Evaluation of MRI to Ultrasound Registration Methods for Brain Shift Correction: The CuRIOUS2018 Challenge

Yiming Xiao [Member, IEEE],
the Robarts Research Institute, Western University, London, ON N6A 5B7, Canada

Hassan Rivaz [SeniorMember, IEEE],
the PERFORM Centre, Concordia University, Montreal, QC H3G 1M8, Canada, and also with the Department of Electrical and Computer Engineering, Concordia University, Montreal, QC H3G 1M8, Canada

Matthieu Chabanas,
the School of Computer Science and Applied Mathematics, Grenoble Institute of Technology, 38031 Grenoble, France, and also with the TIMC-IMAG Laboratory, University of Grenoble Alpes, 38400 Grenoble, France

Maryse Fortin,
the PERFORM Centre, Concordia University, Montreal, QC H3G 1M8, Canada, and also with the Department of Health, Kinesiology and Applied Physiology, Concordia University, Montreal, QC H3G 1M8, Canada

Ines Machado,
the Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115 USA

Yangming Ou,
the Department of Pediatrics and Radiology, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115 USA

Mattias P. Heinrich,
the Institute of Medical Informatics, University of Lübeck, 23538 Lübeck, Germany

Julia A. Schnabel
the School of Biomedical Engineering and Imaging Sciences, King's College London, London WC2R 2LS, U.K

Yiming Xiao: yxiao286@uwo.ca; Hassan Rivaz: hrivaz@ece.concordia.ca; Matthieu Chabanas: matthieu.chabanas@univ-grenoble-alpes.fr

Xia Zhong, **Andreas Maier**
the Pattern Recognition Lab, Friedrich–Alexander-Universität Erlangen–Nürnberg, 91054 Erlangen, Germany, and also with the Department of Computer Science, Friedrich–Alexander-Universität Erlangen–Nürnberg, 91054 Erlangen, German

Wolfgang Wein
ImFusion GmbH, 80992 Munich, Germany

Correspondence to: Yiming Xiao, yxiao286@uwo.ca.

Roozbeh Shams, **Samuel Kadoury**

the CHUM Research Centre, Montreal, QC H2X 0A9, Canada, and also with the Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC H3T 1J4, Canada

David Drobny,

the Wellcome/EPSRC Cenre for Interventional and Surgical Sciences, University of College London, London, U.K., also with the School of Biomedical Engineering and Imaging Sciences, King's College London, London, U.K., and also with King's Health Partners, St Thomas' Hospital, London SE1 7EH, U.K

Marc Modat,

the School of Biomedical Engineering and Imaging Sciences, King's College London, London, U.K., and also with King's Health Partners, St Thomas' Hospital, London SE1 7EH, U.K

Ingerid Reinertsen

SINTEF, 7465 Trondheim, Norway

Ingerid Reinertsen: ingerid.reinertsen@sintef.no

Abstract

In brain tumor surgery, the quality and safety of the procedure can be impacted by intra-operative tissue deformation, called brain shift. Brain shift can move the surgical targets and other vital structures such as blood vessels, thus invalidating the pre-surgical plan. Intra-operative ultrasound (iUS) is a convenient and cost-effective imaging tool to track brain shift and tumor resection.Accurate image registration techniques that update pre-surgical MRI based on iUS are crucial but challenging. The MICCAI Challenge 2018 for Correction of Brain shift with Intra-Operative UltraSound (CuRIOUS2018) provided a public platform to benchmark MRI-iUS registration algorithms on newly released clinical datasets. In this work, we present the data, setup, evaluation, and results of CuRIOUS 2018, which received 6 fully automated algorithms from leading academic and industrial research groups. All algorithms were first trained with the public RESECT database, and then ranked based on a test dataset of 10 additional cases with identical data curation and annotation protocols as the RESECT database. The article compares the results of all participating teams and discusses the insights gained from the challenge, as well as future work.

Index Terms

Registration; brain; ultrasound; MRI; brain shift; tumor

I Introduction

GLIOMAS are the most common brain tumors in adults, and are categorized into grade I-IV by the World Health Organization (WHO). Low-grade gliomas (LGG, grade I and II) are less aggressive and have slower progression than high-grade gliomas (HGG, grade III and IV), but will eventually undergo malignant transformation into high-grade tumors. Evidences [1], [2] have shown that early tumor resection can effectively improve the patient's survival rate. Imageguidance can be a useful tool to assist the surgeon in obtaining

a maximal safe resection of the tumor. Image-guidance based on pre-operative MR images is in routine clinical use worldwide. These systems, however, do not account for the tissue shift and deformations that occur as the resection progresses. Due to brain shift, the surgical target and other vital structures (e.g., blood vessels and ventricles) will be displaced relative to the pre-surgical plan, resulting in inaccurate image-guidance. Multiple factors can contribute to brain shift, including but not limited to drug administration, intracranial pressure change, tissue resection. Often such tissue shift is not directly visible by the surgeon. Both intra-operative ultrasound (iUS) and intra-operative magnetic resonance imaging (iMRI) have been employed to track tissue deformation and surgical progress. Intra-operative US has gained popularity thanks to its low cost, high portability and flexibility. However, limited field of view and challenging image interpretation remain obstacles for widespread use. Together with iUS, automatic image registration algorithms can be used to update the surgical plan based on pre-operative MRI by re-aligning the pre-operative images with intra-operative images and offer more intuitive assessments of the extent of resection.

Previously, a number of algorithms and strategies [3]–[8] have been developed to address iUS-MRI registration for brain shift correction. They range from new strategies to map image features to similar domains [3], [4] to novel cost function [6], [7], and from different deformation models [5], [6] to improved optimization procedures [8]. However, partially due to the lack of relevant clinical datasets, it has been difficult to directly compare different algorithms, thus potentially slowing the speed of knolwedge translation to benefit surgeons and patients. The MICCAI Challenge 2018 for Correction of Brain shift with Intra-Operative UltraSound (CuRIOUS2018) was launched as the first public platform to benchmark the latest image registration algorithms for the task, and to bring the researchers together to discuss the technical and clinical challenges in iUS-guided brain tumor resection. For the first edition of the challenge, we focused on MRI-iUS registration to correct pre-resection deformation after craniotomy, as it typically sets the tone of brain shift for the rest of the surgery.

The challenge was divided into the training and testing phases. While the publicly available REtroSpective Evaluation of Cerebral Tumors (RESECT) database [9] was used in the training phase for algorithm development, in the testing phase, a private testing database was distributed to assess the participating teams. The distances between homologous anatomical landmarks between iUS and MRI were used to assess and rank the registration quality. The CuRIOUS2018 challenge received 8 initial submissions [10]–[17]. Seven teams validated their methods on the testing data, and six participated in the final ranking. The submissions cover a wide variety of approaches, including the latest registration metrics [7], [13], optimization approaches [12], and deep learning techniques [17].

This paper describes the organization, submitted algorithms, and results for the challenge, and further discusses the current challenges and potential future directions of tissue shift correction in US-guided brain tumor surgery.

II Materials

Two datasets were included for the training and testing phases of the CuRIOUS2018 challenge. The RESECT database [9] was provided to the participants as the training dataset for development and fine-tuning of the algorithms. The database contains pre-operative MR and pre-resection iUS images from 22 patients who have received LGG resection surgeries at St. Olavs University Hospital, Trondheim, Norway. The testing dataset was comprised of imaging data from 10 additional patients with LGG obtained in the same setting as the RESECT database. The collection and distribution of both datasets were approved by the Regional Committee for Medical and Health Research Ethics of Central Norway, and all patients signed written informed consent.

For both training and testing databases, Gd-enhanced T1w MRI and T2w fluid-attenuated inversion recovery (FLAIR) MRI scans were acquired for each patient before surgery. Five fiducial markers were glued to the patient's head prior to scanning. The T1w and T2w MRIs were rigidly coregistered, and aligned to the patient's head position on the operating table via a fiducial-based image-to-patient registration. The position-tracked 3D iUS scans were acquired with the Sonowand Invite neuronavigation system (Sonowand AS, Trondheim, Norway), with either the 12FLA-L linear transducer or the 12FLA flat linear array transducer for smaller superficial tumors. 3D volumes were reconstructed from the raw iUS data using the built-in proprietary reconstruction method in the Sonowand Invite system, with a reconstruction resolution in the range of $0.14 \times 0.14 \times 0.14$ mm^3 to $0.24 \times 0.24 \times 0.24$ mm^3 depending on the probe types and imaging depth. Both ultrasound transducers were factory calibrated and equipped with removable sterilizable reference frames for optical tracking. A Polaris camera (NDI, Waterloo, Canada) built in the Sonowand system was used to obtain the position and pose of the ultrasound probe. Therefore, the iUS volumes reveal tissue position and deformation in the patient's head on the operating table.

Homologous anatomical landmarks manually labeled by two raters (authors YX and MF as *Rater 1* and *2*, respectively) were provided to assess registration quality, using the software 'register' included in the MINC toolkit (http://bic-mni.github.io). Typical landmarks include the edge of the tumor, deep grooves of sulci, corners of sulci, convex points of gyri and the horns of the lateral ventricles. After *Rater 1* defined the landmarks in the T2w FLAIR MRIs as the references, *Rater 1* and *Rater 2* then tagged the corresponding landmarks independently within the corresponding US volumes twice. A 1~2-week interval was ensured between the repetitions. The final landmarks in both training and testing database were provided as the averaged results of two trials of landmark tagging by both raters (four 3D points for each landmark). The details of the landmarks are listed in Table I for the testing datasets. Similar details for the training dataset can be found in the original publication for the RESECT database [9]. For both sets, a wide range of brain shifts measured as mean initial distances between corresponding landmarks were included to properly examine the performance of registration algorithms.

