cui, Ajay Malhotra, Antonio Omuro, Bernd Turowski, and Mariam S. Aboian

Brain Tumor Research Group, Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, Connecticut, USA (N.T., A.E.L., G.C., S.M., T.V., T.Z., L.S., H.S., R.C.B., W.B., J.L., L.J., S.P., I.I., M.L., M.H.J., A.M., M.S.A.); Visage Imaging, Inc., San Diego, California, USA (M.L.); Visage Imaging, GmbH, Berlin, Germany (K.B.); Harvey Cushing/John Hay Whitney Medical Library, Yale University, New Haven, Connecticut, USA (A.B.); Department of Pathology, Boston Children's Hospital, Boston, Massachusetts, USA (J.C.); Department of Neurology and Yale Cancer Center, Yale School of Medicine, New Haven, Connecticut, USA (A.O.); University Dusseldorf, Medical Faculty, Department of Diagnostic and Interventional Radiology, Dusseldorf, Germany (N.T., B.T.)

Corresponding Author: Mariam S. Aboian, MD, PhD, 789 Howard Avenue (CB30), PO Box 208042, New Haven, CT 06520, USA (mariam.aboian@yale.edu).

Abstract

Background. While there are innumerable machine learning (ML) research algorithms used for segmentation of gliomas, there is yet to be a US FDA cleared product. The aim of this study is to explore the systemic limitations of research algorithms that have prevented translation from concept to product by a review of the current research literature.

Methods. We performed a systematic literature review on 4 databases. Of 11 727 articles, 58 articles met the inclusion criteria and were used for data extraction and screening using TRIPOD.

Results. We found that while many articles were published on ML-based glioma segmentation and report high accuracy results, there were substantial limitations in the methods and results portions of the papers that result in difficulty reproducing the methods and translation into clinical practice.

Conclusions. In addition, we identified that more than a third of the articles used the same publicly available BRaTS and TCIA datasets and are responsible for the majority of patient data on which ML algorithms were trained, which leads to limited generalizability and potential for overfitting and bias.

Key Points

- Most algorithms are trained on low patient number or highly curated datasets.
- Most studies fail to describe their algorithm and underlying work properly.

Gliomas account for 31% of all brain and central nervous system tumors in the United States and occur with an age standardized incidence rate of 5.3 per 100 000 persons in North America.[1] Gliomas are classified into different histological subgroups according to the World Health Organization (WHO) classification of tumors of the central nervous system, with new guidelines reported in 2021. Glioma diagnosis on initial imaging is not always accurate and is dependent on the level of expertise by neuroradiologist and neuro-oncologists in complex cases. In addition, the definition of glioma margins may be dependent on the expertise of the neuroradiologists evaluating the study and amino

© The Author(s) 2022. Published by Oxford University Press, the Society for Neuro-Oncology and the European Association of Neuro-Oncology.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

acid. Positron emission tomography (PET) has shown tremendous progress in delineating glioma margins as compared to MRI.[2–5] Machine learning (ML) has demonstrated tremendous progress in predicting glioma grade and molecular subtypes based on radiomic analysis of magnetic resonance (MR) images, but the rate-limiting step in the development of classification algorithms is the generation of ground base segmentations of the tumors that provide volumetric information and radiomic features of tumors.[6] MR imaging is the standard imaging method for brain tumors and for radiomic analysis it often includes T1 ± contrast, T2, and fluid attenuated inversion recovery (FLAIR). The length of time that it takes for manual contouring of gliomas and their different parts is significant and has historically led to the generation of limited datasets. Individual segmentations of enhancing portions, necrotic portions, and nonenhancing portions of tumors and edema can take up to an hour if the segmentations are then transported to different sequences on the MRI. In addition, manual contouring is associated with a wide variability and low uniformity among different users, which we will call "raters" in the current review. According to Bondiau et al,[7] the mean time for the analysis and manual delineation of brain structures on a typical MRI study is 86 min.[8] Therefore, there is a critical need to develop automatic segmentation algorithms that are accurate for implementation into research and clinical practice. The purpose of this study is to present a comprehensive systematic review of applications of AI in the segmentation of gliomas whereby several common limitations were identified that have impacted clinical translation of ML algorithms into routine clinical practice.

Methods

This is an IRB approved study. According to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA),[9] a literature review was performed by university librarian and reviewed by second librarian on 4 databases, Ovid Embase, Ovid MEDLINE, Cochrane trials (CENTRAL), and Web of science core-collection first in October 2020 and for a second time in February 2021. The search strategy included both keywords and controlled vocabulary combining the terms for: artificial intelligence, machine learning, deep learning, radiomics, magnetic resonance imaging, glioma, as well as related terms.

Systematic review was performed in Covidence (Melbourne, Australia) with 11 727 articles identified in 4 databases with the search performed and verified by librarians at Yale School of Medicine Library. Of our 1135 articles that qualified for full-text review, 695 articles were used for data extraction. Indication of studies was extracted for these articles and 58 articles were related to segmentation methods applied to glioma datasets (Figure 1 and Supplementary Data 1).

The search strategy was independently reviewed by a second institutional librarian. All publications were screened in Covidence software by a neuroradiology assistant professor, radiology resident and an artificial intelligence graduate student. Three reviewers consisting of an assistant professor of radiology, a medical student, and an undergraduate student evaluated eligible ML performance studies. When questions regarding the inclusion of studies arose, they were resolved by radiology assistant professor. Studies using only logistic regression methods were excluded. All 3 reviewers extracted data using predetermined parameters such as title, author, year of publication, patient characteristics, datasets, modes of ML, gold standard for accuracy, imaging features, magnetic field strength (Tesla) of the MR scanners, sequences used and reported statistics.

