Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study

Philipp Kickingereder, Fabian Isensee*, Irada Tursunova, Jens Petersen, Ulf Neuberger, David Bonekamp, Gianluca Brugnara, Marianne Schell, Tobias Kessler, Martha Foltyn, Inga Harting, Felix Sahm, Marcel Prager, Martha Nowosielski, Antje Wick, Marco Nolden, Alexander Radbruch, Jürgen Debus, Heinz-Peter Schlemmer, Sabine Heiland, Michael Platten, Andreas von Deimling, Martin J van den Bent, Thierry Gorlia, Wolfgang Wick, Martin Bendszus†, Klaus H Maier-Hein†*

Lancet Oncol 2019; 20: 728–740

Published Online April 2, 2019 http://dx.doi.org/10.1016/S1470-2045(19)30098-1

*Joint first authors

†Joint senior authors

Department of Neuroradiology (P Kickingereder MD, I Tursunova MD, J Petersen MSc, U Neuberger MD, G Brugnara MD, M Schell MD, M Foltyn MD, I Harting MD, M Prager PhD, A Radbruch MD, Prof S Heiland PhD, Prof M Bendszus MD), **Neurology Clinic** (T Kessler MD, M Nowosielski MD, A Wick MD, Prof M Platten MD, Prof W Wick MD), **Department of Neuropathology, Institute of Pathology** (F Sahm MD, Prof A von Deimling MD), **and Department of Radiation Oncology** (Prof J Debus MD), **Heidelberg University Hospital, Heidelberg, Germany; Medical Image Computing** (F Isensee MSc, J Petersen, M Nolden PhD, K H Maier-Hein PhD), **Department of Radiology** (D Bonekamp MD, A Radbruch, Prof H-P Schlemmer MD), **and German Cancer Consortium (DKTK)** (T Kessler, F Sahm, Prof M Platten, Prof A von Deimling, Prof W Wick) **German Cancer Research Center (DKFZ), Heidelberg, Germany; Department of Neurology, Medical University Innsbruck, Innsbruck, Austria** (M Nowosielski); **Heidelberg Institute of Radiation Oncology, Heidelberg, Germany** (Prof J Debus); **Heidelberg Ion-Beam Therapy Center, Heidelberg, Germany** (Prof J Debus); **Department of Neurology, Mannheim Medical Center, University of Heidelberg, Mannheim, Germany** (Prof M Platten);

Summary

Background The Response Assessment in Neuro-Oncology (RANO) criteria and requirements for a uniform protocol have been introduced to standardise assessment of MRI scans in both clinical trials and clinical practice. However, these criteria mainly rely on manual two-dimensional measurements of contrast-enhancing (CE) target lesions and thus restrict both reliability and accurate assessment of tumour burden and treatment response. We aimed to develop a framework relying on artificial neural networks (ANNs) for fully automated quantitative analysis of MRI in neuro-oncology to overcome the inherent limitations of manual assessment of tumour burden.

Methods In this retrospective study, we compiled a single-institution dataset of MRI data from patients with brain tumours being treated at Heidelberg University Hospital (Heidelberg, Germany; Heidelberg training dataset) to develop and train an ANN for automated identification and volumetric segmentation of CE tumours and non-enhancing T2-signal abnormalities (NEs) on MRI. Independent testing and large-scale application of the ANN for tumour segmentation was done in a single-institution longitudinal testing dataset from the Heidelberg University Hospital and in a multi-institutional longitudinal testing dataset from the prospective randomised phase 2 and 3 European Organisation for Research and Treatment of Cancer (EORTC)-26101 trial (NCT01290939), acquired at 38 institutions across Europe. In both longitudinal datasets, spatial and temporal tumour volume dynamics were automatically quantified to calculate time to progression, which was compared with time to progression determined by RANO, both in terms of reliability and as a surrogate endpoint for predicting overall survival. We integrated this approach for fully automated quantitative analysis of MRI in neuro-oncology within an application-ready software infrastructure and applied it in a simulated clinical environment of patients with brain tumours from the Heidelberg University Hospital (Heidelberg simulation dataset).

Findings For training of the ANN, MRI data were collected from 455 patients with brain tumours (one MRI per patient) being treated at Heidelberg hospital between July 29, 2009, and March 17, 2017 (Heidelberg training dataset). For independent testing of the ANN, an independent longitudinal dataset of 40 patients, with data from 239 MRI scans, was collected at Heidelberg University Hospital in parallel with the training dataset (Heidelberg test dataset), and 2034 MRI scans from 532 patients at 34 institutions collected between Oct 26, 2011, and Dec 3, 2015, in the EORTC-26101 study were of sufficient quality to be included in the EORTC-26101 test dataset. The ANN yielded excellent performance for accurate detection and segmentation of CE tumours and NE volumes in both longitudinal test datasets (median DICE coefficient for CE tumours 0·89 [95% CI 0·86–0·90], and for NEs 0·93 [0·92–0·94] in the Heidelberg test dataset; CE tumours 0·91 [0·90–0·92], NEs 0·93 [0·93–0·94] in the EORTC-26101 test dataset). Time to progression from quantitative ANN-based assessment of tumour response was a significantly better surrogate endpoint than central RANO assessment for predicting overall survival in the EORTC-26101 test dataset (hazard ratios ANN 2·59 [95% CI 1·86–3·60] *vs* central RANO 2·07 [1·46–2·92]; p<0·0001) and also yielded a 36% margin over RANO (p<0·0001) when comparing reliability values (ie, agreement in the quantitative volumetrically defined time to progression [based on radiologist ground truth *vs* automated assessment with ANN] of 87% [266 of 306 with sufficient data] compared with 51% [155 of 306] with local *vs* independent central RANO assessment). In the Heidelberg simulation dataset, which comprised 466 patients with brain tumours, with 595 MRI scans obtained between April 27, and Sept 17, 2018, automated on-demand processing of MRI scans and quantitative tumour response assessment within the simulated clinical environment required 10 min of computation time (average per scan).

Interpretation Overall, we found that ANN enabled objective and automated assessment of tumour response in neuro-oncology at high throughput and could ultimately serve as a blueprint for the application of ANN in radiology to improve clinical decision making. Future research should focus on prospective validation within clinical trials and application for automated high-throughput imaging biomarker discovery and extension to other diseases.

Funding Medical Faculty Heidelberg Postdoc-Program, Else Kröner-Fresenius Foundation.

Brain Tumor Center at Erasmus MC Cancer Institute, Rotterdam, Netherlands (Prof M J van den Bent MD); and European Organisation for Research and Treatment of Cancer, Brussels, Belgium (T Gorlia PhD)

Correspondence to:
Dr Philipp Kickingereder, Department of Neuroradiology, Heidelberg University Hospital, Heidelberg 69120, Germany
philipp.kickingereder@med.uni-heidelberg.de

Introduction

The development of novel therapies in neuro-oncology requires reliable and accurate endpoints for the assessment of treatment efficacy since both underestimation and overestimation of efficacy restricts trial proficiency. Although overall survival is the most definitive and objective endpoint to assess the efficacy of an investigational treatment, it is sensitive to other interventions (including crossover treatments) that might influence its applicability as an appropriate endpoint.[1,2] To overcome these limitations, objective responses and progression-free survival are considered endpoints that reliably assess the efficacy of an investigational treatment, specifically in small cohorts and diseases with multiple lines of treatment.[3] In neuro-oncology, progression-free survival is assessed according to the Response Assessment in Neuro-Oncology (RANO) working group criteria,[3,4] which are widely accepted for use in clinical trials[5] and increasingly used in routine clinical practice to determine treatment response.[6] These criteria mainly rely on the assessment of treatment response by use of MRI, which can enable both qualitative and quantitative assessment of tumour burden before, during, and after therapy. Underlying the use of RANO is the assumption that tumours grow in spherical shapes and that the two-dimensional (2D) measurement of a lesion's largest diameter on MRI is a surrogate marker of tumour volume.[5] However, in clinical practice, this assumption is not always accurate, since brain tumours frequently display very complicated shapes and anisotropic growth, influenced in part by the surrounding anatomic boundaries, host tissue–tumour interface, or treatment-related effects (eg, areas of necrosis and surgical cavities).[7,8] Consequently, volumetric or three-dimensional (3D) assessment of tumour burden has been of longstanding interest,[3,4,9] with studies indicating that volumetric measurements might be more reliable and accurate than 2D measurements.[10,11] Nevertheless, although volumetric assessment might arguably be one of the most quintessential parameters for accurate assessment of tumour burden and response,[8] it lacks practicability in a clinical setting. Whereas 2D measurements of tumour diameter can be done quickly and without dedicated software, volumetric measurements require sophisticated and time-consuming postprocessing of MRI data with dedicated software.[12,13]

Here, we present our development of a comprehensive, scalable, and validated approach, relying on artificial neural networks (ANNs), that we implemented in an

Research in context

Evidence before this study

MRI is a key method for detection, staging, and evaluation of response to treatment in neuro-oncology. The Response Assessment in Neuro-Oncology (RANO) criteria have been introduced to standardise assessment of patients with neuro-oncological tumours in both clinical trials and daily clinical practice. However, these criteria primarily rely on manual two-dimensional measurements of target lesions, which restricts both the reliability and accuracy of assessment of tumour burden and treatment response. Consequently, longstanding interest has existed in the use of volumetric assessment of tumour burden. We searched PubMed on Oct 31, 2018, with no date restrictions for publications in English using the terms ("neuro-oncology" OR "brain tumor" OR "brain tumour" OR "glioma" OR "glioblastoma") AND ("volumetry" OR "volumetric"). Our search did not identify any articles that assessed the use of automated quantitative assessment of tumour response in comparison with RANO assessment. Previous studies have indicated that volumetric measurements might be more reliable and accurate than manual two-dimensional measurements. However, non-automated volumetric measurements are not practical in a clinical setting and have previously been cited as a labour-intensive, time-consuming, and complex task that prevents clinical translation. We aimed to develop a comprehensive and scalable approach, relying on artificial neural networks (ANNs), that enables fully automated quantitative assessment of tumour burden by use of MRI in neuro-oncology.

Added value of this study

We showed that automated volumetric quantification of tumour burden is highly accurate on independent large-scale datasets. Application of this method for automated quantitative identification of disease progression in a multicentre clinical trial dataset outperformed RANO both in terms of reliability and as a surrogate endpoint for predicting overall survival. To facilitate clinical translation, we integrated this approach for fully automated quantitative analysis of MRI in neuro-oncology into an application-ready software infrastructure.

Implications of all the available evidence

Our results show that automated quantitative analysis of MRI using a comprehensive deep-learning approach with ANN could allow radiologists and clinicians to overcome the inherent limitations of manual assessment of tumour burden. This approach could greatly improve and standardise assessment of tumour response in routine clinical practice and clinical trials and might become a valuable asset for decision making in neuro-oncology.

application-ready software infrastructure that enables fully automated quantitative analysis of MRI in neuro-oncology. Specifically, we aimed to investigate the potential of this approach for automated quantitative assessment of tumour response to overcome the inherent limitations of manual assessment of tumour burden.

Methods

Study design and participants

In this multicentre, retrospective study, we analysed MRI data from patients with brain tumours that were acquired at the Heidelberg University Hospital (Heidelberg, Germany) or as part of the European Organisation for Research and Treatment of Cancer (EORTC)-26101 trial, which was run at 38 institutions in Europe, to develop, train, and test an ANN for automated interpretation of MRI data in clinical settings and prediction of time to progression in these patients.

First, we created a training dataset for the ANN from a non-consecutive cohort of adult patients (aged ≥18 years) with histologically confirmed glioblastoma or lower-grade glioma (including diffuse astrocytic and oligodendroglial WHO grade II and III tumours as defined by the 2016 WHO Classification of Tumors of the Central Nervous System[14]) being treated at the Heidelberg University Hospital (Heidelberg training dataset). Specifically, the Heidelberg training dataset consisted of MRI data from a single timepoint (one MRI scan per patient) either preoperatively from initial diagnosis, or early postoperatively (<72 h after surgery) or at follow-up, and was specifically ensembled to represent the broad phenotypic appearance of brain tumours on MRI during disease evolution. Appropriate MRI scans were manually identified to enrich the dataset with comparatively uncommon and difficult cases on the basis of the judgment of the neuroradiologist (PK, DB)—eg, cases with complex resection cavities, extensive alterations after treatment, different CE pattern (faint, dot-like, or multifocal pattern), or presence of non-tumoral contrast enhancement (intratumoral or peritumoral blood vessels). We had no further inclusion or exclusion criteria. Second, for independent testing of the ANN once it has been developed, we selected a non-consecutive cohort of adult patients (aged ≥18 years) with histologically confirmed glioblastoma or lower-grade glioma from the Heidelberg University Hospital, with individual MRI scans from multiple timepoints for each patient (Heidelberg test dataset). We had no further inclusion or exclusion criteria. This cohort was selected in parallel with and independent of the training dataset. This dataset was a longitudinal dataset with preoperative and consecutive follow-up scans. Finally, another cohort of adult patients (aged ≥18 years) with brain tumours undergoing routine MRI scans at the Heidelberg University Hospital was selected for testing of the developed infrastructure for automated tumour segmentation and quantitative assessment of tumour response in a simulated clinical environment (Heidelberg simulation dataset). MRI scans from all three Heidelberg datasets were acquired according to an established protocol as described previously,[12,13,15] and included T1-weighted images before (T1-w) and after (cT1-w) administration of contrast agent, and fluid-attenuated inversion recovery (FLAIR) and T2-weighted (T2-w) images (requiring either 3D or 2D with axial orientation). Retrospective evaluation of imaging data from Heidelberg University Hospital was approved by the local ethics committee of the University of Heidelberg and informed consent was waived.

For independent testing of the ANN, we collected data from the EORTC-26101 study, which was a prospective randomised phase 2/3 trial in patients with first progression of a glioblastoma after standard chemoradiotherapy. In brief, the phase 2 trial[16] assessed the optimal treatment sequence of bevacizumab and lomustine (four treatment groups with single drug *vs* sequential *vs* combination), whereas the subsequent phase 3 trial[7] (two treatment groups) compared patients treated with lomustine alone with those receiving a combination of lomustine and bevacizumab. Overall, the EORTC-26101 study included 596 consecutively recruited patients (n=159 in phase 2, n=437 in phase 3) with 2593 individual MRI scans acquired at 38 institutions in Europe. The MRI scans were acquired with a uniform imaging protocol at baseline and every 6 weeks until week 24, and thereafter once every 3 months until last follow-up. Data we collected were from T1-w, cT1-w, FLAIR, and T2-w images. The study was conducted in accordance with the Declaration of Helsinki and the protocol was approved by local ethics committees and patients provided written informed consent (EudraCT number 2010-023218-30, ClinicalTrials.gov number NCT01290939). Full study design and outcomes have been published previously.[7,16] Access to the EORTC-26101 trial data for the present study was granted through an EORTC external research project. Of the data collected, we excluded MRI scans from the EORTC-26101 test dataset if data were corrupted; in the case of incomplete availability of T1-w, cT1-w, FLAIR, or T2-w sequences; or if there were heavy motion artifacts that also precluded central RANO assessment.

