AI-based decision support improves reproducibility of tumor response assessment in neuro-oncology: an international multi-reader study.

Philipp Vollmuth[1], Martha Foltyn[1], Raymond Y. Huang[2], Norbert Galldiks[3-5], Jens Petersen[6], Fabian Isensee[6], Martin J. van den Bent[7], Frederik Barkhof[8-9], Ji Eun Park[10], Yae Won Park[11], Sung Soo Ahn[11], Gianluca Brugnara[1], Hagen Meredig[1], Rajan Jain[12], Marion Smits[6,13], Whitney B. Pope[14], Klaus Maier-Hein[6], Michael Weller[15], Patrick Y. Wen[16], Wolfgang Wick[17-18], Martin Bendszus[1]

(1) Department of Neuroradiology, Heidelberg University Hospital, Heidelberg, Germany
(2) Department of Radiology, Brigham and Women's Hospital, Boston, MA, USA.
(3) Department of Neurology, Faculty of Medicine, University Hospital Cologne, University of Cologne, Cologne, Germany.
(4) Institute of Neuroscience and Medicine (INM-3, -4), Research Center Juelich, Juelich, Germany.
(5) Center for Integrated Oncology (CIO), Universities of Aachen, Bonn, Cologne, and Duesseldorf, Germany.
(6) Department of Medical Image Computing (MIC), German Cancer Research Center (DKFZ), Heidelberg, Germany
(7) Brain Tumor Center, Erasmus MC Cancer Institute, Rotterdam, Netherlands
(8) Department of Radiology & Nuclear Medicine, Amsterdam UMC, Vrije Universiteit, Amsterdam, Netherlands.
(9) Institutes of Neurology & Centre for Medical Image Computing, University College London, London, UK.
(10) Department of Radiology and Research Institute of Radiology, Asan Medical Centre, University of Ulsan College of Medicine, Seoul, Republic of Korea
(11) Department of Radiology and Research Institute of Radiological Science and Center for Clinical Imaging Data Science, Yonsei University College of Medicine, Seoul, Republic of Korea.
(12) Department of Radiology, New York University School of Medicine, New York, NY, USA.
(13) Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, Netherlands.
(14) Department of Radiological Sciences, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA.
(15) Department of Neurology, University Hospital and University of Zurich, Zurich, Switzerland
(16) Center for Neuro-oncology, Dana-Farber Cancer Institute, Boston, MA, USA.
(17) Neurology Clinic, Heidelberg University Hospital, Heidelberg, Germany
(18) Clinical Cooperation Unit Neurooncology, German Cancer Consortium (DKTK) within the German Cancer Research Center (DKFZ), Heidelberg, Germany

Running title: AI decision support for response assessment in neurooncology

Corresponding Author:

Philipp Vollmuth, MD MBA

Department of Neuroradiology, Heidelberg University Hospital

Im Neuenheimer Feld 400, 69120 Heidelberg, Germany

Email: philipp.vollmuth@med.uni-heidelberg.de

Phone: +49 (0) 6221 56 39069, Fax: +49 (0) 6221 56 4673

Funding: Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the SFB 1389 - UNITE Glioblastoma, Work Package C02 (Project-ID 404521405) and Priority Programme 2177 "Radiomics: Next Generation of Biomedical Imaging" (Project-ID 428223917).

Conflict of Interest: None

Authorship: Conceived and designed the research: Philipp Vollmuth, Martha Foltyn, Martin Bendszus; Performed tumor response assessment: Philipp Vollmuth, Raymond Y. Huang, Norbert Galldiks, Martin J. van den Bent, Frederik Barkhof, Ji Eun Park, Yae Won Park, Sung Soo Ahn, Rajan Jain, Marion Smits, Whitney B. Pope, Michael Weller, Patrick Y. Wen, Wolfgang Wick, Martin Bendszus; Performed XNAT implementation: Jens Petersen, Fabian Isensee, Klaus Maier-Hein, Gianluca Brugnara, Hagen Meredig, Philipp Vollmuth; Analyzed and interpreted the data: all authors; Performed statistical analysis: Philipp Vollmuth, Martha Foltyn; Handled funding and supervision: Philipp Vollmuth; Drafted the manuscript: Philipp Vollmuth, Martha Foltyn; Made critical revision of the manuscript for important intellectual content: all authors

Total Manuscript word count: 5708 words

Abstract

Background: To assess whether AI-based decision support allows more reproducible and standardized assessment of treatment response on MRI in neuro-oncology as compared to manual 2-dimensional measurements of tumor burden using the RANO criteria.

Methods: A series of 30 patients (15 lower-grade gliomas, 15 glioblastoma) with availability of consecutive MRI scans was selected. The time to progression (TTP) on MRI was separately evaluated for each patient by 15 investigators over two rounds. In the 1st round the TTP was evaluated based on the RANO-criteria, whereas in the 2nd round the TTP was evaluated by incorporating additional information from AI-enhanced MRI-sequences depicting the longitudinal changes in tumor volumes. The agreement of the TTP-measurements between investigators was evaluated using concordance correlation coefficients (CCC) with confidence intervals (CI) and p-values obtained using bootstrap resampling.

Results: The CCC of TTP-measurements between investigators was 0.77 (95%CI=0.69,0.88) with RANO alone and increased to 0.91 (95%CI=0.82,0.95) with AI-based decision support (p=0.005). This effect was significantly greater (p=0.008) for patients with lower-grade gliomas (CCC=0.70 [95%CI=0.56,0.85] without vs. 0.90 [95%CI=0.76,0.95] with AI-based decision support) as compared to glioblastoma (CCC=0.83 [95%CI=0.75,0.92] without vs. 0.86 [95%CI=0.78,0.93] with AI-based decision support). Investigators with less years of experience judged the AI-based decision as more helpful (p=0.02).

Conclusions: AI-based decision support has the potential to yield more reproducible and standardized assessment of treatment response in neuro-oncology as compared to manual 2-dimensional measurements of tumor burden, particularly in patients with lower-grade gliomas. A fully-functional version of this AI-based processing pipeline is provided as open-source (https://github.com/NeuroAI-HD/HD-GLIO-XNAT).

Keywords: RANO; tumor response assessment; tumor volumetry; AI-based decision support

Key Points: (248 characters)

1. AI-based decision support improved the concordance of TTP ratings over RANO alone
2. AI-based decision support was more useful for lower-grade gliomas as compared to glioblastoma
3. Less experienced investigators judged the AI-based decision support as more helpful

Importance of the Study:

The RANO criteria are widely adopted in neuro-oncology, however the prescribed manual measurements of tumor burden on MRI may be challenging and potentially limit the reproducibility of the RANO-criteria for reliable assessment of treatment response. There has been long-standing interest in using volumetric assessment of tumor burden with previous studies indicating that volumetric measurements may be more reliable and accurate as compared to 2-dimensional measurements of tumor diameters in arbitrarily chosen slices. The present study demonstrates that AI-based decision support has the potential to yield more reproducible and standardized assessment of treatment response in neuro-oncology as compared to manual 2-dimensional measurements of tumor burden. Particularly the evaluation of lower-grade gliomas where reliable assessment of the TTP may be challenging due to the slow growing nature of these tumors may benefit from AI-based decision support. A fully functional version of this AI-based processing pipeline is provided as open-source (https://github.com/NeuroAI-HD/HD-GLIO-XNAT).

Introduction

Magnetic resonance imaging (MRI) is used extensively in cancer research during drug development, including clinical trials, as well as for the routine management of cancer patients. [1] It is particularly valuable for brain tumors, which are located in one of the most vulnerable and hard-to-reach regions of the human body. However, the assessment of imaging data by radiologists still relies primarily on qualitative (subjective) visual interpretation, which may increase the burden of time and expenditure on clinical trials, and which may also hamper the validity of imaging biomarkers used in clinical trials and clinical practice for assessing treatment response. The criteria for assessing treatment response and efficacy in neuro-oncology are essentially based on longitudinal measurements of the largest diameters of contrast-enhancing target lesions on imaging as formalized by the Response Assessment in Neuro-Oncology (RANO) criteria [2,3]. The RANO criteria are widely adopted in neuro-oncology clinical trials to yield a standardized and reproducible assessment of treatment response. Underlying the use of RANO is the assumption that the two-dimensional measurement of a contrast-enhancing lesion's largest diameter on MRI is a surrogate marker of the overall tumor burden. However, this assumption is not always accurate, since brain tumors frequently display very complex shapes and anisotropic growth, influenced in part by the surrounding anatomic boundaries, host tissue–tumor interface, or treatment related effects (e.g., areas of necrosis and surgical cavities). Consequently, reproducible assessment of tumor burden and treatment response and/or disease progression between different radiologists using RANO criteria may be challenging and thus potentially limiting its value for clinical decision making. Reproducible assessment is further complicated by the assessment of nonenhancing T2/FLAIR lesions as an additional criterion besides contrast-enhancing lesions for evaluating treatment response and/or disease progression.

In the light of that, there has been long-standing interest in using volumetric assessment of tumor burden [3,4] with previous studies indicating that volumetric measurements may be more reliable and accurate as compared to 2-dimensional measurements of tumor diameters in arbitrarily chosen slices (**Figure 1**) [5,6]. In the present study we investigated the clinical utility of AI-based decision support with automated volumetric quantification of tumor burden on MRI in neuro-oncology and evaluated whether it enables more reproducible and standardized assessment of treatment response as compared to manual 2-dimensional measurements of tumor burden using the RANO criteria.

