Learn2Reg - The Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

Use the title to convey the essential information on the challenge mission.

Learn2Reg - The Challenge

Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

L2R

Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Medical image registration plays a very important role in improving clinical workflows, computer-assisted interventions and diagnosis as well as for research studies involving e.g. morphological analysis. Besides ongoing research into new concepts for optimisation, similarity metrics and deformation models, deep learning for medical registration is currently starting to show promising advances that could improve the robustness, computation speed and accuracy of conventional algorithms to enable better practical translation. Nevertheless, there exists no commonly used benchmark dataset to compare state-of-the-art learning based registration among another and with their conventional (not trained) counterparts. With few exceptions (CuRIOUS at MICCAI 2018/2019 and the Continuous Registration Challenge at WBIR 2018) there has also been no comprehensive registration challenge covering different anatomical structures and evaluation metrics. We also believe that the entry barrier for new teams to contribute to this emerging field are higher than e.g. for segmentation, where standardised datasets (e.g. Medical Decathlon, BraTS) are easily available. In contrast, many registration tasks, require resampling from different voxel spacings, affine pre-registration and can lead to ambiguous and error-prone evaluation of whole deformation fields.
We propose a simplified challenge design that removes many of the common pitfalls for learning and applying transformations. We will provide pre-preprocessed data (resample, crop, pre-align, etc.) that can be directly employed by most conventional and learning frameworks. Only displacement fields in voxel dimensions in a standard orientation will have to be provided by participants and python code to test their application to training data will be provided as open-source along with all evaluation metrics. Our challenge will consist of 4 clinically relevant sub-tasks (datasets) that are complementary in nature. They can either be individually or comprehensively addressed by participants and cover both intra- and inter-patient alignment, CT, ultrasound and MRI modalities, neuro-, thorax and abdominal anatomies and the four of the imminent challenges of medical image registration:
1) learning from small datasets
2) estimating large deformations
3) dealing with multi-modal scans

4) learning from noisy annotations
An important aspect of challenges are comprehensive and fair evaluation criteria. Since, medical image registration is not limited to accurately and robustly transferring anatomical annotations but should also provide plausible deformations, we will incorporate a measure of transformation complexity (the standard deviation of local volume change defined by the Jacobian determinant of the deformation). To encourage the submission of learning based approaches that reduce the computational burden of image registration, the run-time computation time will also be included into the ranking by awarding extra points. Due to differences in hardware, the computation time (including all steps of the employed pipeline) will be measured by running algorithms on the grand-challenge.org evaluation platform (where the whole challenge will be hosted) with the potential of using Nvidia GPU backends (this will not be a strict requirement for participants).

Challenge keywords

List the primary keywords that characterize the challenge.

registration, brain, thorax, abdomen, deformable, multimodal, realtime

Year

The challenge will take place in ...

2020

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

If the challenge is part of a workshop, please indicate the workshop.

None, however a more elaborate syllabus than for conventional challenges is planned to widen the scope and learning outcome for participants:
We plan short presentations of best-performing challenge participants (approx. 2 hours in total), a poster session for all challenge entries and two overview talks and discussions (half hour each) on the challenge tasks and results. To complement the algorithmic presentations, we plan up to 3 invited talks (half hour each) on current advances in learning-based registration.
Based on the overwhelming response of our tutorial Learn2Reg at MICCAI 2019 (more than 70 participants, see https://github.com/learn2reg/tutorials2019 for more information), we concluded that there is a need to provide educational content that enables a greater part of the MICCAI community to engage in medical image registration. Hence, an additional hour tutorial lecture will be integrated into the challenge syllabus.

Duration

How long does the challenge take?

Full day.

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

We expect 70-100 attendees (based on our 2019 MICCAI tutorial), with approx. 20 participating teams (based on the EMPIRE10 challenge at MICCAI 2010).

Publication and future plans

Please indicate if you plan to coordinate a publication of the challenge results.

We plan a joint journal publication in a high-impact journal (TMI or MEDIA) to discuss the challenge results, method developments and remaining research gaps. We will encourage all participants to submit a short (4-8 pages) LNCS paper of their contribution in particular of any novel approaches to medical image registration.

Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

We will make use of the grand-challenge.org website to host the challenge, evaluation engine and leaderboards. Nvidia GPU support would be greatly appreciated for participants and to enable evaluation of GPU runtimes. No onsite challenge is planned, but participants will get the opportunity to upload their transformations at least twice to be evaluated on a small subset of the test data in order to avoid any pitfalls (wrong orientation, etc.).