Procedures

Figure 1 depicts the analysis workflow. MRI data from the Heidelberg training dataset, Heidelberg test dataset, and EORTC-26101 test dataset were preprocessed identically (full description in appendix p 2). Briefly, preprocessing included brain extraction (ie, removing the skull and extracranial tissue from images), followed by image registration, calculation of T1-subtraction volumes, and preparation of volumetric tumour segmentation masks for each MRI scan. The preparation of volumetric tumour segmentation masks included volumetric delineation of both contrast-enhancing (CE) tumours and

See Online for appendix

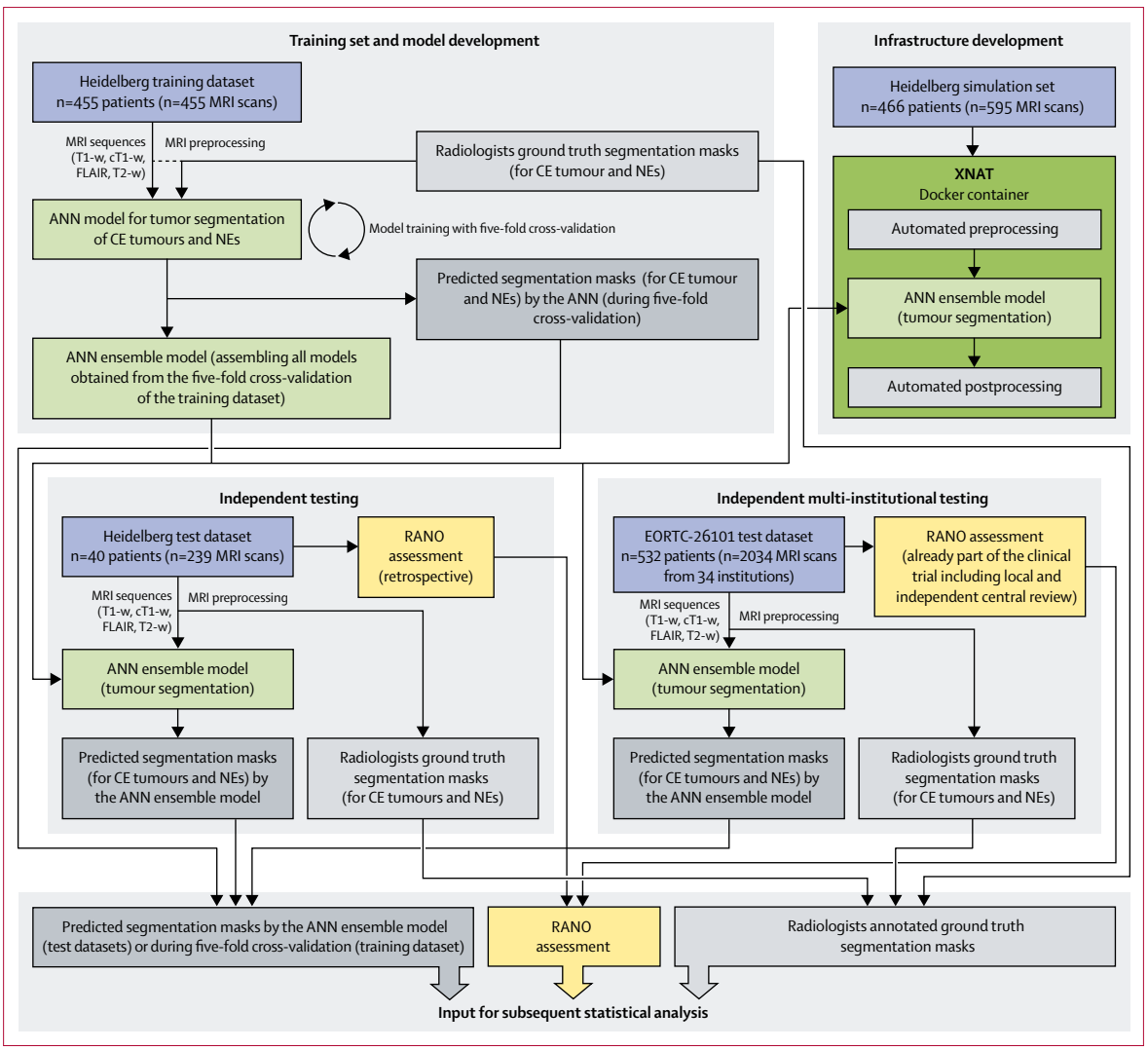


Figure 1: **Flowchart of the procedures for training and model development, testing, statistical analysis, and infrastructure development for automated brain tumour segmentation**

ANN=artifical neural network. CE=contrast enhancing. cT1-w=T1-weighted images after contrast agent. EORTC=European Organisation for Research and Treatment of Cancer. FLAIR=fluid-attenuated inversion recovery. NEs=non-enhancing T2-signal abnormalities. RANO=Response Assessment in Neuro-Oncology. T1-w=T1-weighted images before contrast agent. T2-w=T2-weighted images.

non-enhancing T2-signal abnormalities (NEs; defined as T2-FLAIR hyperintense abnormality excluding the contrast-enhancing and necrotic portion of the tumour, resection cavity, and obvious leukoaraiosis) by experienced neuroradiologists (for the Heidelberg training and test dataset: by PK, a radiology resident with 6 years of experience, and subsequently checked by DB, a board-certified radiologist and neuroradiologist with 17 years of experience in image processing; for the EORTC-26101 test dataset: by IT, a radiology resident with 3 years of experience and subsequently checked by PK. Any discrepancies were resolved through a consensus discussion) using a semi-automated approach as described previously.[12,13] All MRI sequences were normalised independently by subtracting the mean value from each voxel and dividing by its SD. Voxels outside the brain mask were set to zero.

The architecture of our developed ANN was inspired by our work[17] in the Brain Tumor Segmentation (BraTS) challenge[18] and is based on U-Net architecture.[19] The U-Net consists of an encoder and a decoder network that are interconnected with skip connections. Conceptually, the encoder network is used to aggregate semantic information at the cost of decreased spatial information. The decoder is the counterpart to the encoder and reconstructs the spatial information while accounting for the semantic information extracted by the encoder. Skip connections are used to transfer feature maps from the encoder to the decoder to allow for even more precise localisation of the tumour. Our adaptation of the U-Net

(full description of the network architecture, and applied training and testing procedures are in the appendix [pp 2–4, 11]) makes use of residual connections in the encoder[20] while keeping the decoder relatively lightweight. During training of the ANN, it processes large input patches (128×128×128 voxels) to effectively capture as much contextual information as possible. To encourage the training of the bottleneck layers, we made use of auxiliary loss layers deep in the network. We used the Heidelberg training dataset to train and validate the ANN (with five-fold cross-validation). For this training, we provided the ANN with the four different MRI sequences (T1-w, cT1-w, FLAIR, and T2-w sequences) for each MRI scan and the corresponding tumour segmentation masks generated by the radiologists as input. These segmentation masks were used as so-called ground truth measures and allowed the ANN to learn the phenotypic appearance of brain tumours on MRI, and consequently enabled automated identification and volumetric delineation of CE tumours and NEs on MRI. In the Heidelberg training dataset, the tumour segmentation masks were predicted during the five-fold cross-validation procedure. Thereby, the Heidelberg training dataset was randomly partitioned into five equal size subsamples (20% of patients). Of the five subsamples, a single subsample (20% of patients) was retained as the validation data for testing the model, and the remaining four subsamples (80% of patients) were used as training data. The cross-validation process was then repeated five times (the folds), with each of the five subsamples used once as the validation data.

We used the longitudinal Heidelberg test dataset and the longitudinal EORTC-26101 test dataset to independently do large-scale testing of the performance of the ANN. Specifically, to predict the segmentation masks with the CE tumours and NEs on MRI in both test datasets, we used the four different MRI sequences (T1-w, cT1-w, FLAIR, and T2-w sequences) from each MRI scan as input into an ANN ensemble model consisting of the five ANN models obtained during cross-validation of the Heidelberg training dataset. The predicted tumour segmentation masks generated by the ANN within the Heidelberg training dataset, Heidelberg test dataset, and EORTC-26101 test dataset served as a fundamental input for all subsequent analyses.

For both longitudinal test datasets, we quantitatively assessed the volumetric tumour response and calculated the time to progression separately, once on the basis of the ground truth tumour segmentation masks generated by the radiologist and once on the basis of the automatically generated ANN-based tumour segmentation masks. We defined tumour progression as an increase in tumour volume (compared with baseline or best response) in either CE tumour or NEs, beyond a minimal tumour volume of 1 cm^3; or occurrence of a new CE lesion outside of the CE tumour volume from the previous MRI scan (identified automatically using dedicated algorithms with the respective segmentation masks over time as input; extended description is on appendix p 5). For the increase in volume, we applied a uniform threshold of 40% to qualify for progression on the basis of CE tumour, whereas for NE we applied a threshold of 40% for patients with lower-grade glioma and 100% for patients with glioblastoma. These volumetric thresholds are justified by the equivalent 2D threshold of use in the RANO criteria, except for the 100% increase in volume in NEs for patients with glioblastoma, for which the RANO working group has not yet defined a 2D threshold (appendix p 5).[3,4]

Furthermore, in both longitudinal test datasets, we did conventional assessment of tumour response according to RANO criteria.[3,4] In the Heidelberg test dataset, we retrospectively did RANO assessments (PK and DB), with discrepancies resolved through consensus discussion. For the EORTC-26101 test dataset both local assessment and independent central RANO review had already been done as part of the clinical trial, and so we extracted these data from the trial database. We considered the central RANO review to be an unbiased reference standard because two independent expert radiologists did the review, with discrepancies resolved through consensus discussion, and, by contrast with local RANO assessment, they were masked to the type of treatment, neurological status, steroid doses, and the local RANO investigator's assessment. The RANO assessment in both test datasets was only based on imaging criteria and no additional clinical criteria to allow precise comparison with our quantitative assessment method for tumour response.

We developed an application-ready software infrastructure (appendix pp 5, 6) using the XNAT open-source imaging informatics software platform components. We aimed to enable translation and application of our ANN for automated tumour segmentation and quantitative assessment of tumour response in daily clinical practice and clinical trials, with a specific focus on optimising the processing pipeline—ie, that processing is completed in a clinically acceptable timeframe. In routine clinical practice, automated on-demand processing of an MRI scan is triggered after the images have been acquired on the MRI scanner (or alternatively, for example, within clinical trials uploaded to the XNAT server). Processing of the MRI scans is fully automated and does not require any additional manual intervention. The processed results (superimposed tumour segmentation mask on individual MRI sequences and chart depicting longitudinal tumour volume dynamics) are automatically pushed back to the picture archiving and communication system, where they are available for interpretation. We extensively tested the developed infrastructure in a simulated clinical environment (the Heidelberg simulation dataset) with automated (retrospective) processing of all MRI scans.

Outcomes

We had four main objectives. The first objective was to evaluate the accuracy of automated volumetric tumour segmentation by use of the ANN in comparison with radiologist ground truth tumour segmentation. This outcome was to be assessed in the Heidelberg training dataset, Heidelberg test dataset, and the EORTC-26101 test dataset. Our second objective was to assess within both longitudinal test datasets (the Heidelberg test dataset and the EORTC-26101 test dataset) the spatial and temporal tumour volume dynamics in each patient to calculate a quantitative time to progression that is volumetrically defined and investigate the extent of agreement (reliability) of this quantitative volumetrically defined time to progression (ie, time to progression calculated from ANN-based automated tumour segmentation masks *vs* time to progression calculated from radiologist ground truth tumour segmentation masks) and to compare this reliability (in the EORTC-26101 test dataset) with the extent of agreement (reliability) for the time to progression from RANO (ie, time to progression calculated from RANO local assessment *vs* time to progression calculated from RANO central review). Time to progression was calculated from the date of baseline MRI after surgery in the Heidelberg test dataset and from the date of randomisation in the EORTC-26101 test dataset (censored at the date of last MRI if no progression occurred during follow-up). Our third objective was to compare the performance of quantitative volumetrically defined time to progression versus time to progression calculated from RANO central review (unbiased reference standard) as surrogate endpoints for predicting overall survival within the EORTC-26101 test dataset (overall survival and RANO information taken from the EORTC-26101 trial database; and overall survival calculated from the date of randomisation until death or last follow-up).[21] Our final objective was to implement the ANN for automated tumour segmentation and quantitative assessment of tumour response in an application-ready software infrastructure and apply it to the Heidelberg simulation dataset.

Statistical analysis

Cohort size for each of the included datasets were determined by the availability of MRI data and not derived from a power calculation. Detailed information on statistical analyses are in the appendix (pp 4, 5). Briefly, we assessed the accuracy of the automatically generated ANN-based tumour segmentation masks for delineating CE tumours and NEs in comparison with the reference—ie, the ground truth segmentation masks generated by a radiologist—in the Heidelberg training dataset, Heidelberg test dataset, and EORTC-26101 test dataset using DICE similarity coefficients for segmentation agreement and Bland-Altman plots and concordance correlation coefficients for volume agreement. The DICE similarity coefficient is a standard measure to report the performance of a segmentation[18] and measures the extent of spatial overlap between two binary segmentation masks. The DICE similarity coefficient is defined as twice the size of the intersection between the two masks (ground truth [GT] and predicted segmentation mask [PM]), normalised by the sum of their volumes

$$\mathrm{DICE} = \frac{2\,|\mathrm{GT} \cap \mathrm{PM}|}{|\mathrm{GT}| + |\mathrm{PM}|}$$

The DICE coefficient can range between 0 (no overlap) and 1 (perfect agreement). The reported 95% CIs for the median DICE coefficients were calculated using bootstrapping with 1000 iterations.

We calculated the relative and absolute agreement in the time to progression between the different methods (quantitative volumetric assessment [based on the spatial and temporal changes in the radiologist's ground truth *vs* automatically generated ANN tumour volumes] *vs* RANO [based on local assessment *vs* central review]) using the Heidelberg test dataset and EORTC-26101 test dataset (the EORTC-26101 test dataset was limited to a subset of patients with complete data; see appendix p 19). We used a one-tailed two-sample test for equality of proportions to assess whether the agreement in time to progression within the EORTC-26101 test dataset was higher for quantitative volumetric assessment (ie, time to progression from ANN-based automated tumour segmentation *vs* time to progression from radiologist ground truth tumour segmentation) than the RANO assessment (ie, time to progression from RANO local assessment *vs* from central review). We generated Kaplan-Meier plots and log-rank tests to assess whether disagreement in the time to progression on a dataset level was higher for RANO than quantitative volumetric assessment.

In the EORTC-26101 test dataset, we assessed the performance of time to progression from RANO (using central RANO review as an unbiased reference standard) and the time to progression from automated ANN-based assessment as surrogate endpoints for predicting overall survival. Specifically, we generated Cox proportional hazards regression models for overall survival, with the time to progression (either from central RANO review or from automated ANN assessment) included as a time-dependent covariate. Hazard ratios (HRs) indicated risk of death at any time during the study period,[22] and Z values correspond to the ratio of each regression coefficient to its SE. To control for confounding treatment-specific effects in the EORTC-26101 trial, we included the treatment group as a binary covariate (initial treatment containing bevacizumab *vs* no bevacizumab). We assessed the performance of each Cox proportional hazards regression model with Harrell's concordance index (c-index, with 95% CIs calculated using bootstrapping with 1000 iterations), which is a standard performance measure for model assessment in survival analysis.[23] High c-index values indicate better

performance (ie, better discriminative ability) of the model for predicting overall survival.[23] We did an analysis of deviance that allowed us to compare the reduction in the log-likelihood between different Cox proportional hazards regression models using χ^2 test statistics. Specifically, we compared Cox proportional hazards regression models with and without inclusion of time to progression from ANN as additional time-dependent covariate (to a Cox proportional hazards regression model with time to progression from central RANO review as time-dependent covariate and treatment group as a binary covariate).

Finally, we assessed the prognostic relevance of baseline CE tumour and NE volumes and their early changes between baseline and first follow-up in the EORTC-26101 test dataset for predicting overall survival in a multivariable context including both clinical and molecular parameters. Importance of individual covariates in the Cox proportional hazards regression model was assessed by computing the Wald χ^2 statistic and the proportion of the overall model χ^2 that is due to each covariate (full description of analysis is in the appendix p 5).

p values of less than 0·05 were considered significant. We did all statistical analyses using R version 3.5.1.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. PK, FI, MB, and KHM-H had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

The Heidelberg training dataset comprised 455 non-consecutive patients with brain tumours being treated in the Heidelberg University Hospital between July 29, 2009, and March 17, 2017. Of 455 patients, 364 (80%) had histologically confirmed glioblastoma and 91 (20%) lower-grade glioma. Baseline characteristics of these patients are shown in the appendix (p 18). We collected data for one MRI per patient, of which 319 (70%) were preoperative from initial diagnosis and 136 (30%) were early postoperative (<72 h after surgery) or from follow-up scans. These MRI scans, in conjunction with individual radiologist ground truth tumour segmentation masks, were used to train the ANN. Within the Heidelberg training dataset, the ANN acquired the relevant knowledge to accurately delineate the clinically relevant CE tumour and NE compartments (obtained metrics are shown in the appendix [p 9]).

For independent testing and large-scale application of the ANN, two longitudinal testing datasets were compiled: the single-institution Heidelberg test dataset and the multi-institutional EORTC-26101 test dataset. The Heidelberg test dataset comprised 40 non-consecutive patients who were being treated in the Heidelberg University Hospital between July 29, 2009, and March 17, 2017, with histologically confirmed glioblastoma (n=25 [63%]), or lower grade glioma (n=15 [38%]). These 40 patients had MRI data from 239 scans, with data from multiple timepoints (median of five scans per patient [IQR four to six]). The EORTC-26101 study included 596 patients (n=159 in phase 2, n=437 in phase 3) with 2593 individual MRI scans acquired at 38 institutions in Europe. We excluded 559 MRI scans because the data were corrupted after conversion of file formats from DICOM to *NIfTI* (due to non-standardised centre-specific anonymisation of DICOM files or corrupt or incomplete DICOM files; n=178); incomplete availability of T1-w, cT1-w, FLAIR, and T2-w sequences (requiring either 3D acquisitions or 2D with axial orientation; n=374); or heavy motion artifacts (also precluding central RANO assessment; n=7). Therefore, the multicentre EORTC-26101 trial dataset comprised 532 patients, all with histologically confirmed glioblastomas, from 34 institutions with 2034 MRI scans (median of four scans per patient [IQR three to five]) done on 16 different MRI scanners from four manufacturers (information not avaialble for 102 MRI scans) between Oct 26, 2011, and Dec 3, 2015 (extended description and full list of MRI scanners in appendix [p 18]).