Methods

Study Design and Participants

This study was institutional review board-approved and informed consent was waived (S-784/2018). For the present study, a non-consecutive series of n=30 adult brain tumor patients (including n=15 glioblastoma WHO °IV and n=15 lower-grade gliomas, the latter encompassing n=2 IDH-mutant astrocytoma WHO °II, n=3 IDH-mutant astrocytoma WHO °III, n=8 IDH-mutant 1p/19q codeleted oligodendroglioma WHO °II and n=2 IDH-mutant 1p/19q codeleted oligodendroglioma WHO °III) previously treated at Heidelberg University Hospital. The selection of patients that were included for the present study was performed on consensus by three local investigators from Heidelberg University Hospital (P.V. , W.W. and M.B) aiming at representing different clinical scenarios from different disease stages in neuro-oncology. The MRI exams were acquired during the period of 09/2009 and 02/2019 with a standardized imaging protocol [7] and included 3D T1-weighted images before (T1-w) and after contrast agent administration (cT1-w) as well as axial 2D FLAIR and T2-weighted (T2-w) images as well as diffusion-weighted MRI with apparent diffusion coefficient (ADC) maps. To increase the diversity of the dataset, longitudinal MRI scans were selected from the primary treatment setting in 12 cases (with the post-radiation MRI scan used as the first imaging timepoint; except in patients that did not receive radiation therapy the post-surgery MRI scan was used as the first imaging timepoint) and the recurrent treatment setting in 18 cases (with the MRI scan prior to change of therapy as the first imaging timepoint). The last imaging timepoint was the MRI scan showing definitive tumor progression with subsequent change of treatment. Specifically, a median of 5 consecutive MRI scans (IQR, 5-10) were selected for each patient with a median of interval of 3.1 months (IQR, 2.5-3.6 months) between the scans. The interval between scans was significantly longer for lower-grade gliomas as compared to glioblastoma ($p<0.0001$) with 3.5 months (IQR, 3.0-5.6 months) for lower-grade gliomas and 2.8 months (IQR, 1.5-3.1 months) for glioblastoma. The **Supplementary Table 1** contains information on patient and treatment characteristics.

The 15 participating investigators were neuroradiologists (namely P.V., R.Y.H., F.B., J.E.P., Y.W.P., S.S.A., R.J., M.S., W.B.P., M.B.) or neuro-oncologists (namely *N.G., M.J.v.d.B, M.W., P.Y.W., W.W.*) and the majority (11/15, 73%) are active members of the RANO working group and/or the brain tumor group (BTG) from the European Organization for Research and Treatment of Cancer (EORTC). The investigators represented 11 institutions from 5 countries (Germany: n=4, USA: n=4, Netherlands: n=3, South Korea: n=3, Switzerland: n=1) and all of them are authors of this article. None of the investigators used AI-based decision support for assessment of treatment response in neuro-oncology prior to this study. Prior to the start of the study, all investigators reached consensus on the number and composition of patients to be included (i.e., n=30 patients including both lower-grade gliomas and glioblastomas from both primary and recurrent treatment settings). This consensus decision enabled that each of the 15 participating investigators would manage to interpret all cases in a reasonable timeframe while still including patients from a broad range of clinical scenarios.

Image interpretation

In the 1st assessment round of the study investigators were provided with the consecutive MRI scans (including T1-w, cT1-w, FLAIR, T2-w, DWI and ADC sequences for each timepoint) as well as relevant clinical information (integrated diagnosis according to the WHO 2016 classification of CNS tumors, current tumor-specific treatment and in case of recurrent tumors the number of recurrences) from each patient. The investigators received no information regarding at what timepoint when treatment was changed during the period covered by the available MRI scans.

All MRI scans were provided in DICOM format, stripped of patient information (with allocated patient identifiers being subject_01 to subject_30), and delivered to investigators. Investigators used their personal workstations with RadiAnt (Medixant, Poland) or OsiriX Lite (Pixmeo, Switzerland) as DICOM viewer. The investigators were asked to assess the timepoint of tumor progression on MRI in each patient by applying the 2D RANO concept with bi-perpendicular measurements of tumor burden as a general guide as outlined in the RANO criteria for high-grade[3] and lower-grade gliomas [4] and complete all reads at a typical clinical pace within a 2-month timeframe. The investigators received no feedback following submission of their readings. An illustrative case depicting the MRI sequences from consecutive timepoints during the 1st round of the assessment are shown in **Supplementary Figure 1**.

The 2nd assessment round of the study started after a 1-month wash out period. Patient identifiers within the image and clinical data were reordered to impede the re-use of the TTP assessment from the 1st round. The investigators received all the information from the 1st assessment round (MRI scans as well as relevant clinical information) as well as additional MRI sequences for each MRI scan generated using a previously developed and validated in-house AI-based processing pipeline, including deep-learning based skull stripping (https://github.com/NeuroAI-HD/HD-BET) and deep-learning based tumor segmentation (https://github.com/NeuroAI-HD/HD-GLIO) as core components [6,8]. To allow unbiased evaluation of the performance we did not perform any manual adjustments to the output of the AI-based processing pipeline (e.g., editing of tumor segmentation masks). Thereby for each MRI scan three additional MRI sequences were provided: (1-2) cT1-w and FLAIR sequences with color-coded overlays that indicate the contrast-enhancing and T2/FLAIR-hyperintense tumor identified by the AI-based processing pipeline, and (3) a DICOM sequence depicting a graph with the absolute and relative change in these tumor volumes over time (plotting contrast-enhancing and T2/FLAIR-hyperintense tumor volumes from the current and all previous MRI scans). Identical to the 1st assessment round the investigators were asked to assess the timepoint of tumor progression on MRI in each patient by incorporating this additional information and complete all reads at a typical clinical pace within a 2-month timeframe. As a general guideline, investigators were told that a 25% increase in the bi-perpendicular diameter would correspond to a 40% increase in the tumor volume (assuming spherical configuration of the tumor), [6] however final judgment (strict adherence to this volumetric threshold vs. subjective interpretation of the growth curve) was done at the discretion of the investigators. An illustrative case depicting the additional AI-enhanced MRI sequences from consecutive timepoints provided during the 2nd round of the assessment are shown in **Supplementary Figure 2**. Moreover, investigators were asked to record whether the additional information from the AI-based decision support was perceived as helpful or not (specified as “1” or “0”) for each of the assessed patients. The investigators received no feedback following submission of their readings. Subsequently, a questionnaire was circulated among the investigators to collect information about the years of their experience with neuro-oncology imaging.

Statistical analysis

All statistical analyses were performed with R version 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria). The agreement as well as disagreement in the individual readings for the timepoint of tumor progression between 1st and 2nd round of the assessment (overall 450 pairs i.e., from 30 patients * 15 investigators) were analyzed with descriptive metrics (absolute and relative agreement); differences were assessed with a 2-sample test for equality of proportions.

The time to progression (TTP) for each reading was calculated from the date of baseline MRI until the timepoint of tumor progression specified by the readings from each of the investigators. For those cases where the investigator did not judge tumor progression until the last MRI timepoint, an interval of 3 months (equivalent to one follow-up interval) was added as a workaround to the TTP measurement prior calculating the concordance correlation coefficient (CCC) and the standard deviation (SD) of the TTP measurements. This procedure allowed that none of the readings needed to be excluded when calculating these metrics while preserving statistical validity. The agreement of the TTP measurements between the investigators (separately for the 1st and 2nd round of the assessment) was evaluated for the whole patient cohort as well as for the glioblastoma and lower-grade glioma subgroups using the CCC [9]. The reported 95% confidence intervals (CI) were calculated using bootstrapping (with n=1000 iterations) with the bias-adjusted and accelerated bootstrap method. Empirical p-values were computed from the bootstrap distribution to assess differences between the CCC from the 1st and 2nd round for the whole patient cohort as well as for the glioblastoma and lower-grade glioma subgroups. The SD of the TTP measurements from all investigators was computed for each patient (separately for the 1st and 2nd round of the assessment) and was used as an additional metric beyond the CCC to evaluate agreement between investigators on a per-patient level. The reported 95% confidence intervals (CI) were calculated using bootstrapping (with n=1000 iterations) with the bias-adjusted and accelerated bootstrap method. A Pearson correlation test was used to evaluate the association (a) between the percentage of investigators judging AI-based decision support as helpful for assessing the TTP in patients and the standard deviation of the TTP measurements in the 2nd round of the assessment, as well as (b) between the percentage of patients where investigators judged AI-based decision as helpful for assessing the TTP and the experience of the investigators with neuro-oncology imaging. P-values <0.05 were considered significant.

Results

The CCC of TTP measurements between investigators was 0.77 (95% CI = 0.69 – 0.88) in the 1st round of the assessment without AI-based decision support and increased to 0.91 (95%CI = 0.82 – 0.95) with AI-based decision support (p=0.005) (**Figure 2**). This effect was more pronounced for patients with lower-grade gliomas, where the CCC was 0.70 (95% CI = 0.56 – 0.85) without AI-based decision support, as compared to 0.90 (95% CI = 0.76 – 0.95) with AI-based decision support (p=0.008). In contrast, for patients with glioblastoma the CCC was 0.83 (95% CI = 0.75 – 0.92) without AI-based decision support, as compared to 0.86 (95% CI = 0.78 – 0.93) with AI-based decision support (p=0.016). Similarly, the median SD for the TTP measurements between the investigators was 6.1 months (95% CI = 4.3 - 9.6 months) without AI-based decision support and decreased to 4.8 months (95% CI = 3.7 - 6.2 months) with AI-based decision support (p=0.004) (**Figure 3)**. Thereby a greater decrease in the SD when using additional AI-based decision support was observed for patients with lower-grade gliomas (-1.7 months [95% CI: -4.2 to -1.1 months]) as compared to glioblastoma (-0.1 months [95% CI: -0.5 to 0.0 months]) (p<0.001). Illustrative cases from two representative cases which demonstrate improved agreement in the TTP among investigators when using additional AI-based decision support are shown in **Figure 4 and 5 and Supplementary Figure 3**.