TASK: CuRIOUS

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Early brain tumor resection can effectively improve the patient's survival rate. However, resection quality and safety can often be heavily affected by intra- operative brain tissue shift due to factors, such as gravity, drug administration, intracranial pressure change, and tissue removal. Such tissue shift can displace the surgical target and vital structures (e.g., blood vessels) shown in pre- operative images while these displacements may not be directly visible in the surgeon's field of view. Intra-operative ultrasound (iUS) is a robust and relatively inexpensive technique to track intra-operative tissue shift and surgical tools, but to help update pre-surgical plans with this information, accurate and robust image registration algorithms are needed to relate pre-surgical MRI to iUS images. Despite the great progress so far, medical image registration techniques still have not made into the surgical room to directly benefit the patients with brain tumors.

Keywords

List the primary keywords that characterize the task.

intra-patient, ultrasound, MRI, registration, multimodal

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

Adrian Dalca (MIT), Yipeng Hu (UCL), Tom Vercauteren (KCL), Mattias Heinrich (Uni Lübeck), Lasse Hansen (Uni Lübeck), Marc Modat (KCL), Bob de Vos (Uni Amsterdam), Yiming Xiao (Western University), Hassan Rivaz (Concordia University), Matthieu Chabanas (University of Grenoble Alpes), Ingerid Reinertsen, (SINTEF), Bennett Landman (Vanderbilt), Jorge Cardoso (KCL), Bram van Ginneken (Radboud, Nijmegen), Alessa Hering (Fraunhofer MEVIS), Keelin Murphy (Radboud, Nijmegen)

b) Provide information on the primary contact person.

Mattias Heinrich (heinrich@imi.uni-luebeck.de)

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event open call.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

learn2reg.grand-challenge.org

c) Provide the URL for the challenge website (if any).

https://curious2019.grand-challenge.org

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Additional public and non-public data can be used to (pre-train) algorithms as long as authors clearly state this in their method description.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

Organisers and team members of organisers may participate and are ranked but cannot win prizes.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

(tbd) We are in discussions with sponsors (among others Nvidia and Medtronic) to provide monetary and hardware prices for challenge winners.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

All results will be announced publicly unless an obvious error was made in data processing. Participating teams may choose to submit multiple methods, given there are sufficiently distinct and not simply other hyper-parameters and provided that each algorithm is described to clarify differences. Organisers reserve the right to remove lower scoring duplicate submissions from the same team of algorithms that are deemed too similar.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

Up to two team members qualify as author for joint publication. Participants are free to publish their own results separately, but can only make reference to overall results once the challenge paper is published (submission to arXiv is considered as a sufficient waiting period).

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Option 1) upload displacement fields (computed offline) to grand-challenge.org evaluation system (zero score for computation time) .
To limit the data transfer volume we will provide a processing script that will compress displacement fields: half precision float16 (for all tasks), zero values outside lungs (for task 2), downsample to half resolution (tasks 1-3). For task 1 the submission of affine matrices will also be possible.
The evaluation of our metrics on these compressed transformations can be checked on training data with our public evaluation tool.
Option 2) upload docker container to evaluation system to be run on cloud hardware.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Participants will get the opportunity to upload their transformations at least twice to be evaluated on a small subset of the test data in order to avoid any pitfalls (wrong orientation, etc.).

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

February 2020: registration to challenge open
February 2020: release of labelled training data (task 1,3 and 4)
May 2020: release of training data for task 2
June 2020: release of test data w/o labels
Mid July 2020: submission of participants test transformations
End July 2020: contributed challenge papers (reporting on training results is sufficient)

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

If necessary (e.g. not the case for already open sourced data) ethics approval will be requested prior to releasing any new data.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY SA.

Additional comments: Similar to the Medical Decathlon we plan to apply the following:
All training data, training labels and pre-processed data will be made available online (if possible with a permissive copyright-license CC-BY-SA 4.0). If data providers can obtain permissive licenses, participants are allowed to share, re-distribute and improved upon the data. All data will labeled and verified by an expert human rater, and with the best effort to mimic the accuracy required for clinical use. Labels for training data will be made available as soon as available, but test labels are kept secrete.

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

The evaluation software, code to evaluate accuracy, complexity etc will be made publicly available.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

Participating teams' are encouraged, but not required to make their code publicly available. We will provide links to available source code on the challenge website.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

Data providers who annotate test cases have access to test labels (of specific sub-task), organisers who implement evaluation metrics and scripts will have partial access to test labels.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Intervention planning, Surgery, Decision support.

Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Registration.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Patients with low-grade gliomas who underwent intra-operative ultrasound-guided tumor resection procedures.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

All scans were acquired for routine clinical care of brain tumor resection procedures at St Olavs University Hospital (Trondheim, Norway). For each clinical case, the pre-operative 3T MRI includes Gadolinium-enhanced T1w and T2 FLAIR scans, and the intra-operative US volume was obtained to cover the entire tumor region after craniotomy but before dura opening, as well as after resection was completed.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Magnetic Resonance Imaging (MRI) & Ultrasound (US)

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

The MRI data were acquired before the surgery, and the ultrasound data were acquired with an optical tracking system during the surgery so that they would truthfully reflect the states of brain shift. For each clinical case, matching anatomical landmarks between MRI and US scans were provided to assess registration quality.

b) ... to the patient in general (e.g. sex, medical history).