Independent testing in the Heidelberg test dataset yielded median DICE coefficients of 0·89 (95% CI 0·86–0·90) for CE tumours and 0·93 (0·92–0·94) for NEs and in the EORTC-26101 test dataset of 0·91 (0·90–0·92) for CE tumours and 0·93 (0·93–0·94) for NEs (figure 2A, table 1). The performance of the ANN remained stable after application to the broad multicentre setting of the EORTC-26101 test dataset, which was coupled with high agreement between the radiologist ground truth tumour volumes and those automatically predicted by the ANN across both test datasets (concordance correlation coefficients for CE and NE each ≥0·98; figure 2B and 2C, table 1). The performance of the ANN for the segmentation of CE tumours in the EORTC-26101 test dataset was significantly improved when using 3D T1 and cT1 sequences compared with corresponding 2D sequences ($p<0·0001$; appendix p 9). These results with an integrative discussion of the individual performance metrics obtained within each dataset are in the appendix (pp 9, 10, 13–16).

By applying the outlined criteria for quantitative identification of disease progression, patients most frequently qualified for progression because of an increase in the CE tumour volume: 19 (48%) of 40 patients in the Heidelberg test dataset and 189 (62%) of 306 patients with complete data in the EORTC-26101 test dataset (appendix p 20). However, beyond disease progression determined via pure volumetric thresholds, a notable number of patients (four [10%] of 40 patients in the Heidelberg test dataset and 26 [8%] of 306 in the EORTC-26101 test dataset) only developed new anatomically distinct CE tumour lesions during follow-up that

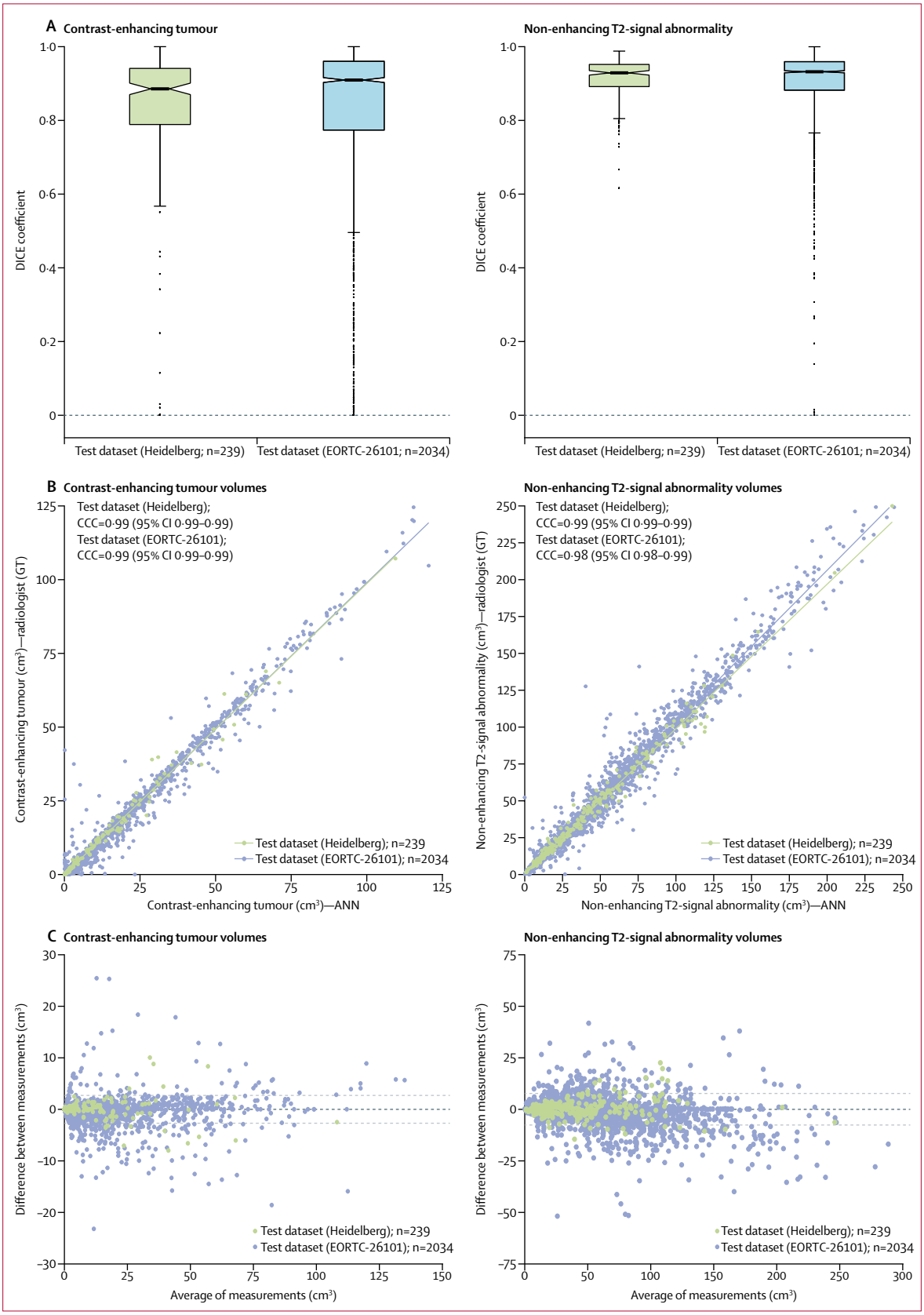


Figure 2: **Agreement between the automated volumetric segmentation with the ANN and the radiologist-generated ground truth for tumour segmentation (A) and tumour volume (B, C)**
(A) Data are median DICE coefficients for tumour segmentation. Blocks show IQR of datapoints, with the horizontal central line showing the median. The sides of blocks are indented and indicate the 95% CI of the median. Whiskers adjacent to the boxes represent 1·5 times the IQR. Dots are outliers. Outliers with DICE coefficients of 0 primarily reflect the uncertainty of accurate tumour segmentation in the post-treatment setting (ie, differentiating true contrast-enhancing tumour from reactive gliosis; more details in appendix pp 9, 10). (B) Data are concordance correlation coefficients (CCCs). (C) Bland-Altman plot. EORTC=European Organisation for Research and Treatment of Cancer. GT=ground truth. ANN=artificial neural network.

would not have qualified for disease progression with volumetric thresholds alone, thereby depicting the relevance of additionally integrating automated screening for newly appearing tumour lesions as a distinct criterion (appendix p 12).

Agreement in quantitative volumetrically defined time to progression (based on radiologist ground truth *vs* automated assessment with ANN) was 90% (36 of 40 patients) in the Heidelberg test dataset and 87% (266 of 306 patients) in the EORTC-26101 test dataset (table 2). With agreement in only 51% (155 of 306) of patients in the EORTC-26101 test dataset, the reference benchmark (ie, agreement in time to progression between local and central RANO assessment) was significantly lower than the agreement in the quantitative volumetric data between the ANN and radiologist ground truth (p<0·0001; table 2). The higher reliability of the quantitative volumetrically defined time to progression than the reference benchmark was also reflected by the corresponding Kaplan-Meier plots for time to progression, which showed no significant difference for the quantitative volumetrically defined time to progression based on ground truth versus ANN on a dataset level (p=0·94 for the Heidelberg test dataset, p=0·77 for the EORTC-26101 test dataset); however, a significant difference in the time to progression based on local versus independent central RANO assessment was seen in the EORTC-26101 test dataset (p<0·0001; figure 3).

We compared the performance of calculation of time to progression determined via quantitative volumetric assessment (using ANN) with those determined via RANO (using central assessment as an unbiased reference standard) as surrogate endpoints for predicting overall survival in the EORTC-26101 test dataset. The Cox regression model for overall survival with the time to progression from central RANO as a time-dependent covariate yielded an HR of 2·07 (95% CI 1·46–2·92) with a Z value of 4·12 and a c-index of 0·57 (95% CI 0·54– 0·61; p<0·0001; table 3). By contrast, the Cox regression model for overall survival with time to progression from ANN as a time-dependent covariate yielded an HR of 2·59 (95% CI 1·86–3·60) with a Z value of 5·64 and a c-index of 0·62 (95% CI 0·59–0·66; p<0·0001; table 3). The treatment regimen in the EORTC-26101 trial had no confounding effect in either model (p=0·34 for both). The inclusion of the time to progression from ANN as an additional time-dependent covariate yielded a significantly improved model fit over a Cox proportional hazards regression model that only included time to progression from central RANO review as time-dependent covariate and the treatment group as a binary covariate (χ^2=21·95; p<0·0001).

The quantitative volumetrically defined criteria for disease progression (40% volume increase in CE tumour for glioblastoma, 40% volume increase in NE for lower-grade glioma, or appearance of a new anatomically distinct lesion) reflect the equivalent (2D) thresholds

	Heidelberg test dataset	EORTC-26101 test dataset
Tumour segmentation agreement		
Contrast-enhancing tumour	0·89 (0·86–0·90)	0·91 (0·90–0·92)
Non-enhancing T2-signal abnormality	0·93 (0·92–0·94)	0·93 (0·93–0·94)
Tumour volume agreement		
Contrast enhancing tumour	0·99 (0·99–1·00)	0·99 (0·99–0·99)
Non-enhancing T2 signal abnormality	0·99 (0·99–0·99)	0·98 (0·98–0·99)

Data are median DICE coefficient for tumour segmentation agreement, and concordance correlation coefficient for tumour volume agreement, with 95% CIs in parentheses. EORTC=European Organisation for Research and Treatment of Cancer. ANN=artificial neural network.

Table 1: **Agreement between tumour segmentation masks and tumour volumes predicted by the ANN and those generated by the radiologist (ground truth)**

	Absolute agreement (both time and event)		Details on disagreement			
	Yes	No	Earlier progressive disease with alternative	Later progressive disease with alternative	No progressive disease with alternative (but with reference)	Progressive disease with alternative (but not with reference)
Heidelberg test dataset (n=40); reference method *vs* alternative method						
Quantitative (radiologist) *vs* quantitative (ANN)	36 (90%)	4 (10%)	2 (5%)	0	2 (5%)	0
Quantitative (radiologist) *vs* RANO*	29 (73%)	11 (28%)	3 (8%)	6 (15%)	2 (5%)	0
EORTC-26101 test dataset (n=306); reference method *vs* alternative method						
Quantitative (radiologist) *vs* quantitative (ANN)	266 (87%)	40 (13%)	23 (8%)	13 (4%)	4 (1%)	0
Quantitative (radiologist) *vs* RANO (central)	156 (51%)	150 (49%)	110 (36%)	17 (6%)	23 (8%)	0
Quantitative (radiologist) *vs* RANO (local)	181 (59%)	125 (41%)	36 (12%)	51 (17%)	28 (9%)	10 (3%)
RANO (central) *vs* RANO (local)	155 (51%)	151 (49%)	15 (5%)	108 (35%)	24 (8%)	4 (1%)

ANN=artificial neural network. RANO=Response Assessment in Neuro-Oncology. EORTC=European Organisation for Research and Treatment of Cancer. *Disagreement between the two RANO readers in five (13%) of 40 patients, resolved through consensus discussion.

Table 2: **Agreement in the time to progression on a patient level between the different methods**

mandated by the RANO criteria.[3,4] We chose the applied threshold of 100% volume increase in NE for glioblastoma independently, because the RANO criteria do not provide an equivalent 2D threshold and only suggest that a significant increase in NE qualifies for disease progression. The basis for our conservative threshold of 100% is the theory that dynamics in NE volume are less specific than for CE tumour volumes to determine tumour burden in patients with glioblastoma. This assumption is supported by our findings in the EORTC-26101 test dataset that both baseline CE tumour volume, and early changes in this volume, were the covariates in the multivariable Cox model for overall survival that showed the greatest contribution to the overall model χ^2 compared with other clinical and molecular parameters (table 4). Specifically, baseline CE tumour volume (HR of 1·02 per 1 cm³, 95% CI 1·01–1·03; p<0·0001) and early change in CE tumour volume (HR of 1·04 per 100% increase, 95% CI 1·02–1·06; p<0·0001), showed the highest χ^2 values (18·87 and 19·88) and contributed 25% and 26% to the overall model χ^2 value of 76·97. The next highest χ^2 values were O⁶-methylguanine-DNA methyltransferase (*MGMT*) promoter methylation status, with an χ^2 value value of 11·42 (HR 0·61, 95% CI 0·46–0·81; p=0·00073) and glucocorticoid intake, with an χ^2 value of 6·64 (HR 1·52, 95% CI 1·11–2·09; p=0·0099); thus contributing 15% and 9% to the overall model χ^2 value (table 4). Moreover, the baseline NE volume, and early changes in this volume, did not show independent significance within this multivariable model (table 4), thereby supporting our chosen conservative threshold of 100% increase in NE tumour volume in patients with glioblastoma.

To facilitate adoption of our approach for automated tumour segmentation and quantitative volumetric assessment of tumour response in clinical trials and routine clinical practice, we developed a fully automated application-ready processing pipeline for MRI scans (schematic illustration of the workflow is in appendix p 17). This approach enables seamless manufacturer neutral integration into existing clinical infrastructures. We applied the approach within a simulated clinical environment and did fully automated processing (including quantitative tumour response assessment) of a simulation dataset drawn from the Heidelberg University Hospital. This simulation dataset comprised 466 patients with primary intra-axial brain tumours undergoing routine MRI (n=241 [52%] glioblastoma, n=177 [38%] lower-grade glioma [diffuse astrocytic and oligodendroglial WHO grade II and III tumours], and n=48 [10%] other histological entities [pilocytic astrocytoma, pleomorphic xanthoastrocytoma, hemangiopericytoma, dysembryoplastic neuroepithelial tumor, ganglioglioma, medulloblastoma, central neurocytoma, and primary central nervous system lymphoma]), with MRI data from 595 scans collected between April 27, and Sept 17, 2018

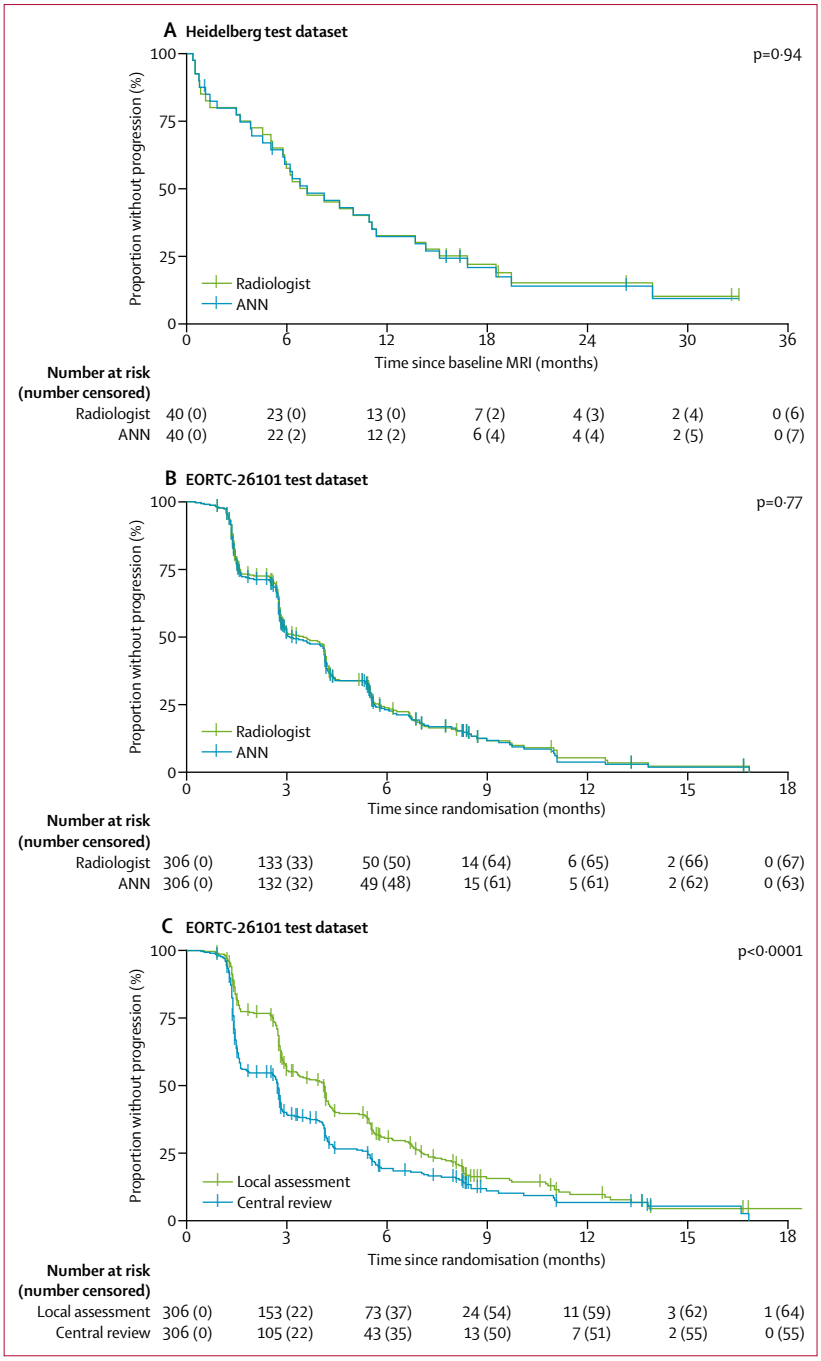


***Figure 3*: Quantitative volumetrically defined time to progression in the Heidelberg test dataset (A) and EORTC-26101 test dataset (B), and RANO-defined time to progression in the EORTC-26101 test dataset (C)**
ANN=artifical neural network. EORTC=European Organisation for Research and Treatment of Cancer. RANO=Response Assessment in Neuro-Oncology.

