ARTICLE | Article in Press

The brain tumor sequence registration (BraTS-Reg) challenge for establishing correspondence between pre-operative and follow-up MRI scans of diffuse glioma patients

Bhakti Baheti · Satrajit Chakrabarty · Hamed Akbari · Michel Bilello · Benedikt Wiestler · Julian Schwarting · Evan Calabrese · Jeffrey Rudie · Syed Abidi · Mina Mousa · Javier Villanueva-Meyer · Brandon K. K. Fields · Florian Kofler · Russell Takeshi Shinohara · Juan Eugenio Iglesias · Tony C. W. Mok · Albert C. S. Chung · Marek Wodzinski · Artur Jurgas · Niccolò Marini · Manfredo Atzori · Henning Müller · Christoph Großbröhmer · Hanna Siebert · Lasse Hansen · Mattias P. Heinrich · Luca Canalini · Jan Klein · Annika Gerken · Stefan Heldmann · Alessa Hering · Horst K. Hahn · Mingyuan Meng · Lei Bi · Dagan Feng · Jinman Kim · Ramy A. Zeineldin · Mohamed E. Karar · Franziska Mathis-Ullrich · Oliver Burgert · Javid Abderezaei · Aymeric Pionteck · Agamdeep Chopra · Mehmet Kurt · Kewei Yan · Yonghong Yan · Zhe Tang · Jianqiang Ma · Sahar Almahfouz Nasser · Nikhil Cherian Kurian · Mohit Meena · Saqib Shamsi · Amit Sethi · Nicholas J. Tustison · Brian B. Avants · Philip Cook · James C. Gee · Lin Tian · Hastings Greer · Marc Niethammer · Andrew Hoopes · Malte Hoffmann · Adrian V. Dalca · Stergios Christodoulidis · Théo Estienne · Maria Vakalopoulou · Nikos Paragios · Daniel S. Marcus · Christos Davatzikos · Aristeidis Sotiras · Bjoern Menze · Spyridon Bakas · Diana Waldmannstetter

Received: 16 January 2026 / Accepted: 28 September 2026
© The Author(s) 2026

Abstract

Registration of longitudinal brain MRI scans containing pathologies remains a challenging problem, due to substantial changes in tissue appearance. While medical image registration methods have advanced considerably, they often lack the required precision and robustness. To address this gap, we introduced the brain tumor sequence registration (BraTS-Reg) challenge, the first public benchmark specifically designed for deformable

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

Bhakti Baheti and Satrajit Chakrabarty equally contributing first authors.

Aristeidis Sotiras, Bjoern Menze, Spyridon Bakas and Diana Waldmannstetter equally contributing senior authors.



https://doi.org/10.1038/s41598-026-74112-3

Published online: 15 November 2026

scientific reports

registration between pre-operative and follow-up scans of patients with adult-type diffuse gliomas. Conducted in conjunction with ISBI-2022 and MICCAI-2022, BraTS-Reg provides de-identified, multi-institutional, multi-parametric MRI data accompanied by expert-annotated landmark points across time. Training data with ground truth annotations were publicly released for algorithm development, while validation data without annotations enabled pre-testing. Submitted containerized algorithms were evaluated on hidden testing cases. Quantitative evaluation was based on median Euclidean error (MEE), Robustness, and the determinant of Jacobian of the displacement field. Top-ranked methods demonstrated comparable performance and shared key features, including pre-alignment of scans, deep neural networks, inverse consistency analysis, and per-case test-time optimization. The best method achieved MEE comparable to inter-rater variability for 63% of landmarks, underscoring the need for further improvements in accuracy and robustness relative to human experts. BraTS-Reg serves as a three-fold active resource for research on image registration, providing a publicly available dataset, a framework for a community benchmark, and online evaluation tools, accessible at https://bratsreg.github.io/.

Keywords BraTS-Reg · Diffuse glioma · Glioblastoma · Longitudinal registration · MRI

Introduction

Registration is a fundamental problem in biomedical image analysis[1,2], aiming to find spatial correspondences between two images for various downstream applications. Accurate longitudinal registration between pre-operative and follow-up scans is crucial for brain tumor patients, aiding in the analysis of healthy tissue, identifying tumor recurrence[3,4], understanding pathophysiological processes, and assessing treatment response. Adult-type diffuse glioma, particularly Isocitrate Dehydrogenase (IDH)-wildtype glioblastoma (CNS WHO grade 4)[5], is the most common and aggressive malignant brain tumor. It heavily infiltrates and deforms surrounding tissue, making early identification of imaging signatures predictive of tumor infiltration and recurrence critical for treatment planning[3,6].

Registering pre-operative and follow-up multi-parametric MRI (mpMRI) scans of diffuse glioma patients is important yet challenging due to: (i) large deformations in the brain tissue caused by tumor mass effect, (ii) missing correspondences between the tumor in pre-operative scans and the resection cavity in follow-ups, and (iii) inconsistent image intensity profiles across scans, as MRI lacks standardized units unlike, for example, Computed Tomography (CT) that depends on Hounsfield units. Figure 1 illustrates these challenges, showing baseline and follow-up scans of the same glioma patient with marked regions of tumor, edema, and resection cavity. Peritumoral tissue labeled as "edema" in pre-operative scans often includes infiltrated tumor cells, potentially leading to recurrence. Thus, corresponding brain tissue regions can have highly heterogeneous intensity profiles across the same brain. These challenges necessitate accurate deformable registration algorithms to establish spatial correspondences between pre-operative and follow-up scans. Such algorithms enable mapping follow-up information onto baseline scans, facilitating the detection of imaging signatures for tumor recurrence[7]. Key complexities include handling tumor-induced deformations, post-resection tissue relaxation, radiation effects, and local tumor progression. Despite decades of research, robust solutions remain elusive[2,6,8,9]. Prior work has explored modifying standard registration algorithms with cost-function masking (i.e., masking out the tumor or cavity areas from the alignment process) and interpolating displacements of the underlying tissue from nearby healthy tissue. However, comprehensive analyses of these approaches and efforts to measure local alignment errors are lacking.

This paper introduces the brain tumor sequence registration (BraTS-Reg) challenge, providing a substantial collection of retrospective multi-institutional mpMRI data acquired from diverse patient demographics to benchmark deformable registration algorithms. The challenge aimed to: (i) identify the most effective registration algorithm for this task, and (ii) establish a quantitative baseline by assessing the accuracy of current state-of-the-art algorithms. The rest of the manuscript provides an overview of prior work in this field, followed by a complete

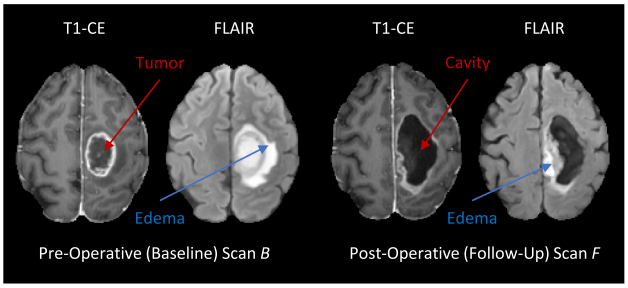


Fig. 1 Example of a pre-operative and follow-up MRI scan pair. The contrast-enhanced T1-weighted (T1-CE), and the T2 fluid attenuated inversion recovery (FLAIR) baseline scans show the tumor and the edema, respectively. Similarly, the T1-CE follow-up scan shows the resection cavity, whereas the edema is visible in the FLAIR scan.

summary of the BraTS-Reg challenge conducted in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI) 2022 and the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2022. The manuscript provides insights into data, evaluation strategies, participating methods, and a meta-analysis of results.

Prior work

Approaches for baseline-to-follow-up registration

Few approaches have been proposed to tackle the task of image registration between pre-operative and follow-up mpMRI scans. In 2013, Kwon et al.[8] developed *PORTR*, a registration approach combining registration and segmentation through an expectation-maximization optimization approach coupled with a biophysical tumor growth modeling framework. They continued their research using patient-specific templates to estimate brain anatomy prior to tumor evolution[6]. In[10], the authors coupled image registration and automatic detection of regions with missing correspondences and[11] modeled region correspondences between images using graph matching. An approach for combining image registration with reconstruction was presented in[9] and a method for finite element biomechanical modeling for registration using image-guided neurosurgery was developed in[12]. Biomechanical models can benefit brain deformation analysis as well as tumor progression evaluation, as Lipkova et al.[13] and Ezhov et al.[14] demonstrated. In[15], the authors utilized deep reinforcement learning for re-detecting landmarks in pre- and post-operative brain scans. There has also been some development in the field of deformable registration of images with pathologies in general, including the simulation of tumor mass effect like in[16], where patient-specific brain anatomy is imitated and in[4], where a deep network is utilized to reconstruct images with pathologies by disentangling the tumor mass effect from the reconstruction of quasi-normal images. Integrating image registration with biophysical modeling is a field with numerous advances[6,8,17]. Nevertheless, there is still research required in this field, especially when it comes to fully automatic registration algorithms. Moreover, most of the referenced methods have been tested on limited amounts of local data and have not been benchmarked to demonstrate stable performance across common multi-institutional datasets.

Related challenges and benchmarks

In the last few years, multiple challenges and benchmarks related to image registration, or brain tumors, have taken place. Starting in 2012 and ongoing to this date, the *Brain Tumor Segmentation (BraTS) Benchmark*[18–31] has shown a remarkable impact on the development of algorithms for the segmentation of brain tumors, and its evolution to a MICCAI 2025 Lighthouse Cluster of Challenges covering workloads beyond segmentation[32–34], to classification[35], and image synthesis[36], redefined BraTS as 'Brain TumorS' challenge. For a subset of the training and testing data of BraTS, several scans are available acquired at pre- and post-operative timepoints. In the field of image registration, multiple challenges and benchmarks have been designed. *Learn2Reg*[37] yearly organizes

data for several registration tasks. *The Continuous Registration Challenge*[38] provides a platform for benchmarking registration methods for the registration of lung CT and brain MRI scans in a fully automated way. Various other challenges exist, addressing the evaluation of MRI-to-ultrasound registration methods for brain shift correction (*CuRIOUS*)[39], the automated registration of histological images (*ANHIR*)[40], the automated registration of breast cancer tissue (*ACROBAT*)[41], and the evaluation of registration methods on thoracic CT (*EMPIRE10*)[42].

Existing datasets for deformable registration and baseline-to-follow-up registration

Several datasets suitable for the task of deformable, and baseline-to-follow-up registration, are publicly available.

QIN GBM Treatment Response: MRI data from 54 newly diagnosed glioblastoma patients, including pre- and post-operative images. Post-operative scans were acquired after surgery, but prior to therapy start[43–45].

IvyGap: 390 studies for 39 patients including pre-operative, post-operative and follow-up mpMRI scans. Genomic information and digitized histopathology tissue sections are also provided[46,47].

UPenn-GBM: MRI data from 60 patients with pre-operative and follow-up scans, which were acquired after a second resection. The dataset includes information on genomics, radiomics, and digitized tissue sections for a comprehensive analysis[48,49].

LUMIERE: Longitudinal MRI data of 91 glioblastoma patients with pre-operative and follow-up scans. Additional provided information includes expert ratings according to the RANO guidelines, patient age at the time of diagnosis, sex, overall survival time, tumor segmentations, radiomics, and pathology information[50].

Nevertheless, none of these listed datasets include expert landmark annotations and the amount of patient data including pre- and post-operative scans offered by each of them is limited (< 100 patients). This is particularly limiting for deep learning algorithms, which typically require large annotated data for effective training and evaluation.