Patients with low-grade gliomas who underwent intra-operative ultrasound-guided tumor resection procedures.

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

The MRI data of the brain was acquired before the surgery, and the ultrasound data were acquired with an optical tracking system during the surgery so that they would truthfully reflect the states of brain shift.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Tissue deformation in low-grade brain gliomas resection.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Robustness, Reliability, Accuracy, Runtime.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Preoperative MR images were acquired on a 1.5T Magnetom Avanto and 3T Magnetom Skyra MRI scanner (Siemens, Erlangen, Germany).
The 3D iUS scans were acquired with the Sonowand Invite neuronavigation system (Sonowand AS, Trondheim, Norway). A Polaris camera (NDI, Waterloo, Canada) built in the Sonowand system was used to obtain the position and pose of the ultrasound probe.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

The MR protocol included a sagittal T1w Gdenhanced sequence (TE = 2.96 ms, TR = 2000 ms, flip angle = 8°, 192 sagittal slices, acquisition matrix = 256 × 256, voxel size = 1.0 × 1.0 × 1.0 mm3), and a sagittal T2w fluidattenuated inversion recovery, or FLAIR (TE = 388 ms, TR = 5000 ms, flip angle = 120°, 192 sagittal slices, acquisition matrix = 256 × 256, voxel size = 1.0 × 1.0 × 1.0 mm3) sequence both acquired on a 3T Magnetom Skyra MRI scanner (Siemens, Erlangen, Germany) with a 20channel head coil. The combined imaging time for both T1w and T2 FLAIR scans was 12 min. For patients 2, 14 and 15, preoperative MR images were acquired on a 1.5T Magnetom Avanto (Siemens, Erlangen, Germany) MRI scanner with a 12channel head coil, and included a sagittal T1w Gdenhanced sequence (TE = 2.30 ms, TR = 2500 ms, flip angle = 7°, 176 sagittal slices, slice thickness = 1 mm, acquisition matrix = 512 × 496, inplane resolution = 0.5 × 0.5 mm2), and a sagittal FLAIR (TE = 333 ms, TR = 6000 ms, flip angle = 120°, 176 sagittal slices, slice thickness = 1 mm, acquisition matrix = 256 × 224, inplane resolution = 1.0 × 1.0 mm2).

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 0.14x0.14x0.14 mm3 to 0.24x0.24x0.24 mm3 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.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

All data was acquired at St. Olavs University Hospital, Trondheim, Norway and is provided by the organizers of the CuRIOUS challenge.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Not applicable.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

A case refers to an MRI iUS image pair of a single patient. Both, training and test images, are labeled with homologous anatomical landmarks.

b) State the total number of training, validation and test cases.

Training: 22 cases
Test: 10 cases

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

Providing corresponding landmarks in multimodal scans is very time consuming and only possible for domain experts hence larger-scale databases for this task are not available.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

There are slight difference in characteristics between training and test cases (but inter-/intra-rater agreement on annotations was similar).

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

The reference anatomical landmarks were selected by Rater 1 in the ultrasound volume before dura is open after craniotomy. Then two raters selected matching anatomical landmarks in the ultrasound volumes obtained after tumor resection using the software 'register' from MINC Toolkit. The ultrasound landmark selection was repeated twice for each rater with a time interval of at least one week. Finally the results (4 points for each landmark location) were averaged. Eligible anatomical landmarks include deep grooves and corners of sulci, convex points of gyri, and vanishing points of sulci.
Same raters produced the anatomical landmarks for both the training and testing data.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

See above.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Two medical imaging experts that were experienced in brain anatomy.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

The final landmarks in both training and testing database were provided as the averaged results of two trials of landmark marked by both raters (four 3D points for each landmark).

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

All images are in the same referential. Common pre-processing to same voxel resolutions and spatial dimensions will be provided to ease the use of learning-based algorithms for participants with little prior experience in image registration.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Train data:
Intra-rater Rater 1: 0.47±0.10 mm
Intra-rater Rater 2: 0.33±0.06 mm
Inter-rater R1 vs. R2: 0.33±0.08

Test data:
Intra-rater Rater 1: 0.21±0.10 mm
Intra-rater Rater 2: 0.48±0.22 mm
Inter-rater R1 vs. R2: 0.42±0.17

b) In an analogous manner, describe and quantify other relevant sources of error.

not applicable

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

1) TRE of a few dozens landmarks
2) -

3) -
4) Robustness: 30% highest TRE of all cases
5) SD (standard deviation) of log Jacobian determinant
6) Run-time computation time (only awarded when evaluation server is used with provided docker container)

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

DSC or TRE respectively measure accuracy; HD95 measures reliability; Outliers are penalised with the robustness score (30% of lowest mean DSC or 30% of highest mean TRE)
The smoothness of transformations (SD of log Jacobian determinant) are important in registration, see references of Kabus and Leow
Run-time computation time is relevant for clinical applications.