We employed the mean Euclidean distance between two sets of corresponding landmark points for each patient to assess the intra- and inter-rater variability. For intra-rater variability, we calculated the metric between two trials of landmark picking for each rater;

for inter-rater variability, the average of two trials for each rater was first computed and used to obtain the value between two raters. The intra- and inter-rater variability evaluations are presented in Table II for both training and testing data.

III Challenge Setup

The CuRIOUS2018 challenge started on April 1st, 2018 when the challenge website went live on curious2018.grand-challenge.org. In the next few days, several groups who were active in the field of MRI-iUS registration were identified by literature search and were invited to participate. The challenge was also widely advertised on mailing lists and on bulletin board of medical imaging conferences held in the first half of 2018. Another factor that leads to a good participation was the incentive of generous support of challenge sponsors, which provided a total of 2100 €for the top winners. The challenge consisted of two phases (see Fig. 1).

In phase I, all the teams were required to submit a short paper that elaborated the technique and results on the 22 patients in the RESECT database. These papers were then peer-reviewed and the final camera-ready conference papers were submitted in July 2018.

Phase II started in August 2018, when all the participants who had submitted reports and results on the training data were provided with MRI and iUS data from 10 additional patients (test data). These datasets had identical data curation and annotation protocols as the RESECT database. The locations of landmarks in the MRI were provided to the teams, and the teams had to return the locations of those landmarks after MRI-iUS registration within 13 days of the data release. All teams presented their methods and results on the training data at the challenge event, which took place in conjunction to MICCAI 2018 in Granada, Spain.

The RESECT database remains public, and has been downloaded 267 times since its release in April 2017. The test datasets were only released to the participants, and the locations of the ground truth landmarks in these datasets remain private. The organizers will continue the challenge in 2019 by adding iUS test data collected during and after tumor resection.

IV Evaluation

The evaluation metric and ranking system are key criteria for the success of a challenge. The metric should reflect the overall quality of the methods and the ranking system should be as fair as possible. It is worth noting that our evaluation method was published on the official website before the challenge took place and was not modified afterwards. Although such transparency in the evaluation process may seem obvious, [18] reported that this transparency was not guaranteed in about 40% of biomedical challenges, which could lead to controversy.

The first component of the evaluation process is a metric to assess the quality of the registration methods. More than 80% of the tasks in biomedical challenges concern segmentation, with the Dice similarity coefficient as the most common evaluation metric [18]. However, challenges with image registration, especially from different modalities, are

rarer and we could not find any standard metrics from these competitions. We thus chose to rely solely on the expert-labeled anatomical landmark pairs, by computing the Euclidian distances between the transformed MRI landmarks, after registration, and the ground-truth landmarks defined in the iUS images.

The second component of the evaluation process concerns how the results for each test case are aggregated to rank the teams. The two main options are 1) aggregate the results on all test cases, then rank; or 2) rank by test case, then aggregate the ranks. In the first scenario, we would have ranked the teams based on the mean distance computed from all landmarks of all cases. Instead, we chose the second scenario because it is better fitted to handle missing cases. For each case, we also ranked fully-automatic methods over semi-automatic methods. To aggregate the case-by-case ranks, we simply computed the mean rank of each team.

The evaluation system was as follows:

1. For each test case and for each team, compute the Euclidian distances between landmark pairs after registration, i.e., between the transformed MRI landmarks and the ground truth iUS landmarks.

2. For each test case, rank teams according to their mean distance between landmark pairs. Exceptions include:

    a. If one team could not provide results for a test case, or if these results could not be processed for any reason, then that team is ranked last for the test case.

    b. If two mean distances differ by less than 0.5 mm, a team with a fully-automatic method is ranked higher than a team with a semi-automatic method.

3. Compute the mean rank of every team, which gives the final ranks of the challenge.

V Challenge Entries

A Team cDRAMMS

Machado *et al.* [13] extended the Deformable Registration via Attribute Matching and Mutual-Saliency Weighting (DRAMMS) algorithm [19], a general-purpose algorithm [20], specifically for the US-MRI registration problem, which they termed as correlation-similarity DRAMMS or cDRAMMS. They released it at https://www.nitrc.org/projects/dramms/ (version 1.5.1). The original DRAMMS has two good properties for US-MRI registration. First, representing each voxel with multi-scale and multi-orientation Gabor attributes in DRAMMS offers richer information than purely image intensities. This helps to establish more reliable voxel correspondences despite the different image protocols and different intensity profiles between US and MRI images. Second, the mutual-saliency module in DRAMMS automatically assigns low confidence or weights to regions that cannot establish reliable or cannot find counterparts across images. This potentially reduces

the negative effects of the missing correspondences between US and MRI images. Different from the original DRAMMS, which uses the sum of square differences (SSD) between attributes for matching, the modified cDRAMMS uses correlation coefficient (CC) [21] and correlation ratio (CR) [22] on attributes for voxel matching. CC and CR on voxel attributes in cDRAMMS establish voxel correspondences at a higher accuracy and higher reliability than SSD in DRAMMS. Free-form deformation and discrete optimization are chosen as the deformation and optimization strategy, respectively.

B Team DeedsSSC

Heinrich *et al.* [14] used DeedsSSC, which comprises a linear and a non-rigid registration that are both based on discrete optimization and modality-invariant image features. Specifically, self-similarity context features (SSC) are extracted for both MRI and ultrasound scans that are matched based on a dense displacement sampling. First, the similarity maps for each considered control point are used to extract correspondences for fitting a linear transform using least trimmed squares, similar as done in blockmatching approaches. Second, new similarity maps are calculated for linearly aligned images and an efficient graphical model based discrete optimization (deeds) is used to estimate a nonlinear displacement field that avoids implausible warps and further improves the registration quality. All computations are performed for scans resampled to isotropic 0.5 mm resolution and using the default parameters (see https://github.com/mattiaspaul/deedsBCV) with an optimization over multiple grid-scales. Finally, the nonlinearly warped landmarks are again constrained to follow a rigid 6-parameter transform for improved robustness. The algorithm is executed within less than 10 seconds per scan pair on a multi-core CPU and ongoing work considers the huge potential for further speed-ups through parallelized GPU computations.

C Team FAX

Zhong *et al.* [17] proposed a learning-based approach to resolve intraoperative brain as an imitation game. This point-based approach predicts the deformation vectors of key points to compensate the non-rigid brain-shift. For each key point, they extract a local 3D patch in iUS and model the key point distribution as the encoding of the current observation. A demonstrator is constructed providing the optimal deformation vector based on the current key point location and the ground truth. An artificial neural network is trained to imitate the behavior of the demonstrator and to predict the optimal deformation vector given current observation. To increase robustness, the proposed technique uses a multi-tasking network with a rigid transformation as auxiliary output. In addition, we use non-rigid deformations to augment the 3D volume and 3D key points to facilitate the training.

D Team ImFusion

The method [16] is based on the multi-modal similarity metric LC^2 [7] and has recently been used in a first live evaluation during surgery [23] (data NOT overlapping with challenge data). The similarity metric was maximized using a non-linear optimization algorithm with a parametric transformation model. In a pre-processing step specific to the challenge data set (cartesian 3D ultrasound volumes compounded by the SonoWand system), the volume sides facing the ultrasound probe is estimated and the outermost 4mm of content are cropped accordingly. The registration algorithm is implemented using GPU acceleration.

For registration, the US volume was employed as the fixed image, and was resampled to 0.5mm, which is half of the MRI resolution. The final selected patch-size for computing similarity metric is 7×7×7 voxels, which was optimized in their prior work [23]. The algorithm employs a two-stage non-linear optimization strategy that successively operates on the transformation parameters. In the first stage, a global DIRECT (DIviding RECTangles) sub-division method [24] searches on translation only, and is followed by a local BOBYQA (Bound Optimization BY Quadratic Approximation) algorithm [25] on all six rigid transformation parameters. Afterwards, the local optimizer conducts another search on full affine parameters in order to accommodate shearing and non-uniform scaling of the data.

E Team MedCAL

Multimodal deformable registration between the MRI and intra-operative 3DUS was achieved with a weighted version of the locally linear correlation metric (LC^2), correlating MRI intensities and gradients with ultrasound, while adapting both hyper-echoic and hypo-echoic regions within the cortex. The method [15] was initialized with a global rotation of the US volume to match the orientation observed on the MRI. This was achieved using a PCA of the extracted inferior skull region, identifying the principal orientation vectors of the head, followed by a scaling and translation correction. This fusion step uses a patch-based approach of the US voxels, comparing intensity and gradient magnitudes extracted from the MRI with a linear relationship. The registration applies sequentially a rigid and non-rigid step, with the latter integrating a weighting term and controlled by a cubic 5×5×5 B-Spline interpolation grid, distributed uniformly in the fan-shaped US volume. The weighting term uses pre-annotated labels on the MRI, representing both the hypoechoic (fluid cavities) and the hyperechoic (ex. choroid plexus) areas observed on ultrasound. This term is added only at the non-rigid step as it is highly specific to the internal areas in US such as the lateral ventricles, requiring a rigid pre-alignment. Registration optimization was performed using BOBYQA, which avoids computing the metric's derivatives.