To assess the quality of reporting in the underlying literature, we used the Transparent Reporting of studies on prediction models for Individual Prognosis Or Diagnosis (TRIPOD).[10,11] Historically the quality of systematic reviews of diagnostic test accuracy reports has been evaluated using QUADAS-2.[12] QUADAS-2 tool is specifically tailored to diagnostic test and procedures and does not address the questions specific to model development studies reported in AI literature. Radiomics quality score is a 2017 established scoring system designed for publications that wish to extract radiomic features and for identifying bias in radiomic studies.[13] The Checklist for Artificial Intelligence in Medical Imaging (CLAIM) is a new bias assessment that was published in 2020 and is targeted toward AI literature. CLAIM is in many ways similar to TRIPOD assessment, although TRIPOD is more comprehensive with up to 91 possible overall scoring items and providing a dedicated TRIPOD adherence form (Supplementary Data 2).[11,14] TRIPOD is a relatively new (2015) quality assessment tool with 22 categories (Supplementary Data 2) applicable to different types of prediction model studies. TRIPOD is an approach, which can be used for developing, validating, or updating prediction models and therefore was most suitable for our review. We used the TRIPOD assessment for the development model, which had 37 scoring items and 65 items overall.[10,11] TRIPOD analysis of our papers was performed by 3 individual reviewers (medical students and an undergraduate student) and required approximately 30 min per paper. Because the overall quality of reporting was rather low, we were not able to perform a risk of bias assessment.[15]

This study was extracted as a diagnostic developmental model. Adherence of a report is calculated per TRIPOD component and its subitems. If the answer to all adherence elements of a particular TRIPOD item is scored with "yes," adherence to that TRIPOD item is scored as "1," and nonadherence as "0." During risk of bias assessment, no exclusion of papers was performed.

Accuracy Reporting

Accuracy analysis was extracted for each combination of imaging features and algorithms. Best accuracy is presented in the figures and our results. We report the Dice coefficient in our report as the accuracy measure, because it was the most consistently reported among all papers. All data extracted from the individual publications was exported into Excel (Microsoft, Redmond, WA), and only the

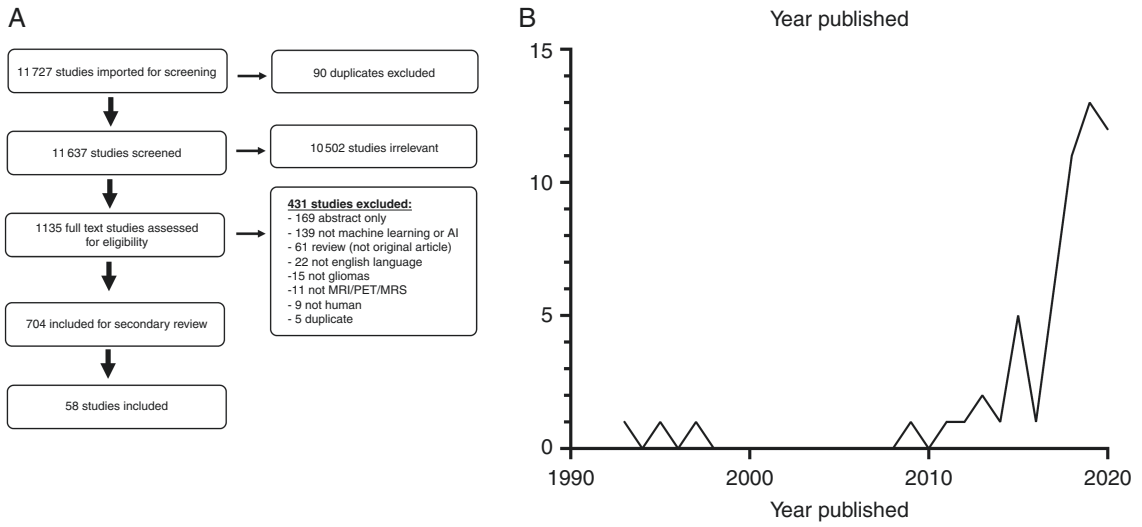

Figure 1. (a) Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) flowchart of the search strategy for the systematic review (created with BioRender.com), verified by librarians at Yale School of Medicine Library. (b) Machine Learning Trends for Glioma Brain Tumor Segmentation, steady until 2018 when a significant rise in papers was observed.

publications with reported Dice scores were further analyzed regarding their accuracy reporting (Figure 5), which consisted of only 50% of papers. From these publications we excluded 1 study, which reported a Dice for the segmentation of 2 different experts and not for an algorithm. We preferably extracted the median Dice similarity coefficient (DSC), but if multiple DSC were reported then the highest fitting value was chosen for extraction.

Results

Study Selection

The systematic review of articles used 4 different databases and identified 11 727 candidate articles. Ninety duplicates were removed and screening of the remaining 11 637 article abstracts was conducted. Abstract review further excluded 10 502 articles that were not neuro-oncology studies. A total of 1135 articles were reviewed at the full-text level. Four hundred and thirty-one articles were excluded for the following reasons: 169 conference abstracts, 139 articles did not use ML, 61 of the articles were review of the literature, 22 of the articles were not in English language, 15 articles did not include glioma or glioblastoma in their analysis, 11 articles did not include imaging either MRI, PET, or MRS, 9 articles involved nonhuman subjects, and 5 articles were found to be duplicates. Seven hundred and four full-text studies were further reviewed and 58 of them included studies that focused on algorithms dedicated to the segmentation of gliomas and their different regions. These 58 studies were analyzed for the systematic review (Figure 1a).