We have considered inverse consistency as additional metric, which is controversial among researchers in medical registration. We decided to not use it as competitive (ranking) metric, but compute it for informative reasons (i.e. the question whether inverse-consistent algorithms are more robust).

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

All metrics but 4) (robustness) use mean rank per case (ranks are normalised to between 0.1 and 1, higher being better). For multi-label tasks the ranks are computed per structure and later averaged. As done in the Medical Segmentation Decathlon we will employ "significant ranks" http://medicaldecathlon.com/files/MSD-Ranking-scheme.pdf
Across all metrics an overall score is aggregated using the geometric mean. This encourages consistency across criteria. The time ranks are only considered with 50% weight (since not all participants are able to use docker containers for evaluation).

b) Describe the method(s) used to manage submissions with missing results on test cases.

Missing results will be awarded the lowest rank (potentially shared and averaged across teams).

c) Justify why the described ranking scheme(s) was/were used.

The geometric mean encourages consistency across criteria. A ten-fold difference between highest and lowest score is fixed to be independent of number of participants.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

Missing data/submission will result in lowest rank for this case.
Ties will result in average rank among all equal participants.

b) Justify why the described statistical method(s) was/were used.

The geometric mean is more robust against outliers, hence methods that perform well on all metrics are encouraged.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

We will provide several baseline algorithms to compare new methods against, there are:
- Voxelmorph (CVPR'18) with and without label supervision
- Deeds
- Elastix
- NiftyReg
- FSL Fnirt
- ANTs

TASK: Lung CT

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Deformable registration of computed tomography (CT) lung volumes has several important clinical applications. In an inter-subject context, it may be used for atlas-based segmentation of the lungs and its lobes to propagate expert segmentations from a database to a new subject. Longitudinal CT scans (inspiration to inspiration) from the same patient can be used to monitor treatment or disease progression, e.g., for lung nodules. Motion estimation of 4DCT scans is now widely used for radiotherapy planning to increase the accuracy of dose delivery. Deformable registration of expiration to inspiration CT scans can also enable direct estimation of lung ventilation to assess patients with breathing disorders, in particular chronic obstructive pulmonary disease, or help to spare well-functioning lung tissue from radiotherapy.

Lung image registration has been an active field of research for decades, also ignited by the very successful EMPIRE challenge at MICCAI 2010. More recently, deep learning based approaches have been successfully employed for image registration, which are in principle well suited for lung registration, because they automatically learn to aggregate relevant information of various complexities in images. New algorithms could provide the potential for higher robustness, because local optima may be of lesser concern and are highly parallelisable enabling very fast implementations and straight-forward execution on GPUs. As a consequence new parallel algorithms (both non-supervised and DL-based) are of great interest for time-critical applications; e.g. image-guided radiotherapy. However, capturing large motion and deformation is still an open challenge. In common iterative image registration approaches, this is typically addressed with iterative multilevel coarse-to-fine registration strategies or discrete displacement settings. To date, even supervised DL-based algorithms have not reached the accuracy of conventional methods for lung registration, which makes this task very relevant for the Learn2Reg registration challenge.

Keywords

List the primary keywords that characterize the task.

intra-patient, CT, lung, inspiration, expiration, deformable registration

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

as above

b) Provide information on the primary contact person.

as above

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event open call.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

learn2reg.grand-challenge.org

c) Provide the URL for the challenge website (if any).

learn2reg.grand-challenge.org

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Additional public and non-public data (e.g. from dir-lab.com) can be used to (pre-train) algorithms as long as authors clearly state this in their method description.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

Organisers and team members of organisers may participate and are ranked but cannot win prizes.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

as above

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

as above

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

as above

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

as above

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

as above

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

as above

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

as above

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC SA.

Additional comments: As above, however, the inspiration / expiration data cannot be redistributed or modified / improved upon.

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

as above

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

as above

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

as above

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Longitudinal study, Treatment planning, Diagnosis, Intervention follow up.

Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Registration.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Patients with clinical routine follow-up CT lung examinations (inspiration-inspiration), that may have been part of lung screening cancer trials (e.g. Nelson study and National Lung Screening Trial) as well as inspiration-expiration CT scan pairs from patients with breathing disorders (COPD, etc.).