Comparison of all available pairs of TTP assessments from the 1st and the 2nd round of the assessment (450 pairs i.e., 30 patients x 15 raters) showed that the assessment performed with RANO alone were kept unchanged with additional AI-based decision support in 251 / 450 instances (56%) and were changed for the remaining 199 / 450 instances (44%) (**Supplementary Table 2**). Thereby, the probability of changing the TTP assessment with additional AI-based decision support was higher for the subset of patients with lower-grade gliomas (114 / 225 [51%]) as compared to glioblastoma (85 / 225 assessments [38%], p=0.008). The AI-based decision support did not systematically shift the judgment of tumor progression towards an earlier or later timepoint, instead the probability between shifting towards an earlier timepoint (105 / 450 instances [23%]) as compared towards shifting to a later timepoint (94 / 450 instances [21%]) with additional AI-based decision support was balanced (p=0.42).

The percentage of patients where individual investigators judged AI-based decision as helpful (median, 57% [IQR, 47-63%]) was negatively correlated with the experience of the investigators with neuro-oncology imaging (median of 19 years [IQR, 12-24 years]; Pearson correlation coefficient = -0.52; p = 0.02) i.e., investigators with less years of experience judged the AI-based decision support as more helpful (**Figure 6a**). Moreover, the percentage of investigators who judged the information provided through AI-based decision support as helpful for assessing the TTP in individual patients (median, 64% [IQR: 45-79%]) was negatively correlated with the SD of the TTP measurements in the 2nd of the assessment (Pearson correlation coefficient = -0.34; p = 0.03) i.e., the more investigators who judged the AI-based decision support to be helpful for a given patient, the better the agreement on TTP measurements for that patient (**Figure 6b**).

A fully functional version of AI-based processing pipeline that was used in the present study (illustrative case shown in the **Supplementary Figure 2**) is provided through https://github.com/NeuroAI-HD/HD-GLIO-XNAT as open-source and allows seamless manufacturer neutral integration into existing radiological infrastructures through the XNAT framework as a Container Service Plugin [10] (**Supplementary Figure 4**).

Discussion

The importance and meaningful clinical use of AI algorithms for automated quantification of tumor burden in neuro-oncology is reflected in the growing body of literature showing that accurate automated delineation of the various tumor sub-compartments can offer the basis for generating quantitative and reproducible imaging endpoints in neuro-oncology [5,6,11-13]. Specifically, AI algorithms for automated volumetric segmentation of tumor burden proved to be highly accurate with spatial overlap agreement between the predicted and the expert ground truth tumor annotation of more than 90% for the segmentation of contrast-enhancing tumor, as well as non-enhancing T2/FLAIR signal abnormality [5,6,11], even when applying the AI algorithm to unseen data from a multicenter phase II/III trial [6]. The findings from the present study now provide additional evidence regarding the clinical value of AI-based decision support towards establishing high-quality imaging endpoints in neuro-oncology. Specifically, we demonstrate within the setting of an international multi-reader study with 15 investigators that automated AI-based volumetric quantification of tumor burden allows to improve the reproducibility and agreement of tumor response assessment measurements as compared to standard RANO criteria within a simulated clinical setting. We demonstrate that particularly lower-grade gliomas where reliable assessment of the tumor progression may be challenging due to their slow growing nature of these tumors may benefit from AI-based decision support with a potentially clinically meaningful and relevant decrease in the SD (by a median of 1.7 months) of the TTP measurements between the investigators. In contrast, when investigators used the AI-based decision support in patients with glioblastoma, there was comparatively less impact on the reproducibility of tumor response assessment. Potentially, this may reflect that tumor growth dynamics are comparatively more robust to discern when assessing tumors with a faster growth trajectory, thereby limiting the impact of AI-based decision support.

The principal benefits of AI-based decision support may be useful not only in a routine clinical scenario, but especially in the context of clinical trials, where the assessment of treatment efficacy on MRI is – besides overall survival – a key endpoint for the approval of new treatment concepts. Therefore, blinded central assessment of treatment efficacy by independent radiologists is frequently requested by regulatory authorities [14] to mitigate over- or under-estimation of the true effect of treatments (i.e., systematic bias) when only relying on the local RANO readings where investigators are not blinded to the patients' treatment assignments and clinical information [15]. Moreover, central RANO reading by expert radiologists is labor and time intensive and thus increases the burden of time and expenditure on clinical trials. Consequently, AI algorithms for automated volumetric delineation of tumor burden and tumor response assessment may assist investigators during central reading of the imaging data to yield high-quality imaging endpoints in neuro-oncology. As part of this study, we provide a fully functional version of the AI-based processing pipeline as open-source, enabling seamless manufacturer neutral integration into existing radiological infrastructures through the XNAT framework [10] (**Supplementary Figure 4**) and thus may hold great promise for enhancing future research efforts in the field of neuro-oncology.

Our study also demonstrates that the information provided through the AI-based decision support is perceived as more helpful by comparatively less experienced investigators. Moreover, perceiving AI-based decision support as helpful by a greater number of investigators for determining the TTP in a specific patient, directly translated into a better agreement in the TTP measurements for this patient. Both findings taken together, highlight (a) that the confidence and validity of tumor response assessment readings could be augmented through AI-based decision support especially for less experienced investigators, and (b) that investigators were able to readily identify appropriate cases where AI-based decision support is helpful and thereby leading to a more reproducible assessment of treatment efficacy in neuro-oncology.

Our study has some limitations. First, we acknowledge the retrospective nature of the study and the selection of a non-consecutive patient series. However, we aimed to simulate a realistic clinical scenario including patients with different tumor subtypes from both primary and recurrent treatment situations and a broad range of treatment scenarios. Although further validation in a prospective clinical scenario is needed to better establish the value of AI-based decision support and to specifically assess whether more reliable surrogate endpoints can be obtained from MRI, it may be challenging to adopt a rigorous prospective multi-reader design with/without AI-based decision support as performed in the present study.

Second, the AI-based processing pipeline applied in the present study makes use of our previously trained and validated artificial neural networks for automated skull-stripping [8] and automated tumor segmentation which has been developed using >3000 MRI examinations from >1400 brain tumor patients [6]. However, the potential underrepresentation of atypical or particularly challenging in these data used for training the artificial neural networks may affect the performance in a real-world clinical scenario and potentially lead to false-positive or false-negative detection of tumor burden. Consequently, this may have negatively affecting the perceived usefulness of the AI-based decision support among the investigators in the present study. Although data sharing initiatives with public deposition of annotated cases (e.g. through collaborative efforts such as the Cancer Genome Imaging Archive [TCIA] or the Brain Tumor Segmentation Challenge [BraTS] for gliomas [11,16]) is a crucial first step to address this limitation, medical data privacy regulations often pose a significant challenge towards establishing a centralized data repository [17]. Recent technical developments in the field of AI, specifically federated learning which allows multiple healthcare institutions to share their data to train an AI model while still guaranteeing medical data privacy, aim to address this challenge [17-20].

Third, the differentiation of T2/FLAIR hyperintensities as well as contrast-enhancing lesions during the follow-up into treatment or tumor-related changes, may still be a challenge in the field of neuro-oncology, particularly with treatment concepts that incorporate immunotherapies or anti-angiogenic drugs [21,22]. Consequently, the future incorporation of advanced MRI modalities such as diffusion or perfusion-weighted imaging [23,24] or metabolic imaging with radiolabeled molecules from positron emission tomography (PET) [25] will be important to overcome limitations of structural MRI and may allow to further optimize the clinical value of the AI-based decision support applied in the present study.

In conclusion, AI-based decision support has the potential to yield more reproducible and standardized assessment of treatment response in neuro-oncology as compared to manual 2-dimensional measurements of tumor burden. Particularly the evaluation of patients with lower-grade gliomas where reliable assessment of the TTP may be challenging due to their slow growing nature of these tumors may benefit from AI-based decision support. To enhance future research efforts in the field of neuro-oncology imaging, we provide a fully functional version of the AI-based processing pipeline as open-source which can readily be integrated into existing radiological (research) infrastructures.

Funding

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 404521405, SFB 1389 - UNITE Glioblastoma, Work Package C02 and Priority Programme 2177 “Radiomics: Next Generation of Biomedical Imaging”.

Acknowledgments

None.

References

1. O'Connor JP, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. *Nature reviews. Clinical oncology.* 2017; 14(3):169-186.

2. Wen PY, Chang SM, Van den Bent MJ, Vogelbaum MA, Macdonald DR, Lee EQ. Response Assessment in Neuro-Oncology Clinical Trials. *J Clin Oncol.* 2017; 35(21):2439-2449.

3. Wen PY, Macdonald DR, Reardon DA, et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J Clin Oncol.* 2010; 28(11):1963-1972.

4. van den Bent MJ, Wefel JS, Schiff D, et al. Response assessment in neuro-oncology (a report of the RANO group): assessment of outcome in trials of diffuse low-grade gliomas. *Lancet Oncol.* 2011; 12(6):583-593.

5. 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.

6. Kickingereder P, Isensee F, Tursunova I, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* 2019; 20(5):728-740.

7. Ellingson BM, Bendszus M, Boxerman J, et al. Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials. *Neuro Oncol.* 2015; 17(9):1188-1198.