(Heidelberg simulation dataset). We yielded an average computational processing time of 10 min 14 s per MRI exam, thus staying well within a clinically acceptable range (individual data are not shown). The processing pipeline can already accommodate three routines running simultaneously and can be scaled up linearly by adding

	Point estimate	Z value	p value
Quantitative ANN assessment			
Time to progression*	HR 2·59 (1·86–3·60)	5·64	<0·0001
Treatment regimen†	HR 1·14 (0·87–1·47)	0·95	0·34
c-index of the model	0·62 (0·59–0·66)	..	..
Central RANO assessment			
Time to progression*	HR 2·07 (1·46–2·92)	4·12	<0·0001
Treatment regimen†	HR 1·14 (0·87–1·47)	0·95	0·34
c-index of the model	0·57 (0·54–0·61)	..	..

ANN=artificial neural network. HR=hazard ratio. RANO=Response Assessment in Neuro-Oncology. Z value is the ratio of each regression coefficient to its SE. 95% CIs are shown in parentheses where appropriate. *Time to progression is included as a time-dependent covariate. †Included as binary covariate (initial treatment containing bevacizumab *vs* no bevacizumab).

Table 3: **Cox proportional hazards regression models for overall survival with time to progression in the EORTC-26101 test dataset by assessment method**

	Hazard ratio	Wald χ^2	p value
Baseline CE tumour volume, cm^3*	1·02 (1·01–1·03)	18·87	<0·0001
Early change in CE tumour volume, %*	1·04 (1·02–1·06)	19·88	<0·0001
Baseline NE volume, cm^3†	1·00 (1·00–1·00)	0·03	0·87
Early change in NE volume, %†	1·14 (0·92–1·41)	1·47	0·22
Age, years‡	1·01 (0·99–1·02)	1·34	0·25
Sex (female *vs* male)	0·92 (0·69–1·23)	0·30	0·59
WHO performance status (>0 *vs* 0)	1·17 (0·85–1·63)	0·92	0·34
MGMT promotor methylation status (methylated *vs* unmethylated)	0·61 (0·46–0·81)	11·42	0·00073
Glucocorticoids intake (yes *vs* no)	1·52 (1·11–2·09)	6·64	0·0099

The Cox model included tumour volumes automatically predicted by the artificial neural network (from baseline MRI and the early change in those volumes between baseline and first follow up MRI as covariates). CE=contrast enhancing. NE=non-enhancing T2 signal abnormality. *MGMT*=O^6-methylguanine-DNA methyltransferase. *Included as continuous variable (hazard ratios correspond to an increase of 1 cm^3). †Included as continuous variable (hazard ratios correspond to an increase of 100%). ‡Included as continuous variable (hazard ratios correspond to an increase of 1 year).

Table 4: **Multivariable Cox proportional hazards regression model for overall survival in the EORTC-26101 test dataset**

additional processing nodes to the cluster without any need to interrupt the existing workflow (appendix p 6).

Discussion

We showed that automated quantitative analysis of MRI using a comprehensive deep-learning approach with ANN could be a valuable tool for clinical decision making in neuro-oncology. Specifically, the standardisation of our approach has great promise to decrease inter-observer variability of assessment of tumour response that often occurs with RANO criteria. We showed robust performance and generalisability of our ANN in the EORTC-26101 trial dataset, which was across 34 institutions including all major MRI manufacturers, with a broad variety of scanner types and field strengths. Moreover, our results suggest superiority of quantitative volumetric assessment of tumour response, both in terms of reliability and performance as a surrogate endpoint for predicting overall survival. These findings point out the inherent limitations of the 2D RANO criteria, which only serve as an imperfect surrogate parameter for the assessment of brain tumours that frequently display complicated shapes and anisotropic growth. Additionally, our implementation into a fully automated application-ready processing pipeline for MRI scans in the open-source XNAT framework holds great promise for standardisation of tumour response assessment in neuro-oncology across institutions and clinical trials. Specifically, this processing pipeline not only allows seamless integration that is manufacturer neutral into routine clinical practice independent of pre-existing infrastructures, but also enables investigators to make use of existing XNAT capabilities to manage and coordinate the analysis of MRI data in large multisite clinical trials.

Although quantitative volumetric assessment of tumour response might arguably be one of the most quintessential parameters for accurate assessment of tumour burden and response,[8,24] it has previously been cited as a labour-intensive, time-consuming, and complex task—even in the case of semi-automated techniques[25]—which ultimately prevents clinical adoption.[3,5,26] Our integration of the robust ANN-based tumour segmentation algorithm into a fully automated application-ready processing pipeline for MRI scans allows investigators to overcome this bottleneck that has previously restricted automated and quantitative analysis of MRIs in neuro-oncology. Although objective and automated assessment of tumour response such as we have presented here is the most evident application of this technology, this technology could also be extended to a broad variety of other applications, including automated high-throughput imaging biomarker discovery (eg, volumetric quantification of advanced MRI parameters, such as apparent diffusion coefficients or relative cerebral blood volumes, and radiomics analysis[27]) or automated contouring of target volumes for radiotherapy treatment planning, all of which require tumour segmentation masks as a fundamental input. For example, in terms of imaging biomarker discovery, we not only confirm the prognostic importance of baseline CE tumour volume,[28] but also show that this parameter outperforms well known molecular (eg, *MGMT* promoter methylation status) or established clinical characteristics within the EORTC-26101 dataset, and consequently provide further rationale to include imaging parameters into clinical trial design.[29]

Extensive investigation into the comparison of 2D measurements (including RANO) with volumetric measurements of tumour burden has been done,[10,11,26,30] and a consensus exists that volumetric measurements are more reliable and accurate than 2D measurements,[4,8,10,11,24] which is also supported by our findings in the EORTC-26101 test dataset (with an increase in reliability from 51% for local *vs* central RANO assessment to 87% for

quantitative volumetric assessment based on ANN *vs* radiologist ground truth). However, we acknowledge that the added value of quantitative volumetric assessment might be less pronounced when comparing two neuroradiologists with extensive RANO experience than when comparing the readings of less experienced RANO readers with those of highly experienced RANO readers, such as in the EORTC-26101 test dataset (with potentially less experienced local RANO readers *vs* highly experienced central RANO readers). Moreover, whether the higher reliability and accuracy of quantitative volumetric measurements than 2D measurements would translate into clinical relevance has been uncertain until now. Indeed, the few studies[26,30] that have compared non-automated volumetric assessment of tumour response with RANO had divergent results, with some studies suggesting added value of volumetric assessment of tumours,[30] whereas others did not find evidence to favour volumetric assessment over RANO as a surrogate endpoint for predicting overall survival.[26] However, all these previous comparisons did not consider that patients might only develop new, anatomically distinct lesions, which would immediately qualify for progression with RANO but might not qualify for progression on the basis of volumetric thresholds because the overall tumour volume could remain below the prespecified threshold. We overcame this limitation and introduced an algorithm to automatically identify the occurrence of new tumour lesions during follow-up. By integrating this algorithm into our automated processing pipeline, we identified that up to 10% of patients fell into this category and would otherwise not have qualified for tumour progression at this timepoint. This finding was of substantial importance for unbiased comparison of the performance of quantitative volumetric assessment of tumour response versus RANO in our study. Consequently, by use of time-dependent Cox regression modelling, our results suggest superiority of time to progression calculated by the automated quantitative ANN-based assessment of tumour response over central RANO assessment as a surrogate endpoint for predicting overall survival.

Our study had some limitations. First, we acknowledge the retrospective nature of the study and the relatively small, single-centre dataset used for training of the ANN. Although we specifically enriched the Heidelberg training dataset with comparatively uncommon and difficult cases, a larger dataset might allow further improvement of the accuracy of the ANN. Second, given the short follow-up period in the Heidelberg simulation dataset, we were unable to investigate the accuracy of automated quantitative volumetric assessment of tumour response in comparison with RANO in this dataset. Third, the suggested added value of automated quantitative volumetric assessment of tumour response compared with RANO, both in terms of reliability and performance, as a surrogate endpoint for predicting overall survival in the EORTC-26101 test dataset requires further validation in a prospective setting. This investigation is currently ongoing via application of the fully automated MRI-processing pipeline within the XNAT infrastructure as part of central neuroradiological assessment for the N²M² umbrella multicentre trial (NCT03158389) in newly diagnosed patients with non-*MGMT* hypermethylated glioblastoma.[31] Moreover, refinement from a methodical perspective will focus on further improving the segmentation performance of the ANN. Specifically, although the ANN and its accompanying training scheme were heavily inspired by our contribution[17] to the BraTS 2017 challenge, we acknowledge that a complete understanding of all design choices and their relative contribution to segmentation performance could point us towards potential further improvements and thus constitutes a valuable topic of research for further projects. From a clinical perspective, we will focus on also including advanced MRI parameters (eg, apparent diffusion coefficients or relative cerebral blood volumes) into the automated analysis workflow, which is of specific importance in the era of immunotherapy—eg, for early separation of pseudoprogression from true progression.[32,33] Finally, the scalability and flexibility of our approach will enable further extension to other disease entities (eg, quantification of lesion load in multiple sclerosis).[34]

Overall, our results suggest that ANN can enable objective and automated assessment of tumour response and imaging biomarker discovery in neuro-oncology at high throughput, and could ultimately serve as a blueprint for the application of ANN in radiology to improve clinical decision making.

Contributors

PK, MB, WW, FI, and KHM-H designed the study. PK, IT, MNow, UN, DB, GB, MS, and MF did quality control of MRI data. PK, IT, DB, UN, GB, MS, and MF preprocessed the MRI data. FI developed, trained, and applied the artificial neural network. PK and FI postprocessed the data generated by the artificial neural network. JP, MNol, PK, MPr, FI, and KHM-H did clinical translation (ie, design and development of the automated postprocessing workflow, integration into the XNAT infrastructure, and application in the simulated clinical environment). WW, MB, IH, MJvdB, TG, FS, AvD, and MPl critically contributed to the primary analysis of the relevant data from the EORTC-26101 trial that were used within this study. FS, TK, and AvD analysed the methylation array data in the EORTC-26101 dataset. PK and TG did statistical analyses. PK, FI, MB, WW, and KHM-H interpreted the findings with essential input from all coauthors. PK, FI, and JP prepared the first draft of the manuscript. All authors critically revised the manuscript for important intellectual content and approved the final version.

Declaration of interests

DB reports personal fees from Profound Medical outside of the submitted work. FS reports personal fees from Agilent and Illumina outside of the submitted work. AR reports grants, personal fees, and other support from Bayer and Guerbet, and personal fees and other support from GE-Healthcare, Bracco, Siemens, Sanofi, and Medscape outside of the submitted work. JD reports grants from ViewRay, CRI—The Clinical Research Institute, Accuray, RaySearch Laboratories, Vision RT, Merck Serono, Astellas Pharma, AstraZeneca, Siemens Healthcare, Solution Akademie, Egomed, Quintiles, Pharmaceutical Research Association, Boehringer Ingelheim, PTW-Freiburg, and Nanobiotix outside of the submitted work. SH reports grants from German Research Council and Dietmar-Hopp Foundation outside of the submitted work. MPl reports non-financial support from Pfizer, and grants and personal fees from Bayer outside of the submitted work. MPl has a patent IDH1 vaccines

licensed, a patent H3 vaccine pending, and a patent AHR inhibitor with royalties paid to Bayer. MJvdB reports personal fees from Roche, Cellgene, Bristol-Myers Squibb, AGIOS, Merck Sharpe & Dohme, and Boehringer Ingelheim; and grants and personal fees from AbbVie outside of the submitted work. WW reports grants from Apogenix, Boehringer Ingelheim, Pfizer; grants and personal fees from Merck Sharp and Dohme and Roche; and personal fees from Bristol-Myers Squibb and Celldex outside of the submitted work. MB reports personal fees from Boehringer Ingelheim, Merck, Bayer, Teva, B Braun, Springer, and Vascular Dynamics; grants and personal fees from Novartis, Codman, and Guerbet; and grants from Siemens, Hopp Foundation, German Research Council, the European Union, Stryker, and Medtronic outside of the submitted work. All other authors declare no competing interests.

Acknowledgments

PK was supported by the Medical Faculty Heidelberg Postdoc-Program and the Else Kröner-Fresenius Foundation (Else-Kröner Memorial Scholarship). FS was supported by the Else Kröner-Fresenius Foundation (EKFS Excellence Scholarship).

References

1 Chinot OL, Wick W, Mason W, et al. Bevacizumab plus radiotherapy-temozolomide for newly diagnosed glioblastoma. *N Engl J Med* 2014; **370:** 709–22.

2 Gilbert MR, Dignam JJ, Armstrong TS, et al. A randomized trial of bevacizumab for newly diagnosed glioblastoma. *N Engl J Med* 2014; **370:** 699–708.

3 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:** 583–93.

4 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:** 1963–72.

5 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:** 2439–49.

6 Thust SC, Heiland S, Falini A, et al. Glioma imaging in Europe: a survey of 220 centres and recommendations for best clinical practice. *Eur Radiol* 2018; **28:** 3306–17.

7 Wick W, Gorlia T, Bendszus M, et al. Lomustine and bevacizumab in progressive glioblastoma. *N Engl J Med* 2017; **377:** 1954–63.

8 Korn RL, Crowley JJ. Overview: progression-free survival as an endpoint in clinical trials with solid tumors. *Clin Cancer Res* 2013; **19:** 2607–12.

9 Yang D. Standardized MRI assessment of high-grade glioma response: a review of the essential elements and pitfalls of the RANO criteria. *Neurooncol Pract* 2016; **3:** 59–67.

10 Chow DS, Qi J, Guo X, et al. Semiautomated volumetric measurement on postcontrast MR imaging for analysis of recurrent and residual disease in glioblastoma multiforme. *AJNR Am J Neuroradiol* 2014; **35:** 498–503.

11 Sorensen AG, Patel S, Harmath C, et al. Comparison of diameter and perimeter methods for tumor volume calculation. *J Clin Oncol* 2001; **19:** 551–57.

12 Kickingereder P, Götz M, Muschelli J, et al. Large-scale radiomic profiling of recurrent glioblastoma identifies an imaging predictor for stratifying anti-angiogenic treatment response. *Clin Cancer Res* 2016; **22:** 5765–71.

13 Kickingereder P, Neuberger U, Bonekamp D, et al. Radiomic subtyping improves disease stratification beyond key molecular, clinical and standard imaging characteristics in patients with glioblastoma. *Neuro Oncol* 2018; **20:** 848–57.

14 Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol* 2016; **131:** 803–20.

15 Ellingson BM, Bendszus M, Boxerman J, et al. Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. *Neuro Oncol* 2015; **17:** 1188–98.

16 Wick W, Stupp R, Gorlia T, et al. Phase II part of EORTC study 26101: the sequence of bevacizumab and lomustine in patients with first recurrence of a glioblastoma. *J Clin Oncol* 2016; **34** (suppl 15): 2019 (abstr).

17 Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH. Brain tumor segmentation and radiomics survival prediction: contribution to the BRATS 2017 challenge. In: Crimi A, Bakas S, Kuijf H, Menze B, Reyes M, eds. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. BrainLes 2017. Lecture notes in computer science, vol 10670. Springer International Publishing, 2018: 287–97.

18 Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging* 2015; **34:** 1993–2024.

19 Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, eds. Medical image computing and computer-assisted intervention—MICCAI 2015. MICCAI 2015. Lecture notes in computer science, vol 9351. Springer International Publishing, 2015: 234–41.

20 He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: Leibe B, Matas J, Sebe N, Welling M, eds. Computer vision—ECCV 2016. ECCV 2016. Lecture notes in computer science, vol 9908. Springer International Publishing, 2016: 630–45.

21 Han K, Ren M, Wick W, et al. Progression-free survival as a surrogate endpoint for overall survival in glioblastoma: a literature-based meta-analysis from 91 trials. *Neuro Oncol* 2014; **16:** 696–706.

22 Sedgwick P, Joekes K. Interpreting hazard ratios. *BMJ* 2015; **351:** h4631.

23 Tripepi G, Jager KJ, Dekker FW, Zoccali C. Statistical methods for the assessment of prognostic biomarkers (part I): discrimination. *Nephrol Dial Transplant* 2010; **25:** 1399–401.

24 Sorensen AG, Batchelor TT, Wen PY, Zhang WT, Jain RK. Response criteria for glioma. *Nat Clin Pract Oncol* 2008; **5:** 634–44.

25 Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 2006; **31:** 1116–28.

26 Gahrmann R, van den Bent M, van der Holt B, et al. Comparison of 2D (RANO) and volumetric methods for assessment of recurrent glioblastoma treated with bevacizumab-a report from the BELOB trial. *Neuro Oncol* 2017; **19:** 853–61.