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Patients with clinical routine follow-up CT lung examinations (inspiration-inspiration), that may have been part of lung screening cancer trials (e.g. Nelson study and National Lung Screening Trial) as well as inspiration-expiration CT scan pairs from patients with breathing disorders (COPD, etc.)
(tbd) the data will be collected from Nijmegen Radboud Medical Centre and the Amsterdam University Medical Centre.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Computed Tomography (CT)

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

Semi-automatic lung lobe segmentations (with manual corrections) will be provided for training data. In addition 20-50 manual selected corresponding landmark pairs (within the same subject, primarily at vessel/airway bifurcations) will be annotated by the challenge organisers (following in principle: K. Murphy et al: "Semi-automatic Reference Standard Construction .." MICCAI 2008) , see also K Murphy, B v. Ginneken et al. TMI 2012.

b) ... to the patient in general (e.g. sex, medical history).

Patients with clinical routine follow-up CT lung examinations (inspiration-inspiration), that may have been part of lung screening cancer trials (e.g. Nelson study and National Lung Screening Trial) as well as inspiration-expiration CT scan pairs from patients with breathing disorders (COPD, etc.).

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

The CT data will show the thorax covering both lungs fully, the registration focusses only on internal lung structures (bones and surrounding organs can be ignored).

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Highly accurate alignment of inner lung structures: lobes, fissures, vessels and airways as well as plausible deformations with spatial smoothness (low standard deviation of Jacobian determinants).

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Robustness, Complexity, Accuracy, Runtime.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

(tbc) Philips Brilliance 16P or Philips Mx8000 IDT 16

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

(tbc) The inspiration scans will be acquired using a low-dose protocol (30 mAs) while the expiration scan with ultra-low-dose (20 mAs). The scanner could e.g. be a Philips Brilliance 16P with slice thickness of 1.00 mm and slice spacing of 0.70 mm. Pixel spacing in the X and Y directions may vary from 0.63 to 0.77 mm with an average value of 0.70 mm.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

(tbc) The data will be provided by Radboud Nijmegen Medical Centre and Amsterdam University Medical Centre, we are also considering to get access to a subset of the COPDgene study data.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Not applicable.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

A case refers to a pair of CT scans from the same patient. All test scan pairs are labelled manually with (sparse) corresponding anatomical landmarks and lung lobe segmentations (see also reference K Murphy MICCAI 2008). For the larger training cohort lung lobe segmentations are provided along with (more densely sampled) automatic correspondences following Ruhaak et al. TMI 2018 "Estimation of Large Motion in Lung CT by Integrating Regularized Keypoint Correspondences ..". These have been shown to nearly match manual annotation accuracy and are beneficial for supervised learning algorithms.

b) State the total number of training, validation and test cases.

Training: 30 cases (20 insp-insp, 10 insp-exp)
Test: 10 cases (5 insp-insp, 5 insp-exp)

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

Inspiration-expiration scan pairs are not normally acquired in clinical practice, inspiration only pairs present are simpler registration task. Data augmentation can help to overcome limitations of small datasets (cf. K Eppenhoff and J Pluim "Pulmonary ct registration through supervised learning with convolutional neural networks" TMI 2018).

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

As mentioned before, more densely but noisier annotations will be provided for the training datasets, which will give exciting insight into the generalisation of learning based methods.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Semi-automatic lung lobe segmentations (with manual corrections) will be provided for training data. In addition 20-50 manual selected corresponding landmark pairs (within the same subject, primarily at vessel/airway bifurcations) will be annotated by the challenge organisers (following in principle: K. Murphy et al: "Semi-automatic Reference Standard Construction .." MICCAI 2008)

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

see above

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Two medical imaging experts that were experienced in lung anatomy.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

The final landmark correspondence for the testing database will be provided as the averaged results of two trials of landmarks marked by both raters (with same reference point).

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Common pre-processing to same voxel resolutions and spatial dimensions as well as affine pre-registration will be provided to ease the use of learning-based algorithms for participants with little prior experience in image registration. Lung masks will be provided to enable the exclusion of outer lung-structures, which increase the problem complexity (e.g. by sliding motion).

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

based on previous studies inter-rater variabilities of 0.5-1.0mm are expected for manual landmarks and lobe segmentations.

b) In an analogous manner, describe and quantify other relevant sources of error.