F Team NiftyReg

Drobny *et al.* [12] suggest a method which uses a block-matching approach to automatically align the pre-operative MRI with the iUS image. The registration algorithm is part of the NiftyReg open-source software package [26]. Their block-matching registration technique iteratively establishes point correspondences between the fixed image and the warped floating image, and finally determines the transformation parameters through least trimmed squares (LTS) regression. A two-level pyramidal strategy was used for coarse-to-fine image registration. For block-matching, both fixed and floating images were divided into uniform blocks of 4 voxel edge length. The top 25% of blocks with the highest intensity variance in the fixed image were used while the rest were discarded. Each of these image blocks was then compared to all floating image blocks that have an overlap of at least one voxel. The best match between the blocks from the fixed and floating images was determined as the pair with maximum absolute normalized cross-correlation (NCC). After establishing the point-wise correspondences, the second step was the update of transformation parameters via LTS regression. At every iteration, the composition of the block-matching correspondence and the transformation of the previous step determined the new transformation by LTS regression.

VI Results

A Phase I: Distances on the Training Data

The results obtained on the training dataset were reported by each team in their respective contribution to the challenge proceedings [27]. Most authors reported their distances obtained after registration for each case, although some reported only averaged values. Table III summarizes the mean distance between landmark pairs after registration, over all landmarks of all cases, computed by each team. All teams but one improved from the initial distances, with three teams achieving a mean distance under 1.75 mm and two more under 3.35 mm. Team cDRAMMS initially reported a mean distance between landmarks of 3.35 ± 1.39mm. With an updated version of their method, this error was later reduced to 2.28 ± 0.71mm. Sun and Zhang [11] provided partial results on 4 cases only, since the other 18 cases were used to train their neural network. This team eventually did not participate in the second phase.

B Phase II: Distances on the Test Data

This section presents the results of the 6 teams that completed phase II of the challenge, on the test dataset. Figure 2 first shows the results per test case, aggregated across all teams. Test cases with the largest initial errors (cases 1, 5, and 3) were the most difficult to treat. Results for test cases with the smallest initial error (4 and 9) were improved on average, but several teams also obtained larger distances after registration. Finally, results were consistently improved for all other cases with an initial error in the 4-6 mm range.

Regarding team-by-team results, mean distances between landmark pairs after registration are summarized in Table III while the distribution of these distances is detailed in Figure 3. Team ImFusion and DeedsSSC obtained a mean distance between landmark pairs well below 2 mm, respectively of 1.57 and 1.87 mm. These excellent results are consistent across all test cases, with a standard deviation around 1 mm for both teams, which confirmed the results reported on the training data set. Team cDRAMMS also consistently obtained very good results, with a mean error of 2.18 mm and a single large residual error of 4.3 mm for case 5. Results of team NiftyReg are more varied. As can be seen on the lower panel of Figure 3, they obtained excellent results for all cases but two, cases 1 and 5, where the distance was reduced from 15.7 to 5.9 mm and 13.2 to 12.8 mm, respectively. Without these two outliers, the mean distance over all cases would be reduced from 3.21 ± 3.57 mm to 1.70 ± 0.91 mm. Team FAX reported the best results on the training set, with a mean distance between landmark pairs of 1.21 ± 0.55 mm. However, this distance leaped to 5.70 ± 0.55 mm on the test data, which potentially shows their deep learning method overfitted the data during the training phase. Finally, team MedICAL obtained few or no improvements from the initial distances between landmark pairs.

C Phase II: Complementary Criteria

All submitted methods were fully automatic. Although it was not a factor in the evaluation, the average computational time per case (excluding image preprocessing) for each team is reported in Table IV, and each team's choice of software implementation and hardware set-

up vary. These values range from 1.8 sec for team FAX with a trained neural network to approximately 450 sec for team cDRAMMS using singe-thread CPU implementation.

D Qualitative Results

To demonstrate the data and the registration task, MR and iUS volumes of one patient was chosen from each of the training and test datasets, and are shown in Figures 4 and 5, respectively. The selected cases have a relatively large initial mTRE and a substantial variability between the teams. Also, note that these two cases do not necessarily directly reflect the overall ranking of the challenge, which was based on averaged rankings of all cases. As no quantitative measures of registration quality are available in a clinical setting, visual inspection of the images is important to obtain an impression of the registration quality. As shown in Figures 4 and 5, the registration accuracy can be evaluated by adapted visualization and identification of homologous features such as sulci, gyri and ventricles in the images.

E CuRIOUS2018 Challenge Ranks

Following the description in Section IV, all teams were ranked independently for each test case, based on the mean distances between landmark pairs. These case-by-case ranks are summarized in Figure 6, with the number of times each team was ranked at the ith place, with i from 1 to 6.

The winner and runner-up are teams ImFusion and DeedsSSC, which are perfectly consistent with their respective results reported in Figure 3. Note that ImFusion obtained the best registration for 6 of the 10 tests cases. Despite a larger mean registration error, team NiftyReg was ranked third before team cDRAMMS as it obtained a better case-by-case rank (3.1 vs 3.4). However, team cDRAMMS also had very good results, but more consistently handled all cases, including the extreme ones. This specific situation pointed out the fact that the challenge metric favors accuracy over precision, with a limited penalty when low quality results are obtained on a single case or two. To overcome this limit, since we consider precision as a crucial factor for the surgeons' acceptance of a method, the challenge's organizers decided to declare a tie for third place. Both NiftyReg and cDRAMMS thus received the same third place prize. Finally, both teams FAX and MedICAL obtained a mean case-by-case rank of 5.3, and were ranked tied at the 5th place in the challenge.

VII Discussion

In this challenge, the focus has been on MR-iUS registration in the context of brain tumor surgery. As both the training and test datasets exclusively contain data from LGG surgeries, there has been a special focus on this tumor type. The resection of LGGs is particularly challenging as the tumor tissue can be very similar to normal brain tissue. In LGG surgery, there are also fewer options for additional guidance as tools like 5-ALA fluorescence are not available. Intraoperative ultrasound is therefore an attractive solution in these cases. The optimization and benchmarking of available registration algorithms on data from these tumors is therefore particularly important for successful future clinical translation. Even though the emphasis has been on LGG, the results from the challenge will generalize well to

other tumor types such as HGGs and metastasis as these tumors are more distinct from normal brain tissue and depict clearer boundaries than LGGs in ultrasound images.

An important obstacle for the widespread use of iUS is the challenging and unfamiliar image interpretation. The integration of iUS into the navigation system and the visualization of corresponding slices in pre-operative MR and iUS makes this interpretation considerably more intuitive. With accurate MR-iUS registration, the surgeon can perform the resection based on the MR images even after brain shift, which makes the neuronavigation accurate and easy to interpret. MR-iUS registration also enables correction of other types of pre-operative MR data such as fMRI and DTI [28].

Image registration techniques tailored for MRI-iUS registration in this challenge were landmark-, intensity- or learning-based. The performance of landmark-based methods in non-linear image registration depends on both finding enough landmarks that cover the entire volume, and correctly finding their corresponding landmarks in the second volume. The voxel-wise attribute-based method of Machado *et al.* [13] (team cDRAMMS) did relatively well despite the fact that iUS and MRI have drastically different salient features, and ranked third in a tie with Drobny *et al.* [12] (team NiftyReg). The top three algorithms in this challenge [12]–[14] were all intensity-based techniques, which calculated a dense transformation map by utilizing intensity values at all locations.

Deep learning has been successfully applied to image registration [29], [30], but in nonlinear image registration tasks with high accuracy requirement, such as brain shift correction, further exploration is still needed [31]. The two submissions that used DL in this challenge were from Sun and Zhang [11], who did not participate to the second phase, and Zhong *et al.* [17] (team FAX), who ranked first in the results reported on the training database. However, their method did not work well on the test database. A common culprit for such behavior is overfitting, where the model overfits the training data and therefore performs poorly on the unseen test data. As more training data becomes available, this type of method is expected to perform better in the future.

Symmetric image registration techniques provide unbiased estimates of the transformation field and are known to generally outperform their asymmetric counterparts [32]. Two of the top three methods in this challenge [12], [14] compared the performance of their techniques in symmetric and asymmetric settings. They both concluded that asymmetric transformations lead to a superior performance in this challenge. This is an interesting finding and is likely due to the vast differences in physics of US and MR imaging modalities.

It was also noticed by some of the challenge participants and discussed during the event that in some cases affine transformations outperformed non-linear elastic transformations. This might seem surprising as brain shift is often described as a non-uniform deformation. However, before resection a large component of the experienced mismatch between MRI and iUS is often due to inaccurate patient-MRI registration. This is a rigid registration most often based on anatomical landmarks, fiducials, surfaces or a combination of these. Consequently, an affine transformation might be sufficient to correct for most of the

misalignment. After resection, the situation will be different with larger and highly non-linear deformations and affine transformations will likely not be sufficient to register the images. Some teams initialize their nonlinear registration with affine alignments first. This approach can ensure the overall registration robustness when only focusing on relatively local tissue deformation. However, in cases of updating other intra-operative data, such as pre-operative tractography [28], direct estimation of nonlinear deformation may be preferred.