Challenge description

Overview

We organized the first BraTS-Reg challenge focusing on estimating correspondences between baseline pre-operative and follow-up scans of patients with brain gliomas and intended to establish a benchmark environment for deformable registration algorithms. The challenge was first organized in March 2022, and conducted as a virtual event in conjunction with the IEEE ISBI conference. It was later held as an in-person event in September 2022, in conjunction with the MICCAI conference in Singapore. Interested participants were required to register on the Image Processing Portal (IPP) of the Center for Biomedical Image Computing & Analytics (CBICA) (https://ipp.cbica.upenn.edu/). This web-based online evaluation portal was also previously utilized as the evaluation platform for the BraTS challenge from 2017 to 2020[18–21]. Upon registration, participants were granted access to (i) training data including GT annotations, and ii) validation data without GT annotations. They were also provided with the necessary resources for the quantitative evaluation of their algorithms[51]. The portal automatically computed and returned detailed case-wise performance scores to participants for their submissions in training and validation phase. We further maintained an unranked live leaderboard (https://www.cbica.upenn.edu/BraTSReg22/) for the training and validation phases. Participants were given 3.5 months from the challenge start date to submit the containerized algorithm and short paper describing their method and results. The organizers evaluated the containers on testing data and reviewed the papers to ensure it contained sufficient details required to understand and reproduce the algorithm.

The participants who secured top rankings in the testing phase of both the ISBI and MICCAI challenges were invited to orally present their methods. The final challenge rankings were officially announced during the respective conferences. Participants from both the ISBI and MICCAI challenges who submitted functioning

containerized algorithms were given the opportunity to extend their papers. Accepted papers, after a double-blind review process, were published as the BraTS-Reg proceedings[52].

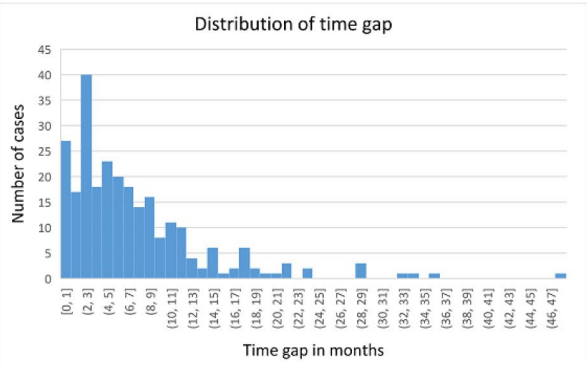
Multi-institutional data sources

The mpMRI data used in the BraTS-Reg challenge consisted of curated and pre-processed retrospective multi-institutional data obtained from routine clinical practice. Specifically, the BraTS-Reg data comprised 259 adult-type diffuse glioma patients from the organizers' affiliated institutions and from the publicly available The Cancer Imaging Archive (TCIA)[53] collections of: (i) TCGA-GBM[54], (ii) TCGA-LGG[55], (iii) IvyGAP[46,56], and (iv) CPTAC-GBM[57,58]. Publicly available datasets were used in accordance with their respective data usage agreements. All private institutional datasets were released following approval from the institutional review boards (IRBs) of the respective contributing institutions. Written informed consent was obtained for all subjects included in the out-of-distribution set. For the remaining private institutional datasets, informed consent was not required, as all imaging data were fully de-identified and originated from patients who were already deceased. No protected health information or other identifiable data were included in the released dataset. The complete dataset consisted of 259 pairs of pre-operative baseline and follow-up brain mpMRI scans, with each pair corresponding to the same patient diagnosed and treated for glioma (WHO CNS grades 3–4). The specific mpMRI sequences at each time-point were (i) native T1-weighted (T1), (ii) contrast-enhanced T1 (T1-CE), (iii) T2-weighted (T2), and (iv) T2 Fluid Attenuated Inversion Recovery (FLAIR).

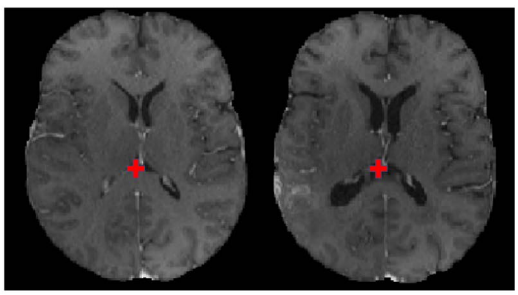
Following close coordination with the clinical experts of the organizing committee, the time-window between the two paired scans of each patient was decided to be selected such that (i) the scans of the two time-points had sufficient apparent tissue deformations, and (ii) confounding effects of surgically induced contrast enhancement[59,60] were avoided. Therefore, after a thorough visual assessment, the identified follow-up data had to be at least 27 days after the initial surgery. Thus, the time-window between all pairs of baseline and follow-up mpMRI scans was in the range of 27 days – 48 months (Fig. 2a).

Data preparation

The complete multi-institutional dataset consisted of scans acquired under routine clinical conditions, and hence reflected very heterogeneous acquisition equipment and protocols, resulting in differences in the image properties. To ensure consistency and keep the challenge focused on the registration problem, all included scans were pre-processed and provided to the participants in the Neuroimaging Informatics Technology Initiative (NIfTI)



(a) Distribution of time gap between baseline and follow-up scans over the BraTS-Reg dataset. Each bin represents one month from 0 to 48.



(b) Example of corresponding landmark points in baseline (left) and follow-up (right) scans, overlaid on a T1-CE slice. This slice contains one of multiple landmarks within the 3D volume.

Fig. 2 Illustrations from the BraTS-Reg challenge dataset. (**a**) Histogram of time gap between baseline and follow-up scans across all cases. (**b**) Example of annotated corresponding landmark points shown on baseline and follow-up T1-CE slices.

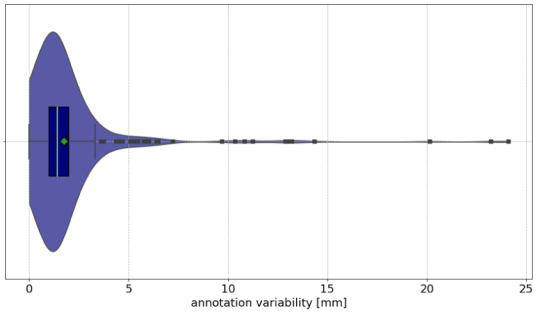
file format[61]. Specifically, three steps of pre-processing were performed for the complete challenge dataset, in line with the BraTS pre-processing protocol[18–20,22]: (i) all scans were first re-oriented into the Left-Post-Superior (LPS) coordinate system and rigidly co-registered to the same canonical anatomical space (i.e., the SRI24 atlas[62]). (ii) all scans were then resampled to the same isotropic resolution (i.e., 1 mm^3). (iii) subsequently, brain extraction was performed to remove non-cerebral tissues like the skull, scalp, and dura from scans[63]. Depending on the data source, these steps have been performed using different tools[51,63–65], all resulting in the same image format. Rigid registration was performed using either "Greedy"[66], a CPU-based C++ implementation of the greedy diffeomorphic registration algorithm[67], or ANTs[68], depending on the institutional source. Greedy shares multiple concepts and implementation strategies within the SyN tool in the ANTs package but focuses on computational efficiency. Brain extraction was performed using Brain Mask Generator (BrainMaGe)[63,69] or HD-BET[70]. BrainMaGe is based on a deep learning segmentation architecture (namely U-Net[71]) and uses a novel training strategy introducing the brain shape as a prior and hence allowing it to be agnostic to the input MRI sequence. HD-BET is another deep learning-based brain extraction method trained on glioma patients. Notably, the complete pre-processing pipeline has been incorporated in the Cancer Imaging Phenomics Toolkit (CaPTk[51]), the BraTS Toolkit[64], and the Integrative Imaging Informatics for Cancer Research: Workflow Automation for Neuro-Oncology (I3CR-WANO[65]) framework, which were actually used at each local institution for the pre-processing of the provided sequences.

Landmark annotation protocol

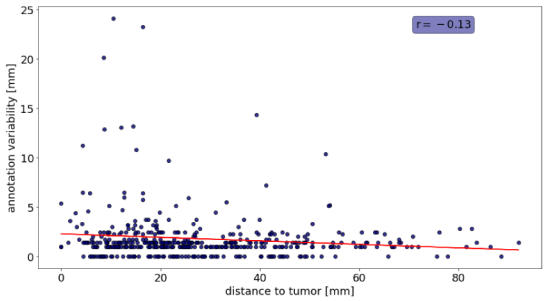
The clinical experts of our team (H.A., M.B., B.W., J.S., E.C., J.R., S.A., and M.M.) were medically trained having several years of experience. Across the dataset, landmark annotations were provided by these eight clinical experts with neuroradiology, neurosurgery or neuro-oncology imaging expertise. They were provided with the segmentation of the tumor core (i.e., the potentially resectable region) and with specific annotation instructions. Specifically, for each pre-operative scan, an expert placed χ number of landmarks near the tumor (within 30 mm) and ψ number of landmarks far from the tumor (beyond 30 mm). Subsequently, the expert identified the corresponding landmarks in the post-operative follow-up scan. For the data used in the longitudinal analyses, the corresponding landmarks were identified by the experts in the second follow-up scans as well. The landmarks were defined on anatomically distinct locations, such as blood vessel bifurcations and anatomical landmarks of the midline of the brain. For each case, experts selected salient locations within the near-tumor and far-from-tumor constraints, emphasizing reliably visible and anatomically distinctive features across the paired scans. Figure 2b shows a sample landmark point marked in the baseline scan along with the location of its corresponding landmark in the follow-up scan. The total number of landmarks ($\chi + \psi$) varied for each case between 6 and 50. The annotators were given the flexibility to use their preferred tool (including MIPAV[72], CaPTk[51], and ITK-SNAP[73]) for making the annotations. Finally, the coordinates of the landmark locations were saved into a common comma-separated value (.csv) format.

Inter-rater annotation variability

To enable the assessment of variance in our annotation process, follow-up-scan landmarks were independently re-annotated through a cross-institutional workflow for a subset of 21 testing-cohort subjects prepared for inter-rater analysis, leading to 446 re-annotated landmarks in total. The subset comprised cases from UCSF, UPenn, and WashU, with re-annotations distributed across these contributing institutions. The re-annotation allowed us to assess the annotation variability (AV) and provided a baseline for comparing differences in tracking errors between various methods. Figure 3a illustrates the distribution of inter-rater variability for the selected 21 samples. For the majority of evaluated landmarks, the AV between clinical experts lay in the range of 1–4 mm, sometimes reaching up to 24 mm. To investigate whether landmarks in the vicinity of the tumor are more difficult to annotate, we correlated landmark-wise AV with respective distance to the tumor core (Fig. 3b). The Pearson



(a) Inter-rater annotation variability on a selection of test cases.



(b) Correlation of inter-rater annotation variability to landmark–tumor distance.

Fig. 3 Inter-rater analysis. (**a**) Distribution of annotation variability across test cases, with $median = 1.41$ and $mean = 1.77$. (**b**) Correlation between annotation variability and landmark–tumor distance (blue), along with regression line (red) and Pearson coefficient $r = -0.13$, showing a small correlation.

coefficient was close to zero ($r = -0.13$), indicating only a weak inverse association: landmarks closer to the tumor core showed a slight tendency toward higher AV, but tumor-core distance did not dominate annotation variability.

Data partitioning in distinct independent cohorts

Following the paradigm of algorithmic evaluation in the machine learning domain, the complete dataset was divided randomly and proportionally into training (70% (140/210)), validation (10% (20/210)), and testing/ranking cohorts (20% (50/210)), while taking into consideration the (1) varying deformation levels (as estimated by the GT annotations), (2) time-window between baseline and follow-up scans, and (3) number of landmarks.

Training data

The training cohort consisted of 140 cases, with each case consisting of 4 baseline (B) and 4 follow-up (F) scans, where corresponding four sequences (T1, T1-CE, T2 and FLAIR) have been acquired during the same session. The challenge participants were provided with the mpMRI scans and the unique landmark point locations on the baseline, as well as on the follow-up scan. The participants were expected to use this data for developing and evaluating their image registration methods and subsequently submit results to the online evaluation portal.