Not applicable.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

1) TRE of a few dozens landmarks
2) DSC (Dice similarity coefficient) of lobe segmentations
3) HD95 (95% percentile of Haussdorff distance) of segmentations
4) Robustness: 30% highest TRE of all cases
5) SD (standard deviation) of log Jacobian determinant
6) Run-time computation time (only awarded when evaluation server is used with provided docker container)

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

same as above

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

same as above

b) Describe the method(s) used to manage submissions with missing results on test cases.

same as above

c) Justify why the described ranking scheme(s) was/were used.

same as above

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

same as above

b) Justify why the described statistical method(s) was/were used.

same as above

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

same as above

TASK: Abdominal CT

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

The human abdomen is an essential, yet complex body space. Bounded by the diaphragm superiorly and pelvis inferiorly, supported by spinal vertebrae, and protected by muscular abdominal wall, the abdomen contains organs involved with blood reservation, detoxification, urination, endocrine function, and digestion, and includes many important arteries and veins. Computed tomography (CT) scans are routinely obtained for the diagnosis and prognosis of abdomen-related disease; yet no specific image registration tools for the abdomen have been developed. On abdominal CT, inter-subject variability (e.g., age, gender, stature, normal anatomical variants, and disease status) can be observed in terms of the size, shape, and appearance of each organ. Soft anatomy deformation further complicates the registration by varying the inter-organ relationships, even within individuals (e.g., pose, respiratory cycle, edema, digestive status). The relevance for this task in the Learn2reg challenge is given by aligning several disjunct regions with large inter-subject variations and great variability in volume: from 5 millilitre (adrenal glands) to 1.6 litre (liver). In the future, we would also consider to add simulated transformations to increase the size of training datasets - but for now leave this up to participants to explore on their own (cf. Eppenhof and Pluim TMI 2018)

Keywords

List the primary keywords that characterize the task.

inter-patient, CT, abdomen, deformable registration

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

as above

b) Provide information on the primary contact person.

as above

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event open call.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

learn2reg.grand-challenge.org

c) Provide the URL for the challenge website (if any).

https://www.synapse.org/#!Synapse:syn3193805

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Additional public and non-public data (e.g. from TCIA Pancreas-CT) can be used to (pre-train) algorithms as long as authors clearly state this in their method description.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

Organisers and team members of organisers may participate and are ranked but cannot win prizes.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

as above

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

as above

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

as above

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

as above

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

as above

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

as above

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

as above

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

as above

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

as above

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

as above

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Longitudinal study, Treatment planning, Diagnosis, Intervention follow up.

Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Registration.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

On the one hand patients undergoing image-guided surgical interventions, biopsies or radiotherapy could benefit from a deformable anatomical organ atlas that captures spatial relations of organs-at-risk and target regions. On the other hand, shape analysis over large cohorts could provide insight into epidemiological difference in relation to common disease.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Patients from an colorectal cancer chemotherapy trial, the baseline sessions of the abdominal CT scans were randomly selected from metastatic liver cancer patients; the remaining scans were acquired from a retrospective post-operative cohort with suspected ventral hernias.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Computed Tomography (CT)

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

Thirteen abdominal organs were considered regions of interest (ROI), including spleen, right kidney, left kidney, gall bladder, esophagus, liver, stomach, aorta, inferior vena cava, portal and splenic vein, pancreas, left adrenal gland, and right adrenal gland. Segmentations for these ROIs will be provided for training data.

b) ... to the patient in general (e.g. sex, medical history).

Patients from an colorectal cancer chemotherapy trial, the baseline sessions of the abdominal CT scans were randomly selected from 75 metastatic liver cancer patients; the remaining scans were acquired from a retrospective post-operative cohort with suspected ventral hernias.

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

The CT data will show the abdomen covering thirteen regions of interest (ROI), including spleen, right kidney, left kidney, gall bladder, esophagus, liver, stomach, aorta, inferior vena cava, portal and splenic vein, pancreas, left adrenal gland, and right adrenal gland.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Alignment of thirteen abdominal organs as regions of interest (ROI), including spleen, right kidney, left kidney, gall bladder, esophagus, liver, stomach, aorta, inferior vena cava, portal and splenic vein, pancreas, left adrenal gland, and right adrenal gland as well as plausible deformations with spatial smoothness (low standard deviation of

Jacobian determinants).

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Robustness, Complexity, Accuracy, Runtime.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

(tbc) CT scanners across the Vanderbilt University Medical Center (VUMC)

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

All scans were captured during portal venous contrast phase with variable volume sizes (512 × 512 × 53 ~ 512 × 512 × 368) and field of views (approx. 280 × 280 × 225 mm3 ~ 500 × 500 × 760 mm3). The in-plane resolution varies from 0.54 × 0.54 mm2 to 0.98 × 0.98 mm2, while the slice thickness ranged from 1.5 mm to 7.0 mm.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

The data is provided by the Vanderbilt University Medical Center (VUMC).

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Not applicable.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

A case refers to a pair of CT scans, each from a different patient (inter-patient registration). All CT scans (train and test) are provided together with segmentations from abdominal organs.

b) State the total number of training, validation and test cases.