In both the training and test databases, we selected landmarks that cover a large part of the iUS volume with sufficient distance between neighboring landmarks. This strategy provides a good benchmark for comparing image registration techniques. However, the quality of the alignment closer to the tumor is more clinically important as it better helps the neurosurgeon to optimize the resection size and location. To further investigate the registration performance, registration errors per landmark for each case are analyzed for the testing dataset. More specifically, the registration errors per landmark across all teams for each case were plotted in Figure 7. As expected, a higher variance of registration errors was observed for Case 1 & 5 with large brain shift (>10mm). Case 12 (initial mTRE = 4.89mm) also exhibits a relatively high variance in registration errors. This is likely due to blurry sulci patterns in MRI in comparison to iUS. Furthermore, we also computed the medians, minimums, and variances of the registration errors per landmark across all teams for each clinical case from the testing dataset, and the results are show in Figure 8. With Tukey's method, we identified outlier landmarks based on the computed statistical measures. In general, there are only very small amount of outlier landmarks (<10 out of 173). While the minimums and medians generally point to the same outlier landmarks, variances did not show any. Based on visual inspection, these landmarks are typically anatomical features, which are relatively close to the interface of tissue and transducer, and the border of the iUS volume, as well as sulcus patterns that are less distinguishable in MRI scans in comparison to the iUS due to differences in resolution.

The distance between corresponding landmarks in the two images before and after registration is a well-established metric for evaluation of registration results in the absence of a ground truth. Despite being widely used, this metric has some limitations. First, as there is only a limited number of landmarks associated with each image, the registration error is only evaluated at a limited number of locations and will therefore not capture local displacements and deformations in other locations. Thus, the number of landmarks and their distribution in the image volume are important. The landmarks in both the training and test sets have been carefully placed in order to capture the displacements and deformation as well as possible. However, we noticed that in Test Case 5, the registration results were not accurate by visual inspection even though the mTREs indicated successful alignment. This emphasizes the need for both quantitative and qualitative assessment of registration results. Another limitation of this metric is the localization error associated with manual placement of points. For the landmarks to be valid for evaluation of registration results, this localization error has to be significantly lower than the expected registration errors. We have measured the inter- and intra-rater variability in both the training and test data and shown that these are indeed significantly lower than the registration errors. Even though the landmarks do not

represent the absolute ground truth, they are valid for the evaluating the registration outcomes.

The use of landmarks as the only metric for the challenge also represent a limitation. For implementation in a clinical setting, for example, other characteristics would also be of critical importance, such as registration precision, accurate mapping of tumor boundaries, smoothness of deformation field (no vanishing tissues), and ease of use. However, some of the factors can be very complex and difficult to quantify, and in the challenge setting, the use of a single well-defined metric is advantageous. A single metric enables a straightforward, comprehensible ranking scheme and an open, fair competition. With the use of multiple metrics, there will always be a discussion on the weighting of the different criteria and how to aggregate them. The rules for aggregation of the ranks in this challenge were outlined before the challenge and were not changed at any point. Still, the system used favors accuracy over precision. As discussed during the challenge event, for the clinical users, high precision and high accuracy are equally important and precision can even be more important than accuracy. This point should be re-designed and improved in future editions of the challenge.

VIII   Future Work

In the first edition, the registration task solely focused on MRI-iUS registration before dura-opening and after craniotomy. However, with the progress of tumor resection, tissue deformation is an on-going process, and accurate tracking can ensure the complete removal of cancerous tissues, preventing any additional surgeries. Intended as a recurrent open challenge to further improve the registration algorithms, we expect to introduce multiple sub-challenges in future CuRIOUS challenges to target brain shift correction at different stages of the surgery, especially during and after resection.

For clinical practices, besides accuracy and robustness, processing speed is an imperative factor. In the inaugural edition of the challenge, performance speed was not emphasized in scoring the teams because it can be affected by multiple factors, including implementation platforms, for prototype algorithms. In future challenges, we aim to place discussions and emphasis on this topic, as well as optimization algorithms to direct the results of the challenge towards more realistic clinical implementations.

IX   Conclusion

Holding great clinical values, MRI-iUS registration for correcting tissue shift in brain tumor resection is still a difficult task. As the first public image processing challenge to tackle this clinical problem, the CuRIOUS2018 Challenge provided a common platform to evaluate and discuss existing and emerging registration algorithms on this topic. The results of CuRIOUS2018 provided valuable insights for the current developments and challenges from both the technical and clinical perspectives. This is an important step forward to help translate research-grade automatic image processing into clinical practice to benefit the patients and clinicians.

Acknowledgment

References

[1]. Jakola AS, et al. Comparison of a strategy favoring early surgical resection vs a strategy favoring watchful waiting in low-grade gliomas. Proc JAMA. 2012 Nov; 308(18):1881–1888.

[2]. Schomas DA, et al. Neill, Intracranial low-grade gliomas in adults: 30-year experience with long-term follow-up at mayo clinic. Neuro-Oncol. 2009 Aug; 11(4):437–445. [PubMed: 19018039]

[3]. Arbel T, Morandi X, Comeau RM, Collins DL. Automatic non-linear MRI-ultrasound registration for the correction of intra-operative brain deformations. Comput Aided Surg. 2004; 9(4):123–136. [PubMed: 16192052]

[4]. Reinertsen I, Lindseth F, Unsgaard G, Collins DL. Clinical validation of vessel-based registration for correction of brain-shift. Med Image Anal. 2007 Dec; 11(6):673–684. [PubMed: 17681484]

[5]. Coupé P, Hellier P, Morandi X, Barillot C. 3D rigid registration of intraoperative ultrasound and preoperative MR brain images based on hyperechogenic structures. Int J Biomed Imag. 2012 Oct.2012 531319

[6]. De Nigris D, Collins DL, Arbel T. Multi-modal image registration based on gradient orientations of minimal uncertainty. IEEE Trans Med Imag. 2012 Dec; 31(12):2343–2354.

[7]. Wein, W; Ladikos, A; Fuerst, B; Shah, A; Sharma, K; Navab, N. Global registration of ultrasound to MRI using the LC^2metric for enabling neurosurgical guidance; Proc Int Conf Med Image Comput Comput-Assist Intervent; 2013. 34–41.

[8]. Rivaz H, Chen sJ-S, Collins DL. Automatic deformable mr-ultrasound registration for image-guided neurosurgery. IEEE Trans Med Imag. 2015 Feb; 34(2):366–380.

[9]. Xiao Y, Fortin M, Unsgard G, Rivaz H, Reinertsen I. REtroSpective evaluation of cerebral tumors (RESECT): A clinical database of pre-operative MRI and intra-operative ultrasound in low-grade glioma surgeries. Med Phys. 2017 Jul; 44(7):3875–3882. [PubMed: 28391601]

[10]. Hong, J; Park, H. Non-linear approach for MRI to intra-operative US registration using structural skeleton; Proc Int Workshops, POCUS, BIVPCS, CuRIOUS, CPM, MICCAI; Granada, Spain. 2018 Sep. 138–145.

[11]. Sun, L; Zhang, S. Deformable MRI-ultrasound registration using 3D convolutional neural network; Proc Int Workshops, POCUS, BIVPCS, CuRIOUS, CPM, MICCAI; Granada, Spain. 2018 Sep. 152–158.

[12]. Drobny, D; Vercauteren, T; Ourselin, S; Modat, M. Registration of MRI and iUS data to compensate brain shift using a symmetric block-matching based approach; Proc Int Workshops POCUS, BIVPCS, CuRIOUS, CPM, MICCAI; Granada, Spain. 2018 Sep. 172–178.

[13]. Machado, I; , et al. Deformable MRI-ultrasound registration via attribute matching and mutual-saliency weighting for image-guided neurosurgery; Proc Int Workshops, POCUS, BIVPCS, CuRIOUS, CPM, MICCAI; Granada, Spain. 2018 Sep. 165–171.

[14]. Heinrich, MP. Intra-operative ultrasound to MRI fusion with a public multimodal discrete registration tool; Proc Int Workshops, POCUS, BIVPCS, CuRIOUS, CPM, MICCAI; Granada, Spain. 2018 Sep. 159–164.

[15]. Shams, R; Boucher, M-A; Kadoury, S. Intra-operative brain shift correction with weighted locally linear correlations of 3DUS and MRI; Proc Int Workshops, POCUS, BIVPCS, CuRIOUS, CPM, MICCAI; Granada, Spain. 2018 Sep. 179–184.

[16]. Wein, W. Brain-shift correction with image-based registration and landmark accuracy evaluation; Proc Int Workshops, POCUS, BIVPCS, CuRIOUS, CPM, MICCAI; Granada, Spain. 2018 Sep. 146–151.

[17]. Zhong, X; , et al. Resolve intraoperative brain shift as imitation game; Proc Int Workshops, POCUS, BIVPCS, CuRIOUS, CPM, MICCAI; Granada, Spain. 2018 Sep. 129–137.

[18]. Reinke, A, , et al. How to Exploit Weaknesses in Biomedical Challenge Design and Organization. Springer; New York, NY, USA: 2018. 388–395.