Validation data

The validation data of 20 cases was provided to the participants as scan pairs of B and F with landmarks provided only for the F scan. Participants were requested to submit:

- Coordinates of warped landmark locations in the B scan: coordinates needed to be submitted in x, y, z format and had to be stored in a comma separated value (.csv) file with each row corresponding to a different landmark.
- Determinant of the Jacobian of the displacement field: This needed to be submitted as a NIfTI image with the same geometry as the B scan.

Test data

The test data contained 50 cases and followed the same format and submission requirements as the validation data. Additionally, a second follow-up timepoint (follow-up 2) was obtained for a subset of 9 subjects from the testing cohort to perform a longitudinal performance analysis of the submitted methods. For follow-up 2, we included the same data as the other timepoints, i.e., T1, T1-CE, T2, and FLAIR scans and landmark points

annotated by clinical experts. This dataset was neither released to the participants nor it is intended to be publicly released at any point.

Access to the test data labels was strictly limited to a defined group of individuals, and all access was monitored to ensure the confidentiality, fairness and integrity of the challenge.

Containerization of participants' methods

The participants were requested to containerize their methods using Singularity containers (https://docs.sylabs.io/guides/3.5/user-guide/introduction.html) and submit them to the challenge organizers for evaluation and ranking. The organizers ran all the methods on the test data using their own computational infrastructure. This enabled confirmation of reproducibility and maximization of the use of these algorithms towards offering publicly available solutions for this problem. Detailed and specific instructions for creating singularity containers of uniform Application Programming Interface (API) for this challenge were provided (https://github.com/satrajitgithub/BraTS_Reg_submission_instructions). There were some differences in the testing phase of ISBI and MICCAI as follows:

- ISBI BraTS-Reg Challenge The testing cohort consisted of 43 scan pairs. Participants were asked to containerize their algorithms such that they produced (i) warped landmark locations in the *B* scan and (ii) determinant of the Jacobian of displacement field.

- MICCAI BraTS-Reg Challenge We included seven additional heavily annotated cases (up to 50 landmarks) in the testing set leading to a total of 50 cases. Also, instructions to prepare the containers were updated such that they should produce the following additional outputs along with the (i) and (ii) above:Forward and backward displacement fields.
- Follow-up sequences registered to baseline sequences.
- `=apply_deformation()` python function to apply a user-specified deformation field on any given input image or segmentation mask and return the output image or mask.

Having the additional outputs like the displacement field and the `=apply_deformation()` function gives flexibility in using these containers for additional analyses and applications.

External out of distribution (OOD) test data

Another subset of 49 cases was reserved to evaluate the generalizability performance of various methods on an unseen Out of Distribution (OOD) test data. Notably, this OOD dataset was not used for the ranking of the participants, in line with the original design document of the challenge[74]. The OOD cohort was an external, center-level held-out test set. Unlike the main training, validation, and testing cohorts, which were generated by splitting the pooled multi-source challenge data, the OOD cohort consisted only of cases from a single external contributing institution and had no institutional/source overlap with the remaining challenge splits. This dataset was neither released to the participants nor it is intended to be publicly released at any point.

Participating methods

In the ISBI 2022 challenge, a total of 79 teams registered, allowing them to download the data and submit the results for the training and validation phase. Finally, 6 teams submitted the singularity containers for the testing phase. In MICCAI 2022, the number of registered teams increased to 110 of which 9 teams successfully submitted their singularity containers. Four teams participated in both ISBI and MICCAI challenges. A request from one

of the participating teams to evaluate a second version of their container after the completion of the challenge was accommodated and the method is suffixed with the term post-challenge (pc) to distinguish it from other methods.

After the conclusion of the challenge, multiple external groups working in the image registration domain were invited to develop methods for the BraTS-Reg challenge. These additional invited teams were also required to submit their singularity containers for evaluation on testing data. Brief descriptions of all the evaluated methods can be found in Supplementary Materials: Appendix A along with tabulated summary of key features in each approach (Supplementary Materials: Table S1).

Quantitative performance evaluation and ranking

Evaluation metrics

The assessment of the registration between the two scans was based on manually seeded landmarks (i.e., GT annotations) in both the pre-operative and the follow-up scans, carried out by the expert clinical neuroradiologists. The performance of the developed registration methods was quantitatively evaluated based on distance metrics such as Manhattan distance, Euclidean distance, and Robustness (R). While the Manhattan distance was used for creating the ranking during the ISBI '22 and MICCAI '22 challenges, the Euclidean distance was considered for detailed analysis in this manuscript, to be inline with other challenges on medical image registration[39,40]. All quantitative metrics were applied on the GT annotations in the baseline images. Additionally, the smoothness of the displacement field was assessed as a separate criterion. MEE and Jacobian-based smoothness are established registration evaluation measures, and Robustness follows successful-registration-style measures used in related registration challenges such as ANHIR[40]. The BraTS-Reg score was introduced for this challenge to aggregate case-wise rankings across MEE and Robustness, following case-wise final-ranking-score concepts used in prior BraTS challenge evaluations[20]. The hit-rate analysis described below was adapted as a post-challenge interpretability analysis against inter-rater annotation variability.

Error distance metrics

For each pair $p \in P$ of baseline (B) and follow-up (F) scans, we refer to the landmarks in the B scan as x_l^B and the corresponding landmarks in the F scan as x_l^F. Here, $l \in L$ where L is the total number of landmarks in each scan. The participants were provided with x_l^F and asked to estimate the coordinates of the corresponding landmark points in scan B ($\hat{x}_l^B$). The performance was then evaluated in terms of the Euclidean error (Euclidean distance). The Median Euclidean Error (MEE) between the submitted coordinates ($\hat{x}_l^B$) and the manually defined coordinates (ground truth) (x_l^B) was then defined as:

$$MEE = Median_{\ l \in L}(||x_l^B - \hat{x}_l^B||_2) \tag{1}$$

Robustness (R)

All the participating algorithms were also evaluated according to the metric of robustness (R). Similar to successful-rate measures used in related registration challenges[40], we defined R as a relative value describing how many landmarks improved their MEE after registration, compared to their MEE before registration. Let us call $K^{B,F} \subseteq L^F$ the set of successfully registered landmarks, i.e., those for which the registration error decreased. Where L^F indicates the number of landmarks in the follow up scan. So, we define robustness (R) for a pair of scans (p) as the relative number of successfully registered landmarks:

$$R^{B,F}(p) = \frac{|K^{B,F}|}{|L^F|} \tag{2}$$

Then the average robustness over all scan pairs (P) in the particular cohort is:

$$R = \frac{1}{P}\sum_{p=1}^{P} R^{B,F}(p) \tag{3}$$

Therefore, R is the relative value (in the range of 0 and 1) of how many landmarks across all scan pairs have an improved *MEE* after registration, when compared to the initial *MEE*. When R is equal to 1, the average distance of all the landmarks in the target and warped images is reduced after registration, whereas 0 means that none of the distances is reduced.

Smoothness of the displacement field

We also evaluated the smoothness of the displacement field by calculating its determinant of the Jacobian and examining the number and percentage of voxels with a non-positive Jacobian determinant for each method. These voxels correspond to locations where the deformation is not diffeomorphic. Furthermore, we computed both the minimum and the 99^{th} percentile of the Jacobian determinant (as opposed to the maximum, which is susceptible to noise). These were computed within different regions of interest (e.g., within the tumor core, within 30 mm from the tumor core boundary, outside 30mm from the tumor core boundary, etc.). This metric was assessed as a separate quality criterion rather than as a primary ranking metric during the challenge.

Baseline performance

The baseline performance for registration between the given pre-operative baseline and follow-up scans was established using an affine registration. This decision was driven by the comparative analysis of various methods shown in[9], where every deformable registration method performed better than affine registration but worse than human experts. The baseline registration was performed using the default parameters of an affine ANTs registration (https://github.com/ANTsX/ANTs/blob/master/Scripts/antsRegistrationSyN.sh). Participating methods that performed worse than the affine registration results were considered for the BraTS-Reg challenge, but were required to state this clearly in their manuscript.

Patient-wise ranking strategy and the BraTS-Reg score

We observed variations in the ranking of different teams based on whether the *MEE* or Robustness metric was taken into account. Hence, we introduced a new ranking score termed the "BraTS-Reg score", which consolidates individual rankings. In this process, each participating team was ranked across the two evaluation metrics (*MEE* and R) and for each case from the entire testing set (comprising 50 cases for MICCAI), resulting in 100 rankings per team. The ultimate ranking was determined by the average of these rankings, normalized by the number of teams, and was referred to as the BraTS-Reg score. Lower BraTS-Reg scores indicate better average case-wise rank. Pairwise permutation testing was then applied to the BraTS-Reg scores to assess whether differences between teams were statistically significant.

Statistical analysis

To more accurately quantify differences among the various methods, we employed a statistical analysis procedure akin to those conducted in other challenges like Ischemic Lesion Segmentation (ISLES)[75,76] and BraTS[18–22]. Specifically, we performed pairwise comparisons for significant differences in the relative rankings based on 100,000

permutations to quantify the statistical similarities among the results of different methods. For each team, we started with the list of observed subject-level cumulative ranks, i.e., the actual ranking based on the BraTS-Reg score. For each pair of teams, we repeatedly randomly permuted the cumulative ranks for each subject. For each permutation, we calculated the difference in the BraTS-Reg score between this pair of teams. The statistical significance of the relative rankings was determined by assessing the proportion of occasions when the difference in BraTS-Reg score, calculated using randomly permuted data, surpassed the observed difference in BraTS-Reg score (i.e., using the actual data). This proportion was reported as a p-value.

Longitudinal analysis

The goal of longitudinal analysis was to check the performance robustness of the proposed methods across different timepoints. For this purpose, the following four different registration tasks were performed (where *m* to *f* signifies *m* as the moving scan and *f* as the fixed target scan):

1. Follow-up 1 to baseline (standard task)
2. Follow-up 2 to baseline
3. Follow-up 2 to Follow-up 1
4. Follow-up 2 to Follow-up 1 to baseline

Tasks 1, 2 and 3 could be performed in a single operation whereas task 4 was performed in two stages. In the first stage, the follow-up 2 scans were warped to follow-up 1, and subsequently, in the second stage, these scans were further warped to the baseline scans. Task 4 aimed to identify the potential for error propagation in registration across multiple time points. To that end, the registration performances of the proposed methods were analyzed for the four tasks and the consistency of the different evaluation metrics was compared across these tasks to perform additional meta-analysis after the conclusion of the challenge. Similar to the analysis performed for the ranking evaluation, the longitudinal performances were assessed in terms of MEE, Robustness, and the BraTS-Reg score.

Analysis on the relation of performance and inter-rater annotation variability

For further analysis with respect to the inter-rater annotation variability as described in previous section, we performed two additional analyses, similar to[77]. In the first analysis, we defined a spherical region of interest (ROI) for each reference ground truth landmark (x_l^B) in a selected subset from the dataset that was prepared for inter-rater analysis, including 21 cases with 446 re-annotated landmarks from the testing cohort. Each ROI was sized depending on its respective annotation variability AV_l^B. This variability was therefore defined as the Euclidean error (EE) between the annotations of two raters (x_{r1}^B) and (x_{r2}^B) for the same landmark:

$$AV_l^B = \text{ROI}\,(x_l^B) = EE_{AV}(x_{r1}^B, x_{r2}^B) \tag{4}$$

with $EE_{AV}(x_{r1}^B, x_{r2}^B) = ||x_{r1}^B - x_{r2}^B||_2$. We then denoted a *hit* as the event (success), when the EE between the reference landmark (x_l^B) and the warped landmark $(\hat{x}_l^B)$ (produced by a participating algorithm) was lower than the respective reference landmark's annotation variability:

$$(\hat{x}_l^B) = \begin{cases} 1, & \text{if } \hat{x}_l^B) \in ROI(x_l^B) \\ 0, & \text{if } \hat{x}_l^B) \notin ROI(x_l^B) \end{cases} \tag{5}$$

Then, we computed the ratio of *hits* and *misses*, giving us the *hit rate* of the respective algorithm. As already mentioned above, MEE and Robustness were initially assessed on the ground truth landmarks in the baseline images. For the inter-rater analysis, follow-up images were used. Kindly note that due to the difference in image sources, the results may not be directly comparable.