Training: 30 cases
Test: 20 cases

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

Additional to the image data, segmentations of 13 abdominal organs are provided . As pixel wise annotations are a very time consuming task (especially for 3D volume data) and are only possible for domain experts larger-scale databases for this task are not available.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

In this inter-subject registration task, all potential pairs can be employed for training. To limit the amount of test transformations to be uploaded, we will announce 40 randomly selected pairs of test subjects that should be registered for evaluation. To measure inverse consistency of algorithms, we will include a subset of bi-directive cases).

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Thirteen abdominal organs were considered regions of interest (ROI), including spleen, right kidney, left kidney, gall bladder, esophagus, liver, stomach, aorta, inferior vena cava, portal and splenic vein, pancreas, left adrenal gland, and right adrenal gland. The organ selection was essentially based on [Shimizu A, Ohno R, Ikegami T, Kobatake H, Nawano S, Smutek D. Segmentation of multiple organs in non-contrast 3D abdominal CT images. International Journal of Computer Assisted Radiology and Surgery. 2007;2:135–142.]. As suggested by a radiologist, the heart was excluded for lack of full appearance in the datasets, and instead the adrenal glands were included for clinical interest. These ROIs were manually labeled by two experienced undergraduate students with 6 months of training on anatomy identification and labeling, and then verified by a radiologist on a volumetric basis using the MIPAV software.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

see above

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Two experienced medical undergraduate students with 6 months of training on anatomy identification and labeling.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Each scan was annotated by a single rater, thus no merging was required. The consistency across different raters was measured as inter-rater variability (see respective field).

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Common pre-processing to same voxel resolutions and spatial dimensions as well as affine pre-registration will be provided to ease the use of learning-based algorithms for participants with little prior experience in image registration.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

mean overall DSC overlap between the raters (i.e., inter-rater variability) was 0.87 ± 0.13 (0.95 ± 0.04 when considering only the spleen, kidneys, and liver).

b) In an analogous manner, describe and quantify other relevant sources of error.

Not applicable.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

1) -
2) DSC (Dice similarity coefficient) of segmentations
3) HD95 (95% percentile of Haussdorff distance) of segmentations

4) Robustness: 30% lowest DSC of all cases

5) SD (standard deviation) of log Jacobian determinant

6) Run-time computation time (only awarded when evaluation server is used with provided docker container)

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

same as above

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

same as above

b) Describe the method(s) used to manage submissions with missing results on test cases.

same as above

c) Justify why the described ranking scheme(s) was/were used.

same as above

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

same as above

b) Justify why the described statistical method(s) was/were used.

same as above

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

same as above

TASK: Hippocampus MRI

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Seated in the temporal lobe, the hippocampus is one of the most studied structures in the human brain. In particular, the hippocampus is known to be involved in aging, memory, and spatial navigation. The hippocampus has also been identified as a key structure in the pathophysiology of Alzheimer's disease, schizophrenia, and epilepsy. The challenge in this task is the alignment of two neighboring small structures (hippocampus head and body) with high precision on mono-modal MRI images between different patients.
Compared to the other tasks, this registration is expected to be easier to estimate: there are smaller deformations, the contrast of structures in the MRI is sufficient and the anatomies are usually at least partially overlapping before alignment. We also anticipate new insights into learning based registration, because in contrast to Task 3 this is a large-scale dataset with hundreds of labelled 3D scans.

Keywords

List the primary keywords that characterize the task.

inter-patient, MRI, hippocampus, deformable registration, small structure

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

as above

b) Provide information on the primary contact person.

as above

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event open call.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

learn2reg.grand-challenge.org

c) Provide the URL for the challenge website (if any).

http://medicaldecathlon.com

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Additional public and non-public data can be used to (pre-train) algorithms as long as authors clearly state this in their method description.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

Organisers and team members of organisers may participate and are ranked but cannot win prizes.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

as above

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

as above

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

as above

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

as above

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

as above

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

as above

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

as above

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY SA.

Additional comments: as above

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

as above

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

as above

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

as above

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Longitudinal study, Treatment planning, Diagnosis, Intervention follow up.

Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Registration.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Patients associated with Alzheimer's disease, schizophrenia, and epilepsy.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

The dataset consisted of MRI acquired in healthy adults and adults with a non-affective psychotic disorder taken from the Psychiatric Genotype/Phenotype Project data repository at Vanderbilt University Medical Center (Nashville, TN, USA). Patients were recruited from the Vanderbilt Psychotic Disorders Program and controls were recruited from the surrounding community.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Magnetic Resonance Imaging (MRI)

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

All MRI images (train and test) are provided together with segmentations from the hippocampus formation.

b) ... to the patient in general (e.g. sex, medical history).

The dataset consisted of MRI acquired in 90 healthy adults and 105 adults with a non-affective psychotic disorder (56 schizophrenia, 32 schizoaffective disorder, and 17 schizophreniform disorder) taken from the Psychiatric Genotype/Phenotype Project data repository at Vanderbilt Univer- sity Medical Center (Nashville, TN, USA). Patients were recruited from the Vanderbilt Psychotic Disorders Program and controls were recruited from the surrounding community.