The distribution of publications on segmentation of gliomas over time show that initial research started in 1993 with a relative gap in publications from 1993 to 2010 (Figure 1b). Since 2010 we see a steady increase in publication rate per year in the field of AI in glioma segmentation. Beginning in 2017, the trend becomes obvious with a steady increase of publications per year, reaching a maximum of 13 publications per year in 2019.

Datasets

The most frequently used datasets were single center and multicenter (not BRaTS, or TCIA) which were employed in 53.4% of the studies; single-center data were used in 22 (37.9%) and multicenter data were used in 9 (15.5%) of the studies (Figure 2a). Publicly available datasets such as The Cancer Imaging Archive (TCIA) and Brain Tumor Segmentation challenge (BRaTS) were used in 21 studies (36.2%). Studies with larger numbers of patients primarily relied on either TCIA or BRaTS datasets. There were approximately 8.6% of studies that did not fully describe their source of data. In all of the studies included in our systematic review just 2 studies used external and geographically distant datasets for validation.[16,17] Just 1 study gave insight into how accuracy has changed through validation. The accuracy decreased from 92.7% in the training cohort, to 92.4% in testing and validation cohort to 78.0% in the external validation set.[16] Many other articles elaborated on external validation as an important issue to address in further studies to guarantee the strength of the reported model but most of them have not incorporated such a method in their actual manuscript.

Number of Patients

Figure 2b shows the patient cohort sizes that were used in the reviewed literature. The number of patients per study ranged from 1 to 622. The mean number of patients was

143, with SD being 155. The SD is high because several studies had very high numbers; as an example, 1 study used a combination of BRaTS dataset with 622 patients.[18] In contrast to this, the median number of patients was 56, which better reflects the large number of studies that used ML-based segmentation on very small datasets. The majority of publications on segmentation of gliomas have a patient number below the mean. Among all studies included in this review, 8305 patients were analyzed.

MRI Sequences

The most frequently used MRI sequence was T2 (81%) followed by FLAIR (78%). More advanced imaging techniques such as functional MRI (fMRI) and position emission tomography (PET) were implemented in under 3% of the examined literature (Figure 3).

Algorithms.—Conventional ML was used in 44.83% of the studies (26) and included the following algorithms: support vector machines (SVM), decision tree methods including decision forests (DF) and random forests (RF), fuzzy *C*-means (FCM), virtual rater, and *k*-mean clustering (Figure 4). The most frequently used algorithm in the category of ML was SVM. Deep learning was used in 31 papers (53.45%) and included the following algorithms and models: convolutional neural networks (CNN), U-Net (subset of CNNs), and other deep neural networks. The most frequently used algorithm was CNN. Algorithms without detailed specification were found in 1 paper (1.72%), describing a nonmodel automatic segmentation method.[19]

Accuracy of Segmentations

Dice scores associated with reported algorithms used for segmentation of either whole tumor, enhancing tumor, necrosis, or core segmentation are reported in Figure 5. The highest Dice scores were reported for segmentation of whole tumor, followed by the core portion of the tumor. The exception was the RF algorithm, which provided highest Dice scores for segmentation of enhancing portion of the tumor (mean = 0.850) followed by whole tumor (0.824). Deep learning algorithms such as U-Net and CNN showed similar results to conventional ML. The whole tumor