[19]. Ou Y, Sotiras A, Paragios N, Davatzikos C. DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting. Med Image Anal. 2011 Aug; 15(4):622–639. [PubMed: 20688559]

[20]. 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 Imag. 2014 Oct; 33(10):2039–2065.

[21]. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal. 2008 Feb; 12(1):26–41. [PubMed: 17659998]

[22]. Roche, A, Malandain, G, Pennec, X, Ayache, N. The correlation ratio as a new similarity measure for multimodal image registrationMedical Image Computing and Computer-Assisted Intervention — MICCAI. Springer; New York, NY, USA: 1998. 1115–1124.

[23]. Iversen DH, Wein W, Lindseth F, Unsgard G, Reinertsen I. Automatic intraoperative correction of brain shift for accurate neuronavigation. World Neurosurg. 2018 Dec.120:E1071–E1078. [PubMed: 30213682]

[24]. Jones DR, Pertunnen CD, Stuckman BE. Lipschitzian optimization without the Lipschitz constant. J Optim Theory Appl. 1993; 79(1):157–181.

[25]. Powell MJD. The BOBYQA algorithm for bound constrained optimization without derivatives. Dept Appl Math Theor Phys, Univ Cambridge, Cambridge, U.K., Tech Rep. 2009

[26]. Modat M, Cash DM, Daga P, Winston GP, Duncan JS, Ourselin S. Global image registration using a symmetric block-matching approach. J Med Imag. 2014 Sep.1(2) 024003

[27]. Stoyanov, D, , et al. Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation: International Workshops, POCUS 2018, BIVPCS 2018, CuRIOUS 2018, and CPM 2018, Held in Conjunction With MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings. Springer; 2018.

[28]. Xiao Y, Eikenes L, Reinertsen I, Rivaz H. Nonlinear deformation of tractography in ultrasound-guided low-grade gliomas resection. Int J Comput Assist Radiol Surg. 2018 Mar.13:457–467. [PubMed: 29299739]

[29]. Miao S, Wang ZJ, Liao R. A CNN regression approach for real-time 2D/3D registration. IEEE Trans Med Imag. 2016 May; 35(5):1352–1363.

[30]. Hou B, et al. 3-D reconstruction in canonical co-ordinate space from arbitrarily oriented 2-D images. IEEE Trans Med Imag. 2018 Aug; 37(8):1737–1750.

[31]. Litjens G, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017 Dec.42:60–88. [PubMed: 28778026]

[32]. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage. 2011 Feb; 54(3):2033–2044. [PubMed: 20851191]
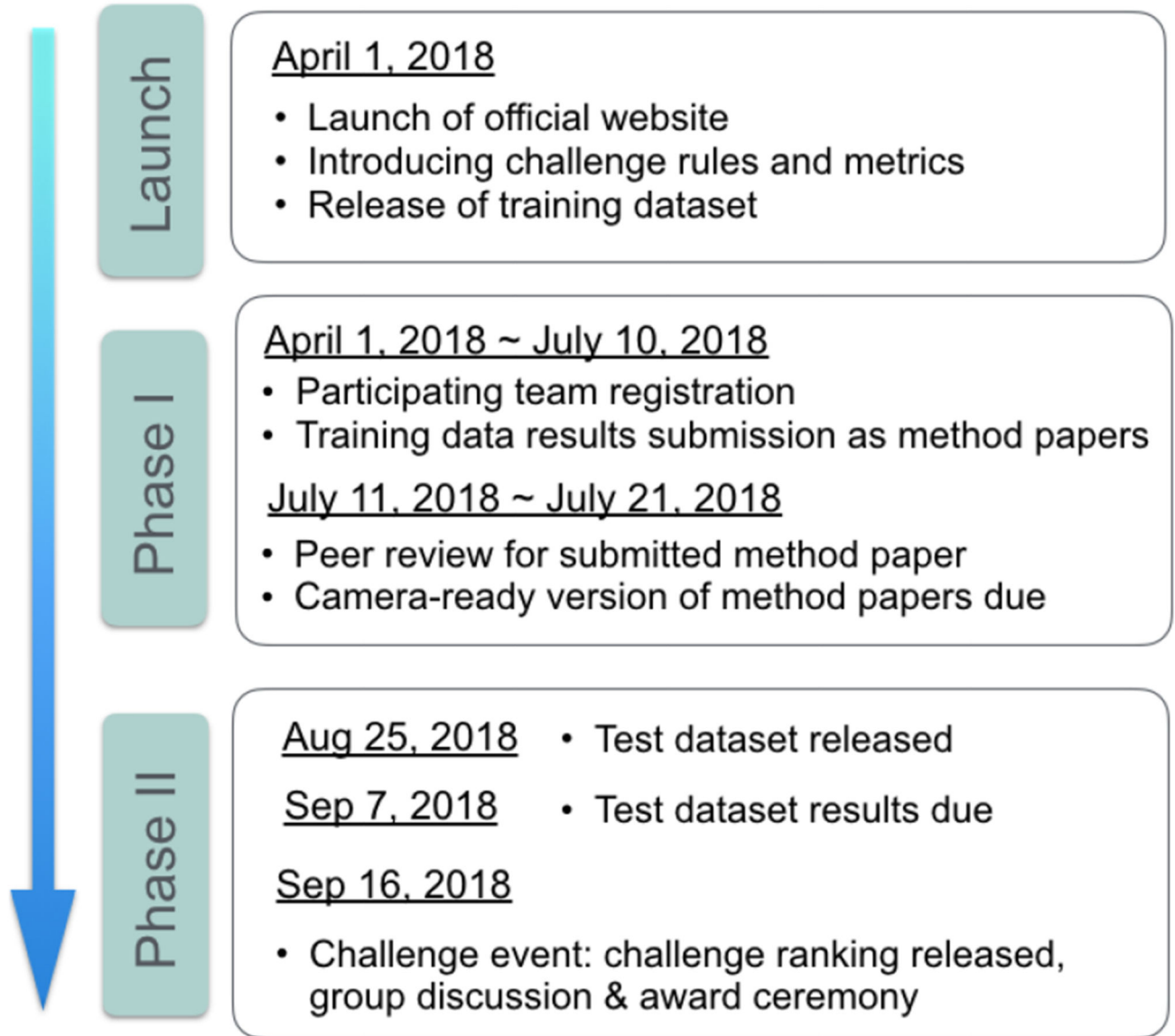
Launch

April 1, 2018

- Launch of official website
- Introducing challenge rules and metrics
- Release of training dataset

Phase I

April 1, 2018 ~ July 10, 2018

- Participating team registration
- Training data results submission as method papers

July 11, 2018 ~ July 21, 2018

- Peer review for submitted method paper
- Camera-ready version of method papers due