8. Isensee F, Schell M, Pflueger I, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Human brain mapping.* 2019.

9. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989; 45(1):255-268.

10. Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics.* 2007; 5(1):11-34.

11. Bakas S, Reyes M, Jakab A, et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *arXiv e-prints.* 2018. https://ui.adsabs.harvard.edu/\#abs/2018arXiv181102629B. Accessed November 01, 2018.

12. Peng J, Kim DD, Patel JB, et al. Deep Learning-Based Automatic Tumor Burden Assessment of Pediatric High-Grade Gliomas, Medulloblastomas, and Other Leptomeningeal Seeding Tumors. *Neuro Oncol.* 2021.

13. Baid U, Ghodasara S, Mohan S, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314.* 2021.

14. Food U, Administration D. Guidance for industry: clinical trial endpoints for the approval of cancer drugs and biologics. *Federal Register.* 2007; 72.

15. Ford R, Schwartz L, Dancey J, et al. Lessons learned from independent central review. *Eur J Cancer.* 2009; 45(2):268-274.

16. Bakas S, Akbari H, Sotiras A, et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data.* 2017; 4:170117.

17. Zerka F, Barakat S, Walsh S, et al. Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. *JCO Clin Cancer Inform.* 2020; 4:184-200.

18. Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation2019; Cham.

19. Rieke N, Hancox J, Li W, et al. The Future of Digital Health with Federated Learning. *arXiv preprint arXiv:2003.08119.* 2020.

20. Pati S, Baid U, Edwards B, et al. Federated Learning Enables Big Data for Rare Cancer Boundary Detection. *arXiv preprint arXiv:2204.10836.* 2022.

21. Okada H, Weller M, Huang R, et al. Immunotherapy response assessment in neuro-oncology: a report of the RANO working group. *Lancet Oncol.* 2015; 16(15):e534-e542.

22. Ellingson BM, Wen PY, Cloughesy TF. Modified Criteria for Radiographic Response Assessment in Glioblastoma Clinical Trials. *Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics.* 2017; 14(2):307-320.

23. Kickingereder P, Park JE, Boxerman JL. Advanced Physiologic Imaging: Perfusion – Theory and Applications. In: Pope WB, ed. *Glioma Imaging: Physiologic, Metabolic, and Molecular Approaches*. Cham: Springer International Publishing; 2020:61-91.

24. LaViolette PS. Advanced Physiologic Imaging: Diffusion – Theory and Applications. In: Pope WB, ed. *Glioma Imaging: Physiologic, Metabolic, and Molecular Approaches*. Cham: Springer International Publishing; 2020:93-108.

25. Galldiks N, Lohmann P, Albert NL, Tonn JC, Langen K-J. Current status of PET imaging in neuro-oncology. *Neuro-Oncology Advances.* 2019; 1(1).

Figure Captions

Figure 1. Use of automated AI-based volumetric quantification of tumor burden to overcome the interrater variability of RANO measurements of tumor diameters towards a more standardized & reproducible assessment of treatment efficacy in neuro-oncology.

Figure 2. Concordance correlation coefficients (CCC) of tumor response assessment between investigators in the 1st round of the study without AI-based decision support (red colored) and the 2nd round of the study with AI-based decision support (green colored). The central line of the boxplot denotes the median and the edges of the boxplot denote the first and the third quartile of the bootstrap distribution of the CCC. The lines extending from the boxes (whiskers) indicating variability outside the upper and lower quartiles. The outliers are denoted by black dots at the end of the whisker lines.

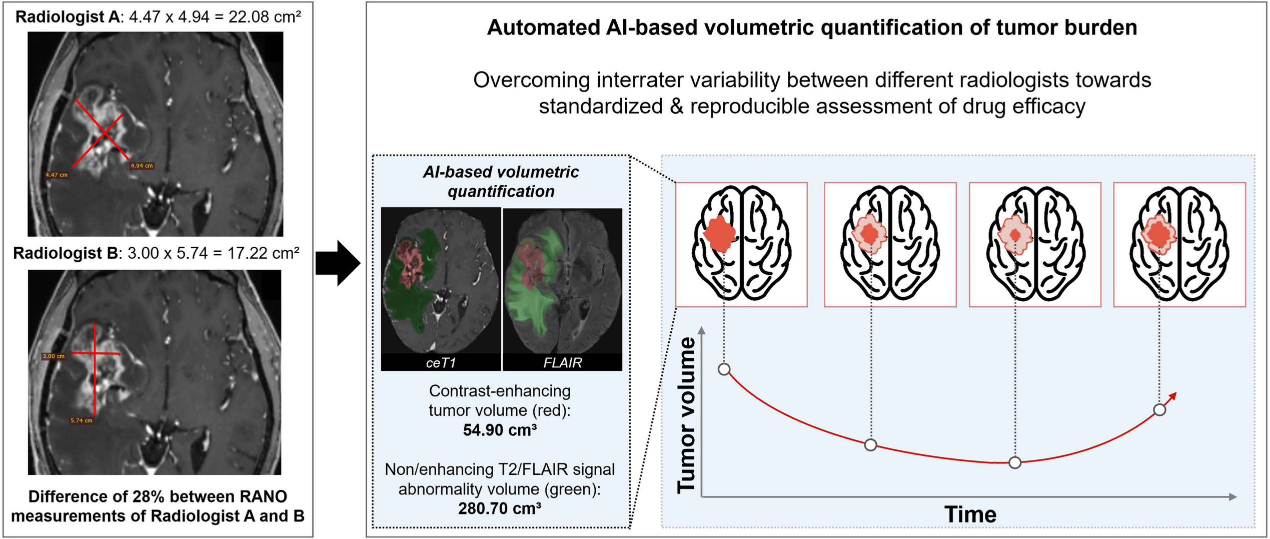
Figure 3. Standard deviation (SD) of tumor response assessment between investigators in the 1st round of the study without AI-based decision support (red colored) and the 2nd round with AI-based decision support (green colored). The difference in the SD between the 1st and the 2nd round is shown in blue.

Figure 4. Illustrative case (patient #17, oligodendroglioma WHO°III) depicting the change in tumor burden over time on cT1-w and FLAIR sequences (1st and 2nd row). The cT1-w overlay and FLAIR overlay sequences (3rd and 4th row) as well as the corresponding tumor volume plot were provided in the 2nd round of the assessment and visualize the contrast-enhancing tumor volumes (red) and T2-w/FLAIR abnormality volumes (green) which were automatically generated by the AI-based decision support for each timepoint. The last row visualizes the time to progression (TTP) measurements from the 15 investigators based on RANO alone (1st round; blue colored boxplot) vs. additional AI-based decision support (2nd round; purple colored boxplot). The boxplots demonstrate higher agreement of the TTP measurements from the 15 investigators with additional AI-based decision support.

Figure 5. Illustrative case (patient #18, astrocytoma WHO°III) depicting the change in tumor burden over time on cT1-w and FLAIR sequences (1st and 2nd row). The cT1-w overlay and

FLAIR overlay sequences (3rd and 4th row) as well as the corresponding tumor volume plot were provided in the 2nd round of the assessment and visualize the contrast-enhancing tumor volumes (red) and T2-w/FLAIR abnormality volumes (green) which were automatically generated by the AI-based decision support for each timepoint. The last row visualizes the time to progression (TTP) measurements from the 15 investigators based on RANO alone (1st round; blue colored boxplot) vs. additional AI-based decision support (2nd round; purple colored boxplot). The boxplots demonstrate higher agreement of the TTP measurements from the 15 investigators with additional AI-based decision support.

Figure 6. (A) Correlation between the percentage of investigators judging AI-based decision support as helpful for assessing the TTP in individual patients and the standard deviation of the corresponding TTP measurements in round 2 (RANO+AI). **(B)** Correlation between the percentage of patients where investigators judged AI-based decision as helpful for assessing the TTP and the corresponding experience of the investigators with neuro-oncology imaging.



Radiologist A: 4.47 x 4.94 = 22.08 cm²

Radiologist B: 3.00 x 5.74 = 17.22 cm²

Difference of 28% between RANO measurements of Radiologist A and B

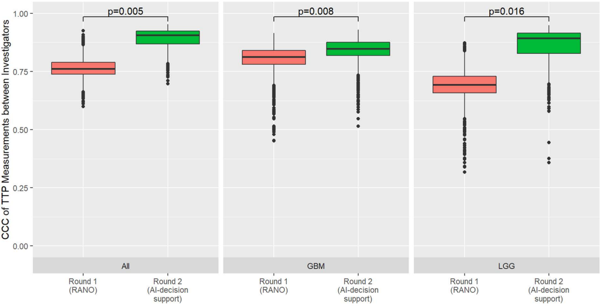
Automated AI-based volumetric quantification of tumor burden

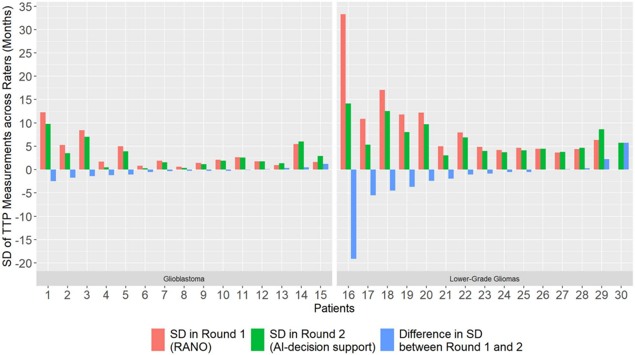
Overcoming interrater variability between different radiologists towards standardized & reproducible assessment of drug efficacy