In a second analysis, we made use of the distribution (D) of the inter-rater annotation variability, as described in previous section and visualized in Fig. 3a. We therefore computed *hit rate* curves[77], by sampling thresholds from D using the formula:

$$median(D) + \delta \cdot median\ absolute\ deviation(D) \tag{6}$$

with $\delta \in [\delta_{min}, \delta_{max}]$ for positive thresholds. $[\delta_{min}, \delta_{max}]$ was therefore defined with respect to the desired evaluation range. Here, δ was chosen to be in $[-1, 10]$ using steps of 0.5, with $median = 1.41mm$ and $median\ absolute\ deviation = 0.77$ mm . The goal of this evaluation was to compare the performance of the automated algorithms against the annotation variability of human experts, representing the "gold standard" in such tasks.

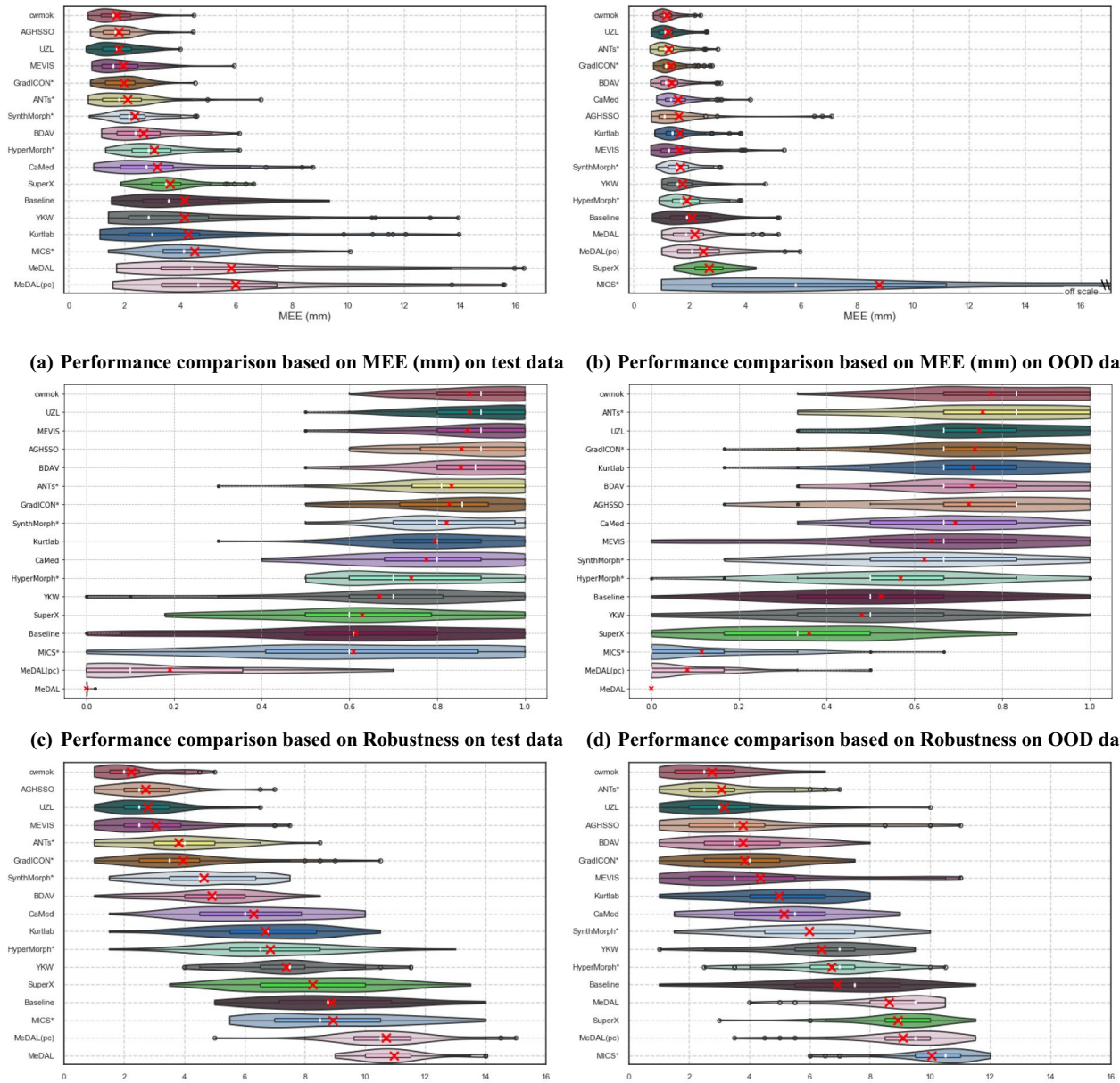
Results

ISBI and MICCAI combined

A detailed overview of different team-wise evaluation metrics is given in Table 1 along with the baseline performance obtained with affine registration using default parameters of *ANTs*. Comparative performance analysis of various participating methods in terms of MEE, Robustness, and BraTS-Reg score in ISBI and MICCAI 2022 along with the invited methods are shown separately on the actual test data and OOD test data in Fig. 4. In all violin plots, the teams are arranged from top to bottom based on their performance, with the order reflecting the spectrum from best to worst in terms of the mean value of the corresponding evaluation metrics. As it can be observed from Fig. 4, the team rankings are not uniform in terms of MEE and R. Team *cwmok* performed best in terms of both MEE and R, clearly indicating the top-ranked team. However, *AGHSSO*, *UZL* and *MEVIS* secured ranks 2, 3, and 4 in terms of MEE, whereas ranks varied in terms of robustness. This made it difficult to finalize the winners. No ties occurred in the official ranking; consequently, displacement-field smoothness was not used to break ties. The final ranking was therefore determined from the BraTS-Reg score generated by the patient-wise

Table 1 Overview of a variety of evaluation metrics per team on the test set. Euclidean Errors (EE) are given in mm and Robustness is given in range [0, 1]. Arrows in column headers indicate the preferred direction of performance. Best values in each column are underlined.

	Median of median EE (↓)	Mean of median EE (↓)	Median of mean EE (↓)	Mean of mean EE (↓)	Median robustness (↑)	Mean robustness (↑)	Mean BraTS-Reg score (↓)
AGHSSO	1.66	1.79	2.26	2.51	<u>0.90</u>	0.86	0.160
ANTs*	1.78	2.10	2.75	2.96	0.81	0.83	0.225
BDAV	2.38	2.68	3.13	3.43	0.89	0.85	0.289
CaMed	2.76	3.17	3.47	4.06	0.80	0.78	0.370
cwmok	1.60	<u>1.73</u>	<u>2.04</u>	<u>2.38</u>	<u>0.90</u>	<u>0.87</u>	<u>0.131</u>
MEVIS	<u>1.59</u>	1.95	2.44	2.75	<u>0.90</u>	<u>0.87</u>	0.179
GradICON*	1.95	1.97	2.61	2.73	0.86	0.83	0.232
HyperMorph*	2.85	3.072	3.60	3.81	0.70	0.74	0.403
Kurtlab	2.99	4.26	3.87	5.07	0.80	0.79	0.394
MeDAL	4.39	5.81	5.21	6.37	0.00	0.00	0.644
MeDAL(pc)	4.62	5.96	5.32	6.47	0.10	0.19	0.629
MICS*	4.12	4.51	4.75	5.15	0.60	0.61	0.525
SuperX	3.46	3.62	4.10	4.39	0.60	0.63	0.486
SynthMorph*	2.15	2.36	2.81	3.06	0.80	0.82	0.274
UZL	1.71	1.79	2.36	2.42	<u>0.90</u>	<u>0.87</u>	0.164
YKW	2.86	4.15	3.68	4.81	0.70	0.67	0.434
Baseline	3.59	4.14	4.48	4.86	0.61	0.61	0.522



(a) Performance comparison based on MEE (mm) on test data

(b) Performance comparison based on MEE (mm) on OOD data

(c) Performance comparison based on Robustness on test data

(d) Performance comparison based on Robustness on OOD data

(e) Performance comparison based on BraTS-Reg score on test data

(f) Performance comparison based on BraTS-Reg score on OOD data

Fig. 4 Comparative performance analysis of various participating methods in terms of median Euclidean error (MEE, mm), Robustness and BraTS-Reg score along with the invited methods (indicated with *). The white line in the violin plots indicates the median, whereas the red cross indicates the mean of the distribution. Panels using the same metric are shown with matched x-axis ranges while preserving the original per-panel sorting; the MEE panels use the test-panel plotting range, the BraTS-Reg score panels use a shared 0–16 x-axis range, and right-edge hash marks with an in-panel note in the OOD MEE panel indicate MEE values beyond the displayed range.

ranking strategy, with pairwise permutation testing used to assess whether differences between teams were statistically significant. Figure 4 also shows the results on OOD data.

Statistical ranking results

Summary of results as shown in Fig. 4 and Table 1 indicates that the performance of some methods was very close to each other. The pairwise permutation analysis described earlier was used to assess whether observed differences in BraTS-Reg scores supported statistically distinct rankings. The resulting p values are reported in an upper triangular matrix in Table 2 for the testing and the OOD data. The statistical evaluation of the top-ranked teams on the testing data revealed that the first three teams were statistically similar to each other ($p > 0.05$). These teams were statistically better than the fourth team ($p = 0.0479$) indicated by a horizontal line in Table 2 Test data results. Similarly, all the participating methods were divided into different tiers and separated by horizontal lines for the test and OOD data in the table.

Longitudinal analysis

The quantitative results of the longitudinal analysis in terms of MEE, Robustness, and the BraTS-Reg score are provided in Supplementary Material (Supplementary Materials: Figure S1). Among the four tasks, MEE values were similar for tasks 1, 2, and 4 across all methods (4.402 ± 3.863, 4.547 ± 3.498, 4.631 ± 3.429) but significantly lower for task 3 (2.155 ± 0.872), likely due to registration between two follow-up scans with similar deformations and structures, making task 3 comparatively easier to the other tasks. However, overall robustness across all methods was generally lower in task 3 (0.543 ± 0.277) compared to tasks 1, 2, and 4 (0.815 ± 0.289, 0.848 ± 0.196, 0.849 ± 0.191 respectively). Additionally, the very similar value ranges of both metrics for tasks 2 and 4 signify that there was no additional error introduced in deforming follow-up 2 to baseline scan, when an intermediate registration step to follow-up 1 scan was involved.

Analysis of top-ranking teams showed that team *UZL* consistently ranked in the top 3 for all tasks in terms of MEE, whereas team *cwmok* ranked in the top 3 for all tasks except task 3. In terms of Robustness, teams *cwmok* and *MEVIS* were consistently among the top 3 across all tasks. For BraTS-Reg score, team *UZL* maintained top 3 rankings for all tasks, and team *cwmok* ranked highly except for task 3, where their drop in MEE ranking impacted the overall score. This disparity highlighted the challenge faced by team *cwmok* in the follow-up to follow-up deformation (task 3) compared to follow-up to baseline scans. Among lower-ranking teams, *MeDAL* consistently ranked in the bottom four across all metrics and tasks.