Phase II

Aug 25, 2018 • Test dataset released

Sep 7, 2018 • Test dataset results due

Sep 16, 2018

- Challenge event: challenge ranking released, group discussion & award ceremony

Fig. 1. Timeline for CuRIOUS 2018 challenge.
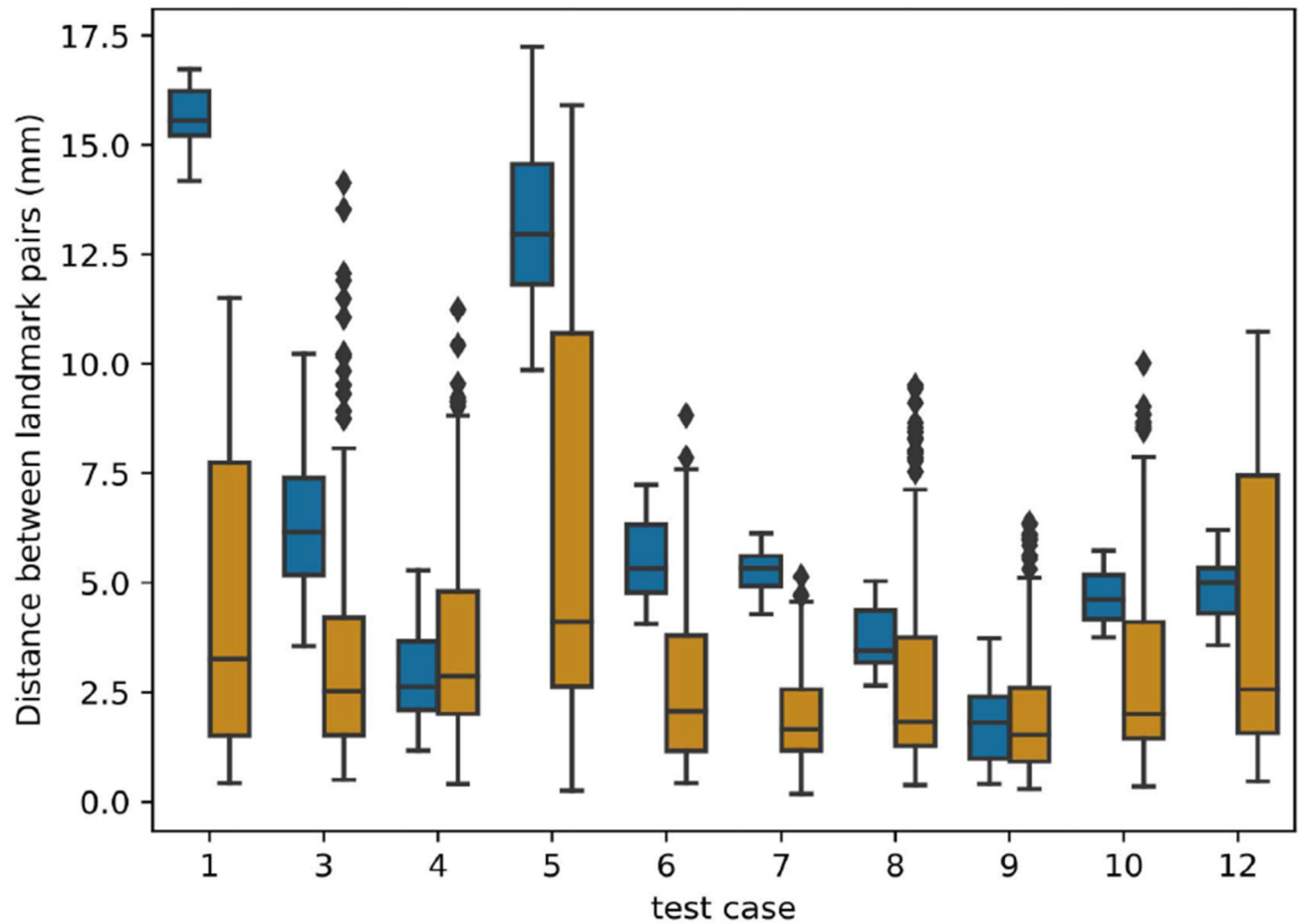
Fig. 2.
Results per test case: box plot distribution of the distances between landmark pairs. For each test case, the left box plot (blue) shows the initial distance before registration while the right box plot (orange) shows the distribution after registration, aggregated over all teams.
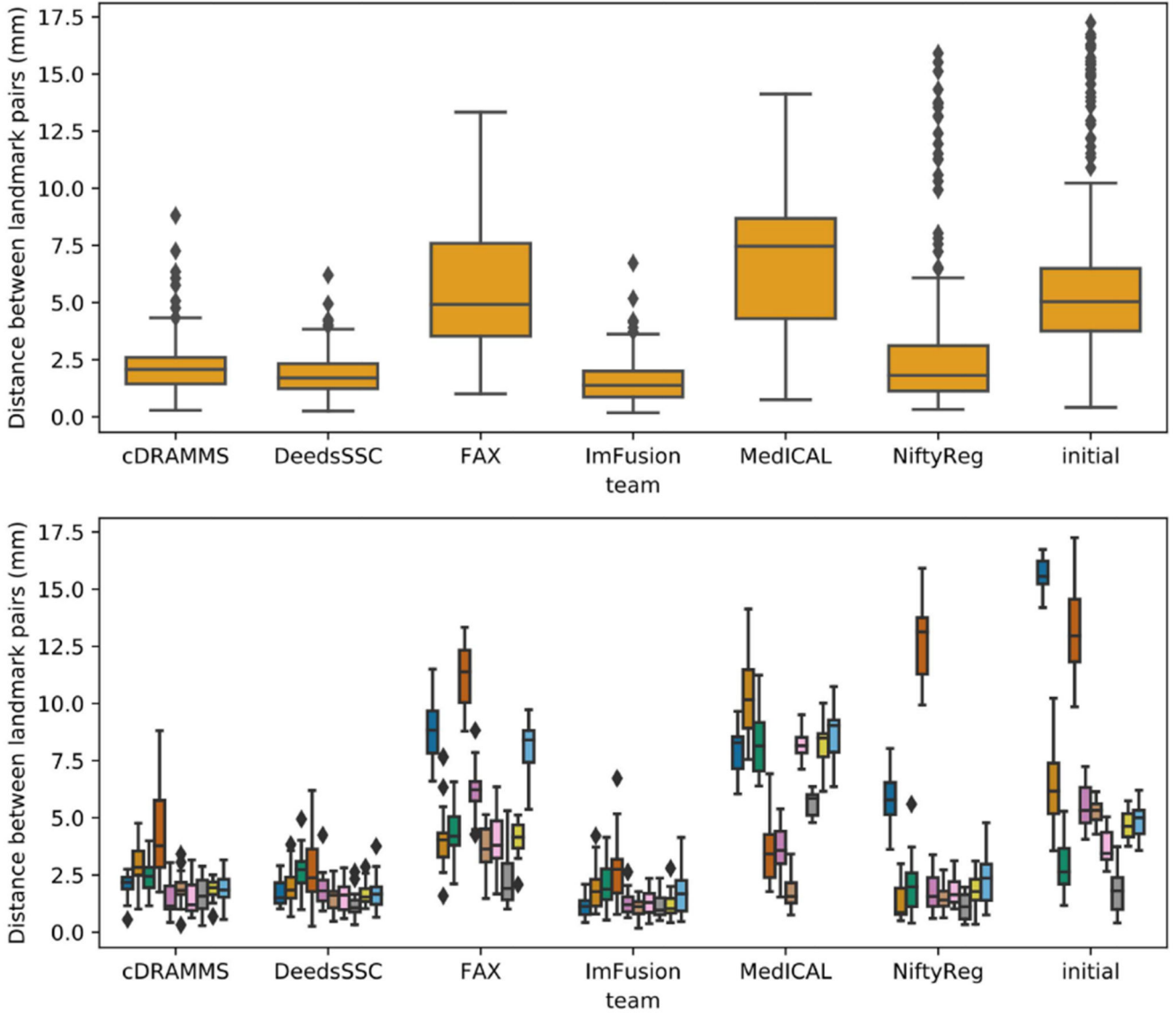
Fig. 3.
Distribution of the distances between landmark pairs obtained by each team after registration, on the test set. For comparison, the last column contains the initial distances before registration. The upper panel shows the global results computed over all landmarks of all test cases. in the lower panel, these results are split by test case.
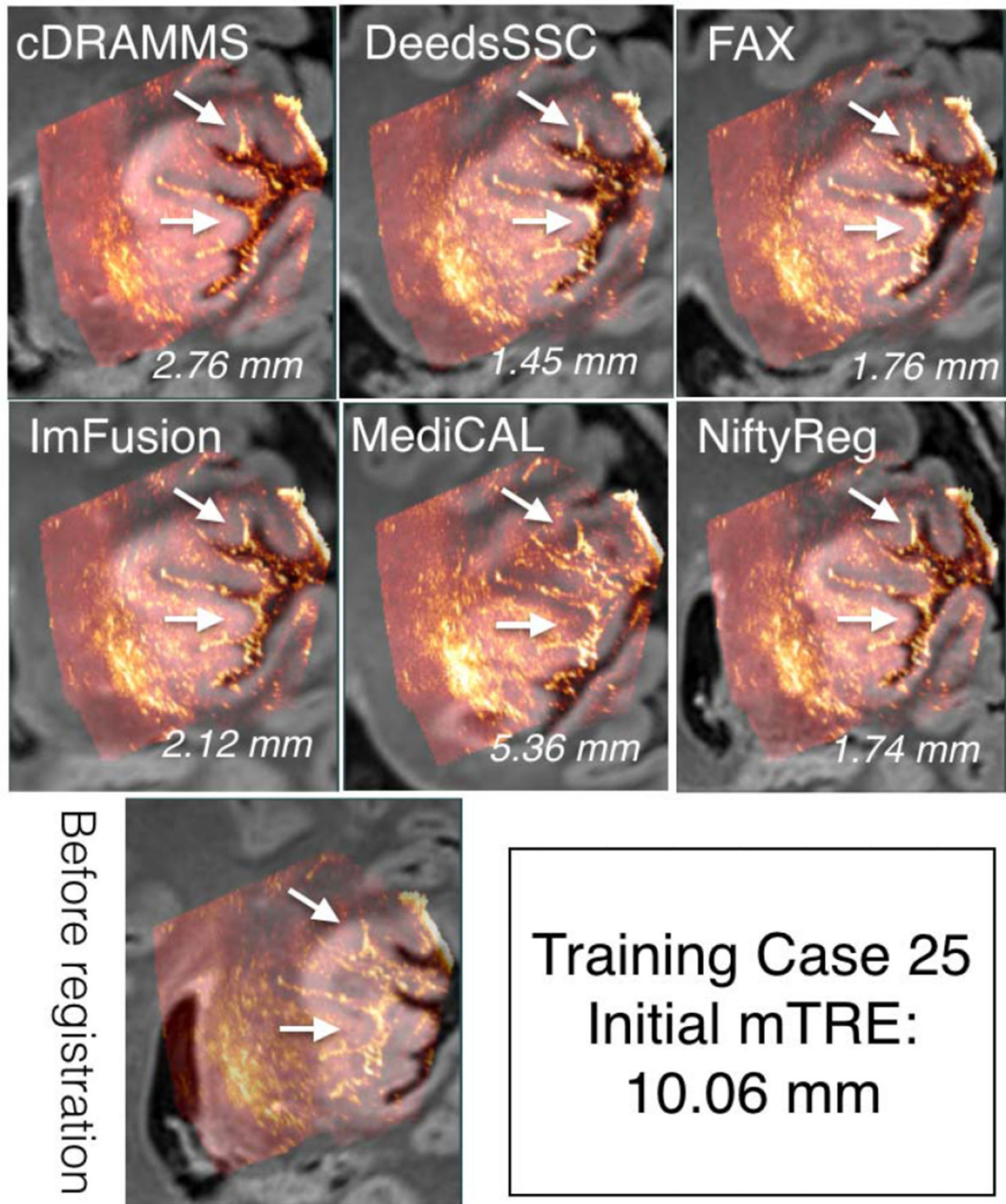
Fig. 4.
Qualitative comparison of registration results for Training Case 25 across different teams. For each team, the ultrasound and deformed FLAIR MRI scan is overlaid together. The mTRE values for each team is listed at the right bottom corner of each image overlay. The arrows point to the sulcus patterns with varied registration quality among teams.