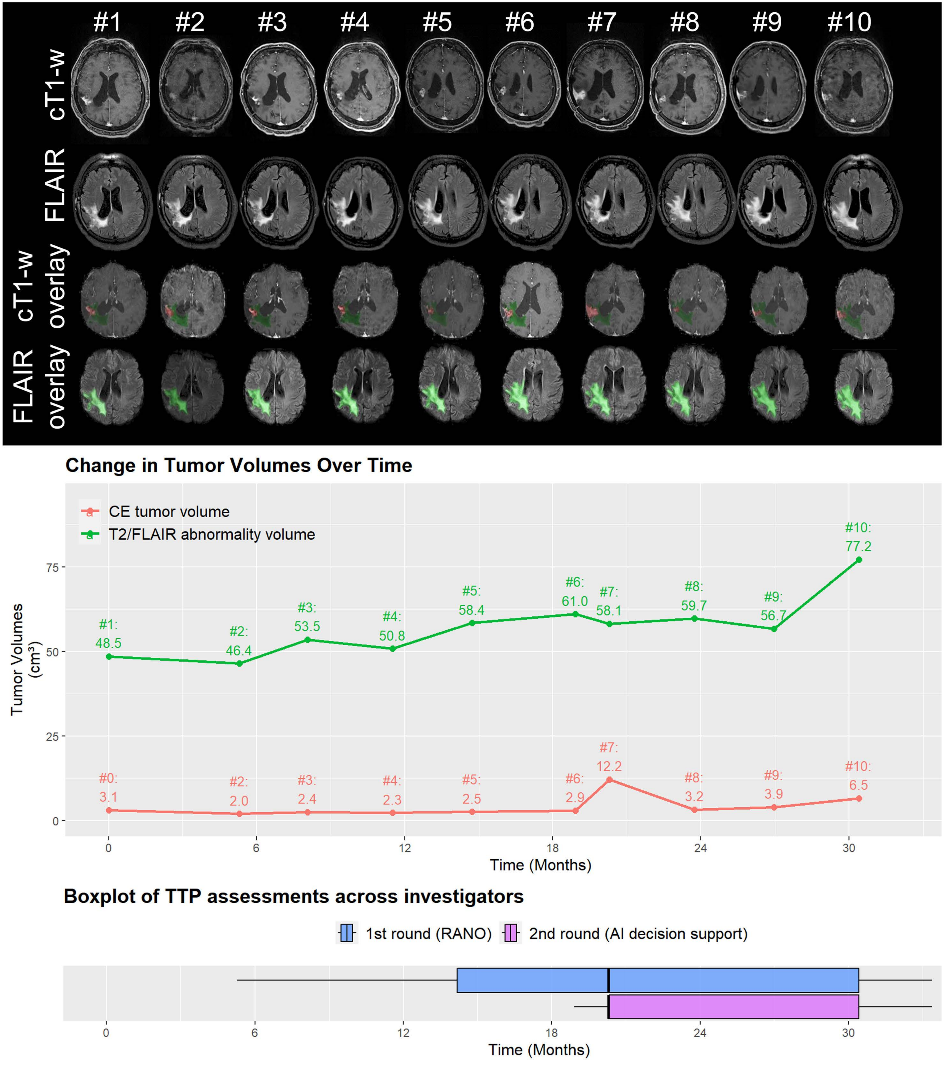
AI-based volumetric quantification

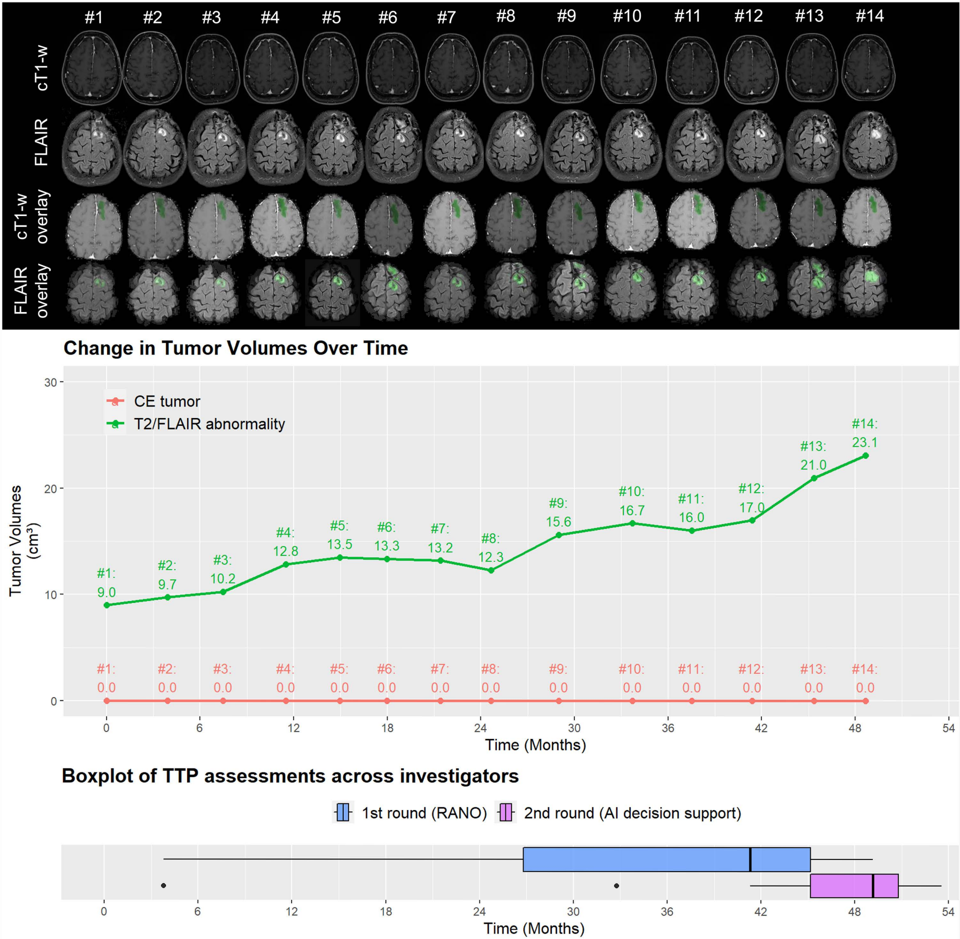
Contrast-enhancing tumor volume (red): **54.90 cm³**

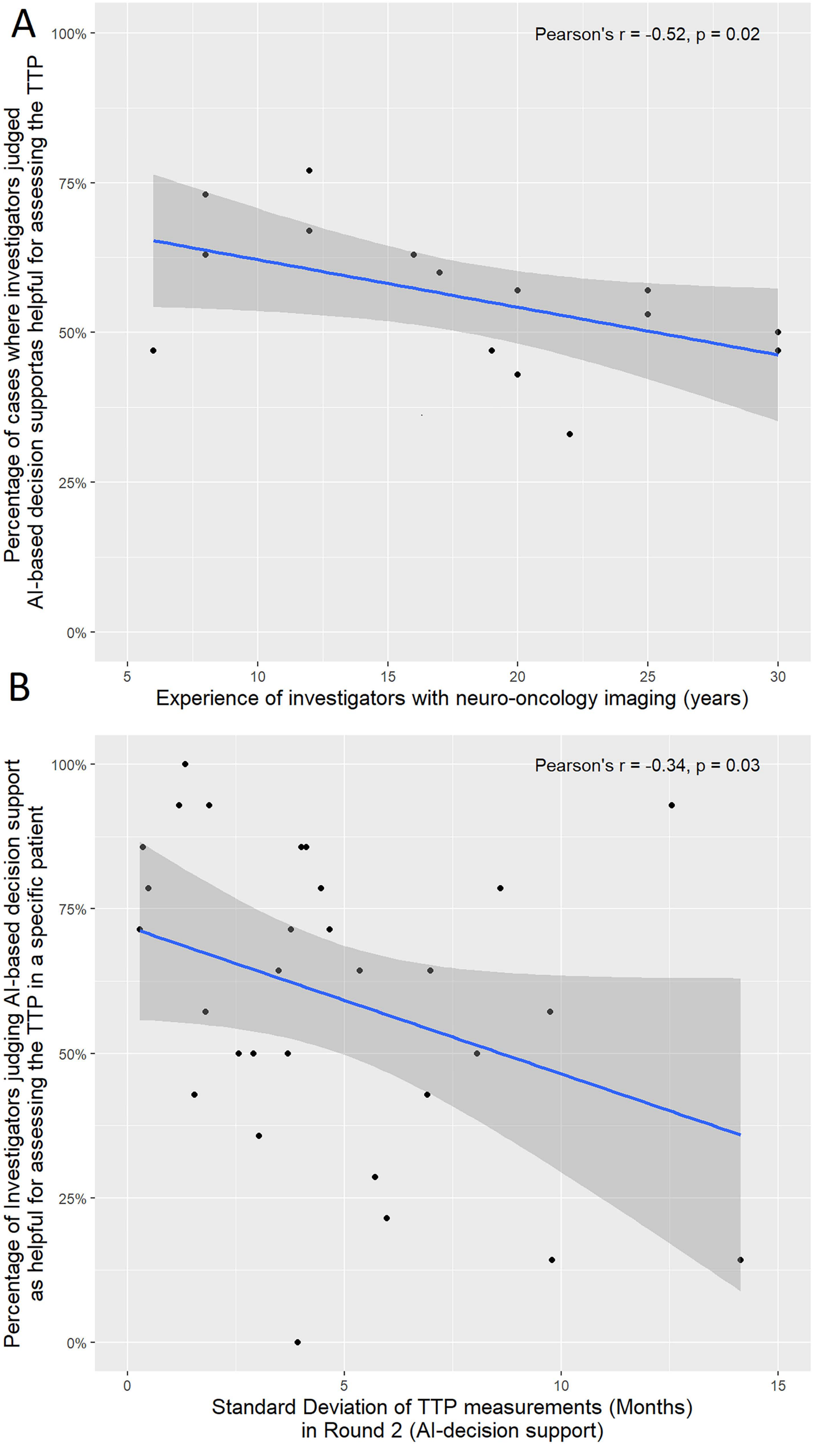
Non/enhancing T2/FLAIR signal abnormality volume (green): **280.70 cm³**


p=0.005
p=0.008
p=0.016
CCC of TTP Measurements between Investigators
1.00
0.75
0.50
0.25
0.00
All
GBM
LGG
Round 1 (RANO)
Round 2 (AI-decision support)
Round 1 (RANO)
Round 2 (AI-decision support)
Round 1 (RANO)
Round 2 (AI-decision support)