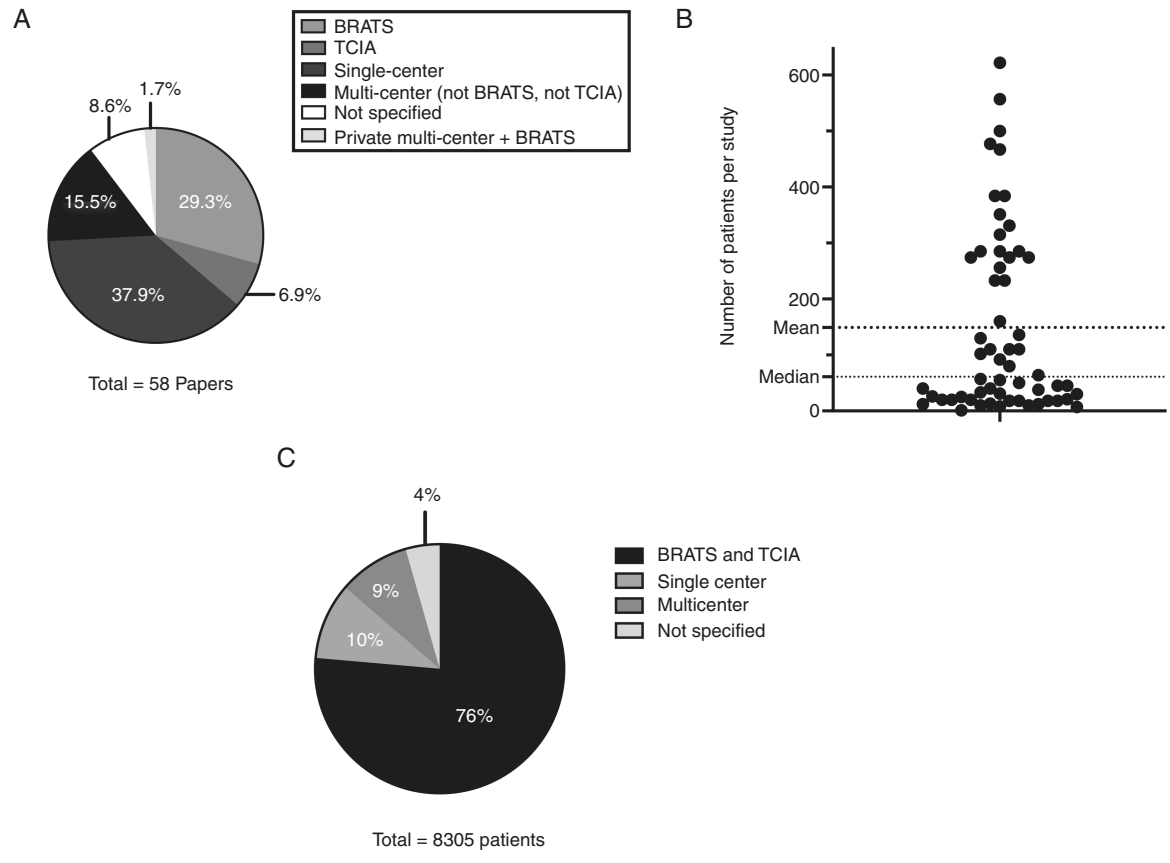

Figure 2. Datasets used in papers evaluating applications of AI in segmentation of gliomas. (a) Percentage of studies that used each dataset type. (b) Range, mean, and median number of patients in the studies. (c) Among all the studies, the percentage of patients that were contributed by different datasets. BRaTS and TCIA include to 76% of the patients. BRaTS, Brain Tumor Segmentation challenge (all years included); TCIA, The Cancer Imaging Archive.

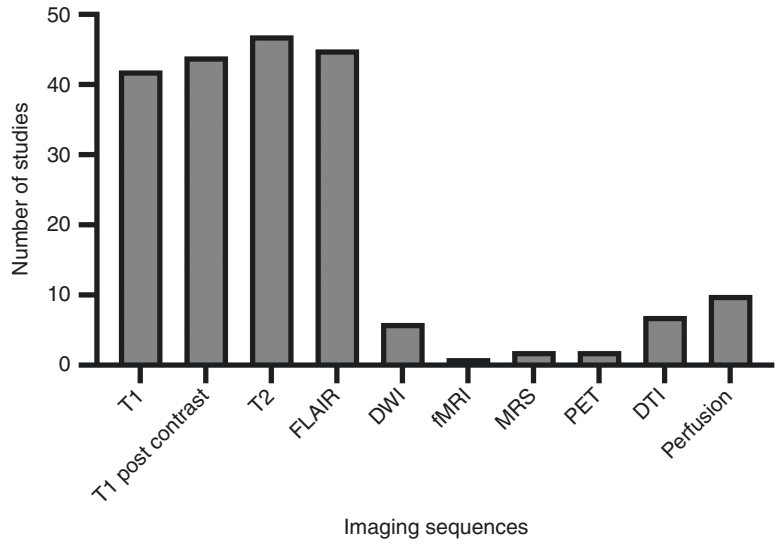

Figure 3. Imaging sequences used for the segmentation of gliomas. Number of studies that used specific imaging sequences. T1 (precontrast) and T1 (postcontrast), T2, and FLAIR were the most common sequences used for tumor segmentation. DTI, diffusion tensor imaging; DWI, diffusion weighted imaging.

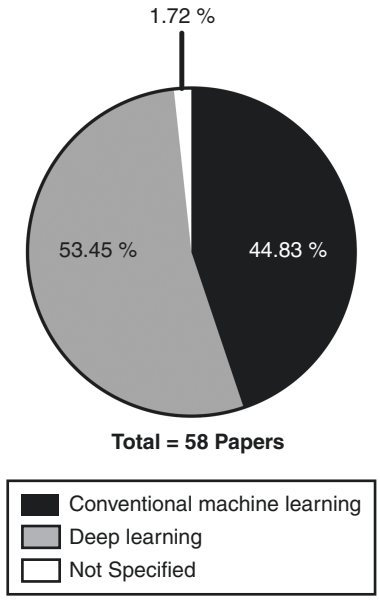

Figure 4. Distribution of machine learning and deep learning algorithms in the extracted publications involved in segmentation of gliomas.

segmentation by U-Net showed a slightly lower mean Dice with 0.865, compared to CNN with a mean Dice of 0.871. However, the SD for CNN is higher (SD ± 0.045) compared to U-Nets (SD ± 0.039), which is interesting regarding 13 papers included with CNN and just 6 papers with U-Net segmentation for whole tumor included. When evaluating core segmentation, the observations are switched, with the mean Dice of U-Net = 0.823 (SD ± 0.118) and mean Dice of CNN = 0.781 (SD ± 0.077). The enhancing tumor segmentation performed equally well. U-Net achieves a Dice of 0.763 (SD ± 0.171) 0.747 (SD ± 0.072) for CNN. Dice scores were reported for only 50% (29 of 58) of papers, with the rest of the papers not assessing accuracy of segmentations with Dice.