27 Kickingereder P, Andronesi OC. Radiomics, metabolic, and molecular MRI for brain tumors. *Semin Neurol* 2018; **38:** 32–40.

28 Ellingson BM, Abrey LE, Nelson SJ, et al. Validation of postoperative residual contrast-enhancing tumor volume as an independent prognostic factor for overall survival in newly diagnosed glioblastoma. *Neuro Oncol* 2018; **20:** 1240–50.

29 Erickson BJ, Galanis E. Where size matters: imaging-based biomarkers for patient stratification. *Neuro Oncol* 2017; **19:** 7–8.

30 Boxerman JL, Zhang Z, Safriel Y, et al. Early post-bevacizumab progression on contrast-enhanced MRI as a prognostic marker for overall survival in recurrent glioblastoma: results from the ACRIN 6677/RTOG 0625 Central Reader Study. *Neuro Oncol* 2013; **15:** 945–54.

31 Wick W, Dettmer S, Berberich A, et al. N2M2 (NOA-20) phase I/II trial of molecularly matched targeted therapies plus radiotherapy in patients with newly diagnosed non-MGMT hypermethylated glioblastoma. *Neuro Oncol* 2019; **21:** 95–105.

32 Antonios JP, Soto H, Everson RG, et al. Detection of immune responses after immunotherapy in glioblastoma using PET and MRI. *Proc Natl Acad Sci USA* 2017; **114:** 10220–25.

33 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:** e534–42.

34 Brugnara G, Isensee F, Wick W, Bendszus M, Maier-Hein KH, Kickingereder P. Automated volumetric assessment of multiple sclerosis disease burden and activity with artificial neural networks. *Insights Imaging* 2019; **10** (suppl 1): 22 (abstr).

THE LANCET
Oncology

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: 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; published online April 2. http://dx.doi.org/10.1016/S1470-2045(19)30098-1.

Supplementary appendix

Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study.

Table of contents:

A) Supplementary Methods

1. Image Preprocessing

DICOM to NifTI conversion (https://nifti.nimh.nih.gov)) was performed with MRIConvert (mcverter, The Lewis Center for Neuroimaging, University of Oregon, United States, https://lcni.uoregon.edu/downloads/mriconvert/mriconvert-and-mcverter). All sequences were reoriented to the standard (MNI) orientation (fslreorient2std, FMRIB software library, FSL, http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL [1, 2]). Brain extraction of individual sequences (T1-weighted (T1-w), contrast-enhanced T1-weighted (cT1-w), FLAIR and T2-weighted (T2-w) image volumes) was followed by registration of the brain-extracted cT1-w, FLAIR and T2-w image volumes to the respective brain extracted T1-w image volume using the linear image registration tool (FMRIB software library, FSL, http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL [1, 2]) with spline interpolation and a 6-degree of freedom transformation. T1 subtraction volumes (subT1) were generated by voxel-wise subtraction of the T1-w from the cT1-w volume (both z-score normalized).
Ground truth tumor segmentation masks were generated for all patients in the HD training and test set (by PK a radiology resident with 6 years of experience and subsequently checked by DB a board-certified radiologist and neuro-radiologist with 17 years of experience in image processing; discrepancies were resolved through a consensus discussion) and *post-hoc* in the EORTC-26101 test set (by IT a radiology resident with 3 years of experience and subsequently checked by PK; discrepancies resolved through a consensus discussion). Specifically, the contrast-enhancing (CE) portion of the whole tumor (on the subT1 images) as well as the associated non-enhancing T2-signal abnormality (NE) (defined as T2/FLAIR hyperintense abnormality excluding the contrast-enhancing and necrotic portion of the tumor, resection cavity and obvious leukoaraiosis) were selected using a region-growing segmentation algorithm implemented in ITK-SNAP (www.itksnap.org [3]), as described previously [4, 5]. Depending on the complexity of the case semi-automated tumor segmentation and manual corrections took about 5 to 20 min per case.

2. Artificial Neural Network

In this section we describe the ANN that was developed for automated brain tumor segmentation (implemented with Python 3.6.3 (www.python.org) using the PyTorch package version 0.4.0 (www.pytorch.org)). The training set of the HD cohort was used to train and validate the ANN (with 5-fold cross-validation) whereas the longitudinal HD test set and the longitudinal multicentric EORTC-26101 cohort were used for independent large-scale testing and application of the ANN (all computations performed with NVIDIA (NVIDIA Corporation, California, United States) Titan Xp graphics processing units). All MRI sequences (T1-w, cT1-w, FLAIR, T2-w) and the corresponding "ground truth" tumor segmentation generated by the radiologist (containing the delineated CE tumor and NE) were resampled to an isotropic spacing of 1x1x1 mm^3. This was followed by MRI intensity normalization, which is a non-trivial issue that has received a lot of attention in the past [6]. In the context of brain tumor segmentation with ANN, however, it has been shown that neural networks have the ability to intrinsically learn such a normalization directly from the data, if necessary. For this reason, just like recent winning contributions to the Brain Tumor Segmentation Challenge (BraTS) [7-11], a simple preprocessing technique was employed: for each modality, the brain region is normalized by subtracting its mean value and dividing by its standard deviation. Voxels outside the brain are set to 0.

2.1. Network Architecture

The network architecture (depicted in Supplementary Figure 1) is inspired by our recent success [9] in the BraTS [7] which in turn was motivated by the U-Net architecture [12] and its 3D derivatives [13-15]. U-Net sets itself apart from other segmentation networks [8, 16-18] by the use of an encoder and a decoder network that are interconnected with skip connections. Conceptually, the encoder network is used to aggregate semantic information at the cost of reduced spatial information. The decoder is the counterpart of the encoder that reconstructs the spatial information while being aware of the semantic information extracted from the encoder. Skip connections are used to transfer feature maps from the encoder to the decoder to allow for even more precise localization of the tumor. This network processes 3-dimensional input patches as large as 128x128x128 voxels during training. Its fully convolutional nature is used to predict entire tumor segmentation mask at once at test time, alleviating the need to stitch patches together.

Heavy encoder, light decoder

Our instantiation of the U-Net utilizes pre-activation residual blocks [19] in the encoder. Contrary to plain convolutions, which learn a nonlinear transformation of the input, residual blocks learn a nonlinear residual that is added to the input. This allows the network by design to learn the identity function and ultimately enables the

utilization of deeper architectures and improves the gradient flow. Here, a residual blocks consists of two 3x3x3 convolutional layers, each of which is preceded by instance normalization and a leaky Rectified Linear Unit (ReLU) non-lineariy. We do not employ residual connections in the decoding pathway. Here, each concatenation is followed by a 3x3x3 convolutional layer that recombines semantic and localization information, followed by a 1x1x1 convolution that halves the number of feature maps. In the decoding pathway, we increase the resolution of the feature maps by means of trilinear upsampling followed by a 3x3x3 convolution that again halves the number of feature maps. This approach allows us to leverage the benefits of convolutional upsampling (typically transposed convolution) without the risk of introducing checkerboard artifacts.

Large Input Patch Size
When designing convolutional neural networks, the amount of available GPU memory is an important hardware constraint. While model parameters occupy only a fraction of this memory, feature map activations (which need to be stored for the backward pass) are typically quite large. This is especially true if using 3D convolutions. Thus, hyperparameters that influence the feature map size, such as the patch size ('how much of a patient does the network see at once'), the batch size ('how many examples does the network see at once') and the number of feature maps must be considered carefully. Keeping our feature maps at a constant 21 (doubling each time we downsample in the encoding pathway, see Summplementary Figure 1), preliminary experiments showed that trading a larger input patch size at the cost of a smaller batch size consistently improved our segmentation performance. This observation is supported by our contribution to the BraTS challenges 2017 [20] and 2018 [11] as well as other winning contributions in medical segmentation [10, 21]. Therefore, in order to maximize the amount of contextual information the encoder can aggregate, we train our network architecture with an input patch size of 128x128x128 voxels. At 1x1x1 mm³ voxel resolution this patch size almost covers an entire patient. Using such a large patch size enables the network to correctly reconstruct the tumor with the help of as much contextual information as possible. Naturally, reducing the batch size can have adverse effects on network training if the network is not adapted to compensate for that. Please refer to the Nonlinearity and Normalization subsection for more information.

Auxiliary Loss Layers
During training, gradients are optimized in a way that most quickly optimizes the loss function. In the case of a U-Net like architecture such as the one presented here, this may lead to too simple decision making in the early stages of the training, i.e. solving most of the segmentation problem by forwarding local structures recognized early in the encoder to the decoder instead of making use of the entire receptive field the network can access. Additionally, gradients at the lower parts of the U shape are typically lower due the nature of the chain rule. As a result, training the lower layers can be slow. We address both of these issues by integrating auxiliary loss layers deep into the network. These layers effectively create smaller versions of the desired segmentation, each of which are trained with its own loss layer and downsampled versions of the reference annotation.

Nonlinearity and Normalization
During model development we continuously observed dying ReLUs which motivated us to replace them with leaky ReLU nonlinearities throughout the network. Due to our small batch size (as described in the *Training Procedure* section below), batch mean and variance are unstable which may be problematic for batch normalization. For this reason we make use of instance normalization, which normalizes each sample in the batch independently of the others and which does not retain moving average estimates of batch mean and variance.

2.2. Training Procedure
Training was done with randomly sampled patches of size 128x128x128 voxels. These patches were cropped randomly from the four possible MRI input modalities (T1-w, cT1-w, FLAIR, and T2-w). The subT1 modality was not used as an additional explicit input, since we found no increase in network performance and instead observed that the network implicitly combines the relevant information from the T1-w and cT1-w modalities during training. Moreover, explicit inclusion of the subT1 modality may introduce errors when a 2D T1-w sequence is subtracted from a 3D cT1-w sequence. The network is optimized using stochastic descent with the Adam algorithm [22] (beta1=0.9, beta2=0.999, initial learning rate=$1e^{-4}$) and a minibatch size of 2. The training took 450 epochs, where we define one epoch as the iteration over 200 training batches. An exponential learning rate decay was included to the training scheme by applying the following learning rate schedule: $\alpha_{epoch} = a_0 * 0.99^{ep}$, where α_{epoch} represents the learning rate used at a specific epoch and $\alpha_0 = 10^{-4}$ is the initial learning rate. Motivated by successful recent work [9, 14, 15, 23, 24] a soft dice loss formulation for training the network was used.

$$l_D(U,V) = -\frac{2}{|K|}\sum_{k\in K}\frac{\sum_i u_{i,k}v_{i,k}}{\sum_i u_{i,k} + \sum_i v_{i,k}}$$

Here, $u \in U$ denotes the voxels of the softmax output and $v \in V$ denotes a one hot encoding of the corresponding ground truth segmentation mask. Both U and V have shape Kx128x128x128 where $k \in K$ are the foreground classes (CE tumor and NE). i is used to index pixels in a patch (discarding spatial information; $i \in 128^3$).

As stated in the previous section, each auxiliary loss layer has its own dice loss term and is trained on a downsampled version of the reference annotation. The global loss is then computed as the weighted sum of these loss terms:

$$l = 0.25 l_{D,\frac{1}{4}} + 0.5 l_{D,\frac{1}{2}} + 1 l_{D,\frac{1}{1}}$$

, where $l_{D,\frac{1}{4}}$ refers to the auxiliary loss layer that processes segmentations at $\frac{1}{4}$ resolution.

2.3. Data Augmentation

Due to their high capacity, neural networks tend to overfit given a limited amount of training data. Besides explicit regularization such as weight decay, stochastic gradient descent and dropout, implicit regularization in the form of data augmentation has proven to be very effective [25]. For this reason we apply a broad range of data augmentation techniques on the fly during training using an in-house developed framework (available via http://github.com/MIC-DKFZ/batchgenerators). Hereby, U(a, b) denotes the uniform distribution on the interval [a, b].

- All input patches are mirrored randomly along all axes with probability 50%.
- 66.7% of patches are augmented with spatial transformations. These transformations include scaling, rotation and elastic deformation. Scaling is applied with a random scaling factor sampled from U(0.75, 1.25). Rotation is performed around all three axes with a random angle sampled from U(-25°, 25°) for each axis. Elastic deformation is implemented by sampling a grid of random, Gaussian distributed displacement vectors (μ=0, σ=1) which is then smoothed by a Gaussian smoothing filter with σ sampled uniformly from U(9, 13) and finally scaled by a randomly chosen scaling factor sampled uniformly from U(0, 900). We then apply the smoothed rescaled displacement vector field to the image and the corresponding segmentation via third order spline interpolation and nearest neighbor interpolation, respectively.
- Finally, we apply gamma augmentation to 50% of the patches. Gamma augmentation is done by transforming the voxel intensities to the interval [0, 1] and then applying the following equation for each voxel I.

  $$I_{transformed} = I^{\gamma}$$

  γ is hereby sampled from U(0.8, 1.5) once for each modality.
- Since the gamma augmentation alters the mean and standard deviation of the patches during training, whereas the network will only be presented z-score normalized inputs at test time, patches are renormalized to their original mean and standard deviation before being fed into the network.

2.4. Evaluation

For evaluation, we use the same preprocessing technique as applied for the training data (resampling to 1x1x1mm³ isotropic voxel spacing followed by intensity normalization as described above). We make use of the fully convolutional nature of our ANN to predict an entire patient at once, thus alleviating the need to stitch patched together.

To increase the quality of our segmentations, we apply test time data augmentation in the form of mirroring the data along all three axes. This results in eight predictions for each patient which are then averaged to produce the final segmentation (by means of softmax averaging prior to resampling).

For the prediction of the HD training set we predict the patients of the held out validation set of each fold with its corresponding network. For the predictions of both the HD test set and EORTC-26101 test set we use the five networks obtained from the cross-validation as an ensemble to further increase our segmentation performance. This ensembling together with test time data augmentation results in 40 predictions per patient that are averaged to obtain the final result.

The predicted output tumor segmentation masks generated by the ANN were linearly resampled to the original resolution of the input MRI sequence and were used as further input for the analysis described in the section "Statistical Analysis" of the main manuscript and data supplement below.

3. Postprocessing & Statistical Analysis

3.1. Accuracy of the automatically generated tumor segmentation masks

Agreement between the radiologists-generated ground truth tumor segmentation mask and the automatically generated ANN-based tumor segmentation masks was evaluated separately for CE tumor and NE within each of the three datasets based on the DICE similarity coefficient (which can range from 0, indicating no spatial overlap, to 1, indicating complete overlap of the segmentation masks). Furthermore, agreement in tumor volumes (separately for CE tumor and NE) derived from the ground truth and ANN segmentation masks was assessed with Bland-Altman plots (where the difference between the tumor volumes automatically predicted by the ANN vs. those generated by the radiologist (ground truth) is plotted against the average of both volumes with horizontal solid lines indicating the mean difference and dotted lines indicating the limits of agreement (95% confidence

interval)) and concordance correlation coefficients (which measures precision and accuracy by determining how far the observed data deviates from the line of perfect concordance (that is, the line at 45 degrees on a square scatter plot with perfect agreement at 1)).

3.2. Criteria for quantitative tumor response assessment

For both longitudinal HD test and EORTC-26101 test sets, automated quantitative tumor response assessment was separately performed based on the radiologists-generated ground truth or automatically ANN-based tumor segmentation masks. To qualify for quantitative tumor progression at least one of the following criteria was required to be met: (1) tumor volume increase (as compared to baseline or best response) in either CE or NE or (2) occurrence of a new CE lesion outside the CE tumor volume from the preceding MRI exam.

For criteria 1, we computed the volumes of both the CE tumor and NE parts from the segmentation masks (for both the ground truth masks as well as the masks predicted by our ANN). A volumetric threshold of 40% was applied to qualify for progression of CE (beyond a minimal tumor volume of 1 cm³ i.e. analogously to the differentiation between "measurable" and "non-measurable" CE at a cut-off of 1x1 cm for the biperpendicular diameter performed by the RANO criteria) which is justified by the equivalent 2-dimensional threshold used by the RANO working group criteria (i.e. 25% increase in the product of perpendicular diameters of the CE tumor) [28]. To qualify for progression based on NE we selected a volumetric threshold of (a) 40% for patients with lower-grade glioma, or (b) 100% for patients with glioblastoma. Specifically, the 40% volumetric threshold for NE in patients with lower-grade glioma is again justified by the equivalent 2-dimensional threshold used by the RANO working group criteria (i.e. 25% increase in the product of perpendicular diameters of the NE lesion) [29]. In contrast, for patients with glioblastoma the RANO criteria only suggest that a "significant increase" in NE qualifies for progression, however they have not yet recommended a quantitative 2-dimensional threshold [28]. To account for the lower specificity of NE as compared to CE to reflect tumor burden in patients with glioblastoma we selected a conservative volumetric threshold of 100% to qualify for progression of NE. In summary, the selected thresholds of 40% and 100% tumor volume increase correspond to an increase of 25% and 59% in the product of perpendicular diameters, or to an increase of 12% and 26% in the single diameter assuming spherical configuration of the tumor i.e. $volume = \frac{4}{3} * \pi * \left(\frac{diameter}{2}\right)^3$.