SD of TTP Measurements across Raters (Months)
35
30
25
20
15
10
5
0
-5
-10
-15
-20
Glioblastoma
Lower-Grade Gliomas
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
Patients
SD in Round 1
(RANO)
SD in Round 2
(AI-decision support)
Difference in SD
between Round 1 and 2


#1
#2
#3
#4
#5
#6
#7
#8
#9
#10
cT1-w
FLAIR
cT1-w overlay
FLAIR overlay
Change in Tumor Volumes Over Time
CE tumor volume
T2/FLAIR abnormality volume
Tumor Volumes (cm^3)
75
50
25
0
#1: 48.5
#2: 46.4
#3: 53.5
#4: 50.8
#5: 58.4
#6: 61.0
#7: 58.1
#8: 59.7
#9: 56.7
#10: 77.2
#0: 3.1
#2: 2.0
#3: 2.4
#4: 2.3
#5: 2.5
#6: 2.9
#7: 12.2
#8: 3.2
#9: 3.9
#10: 6.5
0
6
12
18
24
30
Time (Months)
Boxplot of TTP assessments across investigators
1st round (RANO)
2nd round (AI decision support)
0
6
12
18
24
30
Time (Months)


#1
#2
#3
#4
#5
#6
#7
#8
#9
#10
#11
#12
#13
#14
cT1-w
FLAIR
cT1-w overlay
FLAIR overlay
Change in Tumor Volumes Over Time
CE tumor
T2/FLAIR abnormality
Tumor Volumes (cm³)
30
20
10
0
#1: 9.0
#2: 9.7
#3: 10.2
#4: 12.8
#5: 13.5
#6: 13.3
#7: 13.2
#8: 12.3
#9: 15.6
#10: 16.7
#11: 16.0
#12: 17.0
#13: 21.0
#14: 23.1
#1: 0.0
#2: 0.0
#3: 0.0
#4: 0.0
#5: 0.0
#6: 0.0
#7: 0.0
#8: 0.0
#9: 0.0
#10: 0.0
#11: 0.0
#12: 0.0
#13: 0.0
#14: 0.0
0
6
12
18
24
30
36
42
48
54
Time (Months)
Boxplot of TTP assessments across investigators
1st round (RANO)
2nd round (AI decision support)
0
6
12
18
24
30
36
42
48
54
Time (Months)


A
Pearson's r = -0.52, p = 0.02
Percentage of cases where investigators judged
AI-based decision supportas helpful for assessing the TTP
100%
75%
50%
25%
0%
5
10
15
20
25
30
Experience of investigators with neuro-oncology imaging (years)
B
Pearson's r = -0.34, p = 0.03
Percentage of Investigators judging AI-based decision support
as helpful for assessing the TTP in a specific patient
100%
75%
50%
25%
0%
0
5
10
15
Standard Deviation of TTP measurements (Months)
in Round 2 (AI-decision support)

Data Supplement

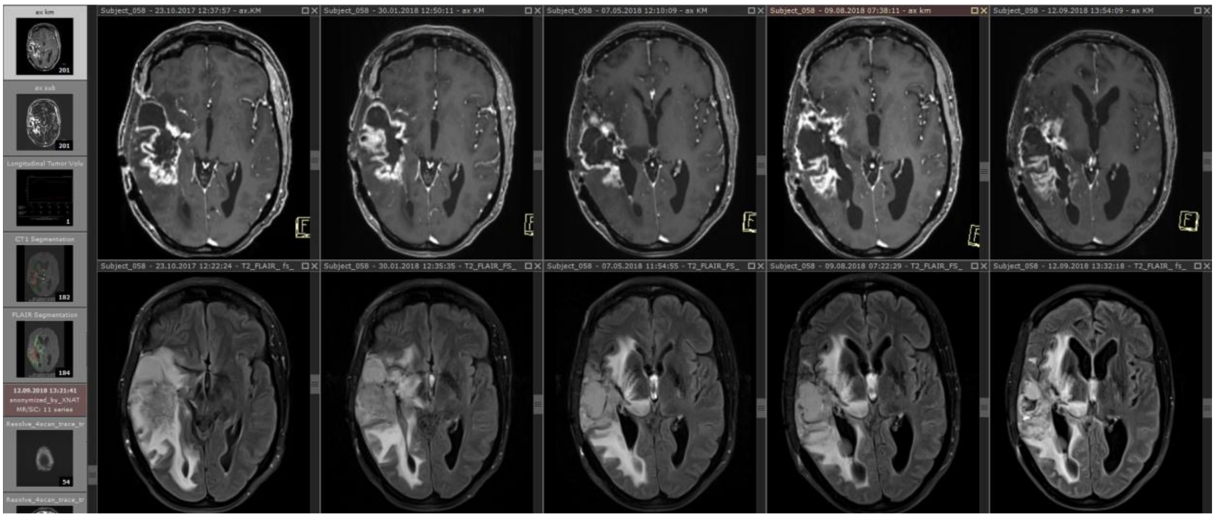
Supplementary Table 1. Characteristics of the patients (integrated diagnosis of the glioma subtype, timepoint of the disease from where MRI scans have been included, current treatment during this timeframe and number of MRI scans sent to the investigators).

ID	Integrated Diagnosis	Timepoint	Current Treatment	No. of MRI scans
1	Glioblastoma (IDH-wt)	newly diagnosed	adjuvant TMZ	19
2	Glioblastoma (IDH-wt)	1st recurrence	bevacizumab + lomustine	9
3	Glioblastoma (IDH-mut)	newly diagnosed	adjuvant TMZ	13
4	Glioblastoma (IDH-wt, MGMT-meth)	newly diagnosed	adjuvant TMZ + Nivolumab	10
5	Glioblastoma (IDH-wt)	1st recurrence	TTF (Tumour Treating Fields)	6
6	Glioblastoma (IDH-wt)	newly diagnosed	adjuvant TMZ	5
7	Glioblastoma (IDH-wt)	1st recurrence	lomustine	5
8	Glioblastoma (IDH-wt)	newly diagnosed	adjuvant TMZ	4
9	Glioblastoma (IDH-wt)	newly diagnosed	adjuvant TMZ	5
10	Glioblastoma (IDH-wt, MGMT-unmeth)	newly diagnosed	adjuvant TMZ	5
11	Glioblastoma (IDH-wt, MGMT-unmeth)	1st recurrence	bevacizumab + lomustine	5
12	Glioblastoma (IDH-wt, MGMT-unmeth)	4th recurrence	bevacizumab	4
13	Glioblastoma (IDH-wt, MGMT-meth)	2nd recurrence	CCNU	4
14	Glioblastoma (IDH-wt, MGMT-meth)	newly diagnosed	adjuvant TMZ	8
15	Glioblastoma (IDH-wt, MGMT-meth)	1st recurrence	bevacizumab + lomustine	6
16	Oligodendroglioma (WHO °II, IDH-mut, 1p19q codel)	1st recurrence	proton beam irradiation	22
17	Oligodendroglioma (WHO °III, IDH-mut, 1p19q codel)	3rd recurrence	CCNU	10
18	Astrocytoma (WHO °III, IDH-mut)	1st recurrence	TMZ	14
19	Oligodendroglioma (WHO °II, IDH-mut, 1p19q codel)	newly diagnosed	no therapy after resection	12
20	Astrocytoma (WHO °III, IDH-mut)	newly diagnosed	TMZ + peptide vaccine	10
21	Oligodendroglioma (WHO °II, IDH-mut, 1p19q codel)	2nd recurrence	TMZ	6
22	Oligodendroglioma (WHO °II, IDH-mut, 1p19q codel)	newly diagnosed	no therapy after resection	8
23	Oligodendroglioma (WHO °II, IDH-mut, 1p19q codel)	newly diagnosed	TMZ	13
24	Oligodendroglioma (WHO °II, IDH-mut, 1p19q codel)	1st recurrence	proton beam irradiation	5
25	Oligodendroglioma (WHO °II, IDH-mut, 1p19q codel)	3rd recurrence	proton beam irradiation	5
26	Astrocytoma (WHO °II, IDH-mut)	2nd recurrence	CCNU/VP16	6
27	Oligodendroglioma (WHO °III, IDH-mut, 1p19q codel)	2nd recurrence	PCV	6
28	Astrocytoma (WHO °II, IDH-mut)	5th recurrence	bevacizumab	7
29	Oligodendroglioma (WHO °II, IDH-mut, 1p19q codel)	1st recurrence	adjuvant TMZ	6
30	Astrocytoma (WHO °III, IDH-mut)	3rd recurrence	TMZ	6