Algorithm performance versus inter-rater variability

We further evaluated the algorithms' results using *hit rates*, as described in the earlier section. We compared the performance of all participating algorithms by calculating the *hit rates* when evaluated against the respective landmark-wise annotation variability. Since inter-rater analysis was performed on voxels with 1 mm resolution, only landmarks with an AV of not less than 1 mm (81\% (361/446)) were considered for evaluation, to ensure a fair comparison against automated algorithms. From Fig. 5a, we observe that the best-performing algorithm achieved *hits* 63.2\% (228/361) of all evaluated landmarks, showing that there is still room for improvement in terms of registration accuracy. Additionally, we evaluated the methods by increasing the error threshold in range 0-10 mm and computing *hit rate* curves, as illustrated in Fig. 5b. This provides an overview of the characteristics of the different algorithms under increasing error tolerance.

Discussion and conclusion

This manuscript presents the BraTS-Reg Challenge, a comprehensive benchmark designed to evaluate and advance deformable registration methods for aligning pre-operative and follow-up brain MR scans in patients with adult-type diffuse glioma. By providing a large multi-institutional dataset, expert-annotated landmarks, and standardized evaluation metrics, the study establishes a valuable framework for assessing algorithmic performance in this complex clinical context. The results highlight encouraging progress, with some methods achieving near human-level accuracy. However, significant challenges remain in developing robust and generalizable registration approaches across diverse tumor presentations and imaging conditions.

The best-performing BraTS-Reg methods achieved median-of-median Euclidean errors of approximately 1.6 mm and mean Robustness values of approximately 0.86–0.87 on the test set, within the broad millimeter-scale range reported for longitudinal brain-tumor registration studies that use landmark- or target-registration-error measures[78]. Direct comparison with nonsurgical GBM follow-up registration is limited, however, because independent studies in that setting generally do not include the pre-operative-to-post-resection change that defines BraTS-Reg and often evaluate registration through downstream response mapping rather than landmark-based Euclidean error[79]. BraTS-Reg instead addresses the more difficult pre-operative-to-follow-up setting, with tumor-to-cavity non-correspondence, post-resection tissue relaxation, radiation effects, and local progression. Consistent with this, our longitudinal analysis showed lower MEE for follow-up-to-follow-up registration than for follow-up-to-baseline tasks, likely reflecting greater anatomical similarity between follow-up scans.

Reference standard considerations while establishing a benchmark environment

The inaugural BraTS-Reg challenge demanded substantial annotation effort involving multiple clinical experts, leading to variability in both landmark number and placement per case. While inter-rater analysis revealed high agreement in most cases, generalizability would benefit from a broader, more standardized annotation protocol. The error in the annotations is inherently biased, as landmarks are selected at salient locations by experts[80]. Including stricter guidelines and more comprehensive annotations, along with additional qualitative metrics—such as displacement field smoothness or assessment of near tumor regions—would enhance evaluation. The top-ranked method, *cwmok*, achieved MEE values equal to or below inter-rater variability for 63% (228/361) of landmarks. Nonetheless, there is room for improvement in accuracy and robustness compared to human experts. Although most algorithms were not designed specifically for this challenge, a few (*cwmok*, *AGHSSO*) addressed missing correspondences by incorporating inverse consistency. Only a limited number of methods leveraged the provided landmarks for training, and none reached the top ranks. Even with atlas space alignment, the use of rigid or affine pre-alignment further improved outcomes for some methods. Rankings varied between the test and OOD datasets, likely reflecting both the origins of the datasets and differences in annotator expertise. Despite this, the leading team (*cwmok*) remained consistent, while the lowest-ranked methods showed little variation across test sets and metrics.

For each case, annotators selected salient anatomical features within the prescribed near-tumor and far-from-tumor regions, prioritizing structures that were clearly visible in the follow-up scans. This case-specific selection accommodated tumor-related distortion and post-treatment anatomy while preserving the intended anatomical context for registration assessment. Inter-rater variability was then summarized across selected landmarks and related to distance from the tumor core (Fig. 3b). Most landmark-wise annotation variability values lay in the range of 1–4 mm, with a median of 1.41 mm and mean of 1.77 mm, although occasional values reached 24 mm. The Pearson coefficient was close to zero ($r = -0.13$), indicating only a weak inverse association: landmarks nearer the tumor core showed a slight tendency toward higher variability, but tumor proximity did not dominate the reliability of the selected landmarks.

https://doi.org/10.1038/s41598-026-74112-3

Table 2 Statistical significance analysis based on p value calculation on the test data, showing different tiers between the participating teams, demarcated by horizontal lines.

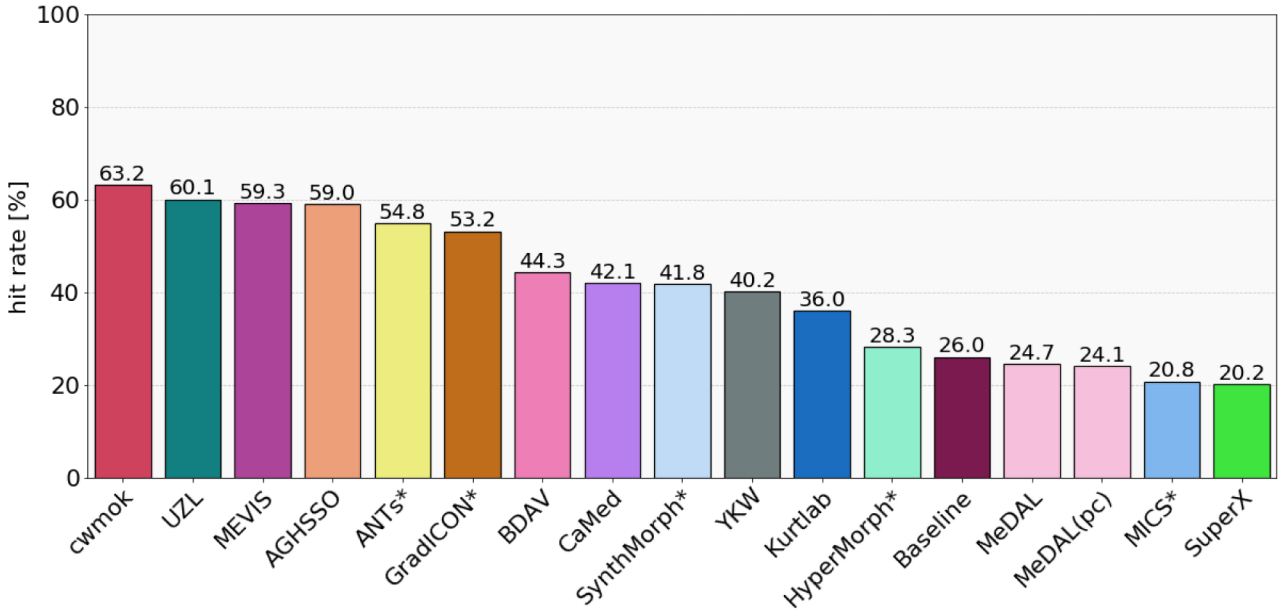
	cwmok	AGHSSO	UZL	MEVIS	ANTs*	Grad ICON*	Synth Morph*	BDAV	CaMed	Kurtlab	Hyper Morph*	YKW	SuperX	Baseline	MICS*	MeDAL (pc)	MeDAL
Test data																	
cwmok	–	0.3385	0.1924	0.0479	0.0051	0.0144	0	0.0007	0	0.0003	0	0	0	0	0	0	0
AGHSSO	–	–	0.2945	0.1648	0.0099	0.0299	0.0018	0	0	0.0004	0.0002	0	0	0	0	0	0
UZL	–	–	–	0.2375	0.0113	0.0293	0.001	0.0003	0.0001	0	0.0003	0	0	0	0	0	0
MEVIS	–	–	–	–	0.0716	0.2936	0.0299	0.0199	0	0.0014	0.0002	0	0	0	0	0	0
ANTs*	–	–	–	–	–	0.6445	0.1697	0.0868	0.0002	0.0153	0.0069	0	0	0	0	0	0
Grad ICON*	–	–	–	–	–	–	0.0556	0.0424	0.0003	0.0096	0.003	0.0003	0	0	0	0	0
Synth Morph*	–	–	–	–	–	–	–	0.383	0.0005	0.0329	0.0067	0.0003	0	0	0	0	0
BDAV	–	–	–	–	–	–	–	–	0.0045	0.056	0.0246	0.0003	0	0	0	0	0
CaMed	–	–	–	–	–	–	–	–	–	0.8653	0.9335	0.276	0.0593	0	0.0141	0.0009	0
Kurtlab	–	–	–	–	–	–	–	–	–	–	0.4164	0.0436	0.0148	0.0005	0.0032	0	0
Hyper Morph*	–	–	–	–	–	–	–	–	–	–	–	0.0311	0.0022	0	0.0003	0	0
YKW	–	–	–	–	–	–	–	–	–	–	–	–	0.2106	0.0045	0.0651	0.0019	0
SuperX	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0194	0.1103	0.0064	0.0003
Baseline	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.8037	0.0722	0.0005
MICS*	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0089	0.0002

	cwmok	ANTs*	UZL	AGHSSO	BDAV	Grad ICON*	MEVIS	Kurtlab	CaMed	Synth Morph*	YKW	Hyper Morph*	Baseline	MeDAL	SuperX	MeDAL (pc)	MICS*
OOD data																	
cwmok	–	0.4913	0.6091	0.1707	0.0545	0.0398	0.0014	0.0156	0.0352	0	0.0001	0	0.0008	0	0	0	0
ANTs*	–	–	0.5973	0.1447	0.0215	0.0364	0.0014	0.0081	0.0227	0.0009	0.0004	0.0002	0.0005	0	0	0	0
UZL	–	–	–	0.1494	0.0108	0.0186	0.0002	0.0032	0.0077	0	0.0001	0	0	0	0	0	0
AGHSSO	–	–	–	–	0.2688	0.4315	0.0189	0.1476	0.1993	0.0077	0.0019	0.0005	0.0025	0	0	0	0
BDAV	–	–	–	–	–	0.7444	0.0195	0.1983	0.365	0.0065	0.0025	0.0003	0.0009	0	0	0	0
Grad ICON*	–	–	–	–	–	–	0.0094	0.1157	0.2037	0.001	0.0006	0	0.0021	0	0	0	0
MEVIS	–	–	–	–	–	–	–	0.959	0.94	0.5991	0.3582	0.2576	0.3339	0.015	0.0087	0.0076	0.0031
Kurtlab	–	–	–	–	–	–	–	–	0.6073	0.0239	0.0016	0.0014	0.0031	0	0	0	0
CaMed	–	–	–	–	–	–	–	–	–	0.0234	0.002	0	0.0002	0	0	0	0
Synth Morph*	–	–	–	–	–	–	–	–	–	–	0.1826	0.1081	0.1335	0.0002	0.0001	0	0.0001
YKW	–	–	–	–	–	–	–	–	–	–	–	0.3261	0.4254	0.0007	0.0001	0.0012	0.0001
Hyper Morph*	–	–	–	–	–	–	–	–	–	–	–	–	0.5745	0.0104	0.0005	0.0034	0.0002
Baseline	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0036	0.0004	0.0011	0.0001
MeDAL	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.1642	0.0152	0.0209