For criteria 2, to enable automated identification of a new CE lesion during follow-up, all MRI exams from each patient were spatially aligned in the same image space by registration of all brain extracted follow-up T1-w volumes to the respective brain extracted baseline T1-w volume using the linear image registration tool of the FMRIB software library [1, 2] with spline interpolation and a 6-degree of freedom transformation. The derived registration matrices were then used for spatial alignment of the respective ground truth and ANN based tumor segmentation masks (using nearest neighbor interpolation). Next, a connected component analysis (implemented in Python 3.6.3 (www.python.org) with the scikit-image library (https://scikit-image.org)) was used to identify the occurrence of a new CE lesions within the current MRI exam located outside of the anatomical boundaries of the CE volume from the preceding MRI exam (to disregard those cases where a tumor shrinks into several smaller sub-volumes) (see Supplementary Figure 2 for schematic illustration).

3.3. Association of baseline tumor volumes and its early change with overall survival (OS)

We evaluated the relevance of CE tumor and NE volumes automatically generated by the ANN at baseline and the early percent-wise change in those volumes (between baseline and 1st follow-up MRI) for predicting OS within the EORTC-26101 trial. For this purpose multivariable Cox proportional hazards regression models for OS with baseline (CE and NE) tumor volumes and the early change in those tumor volumes were selected as covariates. Both models were adjusted for potential confounders by including the patient's age, sex (female vs. male), WHO performance status (>0 vs. 0) at baseline, steroid administration (yes vs. no) at baseline and *O⁶-methylguanine–DNA methyltransferase* (*MGMT*) promoter methylation status (methylated vs. unmethylated) as additional covariates (limited to a subset of 261 patients with complete data, see Supplementary Table 2c). Specifically, *MGMT* promoter methylation status was assessed with the Illumina Infinium HumanMethylation450 array based on the MGMT-STP27 model, as described previously [30, 31]. Importance of covariates in the model was assessed by computing the Wald χ^2 statistic and the proportion of the overall model χ^2 that is due to each covariate (using the anova.rms function of the rms package in R). All statistical analyses were performed with R version 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria)). P-values <0.05 were considered significant.

4. Clinical Integration

We developed an infrastructure based entirely on open source software components that allows clinical translation and application of our ANN for automated tumor segmentation and tumor response assessment (schematic illustration, Supplementary Figure 6). So far, common clinical practice is to send MR examinations after their acquisition from the MR scanner directly to the in-house PACS (picture archiving and communication system) for storage, retrieval and interpretation. To allow completely automated, vendor-neutral, on-demand processing of MR examinations in an end-to-end fashion we make use of the XNAT open-source imaging informatics software

platform (https://www.xnat.org) [33]. This allows that MR examinations may also be sent from the MR scanner to a dedicated XNAT postprocessing server (or alternatively e.g. within clinical trials upload previously acquired MR examinations directly to the XNAT server). Incoming MR examinations automatically trigger execution of a Docker container (https://www.docker.com) via the XNAT Container Service Plugin (https://github.com/NrgXnat/container-service). The Docker container encompasses our developed processing pipeline that executes the following steps in a completely automated fashion: (1) parallelized conversion of DICOM images to NIfTI (https://nifti.nimh.nih.gov) format, (2) parallelized reorientation of all sequences to the standard (MNI) orientation (fslreorient2std, FMRIB software library, FSL, http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL [1, 2]), (3) ANN based brain extraction of all sequences (T1-w, cT1-w, FLAIR and T2-w) with an in-house developed algorithm (https://github.com/MIC-DKFZ/HD-BET) optimized for processing heterogeneous MRI data with varying degree of pathologies or post-treatment alterations (performed on the GPU using NVIDIA-Docker (version 2.0, https://github.com/NVIDIA/nvidia-docker), (4) parallelized registration of individual MR sequences (T1-w, cT1-w, FLAIR and T2-w) to the respective T1-w image space and calculation of subT1 maps, (5) tumor segmentation through the ANN described within this manuscript (also performed on the GPU), (6) comparison of CE tumor and NE volumes with those from previous MR examinations and construction of a chart depicting the longitudinal tumor volume dynamics, (7) parallelized semi-transparent superimposition of tumor segmentation mask on individual MR sequences, and finally (8) submission of the processed results (chart depicting the longitudinal tumor volume dynamics as well as semi-transparent superimposed tumor segmentation mask on individual MR sequences) in DICOM format to the PACS where they are (along with the regular MRI sequences) available for interpretation and further application. Data transfer to and from XNAT uses the DICOM protocol (http://www.dicomstandard.org) requiring no changes made to the infrastructure used in a radiology department.
Processing per MRI exam with this workflow takes an average of 10min14s on a machine with an Intel (Intel Corporation, Santa Clara, California, United States) Xeon(R) E5-1650 v3 central processing unit (CPU) (12 cores at 3.5GHz each) and a NVIDIA Titan Xp GPU. This setup can already accommodate 3 routines running simultaneously. We use Docker Swarm (https://github.com/docker/swarm) as a scheduler, so if the need for larger scale processing arises, the system can be scaled up linearly by adding additional processing nodes to the cluster without any need to interrupt the existing workflow.

References:

1. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage. 2002;17(2):825-41. (last accessed on 28.12.2018)
2. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Medical image analysis. 2001;5(2):143-56. (last accessed on 28.12.2018)
3. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage. 2006;31(3):1116-28. (last accessed on 28.12.2018)
4. Kickingereder P, Neuberger U, Bonekamp D, Piechotta PL, Gotz M, Wick A, et al. Radiomic subtyping improves disease stratification beyond key molecular, clinical and standard imaging characteristics in patients with glioblastoma. Neuro Oncol. 2018;20(6):848-57. (last accessed on 28.12.2018)
5. Kickingereder P, Götz M, Muschelli J, Wick A, Neuberger U, Shinohara RT, et al. Large-scale Radiomic Profiling of Recurrent Glioblastoma Identifies an Imaging Predictor for Stratifying Anti-Angiogenic Treatment Response. Clin Cancer Res. 2016;22(23):5765-71. (last accessed on 28.12.2018)
6. Shinohara RT, Sweeney EM, Goldsmith J, Shiee N, Mateen FJ, Calabresi PA, et al. Statistical normalization techniques for magnetic resonance imaging. NeuroImage: Clinical. 2014;6:9-19. (last accessed on 28.12.2018)
7. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging. 2015;34(10):1993-2024. (last accessed on 28.12.2018)
8. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical image analysis. 2017;36:61-78. (last accessed on 28.12.2018)
9. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH. Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge. In: Crimi A., Bakas S., Kuijf H., Menze B., Reyes M. (eds) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2017. Lecture Notes in Computer Science, vol 10670. Springer, Cham. (last accessed on 28.12.2018)
10. Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. arXiv preprint arXiv:181011654. 2018. (last accessed on 28.12.2018)
11. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH. No New-Net. arXiv preprint arXiv:180910483. 2018. (last accessed on 28.12.2018)
12. Ronneberger O, Fischer P, Brox T, editors. U-net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N., Hornegger J., Wells W., Frangi A. (eds) Medical Image Computing and Computer-

Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham. (last accessed on 28.12.2018)
13. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O, editors. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin S., Joskowicz L., Sabuncu M., Unal G., Wells W. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. MICCAI 2016. Lecture Notes in Computer Science, vol 9901. Springer, Cham. (last accessed on 28.12.2018)
14. Milletari F, Navab N, Ahmadi S-A, editors. V-net: Fully convolutional neural networks for volumetric medical image segmentation. arXiv preprint arXiv: 1606.04797v1. 2016. (last accessed on 28.12.2018)
15. Kayalibay B, Jensen G, van der Smagt P. CNN-based segmentation of medical imaging data. arXiv preprint arXiv:170103056. 2017. (last accessed on 28.12.2018)
16. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with deep neural networks. arXiv preprint arXiv: 1505.03540v3. 2016. (last accessed on 28.12.2018)
17. Kleesiek J, Urban G, Hubert A, Schwarz D, Maier-Hein K, Bendszus M, et al. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. NeuroImage. 2016;129:460-9. (last accessed on 28.12.2018)
18. Zhao G, Liu F, Oler JA, Meyerand ME, Kalin NH, Birn RM. Bayesian convolutional neural network based MRI brain extraction on nonhuman primates. Neuroimage. 2018;175:32-44. (last accessed on 28.12.2018)
19. He K, Zhang X, Ren S, Sun J, editors. Identity mappings in deep residual networks. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9908. Springer, Cham. (last accessed on 28.12.2018)
20. Isensee F, Kickingereder P, Bonekamp D, Bendszus M, Wick W, Schlemmer HP, et al. Brain tumor segmentation using large receptive field deep convolutional neural networks . In: Maier-Hein, geb. Fritzsche K., Deserno, geb. Lehmann T., Handels H., Tolxdorff T. (eds) Bildverarbeitung für die Medizin 2017. Informatik aktuell. Springer Vieweg, Berlin, Heidelberg. (last accessed on 28.12.2018)
21. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, et al. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. arXiv preprint arXiv:180910486. 2018. (last accessed on 28.12.2018)
22. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014. (last accessed on 28.12.2018)
23. Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In: Cardoso M. et al. (eds) Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA 2017, ML-CDS 2017. Lecture Notes in Computer Science, vol 10553. Springer, Cham (last accessed on 28.12.2018)
24. Drozdzal M, Vorontsov E, Chartrand G, Kadoury S, Pal C. The importance of skip connections in biomedical image segmentation. arXiv preprint arXiv: 1608.04117v2. 2016. (last accessed on 28.12.2018)
25. Hernández-García A, König P. Data augmentation instead of explicit regularization. arXiv preprint arXiv:180603852. 2018. (last accessed on 28.12.2018)
26. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-10. (last accessed on 28.12.2018)
27. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1):255-68. (last accessed on 28.12.2018)
28. Wen PY, Macdonald DR, Reardon DA, Cloughesy TF, Sorensen AG, Galanis E, et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. J Clin Oncol. 2010;28(11):1963-72. (last accessed on 28.12.2018)
29. van den Bent MJ, Wefel JS, Schiff D, Taphoorn MJ, Jaeckle K, Junck L, 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):583-93. (last accessed on 28.12.2018)
30. Bady P, Delorenzi M, Hegi ME. Sensitivity Analysis of the MGMT-STP27 Model and Impact of Genetic and Epigenetic Context to Predict the MGMT Methylation Status in Gliomas and Other Tumors. J Mol Diagn. 2016;18(3):350-61. (last accessed on 28.12.2018)
31. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. DNA methylation-based classification of central nervous system tumours. Nature. 2018. (last accessed on 28.12.2018)
32. Devarajan K, Ebrahimi N. Testing for Covariate Effect in the Cox Proportional Hazards Regression Model. Communications in statistics: theory and methods. 2009;38(14):2333-47. (last accessed on 28.12.2018)
33. 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. (last accessed on 28.12.2018)
34. Kamnitsas K, Bai W, Ferrante E, McDonagh S, Sinclair M, Pawlowski N, et al., editors. Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Crimi A., Bakas S., Kuijf H., Menze B., Reyes M. (eds) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2017. Lecture Notes in Computer Science, vol 10670. Springer, Cham. (last accessed on 28.12.2018)
35. Isensee F, Jaeger PF, Full PM, Wolf I, Engelhardt S, Maier-Hein KH, editors. Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features. In: Pop M. et al. (eds) Statistical

Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. STACOM 2017. Lecture Notes in Computer Science, vol 10663. Springer, Cham. (last accessed on 28.12.2018)

B) Supplementary Results

1. Performance metrics obtained in the training set

All metrics in the HD training set were obtained from 5-fold cross-validation. Thereby the tumor segmentation agreement between the automated volumetric segmentation with the artificial neural network (ANN) and the radiologist-generated ground truth tumor segmentation using DICE coefficients was on median 0·883 (95% CI, 0 ·876-0·892) for CE tumor and 0·905 (95% CI, 0·899-0·914) for NE. Corresponding median concordance correlation coefficient (CCC) were 0·986 (95% CI, 0·983 - 0·988) for CE tumor and 0·978 (95% CI, 0·974 - 0·982) for NE (Supplementary Figure 4 – 1st row). Visual inspection of Bland-Altman plots reveals excellent agreement between the radiologist-generated ground truth and the automatically-generated ANN tumor volumes (Supplementary Figure 5 – 1st row).

2. Impact of 2D vs. 3D MRI sequence acquisition on tumor segmentation performance

Comparison of 2D vs. 3D T1/cT1 acquisition on the performance for segmenting CE tumor in the EORTC-26101 dataset demonstrated higher agreement (by means of the DICE coefficient) between the artificial neural network (ANN) and the radiologist-generated ground truth segmentation when using 3D T1/cT1 acquistion. Specifically, median DICE values for CE were 0.914 (95% CI, 0.904-0.920) with 3D T1 / cT1 acquisition versus 0.838 (95% CI, 0.787-0.872) with 2D T1 / cT1 acquisition. Wilcoxon-signed rank sum test demonstrated a significant positive location shift when using 3D T1 / cT1 acquisition ($p<0.0001$).
In contrast, there was no significant difference between 2D vs. 3D FLAIR acquisition on the performance for segmenting NE in the EORTC-26101 dataset. Specifically, median DICE values for NE were 0.925 (95% CI, 0.919-0.933) with 3D FLAIR acquisition versus 0.934 (95% CI, 0.930-0.938) with 2D FLAIR acquisition. Wilcoxon-signed rank sum test demonstrated no significant positive location shift when using 3D FLAIR acquisition ($p=0.058$).

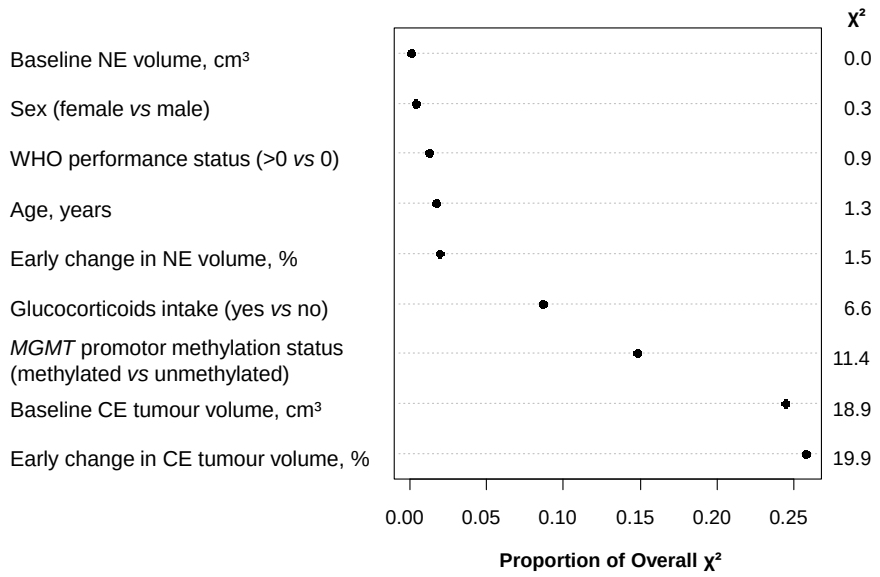
3. Individual visualization & integrative discussion of performance metrics obtained across all dataset

Performance metrics are individually visualized for each of the three datasets (HD training set, HD test set, EORTC-26101 test set), including DICE similarity coefficient for tumor segmentation agreement (Supplementary Figure 3, Supplementary Table 5) and concordance correlation coefficients and Bland-Altman plots for tumor volume agreement (Supplementary Figure 4-5) between the artificial neural network (ANN) and the radiologist-generated ground truth. This allows to visually compare the performance metrics across the different datasets. Of note are the somewhat higher DICE coefficients in both test sets (HD test set, EORTC-26101 test set) as compared to the HD training set. Specifically, median DICE coefficients for segmentation of CE increased by +0.002 in the HD test set and by +0.023 in the EORTC-26101 test set. Similarly, median DICE coefficients for segmentation of NE increased by +0.024 in the HD test set and by +0.027 in the EORTC-26101 test set. Several reasons may accounted for this: First, the use of an ensemble for predicting test set cases (i.e. by ensembling all models obtained from the 5-fold cross-validation of the training set) has previously been shown to systematically increase DICE scores [10, 34], and sometimes even lead to higher test set scores as compared to the corresponding training set cross-validation [21, 35]. Second, the HD training set – as indicated in the Materials & Methods (main manuscript) – was established in a non-consecutive fashion and enriched for comparatively uncommon and difficult cases based on the judgment of the neuroradiologists who established the training set (e.g. cases with complex resection cavities, extensive post-treatment alterations, or dot-like contrast enhancing tumor spots). The rationale for this was to encourage the ANN to learn the tumor segmentations as well as possible from a reasonably sized training cohort (455 MRI scans from 455 individual patients) and to cover for eventualities (i.e. difficult cases) that must be expected in larger-scale test cohorts. In contrast to the HD training set, both test sets (HD test set, EORTC-26101 test set) were longitudinal datasets. The 40 individual patients in the HD test (239 MRI scans) were also selected on a non-consecutive basis, but the criteria were different when compared to HD train. Here, only patients with sufficient longitudinal follow-up i.e. multiple MRI scans throughout their disease were included to enable the measurement of tumor volume change over time. Compared to the 455 individual and specifically selected patients in the HD training set, this resulted in a different distribution of MRI scans (lower variability) and is of importance when considering the DICE score distributions in the HD training set as compared to the HD test set. Finally, the EORTC-26101 test set represented a "real world" clinical trial dataset with longitudinal data of 532 patients (2034 MRI scans). Again this dataset was not specifically enriched for complex cases, which is again of importance when considering the somewhat higher DICE scores in the EORTC-26101 test set as compared to the HD training set. Note the expected broader distribution of DICE coefficients in the EORTC-26101 test set as compared to the HD test set (see Supplementary Figure 3 and Supplementary Table 5 with the broader interquartile range in the boxplot).