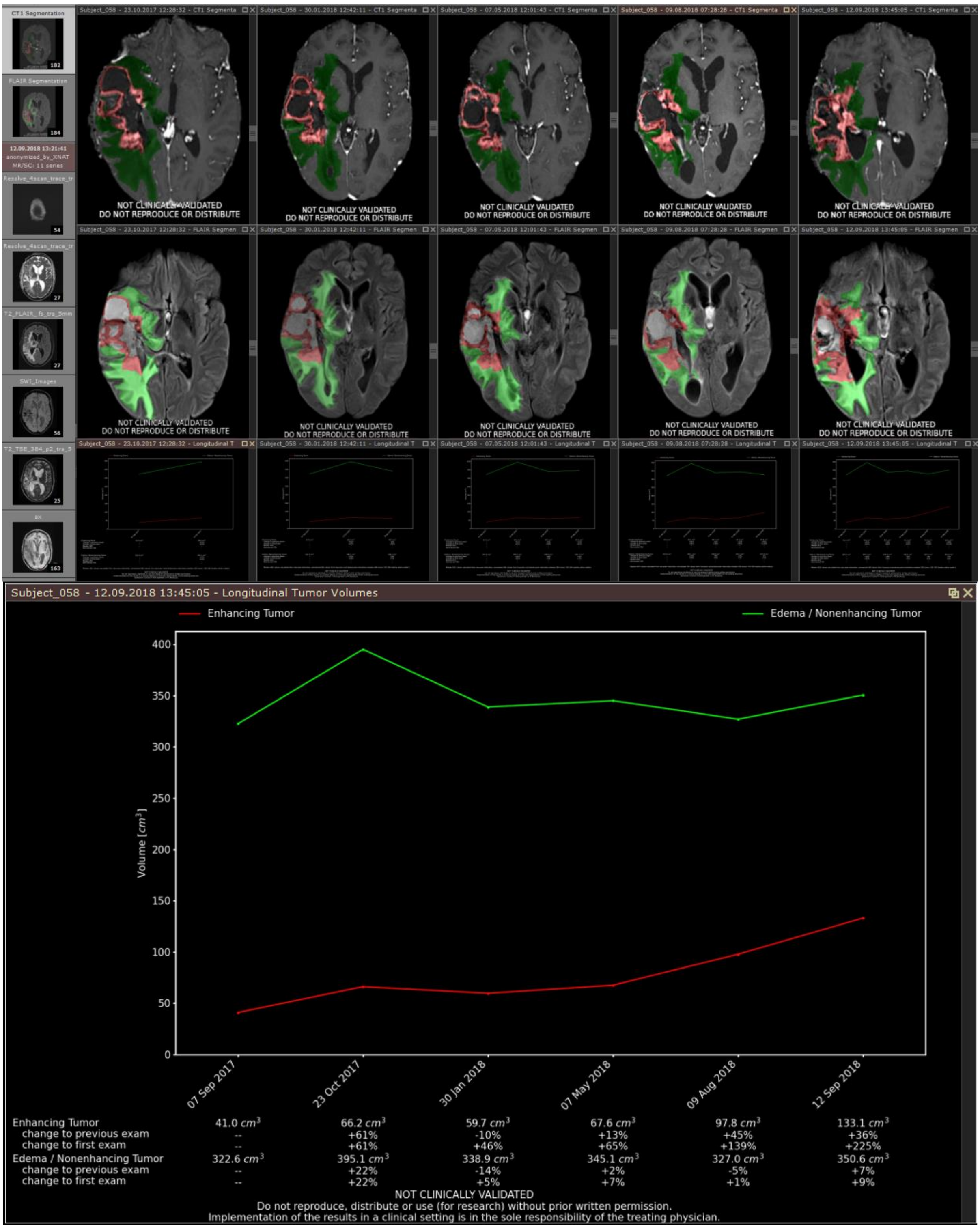
Supplementary Table 2. Details on agreement and disagreement rates across all the time to progression readings (total of 450 readings i.e., readings from 30 patients by 15 investigators) performed during the 1st round with RANO alone and 2nd round with additional AI-based decision support.

	All (n=450)		**LGG (n=225)**		**GBM (n=225)**	
	n	(%)	n	(%)	n	(%)
Absolute agreement in TTP values between 1st and 2nd round	251	56%	111	49%	140	62%
Disagreement in TTP values between 1st and 2nd round	199	44%	114	51%	85	38%
Shorter TTP with AI-based decision support (as compared to RANO alone)	105	23%	55	24%	50	22%
Tumor progression only identified with AI-based decision support (but not with RANO alone)	17	4%	8	4%	9	4%
Longer TTP with AI-based decision support (as compared to RANO alone)	94	21%	59	26%	35	16%
Tumor progression only identified with RANO alone (but not with additional AI-based decision support)	17	4%	10	4%	7	3%

Supplementary Figure 1. Screenshot of DICOM images provided during the 1st round of the assessment, depicting consecutive MRI scans in a representative patient with 2nd recurrence of an (initially diagnosed) IDH-mutant astrocytoma WHO °II. The investigators evaluated the timepoint of tumor progression based on the response assessment in neuro-oncology (RANO) criteria (exemplarily shown here are the post-contrast T1-weighted sequences in the 1st row and fluid attenuated inversion recovery [FLAIR] sequences in the 2nd row from 5 consecutive timepoints).


ax km
201
ax sub
201
Longitudinal Tumor Volu
1
CT1 Segmentation
182
FLAIR Segmentation
184
12.09.2018 13:21:41
anonymized_by_XNAT
MR/SC: 11 series
Resolve_4scan_trace_tr
54
Resolve_4scan_trace_tr
Subject_058 - 23.10.2017 12:37:57 - ax.KM
Subject_058 - 30.01.2018 12:50:11 - ax KM
Subject_058 - 07.05.2018 12:10:09 - ax KM
Subject_058 - 09.08.2018 07:38:11 - ax km
Subject_058 - 12.09.2018 13:54:09 - ax KM
Subject_058 - 23.10.2017 12:22:24 - T2_FLAIR_ fs_
Subject_058 - 30.01.2018 12:35:35 - T2_FLAIR_FS_
Subject_058 - 07.05.2018 11:54:55 - T2_FLAIR_FS_
Subject_058 - 09.08.2018 07:22:29 - T2_FLAIR_FS_
Subject_058 - 12.09.2018 13:32:18 - T2_FLAIR_ fs_
F
F
F
F
F

Supplementary Figure 2. Screenshot of DICOM images that were additionally provided during the 2nd round of the assessment for each of the MRI scans (same patient & timepoints as shown in Supplementary Figure 1). Specifically, for each MRI scan three additional MRI sequences were provided: skull-stripped and co-registered post-contrast T1-weighted sequences (1st row) and fluid attenuated inversion recovery (FLAIR) sequences (2nd row) with color-coded overlays that indicate the contrast-enhancing and T2/FLAIR-hyperintense tumor volumes (red and green colored) identified by the AI-based processing pipeline, and a DICOM sequence depicting a graph with the absolute and relative change in these tumor volumes over time (3rd row; plotting contrast-enhancing and T2/FLAIR-hyperintense tumor volumes from the current and all previous MRI scans; a magnified version of the graph from the last MRI scan is shown in the 4th row).


CT1 Segmentation
FLAIR Segmentation
12.09.2018 13:21:41
anonymized_by_XNAT
MR/SC: 11 series
Subject_058 - 23.10.2017 12:28:32 - CT1 Segmenta
Subject_058 - 30.01.2018 12:42:11 - CT1 Segmenta
Subject_058 - 07.05.2018 12:01:43 - CT1 Segmenta
Subject_058 - 09.08.2018 07:28:28 - CT1 Segmenta
Subject_058 - 12.09.2018 13:45:05 - CT1 Segmenta
Subject_058 - 23.10.2017 12:28:32 - FLAIR Segmen
Subject_058 - 30.01.2018 12:42:11 - FLAIR Segmen
Subject_058 - 07.05.2018 12:01:43 - FLAIR Segmen
Subject_058 - 09.08.2018 07:28:28 - FLAIR Segmen
Subject_058 - 12.09.2018 13:45:05 - FLAIR Segmen
Subject_058 - 23.10.2017 12:28:32 - Longitudinal T
Subject_058 - 30.01.2018 12:42:11 - Longitudinal T
Subject_058 - 07.05.2018 12:01:43 - Longitudinal T
Subject_058 - 09.08.2018 07:28:28 - Longitudinal T
Subject_058 - 12.09.2018 13:45:05 - Longitudinal T
NOT CLINICALLY VALIDATED
DO NOT REPRODUCE OR DISTRIBUTE
Subject_058 - 12.09.2018 13:45:05 - Longitudinal Tumor Volumes
Enhancing Tumor
Edema / Nonenhancing Tumor
Volume [cm^3]
0
50
100
150
200
250
300
350
400
07 Sep 2017
23 Oct 2017
30 Jan 2018
07 May 2018
09 Aug 2018
12 Sep 2018