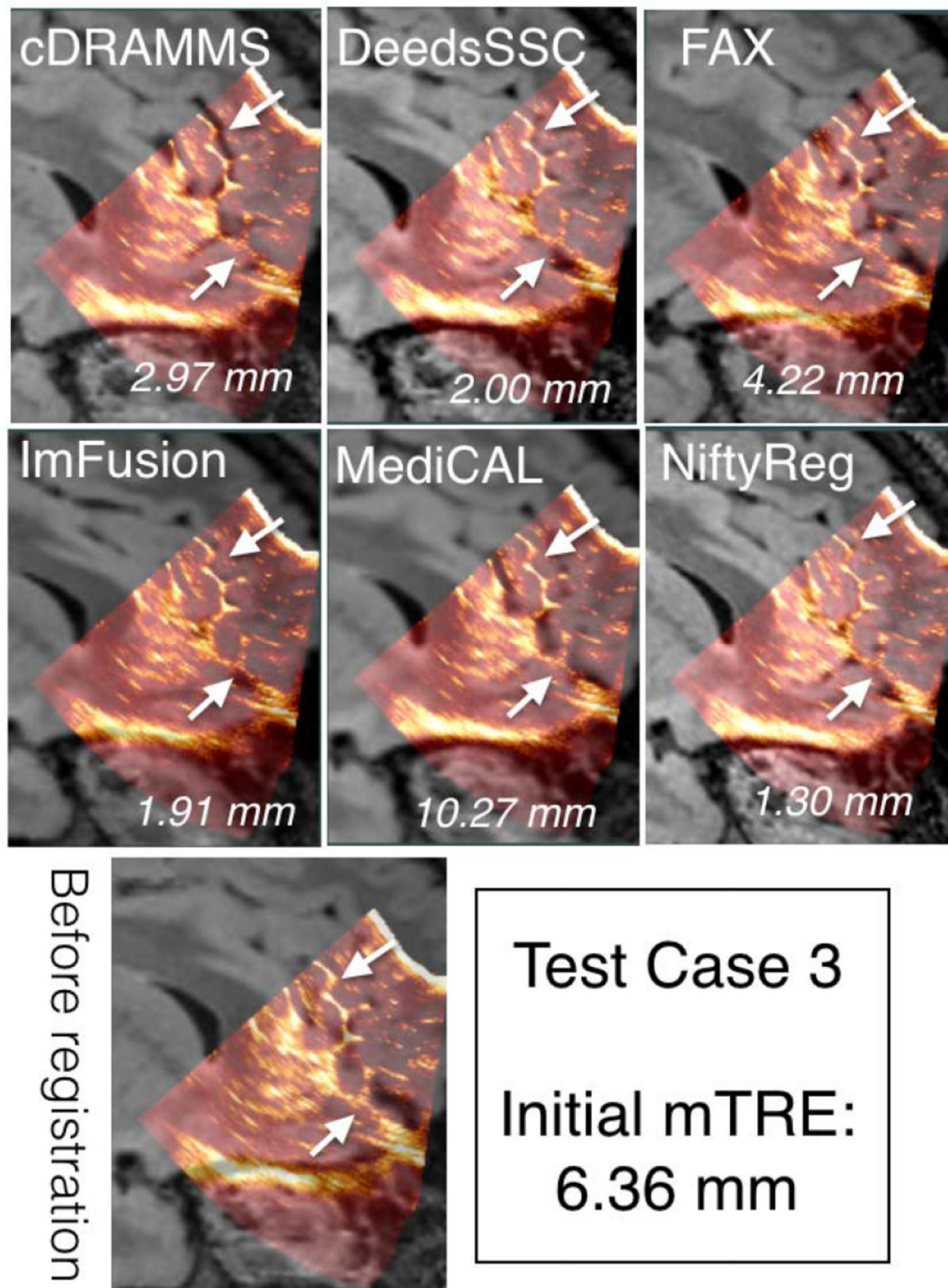
Fig. 5.
Qualitative comparison of registration results for Test Case 3 across different teams. For each team, the ultrasound and deformed FLAIR MRI scan is overlaid together. The mTRE values for each team is listed at the right bottom corner of each image overlay. The arrows point to the sulcus patterns with varied registration quality among teams.
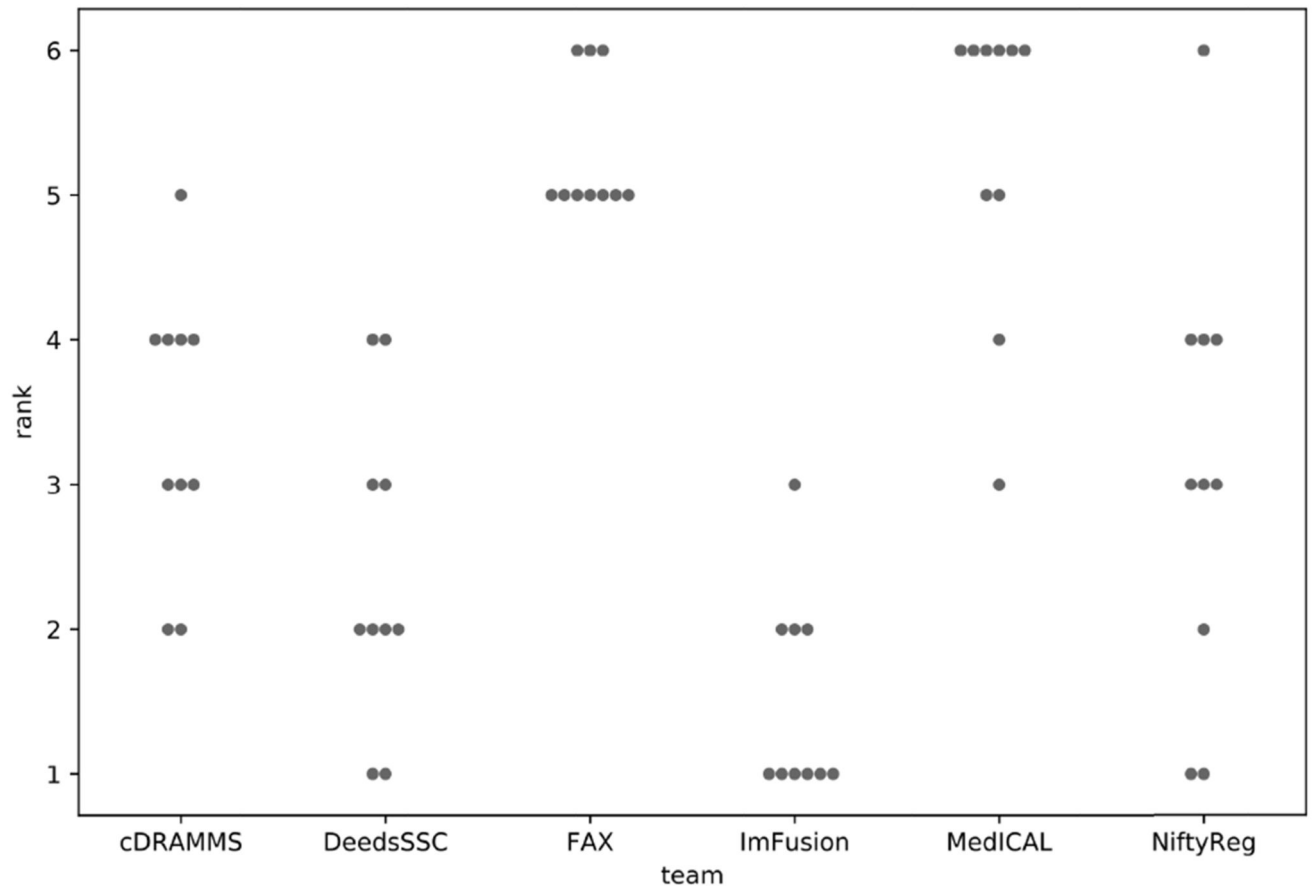
Fig. 6.
Case-by-case ranks for each team. For example: over the ten test cases, team ImFusion was ranked first six times, second three times, and third one time.
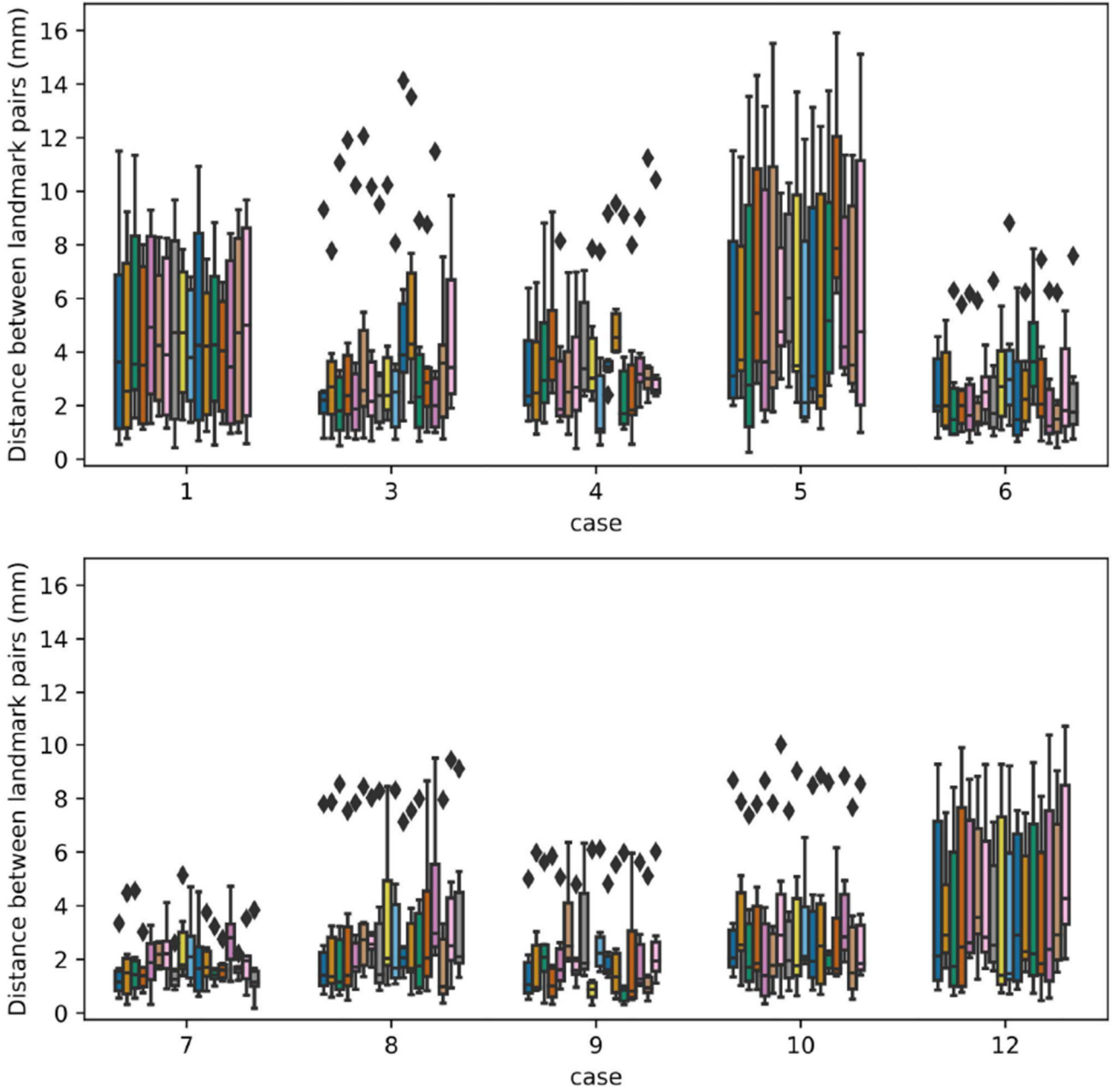
Fig. 7.
Regisstration errors per landmark across all teams shown as boxplots for each case from the testing dataset. The error is measured as the Euclidean distance between the deformed landmark in MRI and the corresponding one in ultrasound, and the landmarks are ordered from the left to the right for each case according to their number provided.
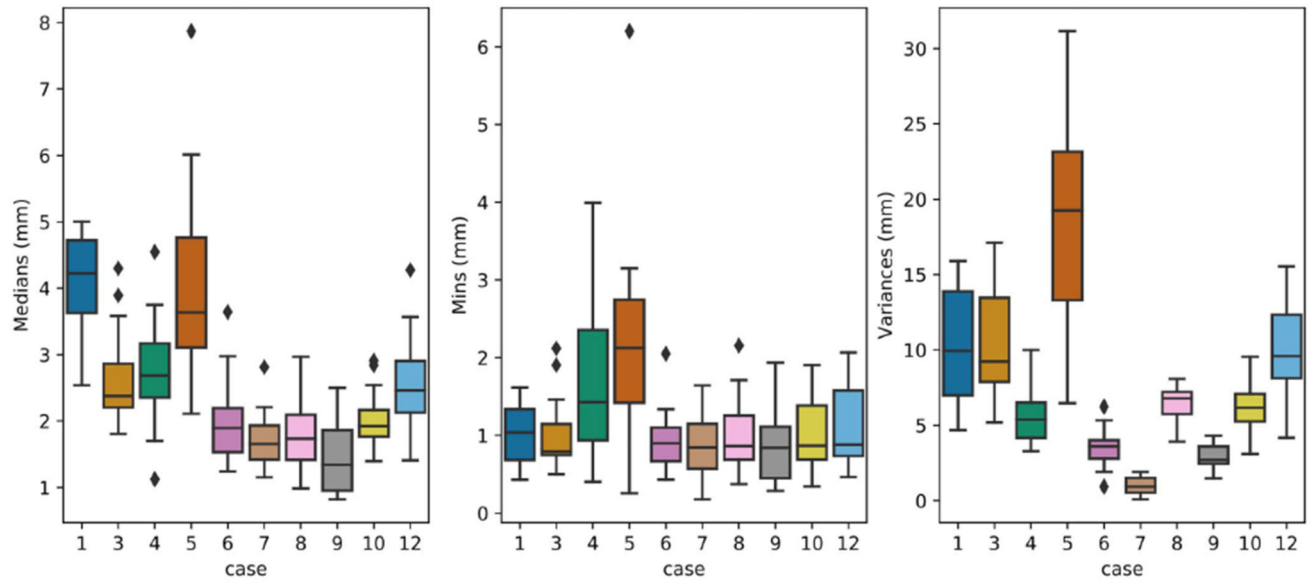
Fig. 8. The medians, minimums, and variances of registration errors for each landmark across all teams, and the results are plotted as boxplots for each case from the testing dataset.