Of note are the outliers in Supplementary Figure 3 with DICE scores of 0 for some of the cases (primarily restricted to the segmentation of CE tumor e.g. occurring in 68 / 2034 of MRI exams (3.3%) in the EORTC-26101 test set as compared to 2 / 2034 MRI exams (0.1%) for segmentation of NE lesions in the EORTC-26101 test set) which predominantly arise in the post-treatment setting of longitudinal datasets, where MRI exams frequently display only very small contrast-enhancing tumor spots (e.g. after gross-total resection where little or no contrast-enhancing tumor is left). Here it is often difficult to finally judge whether a contrast-enhancing spot corresponds to contrast-enhancing tumor or to reactive gliosis (e.g. at the borders of the resection cavity) and frequently only further follow-up imaging will allow to differentiate. The DICE scores of 0 reflect the uncertainly in this setting i.e. the discrepancy between the radiologist-generated ground truth tumor segmentations and those predicted by the ANN. Specifically, even a single falsely segmented voxel by the ANN results in a DICE score of 0 if the ground truth segmentation does not contain any CE at all. Moreover we acknowledge that DICE scores of 0 may arise in those cases where a blood vessel that passes through the tumor is wrongly segmented as CE. The implications of DICE scores of 0 for downstream-analysis are however limited since we apply a uniform threshold of 1 cm^3 beyond which either CE tumor and NE has to increase to qualify for tumor progression (see Supplementary Methods, section 3.2 or main manuscript, Materials and Methods, Statistical analysis section).

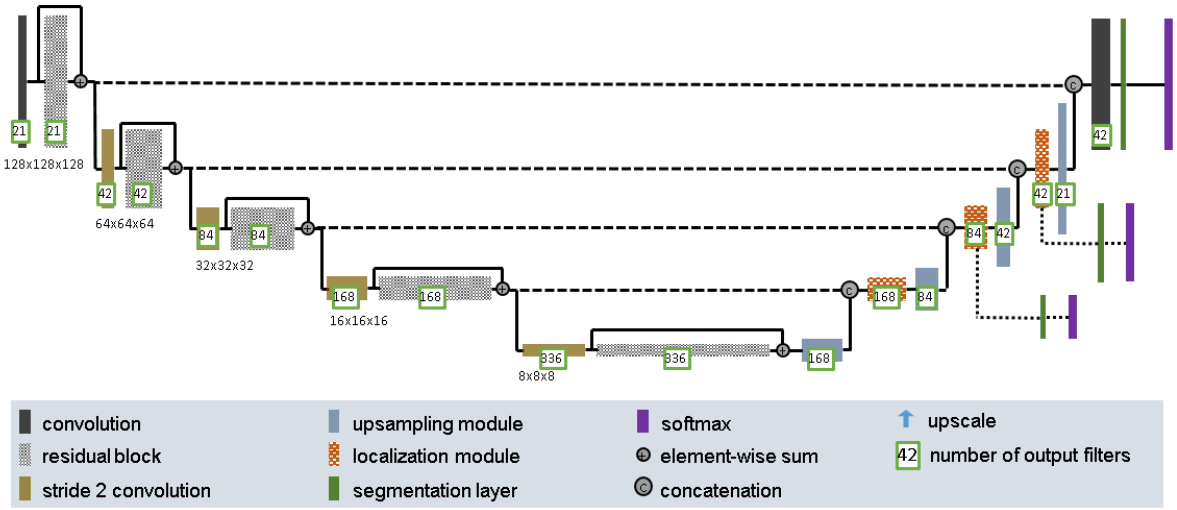
4. Importance of CE tumour and NE volumes and their early changes between baseline and first follow-up for predicting overall survival

We evaluated the relevance of CE tumor and NE volumes automatically generated by the ANN at baseline and the early percent-wise change in those volumes (between baseline and 1st follow-up MRI) in addition to molecular characteristics (MGMT promoter methylation status) and clinical characteristics for predicting OS within the EORTC-26101 trial. The relative importance of each covariate in the multivariate Cox proportional hazards regression model (assessed by computing the Wald χ^2 statistic and the proportion of overall model χ^2 that is due to each covariate) is illustrated in the chart below. Specifically, baseline CE tumour volume and early change in CE tumour volume showed the highest χ^2 values (18·87 and 19·88) and contributed 25% and 26% to the overall model χ^2 of 76·97. This was followed by *MGMT* promoter methylation status with an χ^2 value of 11·42 and glucocorticoid intake with an χ^2 value of 6·64 thus contributing 15% and 9% to the overall model χ^2. All of the remaining covariates in the model were not significant (p>0·05 each).

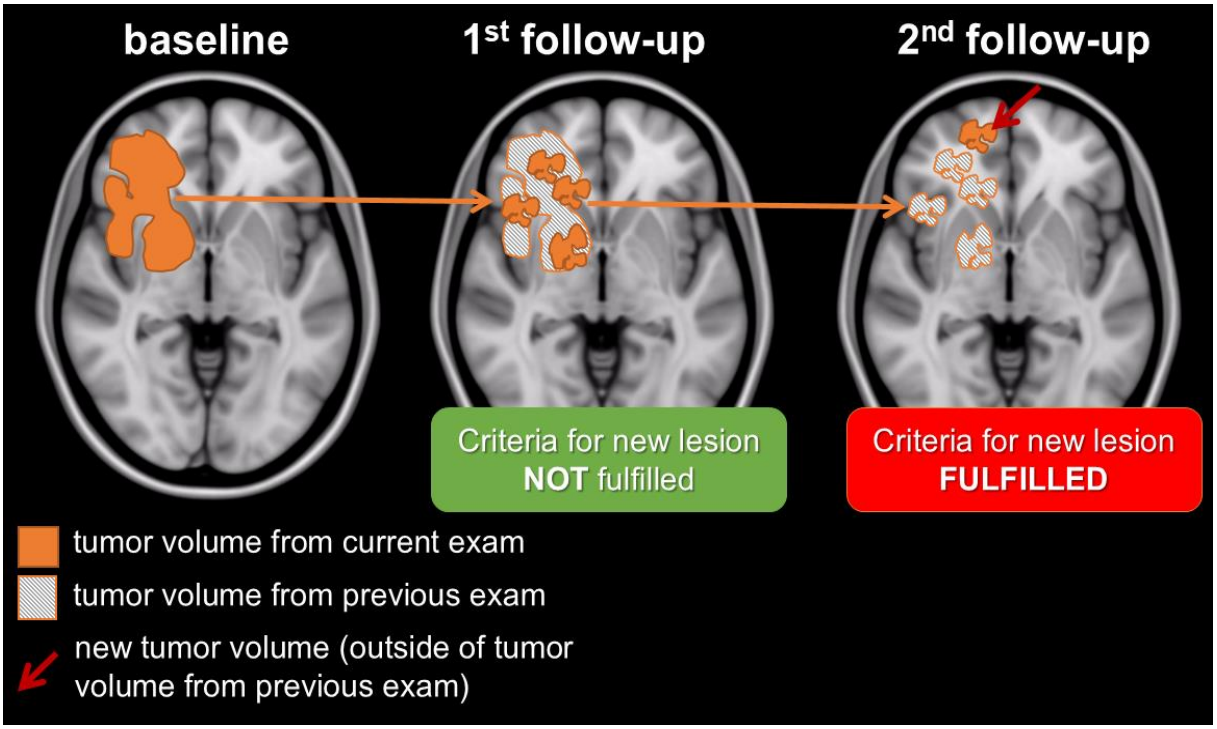


C) Supplementary Figures

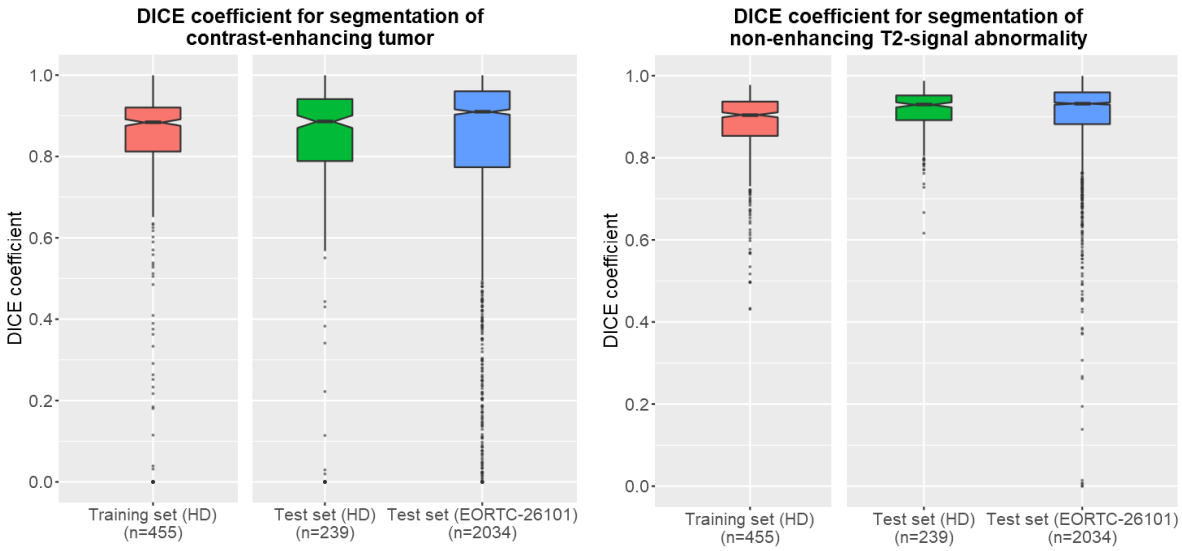
Supplementary Figure 1. Artificial neural network (ANN) architecture for automated tumor identification and segmentation on MRI. Our network architecture makes use of the encoder-decoder paradigm first introduced by the U-Net. We use residual connections in the encoder while keeping the decoder as lightweight as possible. Auxiliary segmentation outputs inject gradients deep into the network and facilitate the training of all layers. This network processes 3-dimensional input patches as large as 128x128x128 voxels during training. Its fully convolutional nature is used to predict entire tumor segmentation mask at once at test time, alleviating the need to stitch patches together.



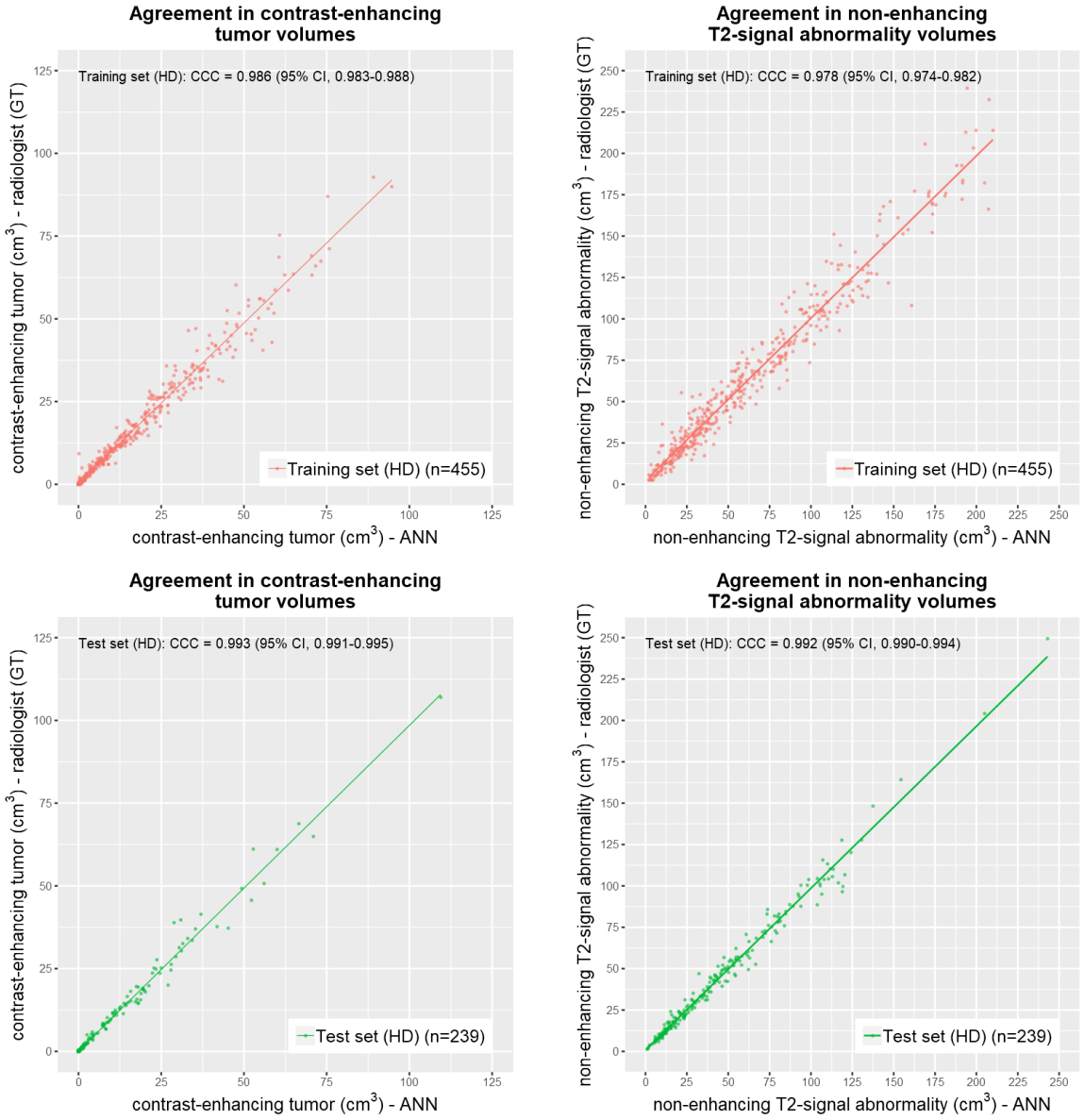
Supplementary Figure 2. Schematic illustration of the connected component analysis to enable automated identification of new contrast-enhancing (CE) tumor lesions during follow-up. Analysis is performed outside the boundaries of the CE tumor volume from the preceding MRI exam to disregard those cases where a tumor shrinks into several smaller sub-volumes.

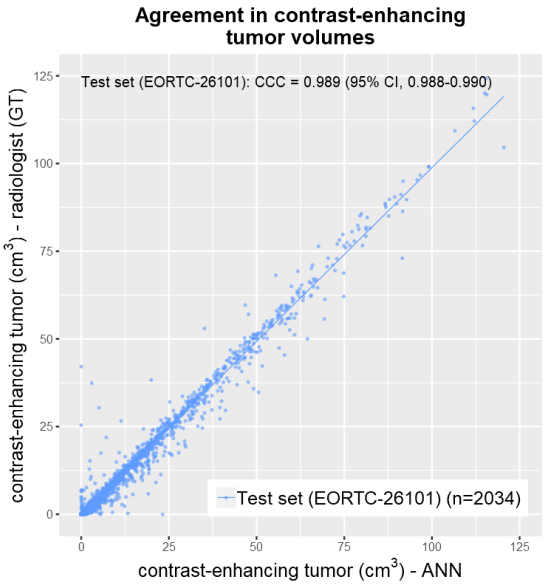


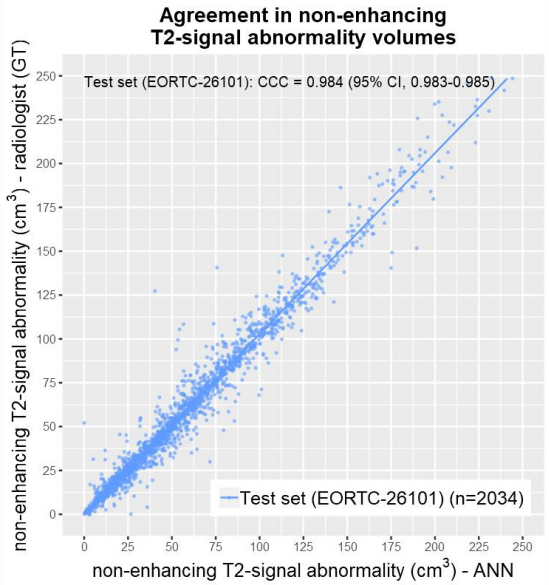
Supplementary Figure 3. DICE coefficients in the HD training set (obtained using 5-fold cross validation) and in both test sets (HD test set, EORTC-26101 test set) for segmentation of contrast-enhancing (CE) tumor (left) and non-enhancing T2-signal abnormality (NE) (right) – also see Supplementary Table 5 for summary statistics.