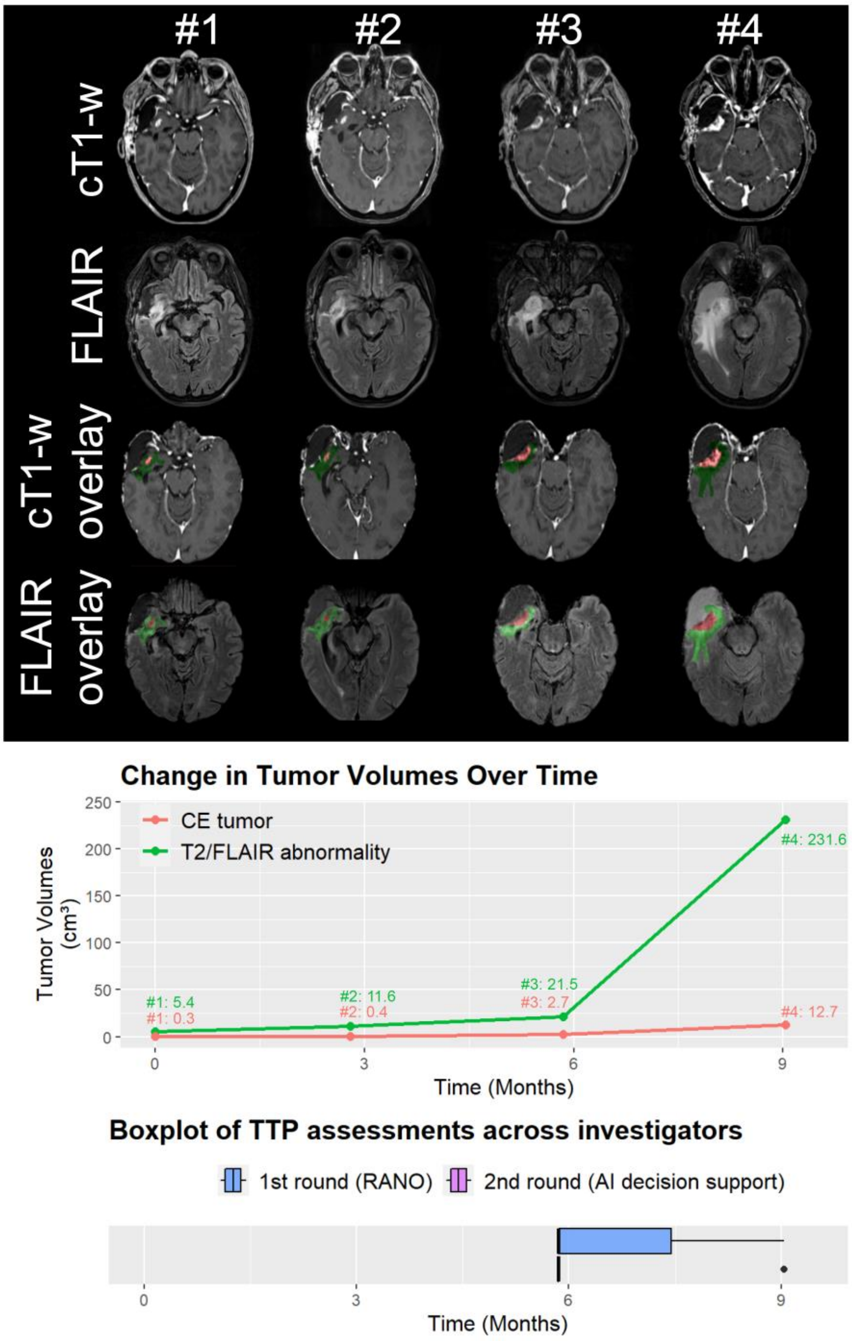
	07 Sep 2017	23 Oct 2017	30 Jan 2018	07 May 2018	09 Aug 2018	12 Sep 2018
Enhancing Tumor	41.0 cm^3	66.2 cm^3	59.7 cm^3	67.6 cm^3	97.8 cm^3	133.1 cm^3
change to previous exam	--	+61%	-10%	+13%	+45%	+36%
change to first exam	--	+61%	+46%	+65%	+139%	+225%
Edema / Nonenhancing Tumor	322.6 cm^3	395.1 cm^3	338.9 cm^3	345.1 cm^3	327.0 cm^3	350.6 cm^3
change to previous exam	--	+22%	-14%	+2%	-5%	+7%
change to first exam	--	+22%	+5%	+7%	+1%	+9%

NOT CLINICALLY VALIDATED

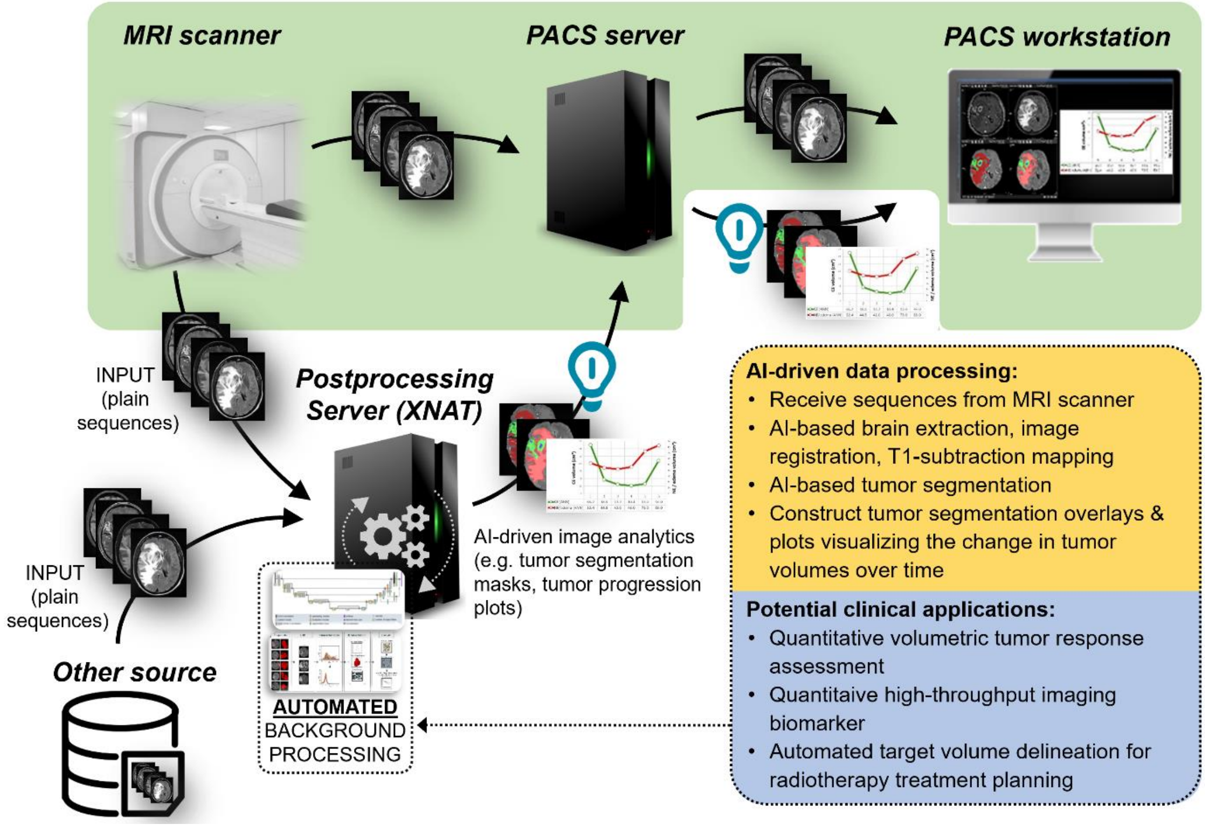
Do not reproduce, distribute or use (for research) without prior written permission.

Implementation of the results in a clinical setting is in the sole responsibility of the treating physician.

Supplementary Figure 3. Illustrative case (patient #8, glioblastoma IDH-wildtype) depicting the change in tumor burden over time on cT1-w and FLAIR sequences (1st and 2nd row). The cT1-w overlay and FLAIR overlay sequences (3rd and 4th row) as well as the corresponding tumor volume plot were provided in the 2nd round of the assessment and visualize the contrast-enhancing tumor volumes (red) and T2-w/FLAIR abnormality volumes (green) which were automatically generated by the AI-based decision support for each timepoint. The last row visualizes the time to progression (TTP) measurements from the 15 investigators based on RANO alone (1st round; blue colored boxplot) vs. additional AI-based decision support (2nd round; purple colored boxplot). The boxplots demonstrate higher agreement of the TTP measurements from the 15 investigators with additional AI-based decision support. Specifically, with additional AI-based decision support 14/15 investigators selected timepoint #3 for tumor progression, which demonstrated an increase in the contrast-enhancing tumor volume from 0.3 cm³ (manually measured biperpendicular diameter: 9 x 5 mm) at baseline to 2.7 cm³ (manually measured biperpendicular diameter: 16 x 8 mm) at timepoint #3, i.e., corresponding to a volume increase of +900%. Without AI-based decision support, only 11/15 investigators selected timepoint #3 for tumor progression, whereas the remaining 4/15 investigators selected timepoint #4 for tumor progression, where contrast-enhancing tumor volume was 12.7 cm³ (manually measured biperpendicular diameter: 22 x 11 mm), i.e., corresponding to a volume increase of +4200% as compared to baseline).



Supplementary Figure 4. Workflow of how the AI-based processing pipeline can be integrated into existing radiological infrastructures through the Extensible Neuroimaging Archive Toolkit (XNAT) framework as a docker container service plugin.



AI-driven image analytics (e.g. tumor segmentation masks, tumor progression plots)

AI-driven data processing:

- Receive sequences from MRI scanner
- AI-based brain extraction, image registration, T1-subtraction mapping
- AI-based tumor segmentation
- Construct tumor segmentation overlays & plots visualizing the change in tumor volumes over time

Potential clinical applications:

- Quantitative volumetric tumor response assessment
- Quantitaive high-throughput imaging biomarker
- Automated target volume delineation for radiotherapy treatment planning