Reporting quality; TRIPOD

The mean TRIPOD score of all 58 publications is 12.5 (43.10%, SD = 2.1) with the highest achievable score being 29. There were 7 categories where none of the papers achieved the objective, including title, abstract, risk groups, participants, model specification, and model performance. The highest scoring items, in which every paper scored a point were predictors model development and discussion and interpretation (Figure 6). The categories background and rationale had an overall adherence score of 95% and 84%, respectively. An average adherence score of 56% was achieved across studies for methods-participants. While outcome definition (including time and method of assessment) was rarely presented (5%), actions to blind assessment of outcome to be predicted had a high score of 97%. Similarly, predictors definitions (including time and method of assessment) had a low score of 7%, with assessments of blinding predictor for outcome and for other predictors scoring 100%. Forty-three percent of all studies explained how sample size was derived. Only 7% of all studies elaborated on missing data and methods for handling it. Statistical analysis methods showed an overall adherence score of 17%. Results and participants scored 41% for general participant information and 0% for reporting background and missing data. Model development categories scored 97% and 100%, respectively.[20] In

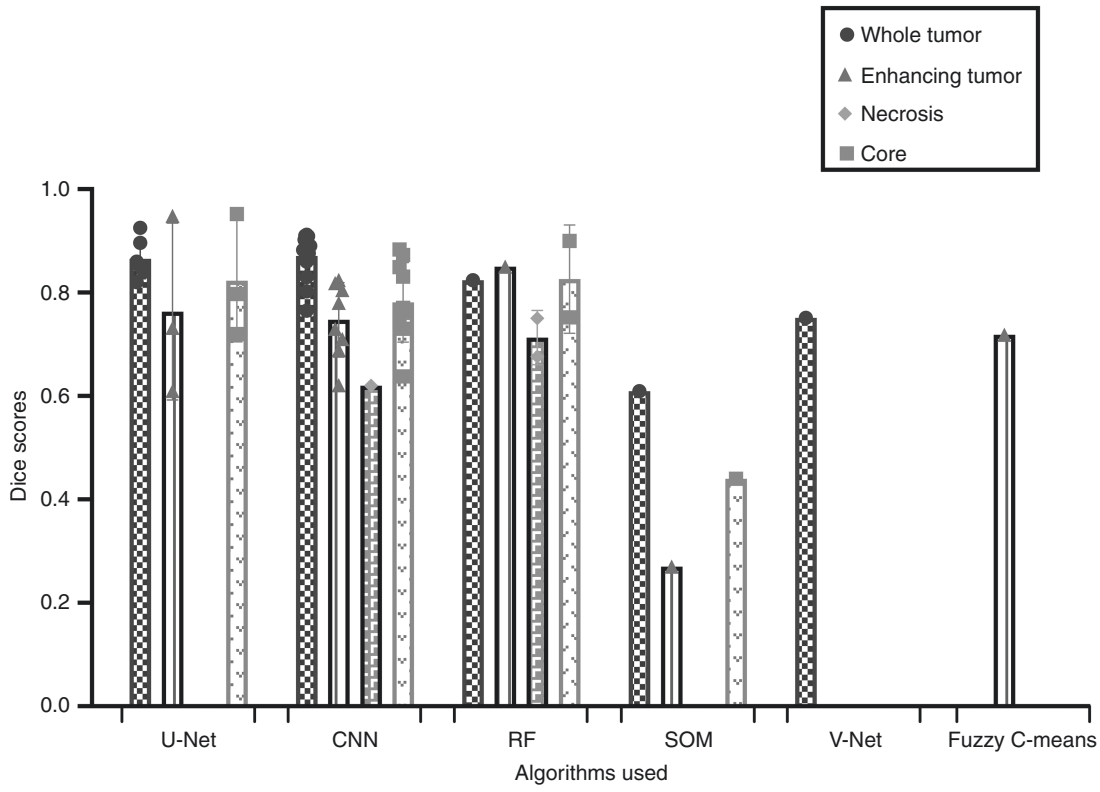

Figure 5. Accuracy of segmentations with reported Dice scores according to segmented region and the associated algorithm. CNN, convolutional neural networks; RF, random forest.

model specification, a regression coefficient was always missing (0%), but 50% of all studies explained ways to use their models for individual predictions. Fifty-nine percent of all studies discussed their limitations and all of them (100%) provided us with an overall interpretation of their results. Only a little more than half of all studies (53%) discussed potential for clinical use and implications for future research. Funding scored 10% across studies.

Discussion

ML approaches for the automatic segmentation of gliomas started appearing in the early 1990s and the field has since significantly expanded, making evaluation of literature complex and difficult to synthesize. We present a systematic review of applications of AI in segmentation of gliomas with focus on characterization of data used for the development of algorithms and identification of the most accurate algorithms that can potentially be used for clinical implementation.

There are different methods available to evaluate the segmentation results of a proposed algorithm.[21] These include the Jaccard index,[22] Sørensen–Dice coefficient, and Hausdorff distance.[23] These indexes vary in the way they evaluate the result of a proposed segmentation. The Jaccard index is often used to assess the intersection over union. It is defined by the size of intersection divided by the size of the union of the sample sets. In contrast, the Dice coefficient, also called the Sørensen–Dice index or DSC, is a method to estimate the overlap of 2 samples. The Hausdorff distance measures the distance between 2 subsets of a metric space and reports the largest distance between 2 subsets of points. In the reported papers, the comparison was made between the gold standard of segmentation (often the neuroradiologist's interpretation and manual delineation) and the proposed segmentation by the algorithm. The Dice numbers reported in Figure 5 are somewhat limited, because only 50% of papers reported a measurable accuracy score. Often, the publications only mention a single data point for Dice which limited evaluation of the quality of segmentations. While DICE coefficient is the most commonly used metric for assessment of segmentation quality in the literature, we recommend to use additional metrics such as Hausdorff distance. Beyond the standardized methods, we recommend to include information such as motion artifact and heterogeneity of protocols. We recently presented this approach at ISMRM which includes assessment of segmentation in the setting of motion artifact or nondegraded studies (K. B. Sara Merkaj, unpublished data, 2022). But there is a need for further research on metrics that measure clinical applicability of algorithms in face of rare relevant false-positive and