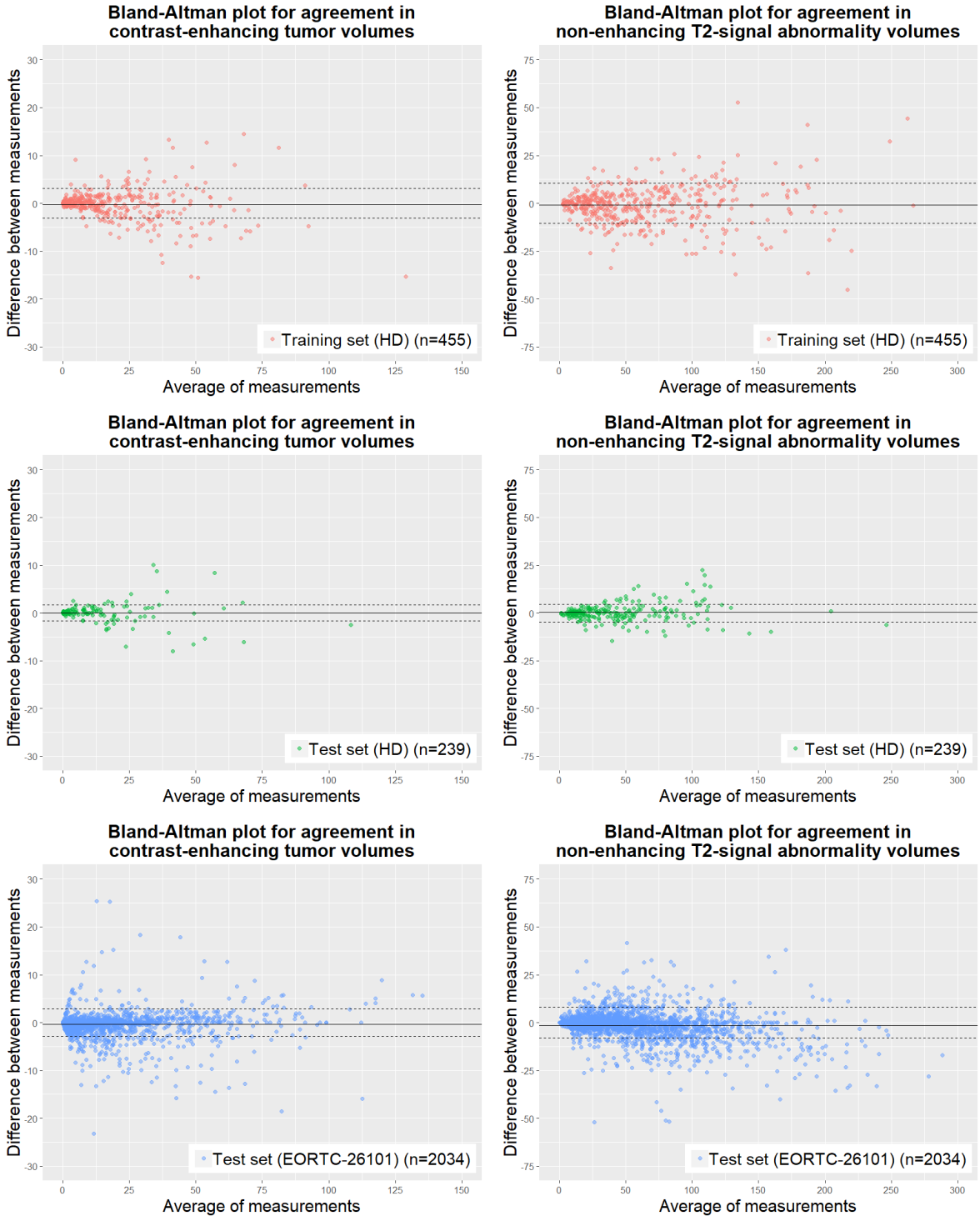
Supplementary Figure 4. Agreement between tumor volumes automatically predicted by the artificial neural network (ANN) and those generated by the radiologist (ground truth) separately illustrated for the HD training set (1st row), HD test set (2nd row) and EORTC-26101 test set (3rd row). Concordance correlation coefficients (CCC), which measure both precision and accuracy by determining how far the observed data deviate from the line of perfect concordance (that is, the line at 45 degrees on a square scatter plot with perfect agreement at 1), ranged from 0·986 to 0·993 for contrast-enhancing (CE) tumor volumes and from 0·978 to 0·992 for non-enhancing T2-signal abnormality (NE) volumes among the different datasets.



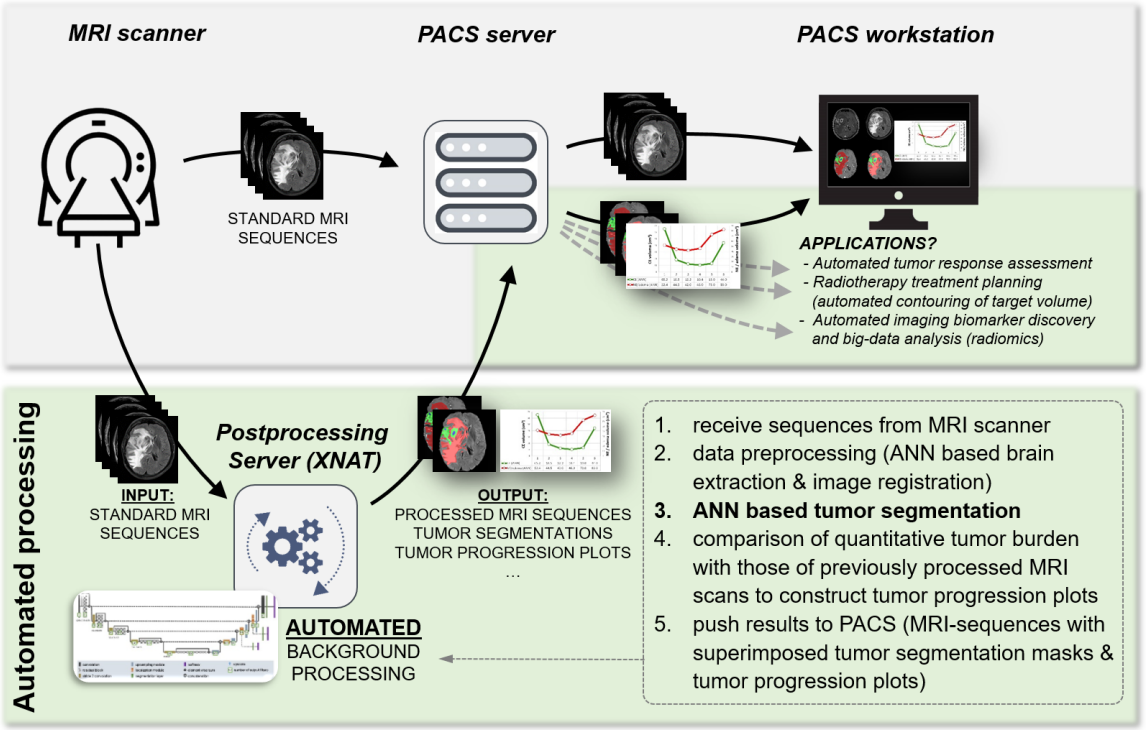

Agreement in contrast-enhancing tumor volumes
Test set (EORTC-26101): CCC = 0.989 (95% CI, 0.988-0.990)
Test set (EORTC-26101) (n=2034)
contrast-enhancing tumor (cm3) - radiologist (GT)
contrast-enhancing tumor (cm3) - ANN
0
25
50
75
100
125


Agreement in non-enhancing T2-signal abnormality volumes
Test set (EORTC-26101): CCC = 0.984 (95% CI, 0.983-0.985)
Test set (EORTC-26101) (n=2034)
non-enhancing T2-signal abnormality (cm3) - radiologist (GT)
non-enhancing T2-signal abnormality (cm3) - ANN
0
25
50
75
100
125
150
175
200
225
250

Supplementary Figure 5. Bland-Altman plots for the agreement on tumor volumes automatically predicted by the artificial neural network (ANN) and those generated by the radiologist (ground truth) separately illustrated for the HD training set (1st row), HD test set (2nd row) and EORTC-26101 test set (3rd row). The difference between ground truth and ANN tumor volumes (cm³) is plotted against the average of both volumes (cm³). Horizontal solid lines are drawn at the mean difference and dotted lines at the limits of agreement (95% confidence interval). Visual inspection reveals excellent agreement between ground truth and ANN tumor volumes.



Supplementary Figure 6. The developed artificial neural network (ANN) for tumor segmentation and quantitative volumetric tumor response assessment is part of a scalable and fully automated processing pipeline for MRI exams implemented within the XNAT open-source imaging informatics software platform (www.xnat.org). This approach enables seamless, vendor-neutral integration independent of preexisting infrastructures, but also enables to make use of the existing XNAT capabilities to manage and coordinate the analysis of MRI data in large multi-site clinical trials. Automated on-demand processing is triggered after the images have been acquired on the MRI scanner (or alternatively e.g. within clinical trials uploaded to the XNAT server). Processing is performed in a fully automated fashion and does not require any additional (manual) intervention. The processed results (e.g. superimposed tumor segmentation mask on individual MR sequences, chart depicting longitudinal tumor volume dynamics) are automatically pushed back to the PACS (picture archiving and communication system) where they are available for interpretation. Overall, this enables objective and automated tumor response assessment and imaging biomarker discovery in neuro-oncology at high throughput.



D) Supplementary Tables

Supplementary Table 1. Characteristics of the single-institutional training set (HD training) as well as the single-institutional longitudinal test set (HD test) and the multi-institutional longitudinal test dataset from the prospective randomized phase II and III EORTC 26101 trial (EORTC-26101 test set).

	Heidelberg (HD) training set	**Heidelberg (HD) test set**	**EORTC-26101 test set**
Institutions (n)	1	1	34
Patients (n)	455	40	532
MRI exams (n)	455	239	2034
MRI exams per patient median (IQR)	1 (1-1)	5 (4-6)	4 (3-5)
Histology			
glioblastoma	364 (80%)	25 (63%)	532 (100%)
lower-grade glioma	91 (20%)[1]	15 (38%)[2]	-
MRI scanners			
Siemens	455 (100%)	239 100%	805 (40%)
Unknown	-	-	68 (3%)
Verio	288 (63%)	170 (71%)	241 (12%)
Aera	-	-	209 (10%)
Avanto	-	-	124 (6%)
Symphony	-	-	103 (5%)
Trio	167 (37%)	69 (29%)	37 (2%)
Skyra	-	-	16 (1%)
Prisma	-	-	7 (0.3%)
Philips	-	-	427 (21%)
Achieva	-	-	211 (10%)
Ingenia	-	-	125 (6%)
Intera	-	-	89 (4%)
Panorama	-	-	2 (0.1%)
GE	-	-	756 (37%)
Signa series	-	-	442 (22%)
Optima (450)	-	-	162 (8%)
Discovery (750)	-	-	146 (7%)
Discovery (450)	-	-	6 (0.3%)
Toshiba	-	-	12 (1%)
Titan	-	-	12 (1%)
Unknown	-	-	34 (2%)
MRI field strength			
1.0 Tesla	-	-	2 (0.1%)
1.5 Tesla	-	-	604 (30%)
3.0 Tesla	455 (100%)	239 (100%)	459 (23%)
1.5 or 3.0 Tesla	-	-	867 (43%)
Unknown	-	-	102 (5%)

Annotation: 1 = including n=20 astrocytoma WHO °II, n=41 astrocytoma WHO °III, n=20 oligodendroglioma WHO °II and n=10 oligodendroglioma WHO °III; 2 = including n=2 astrocytoma WHO °II, n=11 astrocytoma WHO °III and n=2 oligodendroglioma WHO °III

Supplementary Table 2. Number of patients and MRI scans included for the different sub-analyses performed within the EORTC-26101 test set.

A.) Tumor volumetry of individual MRI scans:

	MRI scans	Patients
Total	2593	596
Excluded*	-559	64
Included	**2034**	**532**

B.) Quantitative tumor response assessment and comparison with RANO assessment:

	Patients
Total (from A)	532
Baseline MRI scan not included	-77
No follow-up MRI scan included	-56
Comparison with RANO assessment (local & central) not possible:	
Quantitative assessment not performed for MRI scan from the date of (local or central) RANO disease progression **	-36
Data on RANO assessment (central) not available	-47
Data on RANO assessment (local) not available	-10
Included	**306**

C.) Association of baseline tumor volumes and its early change with overall survival:

	Patients
Total (from A)	532
Baseline MRI scan not included	-77
No follow-up MRI scan included	-56
Incomplete molecular data (MGMT)	-138
Included	**261**

Annotation: * due to (a) incomplete availability of (axial or 3D oriented) pre- and postcontrast T1-weighted, FLAIR and T2-weighted sequences, (b) heavy motion artifacts, or (c) corrupt data; ** unless disease progression from volumetric assessment already occurred before local and central RANO disease progression.

Supplementary Table 3. Type of progression for quantitative assessment (based on the tumor volumes automatically predicted by the artificial neural network (ANN) vs. those generated by the radiologist (ground truth)) in both Heidelberg (HD) test set and EORTC-26101 test set. Type of progression was classified as progression of (a) contrast enhancing (CE) tumor only (b) non-enhancing T2-signal abnormality (NE) only (c) both CE and NE tumor volume, (d) appearance of a new CE lesion, or (d) no progression. Overall, patients most frequently qualified for progression because of an increase in the CE tumor volume (by including both "CE only" and "CE and NE" categories: 47·5% of patients in the HD test set and 61·8% in the EORTC-26101 test set).

Type of progression	**HD test set**		**EORTC-26101 test set**	
	Quantitative (radiologist)	Quantitative (ANN)	Quantitative (radiologist)	Quantitative (ANN)
CE only[1]	35.0% (14)	35.0% (14)	46.1% (141)	48.4% (148)
NE only[2]	27.5% (11)	25.0% (10)	7.8% (24)	9.5% (29)
CE and NE	12.5% (5)	12.5% (5)	15.7% (48)	13.7% (42)
New CE lesion[3]	10.0% (4)	10.0% (4)	8.5% (26)	7.8% (24)
No progression	15.0% (6)	17.5% (7)	21.9% (67)	20.6% (63)

Annotation**:** 1 = beyond a threshold of 40% volume increase as compared to baseline or best response; 2 = beyond a threshold of 40% volume increase for lower-grade glioma or 100% volume increase for glioblastoma as compared to baseline or best response; 3 = not qualifying for disease progression based on criteria #1 but occurrence of a new contrast enhancing (CE) lesion outside the CE tumor volume from the preceding MRI exam.

Supplementary Table 4. Details on imaging, clinical and molecular characteristics included in the Cox proportional hazards regression model for overall survival (OS) in the EORTC-26101 test set.

Imaging parameters:	Quantitative – ANN	
	Median	(IQR)
Baseline CE tumor volume (cm³)	13.2	(5.6 - 26.0)
Baseline NE tumor volume (cm³)	62.2	(31.1 - 103.1)
Early change in CE tumor volume (%)	-42%	(-78% - +27%)
Early change in NE tumor volume (%)	-18%	(-56% - +19%)

Clinical & Molecular parameters:		
Age (years)		
median (IQR)	57.9	(51.2 - 64.6)
Sex (female)		
female	116	(44%)
male	145	(56%)
WHO Performance Status (>0)		
0	100	(38%)
>0	161	(62%)
MGMT promoter methylation status		
methylated	131	(50%)
unmethylated	130	(50%)
Glucocorticoids intake		
yes	111	(43%)
no	150	(57%)

Supplementary Table 5. Summary statistics for the agreement between tumor segmentation masks automatically predicted by the artificial neural network (ANN) and those generated by the radiologist (ground truth) as indicated by the DICE coefficient (median DICE, corresponding 95% confidence interval (CI, calculated using bootstrapping with n=1000 iterations) and interquartile range (IQR)) in each individual dataset.

	Contrast enhancing (CE) tumor			*Non-enhancing T2 signal abnormality (NE)*		
	DICE (median)	*95% CI*	*IQR*	*DICE (median)*	*95% CI*	*IQR*
HD training set	0·883	(0·876 - 0·892)	(0·812 - 0·920)	0·905	(0·899 - 0·914)	(0·853 - 0·937)
HD test set	0·885	(0·864 - 0·902)	(0·789 - 0·941)	0·929	(0·920 - 0·937)	(0·892 - 0·952)
EORTC-26101 test set	0·906	(0·899 - 0·915)	(0·769 - 0·959)	0·932	(0·929 - 0·936)	(0·882 - 0·960)