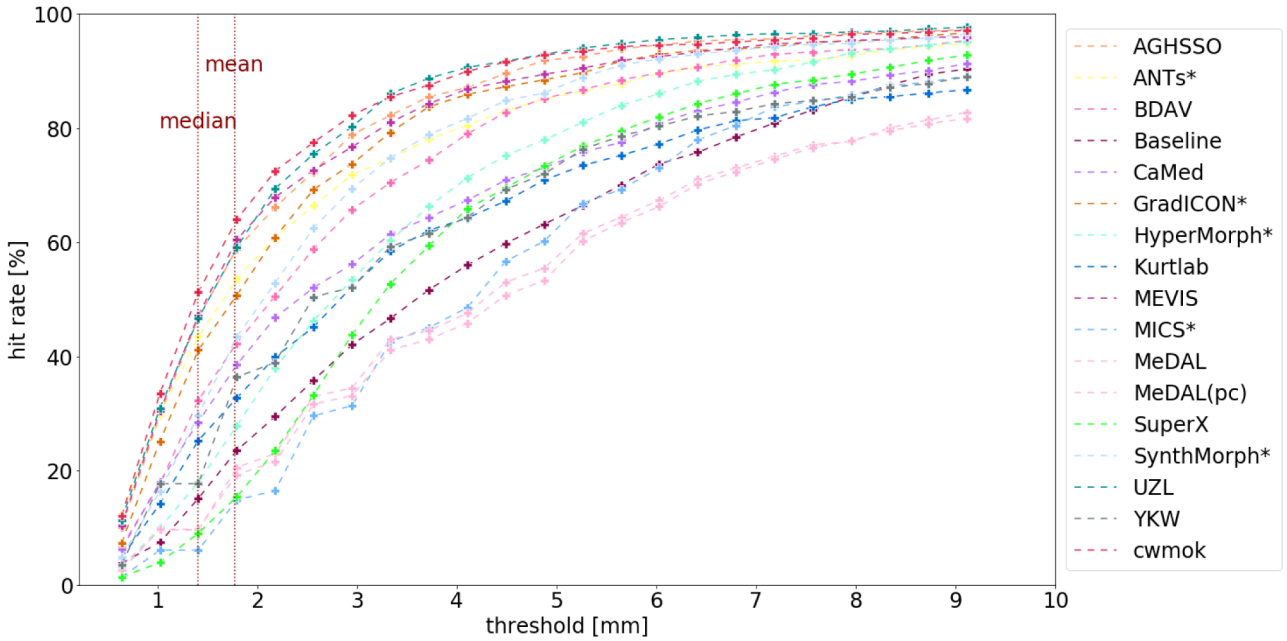
https://doi.org/10.1038/s41598-026-74112-3

Table 2 (continued)

	cwmok	ANTs*	UZL	AGHSSO	BDAV	Grad ICON*	MEVIS	Kurtlab	CaMed	Synth Morph*	YKW	Hyper Morph*	Baseline	MeDAL	SuperX	MeDAL (pc)	MICS*
SuperX	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.4743	0.0223
MeDAL (pc)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0775
MICS*	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–



(a) *Hit rate* per team on the respective landmarks' inter-rater annotation variability



(b) *Hit rate* per team on increasing thresholds

Fig. 5 *Hit rate* analysis. (**a**) gives an overview of the *hit rate* per team using a threshold corresponding to the landmark-specific inter-rater annotation variability. (**b**) shows the *hit rate* per team for increasing thresholds. Percentages in between evaluated thresholds are interpolated. Dotted lines indicate the mean and median inter-rater annotation variability.

Identifying the optimal registration algorithm

The comparative quantitative analysis of all participating methods highlights a diverse set of approaches, each leveraging unique technical principles (see y Materials: Table S1) to address challenges such as multi-modality integration, missing correspondences, robustness, computational efficiency, and adaptability to more complex or multi-sequence/multi-parametric datasets. Overall, the top-ranked methods were closely matched across all

evaluation metrics. This trend was particularly evident among the top three methods and, to a lesser extent, across the top half of the rankings. An examination of the methodologies employed by the top-performing teams revealed the use of pre-alignment, deep neural networks, and inverse consistency analysis. Traditional parameterized pipelines for pre-alignment effectively handle rigid and affine transformations but introduce additional computational overhead. Deep learning-based approaches, utilizing transformers or convolutional architectures, excel at processing multi-sequence/multi-parametric data, capturing spatial correspondences, and extracting modality-specific features. Optimization-based frameworks, relying on gradient-free or convex methods, offer scalability and computational efficiency, though they may compromise precision for intricate deformations. Additionally, instance optimization, which refines registration results during post-processing at test time, proved advantageous.

Methods from teams ANTs and cwmok are highly robust and accurate for complex anatomical deformations, with ANTs excelling in reproducibility and flexibility through parameterized pipelines. SynthMorph and GradICON stand out for their adaptability across modalities and anatomical variations. These methods are particularly effective for high-fidelity alignments but can be computationally intensive. HyperMorph provides a novel approach by enabling hyperparameter tuning during inference, reducing the need for multiple model trainings. For scalable and computationally efficient solutions, *YKW*, *UZL*, and *HyperMorph* provides lightweight frameworks with robust performance. In terms of multi-sequence integration, *kurtlab*, *GradICON*, and *SynthMorph* offers advanced architectures for aligning diverse imaging sequences. *Kurtlab's* transformer-based architecture and *GradICON's* stages training approach enhanced robustness for multi-modal tasks, whereas *SynthMorph* generalizes well across unseen data through its synthetic training paradigm.

For real-time and adaptive applications, *UZL*, *YKW*, and *Camed* features faster processing pipelines with minimal computational overhead, making them ideal for scenarios where speed is prioritized over fine-grained accuracy. In contrast, *ANTs*, *cwmok*, and *GradICON* are better suited for high-precision tasks that could accommodate longer processing times due to their computational demands. *UZL* achieves an excellent balance between accuracy and speed by employing a combination of hand-crafted features, convex optimization, and Adam-based instance optimization. This approach facilitates efficient computation, avoids heavy reliance on pre-training, and produces smooth and plausible deformation fields, making it particularly suitable for large-deformation registration with limited computational resources.

Insights into successful techniques

Analysis of the top-performing methods reveals that inverse consistency and pre-alignment are common strategies for success. For example, *cwmok* and *AGHSSO* employs inverse consistency to improve robustness in cases with missing correspondences, ensuring the integrity of the transformation process. This technique help maintains alignment even when landmark correspondences are incomplete or uncertain, leading to more reliable results. Deep learning-based methods, such as *SynthMorph* and *GradICON*, demonstrates strong generalization across modalities. *SynthMorph* uses synthetic training data to enhance adaptability to unseen data, improving robustness across diverse anatomical structures. *GradICON*, with its staged training approach, refines multi-modal integration, allowing for better handling of complex anatomical variations. Additionally, *MEVIS* adopts a hybrid approach, blending traditional image registration techniques with deep learning models. This allows *MEVIS* to effectively capture complex anatomical variations, while maintaining high alignment accuracy across different modalities. *UZL* balances computational efficiency and accuracy through convex and instance optimization, producing smooth deformation fields suitable for real-time processing. Table 3 summarizes the key techniques employed by the top-ranked methods.

Conclusion

The diversity of methods evaluated for the problem that BraTS-Reg is posed to solve underscores the importance of adopting a hybrid approach that combines the strengths of various techniques to achieve an optimal balance

Table 3 Comparison of top-ranked methods based on key techniques and strengths.

Method	Key techniques	Strengths
AGHSSO	Inverse consistency measurement	Robust to missing correspondences, ensures transformation integrity
ANTs	Traditional techniques, parameterized optimization	Highly robust, accurate for complex deformations, flexible and reproducible
cwmok	Inverse consistency, deep learning, pre-alignment	Handles missing correspondences effectively, high accuracy, robust for complex deformations
GradICON	Staged training, multi-modal integration	Excels at multi-modal integration, handles complex deformations, strong generalization
MEVIS	Hybrid traditional and deep learning methods	Captures complex anatomical variations, high alignment accuracy across modalities
SynthMorph	Synthetic training, multi-modal generalization	Effective for multi-modal integration, adapts well to unseen data, robust across datasets
UZL	Convex optimization, instance optimization, adaptive	Efficient, smooth deformation fields, good balance between speed and accuracy

between robustness, accuracy, speed, and scalability. Hybrid strategies that integrate initial alignment using traditional pipelines, multi-modal integration with learning-based architectures, deformable alignment through coarse-to-fine approaches, and final refinement with smoothness-preserving techniques offer promising potential. A potential hybrid approach might combine the precision and robustness of methods like *ANTs* or *cwmok* with the scalability of *YKW* or *HyperMorph*, and the adaptability of *SynthMorph* or *kurtlab*, tailoring performance to specific use cases while mitigating systematic weaknesses of individual approaches. Researchers should continue to explore advancements in pre-alignment techniques, deep learning, synthetic data generation, and hyperparameter optimization to further enhance performance. Collaborative benchmarks, open-source datasets, and pipelines will continue to play a critical role in fostering innovation while ensuring robust and reproducible results in brain tumor sequence registration.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-74112-3.

Acknowledgements Hamed Akbari, Michel Bilello, Benedikt Wiestler, Julian Schwarting, Evan Calabrese, Jeffrey Rudie, Syed Abidi, Mina Mousa, Javier Villanueva-Meyer, Brandon K. K. Fields, Florian Kofler, Russell Takeshi Shinohara, Juan Eugenio Iglesias, Daniel S. Marcus, Christos Davatzikos, Aristeidis Sotiras, Bjoern Menze, Spyridon Bakas and Diana Waldmannstetter: People involved in organization of the challenge.

Author contributions Bhakti Baheti, Satrajit Chakrabarty, Aristeidis Sotiras, Bjoern Menze, Spyridon Bakas, and Diana Waldmannstetter conceived and designed the study and developed the methodology. Bhakti Baheti, Satrajit Chakrabarty, and Diana Waldmannstetter implemented the software, performed the formal analyses, generated visualizations, and wrote the original draft. Bhakti Baheti, Satrajit Chakrabarty, Hamed Akbari, Michel Bilello, Benedikt Wiestler, Julian Schwarting, Evan Calabrese, Jeffrey Rudie, Syed Abidi, Mina Mousa, Javier Villanueva-Meyer, Brandon K.K. Fields, Florian Kofler, Russell Takeshi Shinohara, Juan Eugenio Iglesias, Daniel S. Marcus, Christos Davatzikos, Aristeidis Sotiras, Bjoern Menze, Spyridon Bakas, and Diana Waldmannstetter contributed to organization of the BraTS-Reg Challenge. Aristeidis Sotiras, Bjoern Menze, and Spyridon Bakas supervised the study, acquired funding, and oversaw research execution. Tony C. W. Mok, Albert C. S. Chung, Marek Wodzinski, Artur Jurgas, Niccolò Marini, Manfredo Atzori, Henning Müller, Christoph Großbröhmer, Hanna Siebert, Lasse Hansen, Mattias P. Heinrich, Luca Canalini, Jan Klein, Annika Gerken, Stefan Heldmann, Alessa Hering, Horst K. Hahn, Mingyuan Meng, Lei Bi, Dagan Feng, Jinman Kim, Ramy A. Zeineldin, Mohamed E. Karar, Franziska Mathis-Ullrich, Oliver Burgert, Javid Abderezaei, Aymeric Pionteck, Agamdeep Chopra, Mehmet Kurt, Kewei Yan, Yonghong Yan, Zhe Tang, Jianqiang Ma, Sahar Almahfouz Nasser, Nikhil Cherian Kurian, Mohit Meena, Saqib Shamsi, Amit Sethi, Nicholas J. Tustison, Brian B. Avants, Philip Cook, James C. Gee, Lin Tian, Hastings Greer, Marc Niethammer, Andrew Hoopes, Malte Hoffmann, Adrian V. Dalca, Stergios Christodoulidis, Théo Estienne, Maria Vakalopoulou, and Nikos Paragios participated in the BraTS-Reg Challenge and contributed to methodology development and result validation. All authors reviewed and approved the final version of the manuscript.