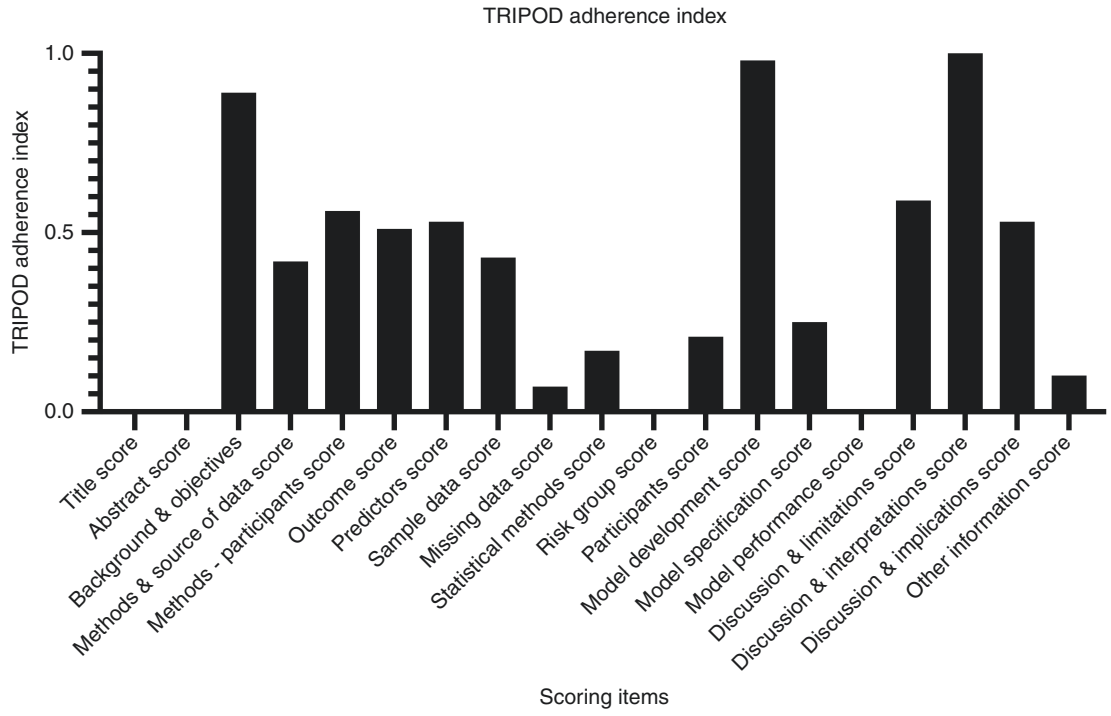

Figure 6. The TRIPOD adherence index, a measure for degree of satisfaction for each main item regardless of the comprised number of subitems, indicating overall strengths and weaknesses in reporting in our study cohort. Notice that some items are not shown within the graph, since they were only pertinent to validation studies (12, 17, subitems 10c/e, 13c, 19a), but not model development studies.

-negative cases that will appear in clinical practice. These problems are accompanied by the fact that most articles rely heavily on radiologists' interpretations of the images as a gold standard for segmentation and multiple radiologists' assessment of segmentation was rare. It is important to include the information between inter-rater variability in segmentations in future work that described novel segmentation algorithms. Development of algorithms has the potential to eliminate the human error with respect of consistency in segmentation of brain tumors in the future. We recommend the literature to pay close attention to this important evaluation metric that goes beyond simple measures as Dice.

Deep learning algorithms have become more common in the last 2 years and are the preferred method for automatic segmentation in the most recent literature. Among the different algorithms, RF was the most common ML method used for segmentation with Dice coefficient scores extending beyond 0.8. Deep learning algorithms, such as CNN and U-Net, were similar in accuracy of segmentation and also demonstrated Dice coefficient scores beyond 0.8. The time for segmentation by these algorithms was not detailed in the majority of papers, making it unclear which algorithm would be best for clinical implementation.

TRIPOD score for our studies was 43.10%. Major deficiencies in title, abstract, statistical methods, risk groups, participants, model specification, and model performance sections of the papers may lead to ambiguity and exclusion of potentially good studies in meta-analysis or future studies. Limitations in the methods (specifically description of predictors and statistical analysis) and results (specifically model development and performance) portions of the paper will result in difficulty reproducing the results. While low reporting scores do not directly affect ability to translate the ML algorithms into clinical practice, the lack of details on these critical aspects of training data, algorithm design, and validation results makes it difficult to reproduce the findings from the literature and thereby move clinical implementation of suitable algorithms forward. To address the current gap between algorithm development and clinical implementation, further research, which adheres to strict reporting guidelines in order to allow for reproducibility, is needed (M. Lin, unpublished data, 2021; S. Ebrahimian et al, unpublished data, 2021). A common deficiency in all of the papers was not reporting the role of the funders in the funding portion, which is relevant for identification of potential biases and conflicts of interest, and for better interpretation of findings. We recommend that all journals require this information to be disclosed in their manuscripts. In addition, researchers and reviewers should keep an eye on the upcoming TRIPOD-AI expansion, which will likely improve the state of the art in reporting ML studies in the medical field (Figure 6).[24] TRIPOD has similar items than CLAIM but is structured in 65 clearly defined items. Nonetheless we highly encourage the use of the upcoming TRIPOD-AI guideline for further

study assessment. Future applicability to clinical neuro-oncology may also benefit from study designs with Minimum Information for AI Reporting (MINIMAR) in mind to include under-represented patients. MINIMAR is a "proposal describing the minimum information necessary to understand intended predictions, target populations, and hidden biases, and the ability to generalize these emerging technologies."[25]