Table I

Details of Inter-Modality Landmarks for Each Patient in the Testing Dataset

Patient ID	# of landmarks MRI vs. before US	Mean initial distance (range) in mm MRI vs. before US
1	17	15.66 (14.19~16.74)
3	17	6.36 (3.57~10.23)
4	17	2.98 (1.17~5.28)
5	17	13.19 (9.86~17.25)
6	18	5.52 (4.07~7.24)
7	18	5.27 (4.28~6.14)
8	18	3.73 (2.66~5.04)
9	17	1.80 (0.41~4.15)
10	17	4.66 (3.76~5.74)
12	17	4.89 (3.58~6.21)
mean±sd	17.3+0.5	6.4l±4.46

The number of landmarks and mean initial Euclidean distances between landmark pairs are shown, and the range (min ~ max) of the distances is shown in parenthesis after the mean value.

Table II

Inter- and Intra-Rater Evaluations With Mean Euclidean Distance Between Landmark Sets

Type	Intra-rater Rater 1	Intra-rater Rater 2	Inter-rater R1 vs. R2
Training data	0.47±0.10 mm	0.33±0.06 mm	0.33±0.08 mm
Testing data	0.2l±0.l0 mm	0.48±0.22 mm	0.42±0.17mm

The results are shown as mean±standard deviation.

Table III

Summary of the Challenge Results

Team	Distances between landmark pairs after registration Mean ± std, in mm Training set / Test set		Mean case-by-case rank	Final challenge rank
cDRAMMS	3.35 ±1.39	2.18 ±1.23	3.4	3 =
DeedsSSC	1.67 ± 0.54	1.87 ± 0.93	2.4	2
FAX	1.21 ± 0.55	5.70 ± 2.93	5.3	5 =
ImFusion	1.75 ± 0.62	1.57 ± 0.96	1.5	1
MedICAL	4.60 ± 3.40	6.59 ± 2.89	5.3	5 =
NiftyReg	2.90 ± 3.59	3.21 ± 3.57	3.1	3 =
Hong et al.	5.60 ± 3.94	6.65 ± 4.55	-	-
*Sun et al.	3.91 ± 0.53	-	-	-
Initial distances	5.37 ± 4.27	6.38 ± 4.36	-	-

For each team, the first columns give the mean distances between landmark pairs after registration, computed over all landmark of all cases, for the training and test sets. The mean case-by-case rank, computed on the test set only, and the final challenge rank are then given. For comparison, the last line contains the mean initial distances, before registration. Teams cDRAMMS and NiftyReg were eventually ranked tied at third (=). Hong et al. sent results on the test data but did not attend the challenge event. Sun et al. sent only partial results (*) on the training set, but did not participate to the second phase of the challenge. These two teams were thus not ranked.

Table IV

Summary of Average Computational Time Per Case for Each Team

Team	Mean comutational time per case	Implementation with GPU/CPU
cDRAMMS	450 sec	CPU
DeedsSSC	25 sec	CPU
FAX	1.8 sec	CPU
ImFusion	20 sec	GPU
MedICAL	103 sec	CPU
NiftyReg	115 sec	GPU