Funding This study was supported in part by the: 1) National Institutes of Health (NIH) National Institute of Neurological Disorders and Stroke (NINDS) R01 grant (R01NS042645), 2) NIH Informatics Technology for Cancer Research (ITCR) program of the National Cancer Institute (NCI) under award numbers U24CA189523, U24CA279629, U01CA242871, 3)

NIH Ruth L. Kirschstein Institutional National Research Service Award (T32EB001631), 4) RSNA Research & Education Foundation under grant number RR2011, and 5) Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE) GSC 81. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of NIH, the RSNA R & E Foundation, or any of the other funding bodies. B.Baheti, H.Akbari, and Spyridon Bakas conducted part of the work reported in this manuscript at their current affiliation, as well as while they were affiliated with the Center for Artificial Intelligence and Data Science for Integrated Diagnostics (AI2D) and the Center for Biomedical Image Computing and Analytics (CBICA) at the University of Pennsylvania, Philadelphia, PA 19104, USA. D. Waldmannstetter conducted most of the work presented in this manuscript at the University of Zurich and the Technical University of Munich, with additional work performed at the Department of Nuclear Medicine and the Medical Data Integration Center (MEDIZUKR), University Hospital Regensburg, Regensburg, Germany. Partially funded by the German Federal Ministry of Research, Technology and space (BMFTR) Network of University Medicine 3.0: "NUM 3.0", Grant No.01KX2524, Project: NUM-DIZ.

Data availability The training and validation data generated for the BraTS-Reg challenge are publicly available at: https://doi.org/10.5281/zenodo.14642405.

Code availability The public BraTS-Reg container implementation and submission instructions, including the Singularity interface provided to challenge participants for container creation, are available at https://github.com/satrajitgithub/BraTS_Reg_submission_instructions. Public participant methods can be implemented by pulling the available BraTS-Reg Singularity containers from Sylabs Cloud Library and following these GitHub instructions. The public containers can be accessed through the BraTS-Reg-specific Sylabs Cloud Library search at https://cloud.sylabs.io/library/search?q=brats_reg. The organizer-side evaluation code used for hidden test and ranking evaluation is not publicly released because public release could compromise the integrity of current and potential future BraTS-Reg challenge evaluations by enabling participants to optimize specifically for the challenge evaluator rather than for generalizable registration performance.

Declarations

Competing interests S.C. reports that the research described in this manuscript was conducted while affiliated with Washington University in St.Louis, St. Louis, MO, USA. Although S.C. is currently employed by GE HealthCare, GE HealthCare had no involvement in thedesign, execution, analysis, or funding of the study. B.K.K.F. reports consulting or advisory relationships with Mendaera and Elsevier Inc., and board membership, funding grants, non-financial support, and travel reimbursement from the Radiological Society of North America. B.K.K.F. also reports board membership with the American Board of Radiology and Frontiers in Radiology, as well as priorresearch grants from the RSNA R&E Foundation (2019–2020 RMS 1909; 2018–2019 RMS #1810), consulting fees fromMendaera, honoraria from Neurodiem (invited author) and Elsevier (book proposal reviews), and RSNA/institutional travelsupport. R.T.S. reports financial support from the National Institutes of Health and consulting or advisory roles with OctaveBioscience and the American Medical Association. M.H. reports financial support from the National Institutes of Health and afunding relationship with GE HealthCare. All other authors declare no competing financial or non-financial interests related to this work.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Sotiras, A., Davatzikos, C. & Paragios, N. Deformable medical image registration: A survey. *IEEE Trans. Med. Imaging* **32**, 1153–1190 (2013).
2. Ou, Y., Akbari, H., Bilello, M., Da, X. & Davatzikos, C. Comparative evaluation of registration algorithms in different brain databases with varying difficulty: Results and insights. *IEEE Trans. Med. Imaging* **33**, 2039–2065 (2014).
3. Akbari, H. et al. Imaging surrogates of infiltration obtained via multiparametric imaging pattern analysis predict subsequent location of recurrence of glioblastoma. *Neurosurgery* **78**, 572–580 (2016).

4. Han, X. A deep network for joint registration and reconstruction of images with pathologies. In *International Workshop on Machine Learning in Medical Imaging* 342–352 (Springer, 2020).
5. WHO Classification of Tumours Editorial Board. *World Health Organization Classification of Tumours of the Central Nervous System* 5th edn. (International Agency for Research on Cancer, Lyon, 2021).
6. Kwon, D., Zeng, K., Bilello, M. & Davatzikos, C. Estimating patient specific templates for pre-operative and follow-up brain tumor registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 222–229 (Springer, 2015).
7. Akbari, H. et al. Histopathology-validated machine learning radiographic biomarker for noninvasive discrimination between true progression and pseudo-progression in glioblastoma. *Cancer* **126**, 2625–2636 (2020).
8. Kwon, D. et al. Portr: Pre-operative and post-recurrence brain tumor registration. *IEEE Trans. Med. Imaging* **33**, 651–667 (2013).
9. Han, X. Patient-specific registration of pre-operative and post-recurrence brain tumor MRI scans. In *International MICCAI Brainlesion Workshop* 105–114 (Springer, 2018).
10. Chen, K., Derksen, A., Heldmann, S., Hallmann, M. & Berkels, B. Deformable image registration with automatic non-correspondence detection. In *International Conference on Scale Space and Variational Methods in Computer Vision*, 360–371 (Springer, 2015).
11. Feng, T., Lv, Y. & Qin, B. Region matching for pre-operative and post-operative brain images. In *2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA)*, 951–954 (IEEE, 2015).
12. Drakopoulos, F. & Chrisochoides, N. P. Accurate and fast deformable medical image registration for brain tumor resection using image-guided neurosurgery. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **4**, 112–126 (2016).
13. Lipková, J., Menze, B., Wiestler, B., Koumoutsakos, P. & Lowengrub, J. S. Modelling glioma progression, mass effect and intracranial pressure in patient anatomy. *J. R. Soc. Interface* **19**, 20210922 (2022).
14. Ezhov, I. et al. Learn-morph-infer: A new way of solving the inverse problem for brain tumor modeling. *Med. Image Anal.* **83**, 102672 (2023).
15. Waldmannstetter, D. Reinforced redetection of landmark in pre-and post-operative brain scan using anatomical guidance for image alignment. In *International Workshop on Biomedical Image Registration* 81–90 (Springer, 2020).
16. Zacharaki, E. I., Shen, D., Lee, S.-K. & Davatzikos, C. Orbit: A multiresolution framework for deformable registration of brain tumor images. *IEEE Trans. Med. Imaging* **27**, 1003–1017 (2008).
17. Scheufele, K. et al. Coupling brain-tumor biophysical models and diffeomorphic image registration. *Comput. Methods Appl. Mech. Eng.* **347**, 533–567 (2019).
18. Menze, B. H. et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imaging* **34**, 1993–2024 (2014).
19. Bakas, S. et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* **4**, 1–13 (2017).
20. Bakas, S. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge (2018) arXiv:1811.02629 arXiv preprint.
21. Bakas, S., Akbari, H., Bilello, J. S., Hall, M. E. & Davatzikos, C. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. *Cancer Imaging Arch. (TCIA)*. https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF (2017).
22. Baid, U. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification (2021) arXiv:2107.02314 arXiv preprint.
23. Adewole, M. The brain tumor segmentation (BraTS) Challenge 2023: Glioma segmentation in sub-Saharan Africa patient population (BraTS-Africa). arXiv arXiv:2305.19369v1 (2023).
24. LaBella, D. et al. Analysis of the brats 2023 intracranial meningioma segmentation challenge. arXiv preprint arXiv:2405.09787 (2024).
25. Moawad, A. W. The brain tumor segmentation—metastases (BraTS-METS) challenge 2023: Brain metastasis segmentation on pre-treatment MRI. arXiv arXiv:2306.00838v3 (2024).
26. LaBella, D. et al. A multi-institutional meningioma MRI dataset for automated multi-sequence image segmentation. *Sci. Data* **11**, 496 (2024).
27. LaBella, D. et al. Brain tumor segmentation (brats) challenge 2024: Meningioma radiotherapy planning automated segmentation. arxiv 2024. arXiv preprint arXiv:2405.18383 (2024).
28. de Verdier, M. C. et al. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment MRI. arXiv preprint arXiv:2405.18368 (2024).
29. Kazerooni, A. F. The brain tumor segmentation in pediatrics (brats-peds) challenge: focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds) (2024) arXiv:2404.15009 arXiv preprint.
30. Maleki, N. Analysis of the miccai brain tumor segmentation-metastases (brats-mets) 2025 lighthouse challenge: Brain metastasis segmentation on pre-and post-treatment mri (2025) arXiv:2504.12527 arXiv preprint.
31. Adewole, M. et al. The brats-Africa dataset: Expanding the brain tumor segmentation data to capture African populations. *Radiol. Artif. Intell.* **7**, e240528 (2025).

32. Amiruddin, R. Training the next generation of physicians for artificial intelligence-assisted clinical neuroradiology: Asnr miccai brain tumor segmentation (brats) 2025 lighthouse challenge education platform (2025) arXiv:2509.17281 arXiv preprint.
33. Kofler, F. The brain tumor segmentation (brats) challenge 2023: Local synthesis of healthy brain tissue via inpainting (2023) arXiv:2305.08992 arXiv preprint.
34. Kofler, F. Brats orchestrator: Democratizing and disseminating state-of-the-art brain tumor image analysis (2025) arXiv:2506.13807 arXiv preprint.
35. Bakas, S. Brats-path challenge: Assessing heterogeneous histopathologic brain tumor sub-regions (2024) arXiv:2405.10871 arXiv preprint.
36. Li, H. B. The brain tumor segmentation (BraTS) challenge 2023: Brain MR image synthesis for tumor segmentation (BraSyn). arXiv arXiv:2305.09011v6 (2024).
37. Hering, A. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. IEEE Trans. Med. Imaging (2022).
38. Marstal, K., Berendsen, F., Dekker, N., Staring, M. & Klein, S. The continuous registration challenge: Evaluation-as-a-service for medical image registration algorithms. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 1399–1402 (IEEE, 2019).
39. Xiao, Y. et al. Evaluation of MRI to ultrasound registration methods for brain shift correction: The curious2018 challenge. *IEEE Trans. Med. Imaging* **39**, 777–786 (2019).
40. Borovec, J. et al. Anhir: Automatic non-rigid histological image registration challenge. *IEEE Trans. Med. Imaging* **39**, 3042–3052. https://doi.org/10.1109/TMI.2020.2986331 (2020).
41. Weitz, P. Acrobat-automatic registration of breast cancer tissue. In *10th International Workshop on Biomedical Image Registration* (2022).
42. Murphy, K. et al. Evaluation of registration methods on thoracic CT: The empire10 challenge. *IEEE Trans. Med. Imaging* **30**, 1901–1920 (2011).
43. Prah, M. et al. Repeatability of standardized and normalized relative CBV in patients with newly diagnosed glioblastoma. *Am. J. Neuroradiol.* **36**, 1654–1661 (2015).
44. Mamonov, A. & Kalpathy-Cramer, J. Data from QIN GBM treatment response. In *The Cancer Imaging Archive* (2016).
45. Mamonov, A. B. & Kalpathy-Cramer, J. Data from QIN GBM treatment response (2016).
46. Puchalski, R. B. et al. An anatomic transcriptional atlas of human glioblastoma. *Science* **360**, 660–663 (2018).
47. Shah, N., Feng, X., Lankerovich, M., Puchalski, R. B. & Keogh, B. Data from IVY GAP (2016).
48. Bakas, S. et al. The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: Advanced MRI, clinical, genomics, & radiomics. *Sci. Data* **9**, 1–12 (2022).
49. Bakas, S. Multi-parametric magnetic resonance imaging (mpMRI) scans for de novo glioblastoma (GBM) patients from the University of Pennsylvania health system (UPENN-GBM, 2021).
50. Suter, Y. et al. The LUMIERE dataset: Longitudinal Glioblastoma MRI with expert RANO evaluation. *Sci. Data* **9**, 1–8 (2022).
51. Davatzikos, C. et al. Cancer imaging phenomics toolkit: Quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *J. Med. imaging* **5**, 011018 (2018).
52. Bakas, S. et al. (eds.). *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries—8th International Workshop, BrainLes 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Revised Selected Papers, Part II*, vol. 14092 of *Lecture Notes in Computer Science* (Springer, 2023). https://doi.org/10.1007/978-3-031-44153-0.
53. Clark, K. et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).
54. Scarpace, L. et al. Radiology data from the cancer genome atlas glioblastoma multiforme [TCGA-GBM] collection. *Cancer Imaging Arch.* **11**, 1 (2016).
55. Pedano, N. Radiology data from the cancer genome atlas low grade glioma [TCGA-LGG] collection. *Cancer Imaging Arch.* **2** (2016).
56. Shah, N., Feng, X., Lankerovich, M., Puchalski, R. B. & Keogh, B. Data from Ivy gap. *Cancer Imaging Arch.* **10**, K9 (2016).
57. TCIA. Radiology data from the clinical proteomic tumor analysis consortium glioblastoma multiforme [CPTAC-GBM] collection. *Cancer Imaging Arch.* **10**, K9 (2016).
58. Wang, L.-B. et al. Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* **39**, 509–528 (2021).
59. Albert, F. K., Forsting, M., Sartor, K., Adams, H.-P. & Kunze, S. Early postoperative magnetic resonance imaging after resection of malignant glioma: Objective evaluation of residual tumor and its influence on regrowth and prognosis. *Neurosurgery* **34**, 45–61 (1994).
60. Wen, P. Y. et al. Updated response assessment criteria for high-grade gliomas: Response assessment in neuro-oncology working group. *J. Clin. Oncol.* **28**, 1963–1972 (2010).