The literature also shows a lack of implementation of algorithms into clinical practice. One of the major limitations of the manuscripts published on development of segmentation algorithms was the limited number of patients used in studies reporting segmentation tools for intracranial malignancies (mean 148.6, median 60.5). The discrepancy between the mean and median is explained by 2 specific studies that were outliers with high patient numbers.[18,26] These studies also were combinations of BRaTS and TCIA databases, which shows that the largest proportion of patients evaluated in our cohort of patients (6124 out of 8604 patients) were highly curated datasets. This makes up 76% of patients overall, even when BRaTS and TCIA combined are just implemented in 37.9% of the publications. BRaTS and TCIA are responsible for the majority of patient data on which algorithms get trained. On the one hand, this is a good publicly available dataset that was used by multiple groups to develop algorithms for tumor segmentations, but there is a substantial risk of overfitting which can explain high accuracy among different reported algorithms.

Most algorithms reported in the literature are trained on MRI sequences that are common in clinical practice, with very few of the papers focusing on specialized perfusion or fMRI sequences. This allows clinical translation of these algorithms into the majority of clinical practices because, due to overall good accuracy results by these algorithms, elaborate scanning techniques are not required for simple volumetric segmentation. On the other hand, 24% of papers used advanced imaging methods, such as fMRI, PET, and perfusion imaging. This suggests that application of AI tools to these modalities are still early in their development or performed at institutions with different equipment (such as intraoperative MRI or newer radiopharmaceuticals), and we are eager to watch for publications with novel ideas on how AI can be applied for these modalities. Notably, most algorithms in the current literature are based on preoperative tumor imaging, whereas most clinical imaging techniques for brain tumors are used after treatment to assess response or to monitor progression. This is a topic beyond the scope of this review, but definitely one that needs to be addressed in further developmental studies. The results of our systematic review lead us to conclude that the implementation of algorithms into overall patient management is critical. This is underscored by our group's current research, and we look forward to sharing this research with the audience in the near future.

Limitations of this systematic review include the exclusion of abstracts and information from segmentation competitions and hackathons. Kaggle and BRaTS challenges played a significant role in identified best segmentation algorithms applied to different imaging challenges. Many of these competition results are published in the literature, but it is possible that not all of the significant advances were advanced into peer-review published papers and are still available in abstract format. These public AI challenges also mandate winning algorithms be made open source for public scrutiny, and therefore commercial entities are absent, making unclear if the best algorithms are open source or commercialized. As we continue to analyze the field of AI applications in segmentation of gliomas, we will start to capture those articles and the advances that are reported in them. Another limitation is the wide search strategy that we performed that resulted in a much larger number of articles in our search strategy than average previously published systematic reviews on AI in neuro-oncology. The time for evaluation of such a long list of articles delayed our data extraction process initially ending October 2020. We had to repeat the search in February 2021 to make sure our results are relevant.

In conclusion, we present a comprehensive systematic review of applications of AI in segmentation of gliomas and have identified several limitations that have impacted clinical translation of ML algorithm that can be avoided in future publications. Since most of the algorithms in research report acceptable accuracy results, it should be possible to use most of the algorithms for clinical implementation as well. Concluding from our literature review deep learning based approaches like U-Net have the most potential for clinical implementation. Based on recent advances in the field, namely the RSNA MICCAI challenge, nn-U-NET architectures should yield the best segmentation performance in the moment.[27–29] But there are a few points that need to be considered when moving forward clinical research on algorithms. At first large databases are needed in order to train the algorithm sufficiently and lower the risk of overfitting. This requirement can be addressed by using either the already existing and publicly available datasets, or by creating hospital datasets, which will be more suitable to the clinical imaging protocols of the hospital and the patient cohort on site. Additionally implementation of multiside validation will make algorithms more robust (M. Lin, unpublished data, 2021; S. Ebrahimian et al, unpublished data, 2021).

Second we need reporting guidelines, in this growing research field, that are mandatory for publishing in peer-reviewed journals to guarantee the high standard of research. This can be achieved by requiring checklists like CLAIM[14] or TRIPOD[11] for publishing AI articles in journals.

Supplementary Material

Supplementary material is available at *Neuro-Oncology Advances* online.

Keywords

artificial intelligence | glioma | machine learning | segmentation

Funding

Biomedical Education Program to S.M.; National Institute of Diabetes and Digestive and Kidney Disease of the National Institutes of Health (T35DK104689) to R.C.B.; American Society of Neuroradiology Fellow Award 2018 to M.S.A.; National Center for Advancing Translational Science components of the National Institutes of Health, and National Institute of Health roadmap for Medical Research (KL2 TR001862 to M.S.A.); National Institute of Health (R01 CA206180 to M.L., K23NS118056 to S.P.); Foundation of American Society of Neuroradiology (1861150721 to S.P.); Doris Duke Charitable Foundation (#2020097 to S.P.); NVIDIA to S.P.

Acknowledgments

We appreciate the support by institutional librarian (Thomas Mead) for search strategy assistance, and Mary Hughes and Vermetha Polite for their technical support. We thank Julia Shatalov for assistance in abstract screening.

Portions of this work were presented in abstract form and in poster form at the annual meeting of the Society of NeuroOncology, Boston, United States, November 19, and the annual meeting of the Radiological Society of North America, Chicago, United States, November 30, 2021.

Conflict of interest statement. None declared.

Authorship Statement. Literature research: Niklas Tillmanns, Avery E. Lum, Gabriel Cassinelli, Sara Merkaj, Tej Verma, Waverly Brim, Jan Lost, Leon Jekel, Harry Subramanian, and Mariam S. Aboian. Writing of final manuscript: Niklas Tillmanns and Avery E. Lum. Reviewing and editing of the final manuscript: Niklas Tillmanns, Avery E. Lum, Gabriel Cassinelli, Sara Merkaj, Tej Verma, Tal Zeevi, Lawrence Staib, Harry Subramanian, Ryan C. Bahar, Waverly Brim, Jan Lost, Leon Jekel, Sam Payabvash, Ichiro Ikuta, MingDe Lin, Khaled Bousabarah, Michele H. Johnson, Jin Cui, Ajay Malhotra, Antonio Omuro, Bernd Turowski, and Mariam S. Aboian.

References

1. Ostrom QT, Gittleman H, Stetson L, Virk SM, Barnholtz-Sloan JS. Epidemiology of gliomas. In: Raizer J, Parsa A, eds. *Current Understanding and Treatment of Gliomas*. Cham, Switzerland: Springer International Publishing; 2015:1–14.
2. Pauleit D, Floeth F, Hamacher K, et al. *O*-(2-[^{18}F]fluoroethyl)-L-tyrosine PET combined with MRI improves the diagnostic assessment of cerebral gliomas. *Brain*. 2005;128(3):678–687.
3. Pafundi DH, Laack NN, Youland RS, et al. Biopsy validation of ^{18}F-DOPA PET and biodistribution in gliomas for neurosurgical planning and radiotherapy target delineation: results of a prospective pilot study. *Neuro Oncol*. 2013;15(8):1058–1067.
4. Galldiks N, Ullrich R, Schroeter M, Fink GR, Kracht LW. Volumetry of [^{11}C]-methionine PET uptake and MRI contrast enhancement in patients with recurrent glioblastoma multiforme. *Eur J Nucl Med Mol Imaging*. 2010;37(1):84–92.
5. Lohmann P, Werner JM, Shah NJ, et al. Combined amino acid positron emission tomography and advanced magnetic resonance imaging in glioma patients. *Cancers (Basel)*. 2019;11(2):153.
6. Lotan E, Jain R, Razavian N, Fatterpekar GM, Lui YW. State of the art: machine learning applications in glioma imaging. *AJR Am J Roentgenol*. 2019;212(1):26–37.
7. Bondiau PY, Malandain G, Chanalet S, et al. Atlas-based automatic segmentation of MR images: validation study on the brainstem in radiotherapy context. *Int J Radiat Oncol Biol Phys*. 2005;61(1):289–298.
8. Ermiş E, Jungo A, Poel R, et al. Fully automated brain resection cavity delineation for radiation target volume definition in glioblastoma patients using deep learning. *Radiat Oncol*. 2020;15(1):100.
9. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. 2015;4(1):1.
10. Heus P, Damen J, Pajouheshnia R, et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. *BMJ Open*. 2019;9(4):e025611.
11. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1–W73.
12. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529–536.
13. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749–762.
14. Mongan J, Moy L, Charles E, Kahn J. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2(2):e200029.
15. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med*. 2015;13(1):55–63.
16. Yan J-L, Li C, van der Hoorn A, et al. A neural network approach to identify the peritumoral invasive areas in glioblastoma patients by using MR radiomics. *Sci Rep*. 2020;10(1):9748.
17. Juan-Albarracín J, Fuster-Garcia E, García-Ferrando GA, García-Gómez J. ONCOhabitats: a system for glioblastoma heterogeneity assessment through MRI. *Int J Med Inform*. 2019;128:53–61.
18. Sharif M, Li J, Khan M, Saleem M. Active deep neural network features selection for segmentation and recognition of brain tumors using MRI images. *Pattern Recognit Lett*. 2019;129:181–189.
19. Lu M, Zhang X, Zhang M, et al. Non-model segmentation of brain glioma tissues with the combination of DWI and fMRI signals. *Biomed Mater Eng*. 2015;26(suppl 1):S1315–S1324.
20. TRIPOD Checklist: Prediction Model Development. 2022. https://www.tripod-statement.org/wp-content/uploads/2020/01/Tripod-Checlist-Prediction-Model-Development.pdf. Accessed April 26, 2022.
21. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015;15:29.
22. Yuan J, Liu L. Brain glioma growth model using reaction-diffusion equation with viscous stress tensor on brain MR images. *Magn Reson Imaging*. 2016;34(2):114–119.

23. Visser M, Petr J, Müller DMJ, et al. Accurate MR image registration to anatomical reference space for diffuse glioma. *Front Neurosci.* 2020;14(585). doi:10.3389/fnins.2020.00585
24. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open.* 2021;11(7):e048008.
25. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc.* 2020;27(12):2011–2015.
26. Ahammed Muneer KV, Rajendran VR, Paul Joseph K. Glioma tumor grade identification using artificial intelligent techniques. *J Med Syst.* 2019;43(5):113.
27. 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.
28. RSNA. Brain Tumor AI Challenge (2021). 2021. https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/brain-tumor-ai-challenge-2021. Accessed May 3, 2022.
29. Baid U, Ghodasara S, Mohan S, et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification, arXiv, arXiv:2107.02314, 2021, preprint: not peer reviewed. doi:10.48550/arXiv.2107.02314