61. Cox, R. A (sort of) new image data format standard: Nifti-1: We 150. *Neuroimage* **22** (2004).
62. Rohlfing, T., Zahr, N. M., Sullivan, E. V. & Pfefferbaum, A. The sri24 multichannel atlas of normal adult human brain structure. *Hum. Brain Mapp.* **31**, 798–819 (2010).
63. Thakur, S. et al. Brain extraction on MRI scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. *Neuroimage* **220**, 117081 (2020).
64. Kofler, F. et al. Brats toolkit: Translating brats brain tumor segmentation algorithms into clinical and scientific practice. *Front. Neurosci.* **14**, 125 (2020).
65. Chakrabarty, S. et al. Integrative imaging informatics for cancer research: Workflow automation for neuro-oncology (i3cr-wano). *JCO Clin. Cancer Inform.* **7**, e2200177 (2023).
66. Yushkevich, P. A. et al. Fast automatic segmentation of hippocampal subfields and medial temporal lobe subregions in 3 tesla and 7 tesla t2-weighted mri. *Alzheimer's Dementia* **7**, P126–P127 (2016).
67. Joshi, S., Davis, B., Jomier, M. & Gerig, G. Unbiased diffeomorphic atlas construction for computational anatomy. *Neuroimage* **23**, S151–S160 (2004).
68. Avants, B. B. et al. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage* **54**, 2033–2044 (2011).
69. Thakur, S. P. Skull-stripping of glioblastoma MRI scans using 3d deep learning. In *International MICCAI Brainlesion Workshop* 57–68 (Springer, 2019).
70. Isensee, F. et al. Automated brain extraction of multisequence mri using artificial neural networks. *Hum. Brain Mapp.* **40**, 4952–4964 (2019).
71. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241 (Springer, 2015).
72. McAuliffe, M. J. *et al.* Medical image processing, analysis and visualization in clinical research. In *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001* 381–386 (IEEE, 2001).
73. Yushkevich, P. A., Gao, Y. & Gerig, G. Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* 3342–3345 (IEEE, 2016).
74. Bhakti, B. et al. The brain tumor sequence registration (BraTS-REG) challenge. https://doi.org/10.5281/ZENODO.6362420 (2022).
75. Maier, O. et al. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Med. Image Anal.* **35**, 250–269 (2017).
76. Winzeck, S. et al. Isles 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI. *Front. Neurol.* **9**, 679 (2018).
77. Waldmannstetter, D. Framing image registration as a landmark detection problem for label-noise-aware task representation (hitr) (2024) arXiv:2308.01318 arXiv preprint.
78. van der Hoorn, A. et al. Validation of a semi-automatic co-registration of mri scans in patients with brain tumors during treatment follow-up. *NMR Biomed.* **29**, 882–889. https://doi.org/10.1002/nbm.3538 (2016).
79. Ellingson, B. M., Cloughesy, T. F., Lai, A., Nghiemphu, P. L. & Pope, W. B. Nonlinear registration of diffusion-weighted images improves clinical sensitivity of functional diffusion maps in recurrent glioblastoma treated with bevacizumab. *Magn. Reson. Med.* **67**, 237–245. https://doi.org/10.1002/mrm.23003 (2012).
80. Peter, L., Alexander, D. C., Magnain, C. & Iglesias, J. E. Uncertainty-aware annotation protocol to evaluate deformable registration algorithms. *IEEE Trans. Med. Imaging* **40**, 2053–2065 (2021).

Publisher"s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Bhakti Baheti[1,2] · Satrajit Chakrabarty[3,4] · Hamed Akbari[5] · Michel Bilello[6] · Benedikt Wiestler[7] · Julian Schwarting[7] · Evan Calabrese[8] · Jeffrey Rudie[9] · Syed Abidi[10] · Mina Mousa[10] · Javier Villanueva-Meyer[11] · Brandon K. K. Fields[11] · Florian Kofler[7,12,13] · Russell Takeshi Shinohara[14,15] · Juan Eugenio Iglesias[16,17,18] · Tony C. W. Mok[19] · Albert C. S. Chung[19] · Marek Wodzinski[20,21] · Artur Jurgas[20,21] · Niccolò Marini[20,22] · Manfredo Atzori[20,23] · Henning Müller[20,24] · Christoph Großbröhmer[25] · Hanna Siebert[26] · Lasse Hansen[25] · Mattias P. Heinrich[26] · Luca Canalini[27,28] · Jan Klein[29] · Annika Gerken[27] ·

Stefan Heldmann[30] · Alessa Hering[30,31] · Horst K. Hahn[27,28] · Mingyuan Meng[32] · Lei Bi[32] · Dagan Feng[32,33] · Jinman Kim[32] · Ramy A. Zeineldin[34] · Mohamed E. Karar[34] · Franziska Mathis-Ullrich[35] · Oliver Burgert[36] · Javid Abderezaei[37] · Aymeric Pionteck[37] · Agamdeep Chopra[37] · Mehmet Kurt[37] · Kewei Yan[38] · Yonghong Yan[38] · Zhe Tang[39] · Jianqiang Ma[39] · Sahar Almahfouz Nasser[40] · Nikhil Cherian Kurian[40] · Mohit Meena[40] · Saqib Shamsi[41] · Amit Sethi[40] · Nicholas J. Tustison[42] · Brian B. Avants[42] · Philip Cook[43] · James C. Gee[43] · Lin Tian[44] · Hastings Greer[44] · Marc Niethammer[44] · Andrew Hoopes[17,45,46] · Malte Hoffmann[45,46,47,48] · Adrian V. Dalca[17,45,46,47] · Stergios Christodoulidis[49] · Théo Estienne[49] · Maria Vakalopoulou[49] · Nikos Paragios[50] · Daniel S. Marcus[10] · Christos Davatzikos[6,14] · Aristeidis Sotiras[10,51] · Bjoern Menze[12,52] · Spyridon Bakas[2,53,54,55] · Diana Waldmannstetter[12,52,56,57]

✉ Satrajit Chakrabarty
satrajit.chakrabarty@wustl.edu

[1] Department of Biomedical Engineering, Emory University, Atlanta, GA, USA

[2] Division of Computational Pathology, Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN, USA

[3] Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO, USA

[4] GE HealthCare, San Ramon, CA, USA

[5] Department of Bioengineering, Santa Clara University, Santa Clara, CA, USA

[6] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[7] Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

[8] Department of Radiology, Duke University Medical Center, Durham, NC, USA

[9] Department of Radiology, University of California, San Diego, San Diego, CA, USA

[10] Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO, USA

[11] Center for Intelligent Imaging (ci2), San Francisco, CA, USA

[12] Department of Computer Science, TUM School of Computation, Information and Technology, Technical University of Munich, Munich, Germany

[13] Helmholtz AI, Neuherberg, Germany

[14] Center for Artificial Intelligence and Data Science for Integrated Diagnostics (AI2D) and Center for Biomedical Image Computing and Analytics (CBICA), Philadelphia, PA, USA

[15] Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics, Philadelphia, USA

[16] Massachusetts General Hospital, Boston, USA

[17] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

[18] Centre for Medical Imaging Computing, University College London, London, United Kingdom

[19] The Hong Kong University of Science and Technology, Hong Kong, China

[20] University of Applied Sciences and Arts Western Switzerland (HES-SO), Delémont, Switzerland

[21] Department of Measurement and Electronics, AGH University of Krakow, Krakow, Poland

[22] Department of Computer Science, University of Geneva, Geneva, Switzerland

[23] Department of Neuroscience, University of Padova, Padova, Italy
[24] Medical Faculty, University of Geneva, Geneva, Switzerland
[25] EchoScout GmbH, Lübeck, Germany
[26] Institute of Medical Informatics, Universität zu Lübeck, Lübeck, Germany
[27] Fraunhofer MEVIS, Institute for Digital Medicine, Bremen, Germany
[28] University of Bremen, Bremen, Germany
[29] Fraunhofer Institute for Digital Medicine MEVIS, Lübeck, Germany
[30] Fraunhofer MEVIS, Institute for Digital Medicine, Lübeck, Germany
[31] Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, Netherlands
[32] School of Computer Science, The University of Sydney, Sydney, Australia
[33] Med-X Research Institute, Shanghai Jiao Tong University, Shanghai, China
[34] Menoufia University, Shebeen El-Kom, Egypt
[35] Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-University Erlangen-Nürnberg, Nürnberg, Germany
[36] Research Group Computer Assisted Medicine (CaMed), Reutlingen University, Reutlingen, Germany
[37] Department of Mechanical Engineering, University of Washington, Seattle, USA
[38] University of North Carolina at Charlotte, Charlotte 28223, USA
[39] Keya Medical Technology Co. Ltd, Beijing, China
[40] Indian Institute of Technology, Bombay, India
[41] Whirlpool, Pune, India
[42] Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, USA
[43] Penn Image Computing and Science Lab, Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[44] University of North Carolina at Chapel Hill, Chapel Hill, USA
[45] Athinoula A. Martinos Center for Biomedical Imaging, 149 13th St, Charlestown, MA 02129, USA
[46] Department of Radiology, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114, USA
[47] Department of Radiology, Harvard Medical School, 25 Shattuck St, Boston, MA 02115, USA
[48] Harvard-MIT Health Sciences and Technology, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA
[49] CentraleSupélec, 3 Rue Joliot Curie, 91190 Gif-sur-Yvette, France
[50] Therapanacea, 7 bis Bd Bourdon, 75004 Paris, France
[51] Institute for Informatics, Data Science and Biostatistics, Washington University School of Medicine, St. Louis, MO, USA
[52] Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland
[53] Indiana University Melvin and Bren Simon Comprehensive Cancer Center, Indianapolis, IN, USA
[54] Departments of Radiology and Imaging Sciences, Neurological Surgery, Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA
[55] Department of Computer Science, Luddy School of Informatics Computing and Engineering, Indianapolis, IN, USA
[56] Department of Nuclear Medicine, University Hospital Regensburg, Regensburg, Germany
[57] Medical Data Integration Center (MEDIZUKR), Regensburg, Germany