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

The MRI data will show parts of the brain covering the hippocampus formation.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

The algorithm targets the alignment of two neighboring small structures (hippocampus head and body) with high precision on mono-modal MRI images between different patients (new insights into learning based registration in contrast to Task 3 due to a large-scale dataset).

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Robustness, Complexity, Accuracy, Runtime.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

All images were collected on a Philips Achieva scanner (Philips Healthcare, Inc., Best, The Netherlands)

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Structural images were acquired with a 3D T1-weighted MPRAGE sequence (TI/TR/TE, 860/8.0/3.7 ms; 170 sagittal slices; voxel size, 1.0 mm3).

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

The data is provided by the Vanderbilt University Medical Center (VUMC).

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Not applicable.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

A case refers to a pair of MRI scans, each image from a different patient (inter-patient registration). All MRI images (train and test) are provided together with segmentations from the hippocampus formation.

b) State the total number of training, validation and test cases.

Training: 263 cases
Test: 131 cases

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

A large-scale database is ideal for learning based registration with deep neural networks.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

In this inter-subject registration task, all potential pairs can be employed for training. To limit the amount of test transformations to be uploaded, we will announce 100 randomly selected pairs of test subjects that should be registered for evaluation. To measure inverse consistency of algorithms, we will include a subset of bi-directive cases).

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Manual tracing of the head, body, and tail of the hippocampus on images was completed following a previously published protocol [Pruessner, J. et al. Volumetry of hippocampus and amygdala with high- resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cerebral cortex 10, 433–442 (2000).; Woolard, A. & Heckers, S. Anatomical and functional correlates of human hippocampal volume asymmetry. Psychiatry Research: Neuroimaging 201, 48–53 (2012).]. For the purposes of this dataset, the term hippocampus includes the hippocampus proper (CA1-4 and dentate gyrus) and parts of the subiculum, which together are more often termed the hippocampal formation [Amaral, D. & Witter, M. The three-dimensional organization of the hip- pocampal formation: a review of anatomical data. Neuroscience 31, 571– 591 (1989)]. The last slice of the head of the hippocampus was defined as the coronal slice containing the uncal apex.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

see above

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Manual tracing of the head, body, and tail of the hippocampus on images was completed following a previously published protocol [Pruessner, J. et al. Volumetry of hippocampus and amygdala with high- resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cerebral cortex 10, 433–442 (2000).; Woolard, A. & Heckers, S. Anatomical and functional correlates of human hippocampal volume asymmetry. Psychiatry Research: Neuroimaging 201, 48–53 (2012).].

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Each scan was annotated by a single rater, thus no merging was required.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Common pre-processing to same voxel resolutions and spatial dimensions as well as affine pre-registration will be provided to ease the use of learning-based algorithms for participants with little prior experience in image registration.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

b) In an analogous manner, describe and quantify other relevant sources of error.

not applicable

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

1) -
2) DSC (dice similarity coefficient) of segmentations
3) HD95 (95% percentile of Haussdorff distance) of segmentations
4) Robustness: 30% lowest DSC of all cases
5) SD (standard deviation) of log Jacobian determinant
6) Run-time computation time (only awarded when evaluation server is used with provided docker container)

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

same as above

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

same as above

b) Describe the method(s) used to manage submissions with missing results on test cases.

same as above

c) Justify why the described ranking scheme(s) was/were used.

same as above

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

same as above

b) Justify why the described statistical method(s) was/were used.

same as above

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

same as above

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

Y Xiao et al.: "Evaluation of MRI to ultrasound registration methods for brain shift correction: The CuRIOUS2018 Challenge" to appear in IEEE TMI 2020
K Murphy et al.: "Semi-automatic reference standard construction for quantitative evaluation of lung CT registration" MICCAI 2008
A Simpson et al.: "A large annotated medical image dataset for the development and evaluation of segmentation algorithms" arXiv 2019
K Murphy, B Van Ginneken et al.: "Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge" TMI 2011
L Mercier, et al.: "Online database of clinical MR and ultrasound images of brain tumors" Medical Physics 2012
AD Leow, et al.: "Statistical properties of Jacobian maps and the realization of unbiased large-deformation nonlinear image registration" TMI 2007
S Kabus, et al.: "Evaluation of 4D-CT Lung Registration" MICCAI 2009
Z Xu, et al.: "Evaluation of six registration methods for the human abdomen on clinically acquired CT" TBME 2016
L Maier-Hein et al.: “Is the winner really the best? A critical analysis of common research practice in biomedical image analysis competitions”, arXiv:1806.02051, 2018 (for significance ranking)
Eppenhof, K. A., & Pluim, J. P.: "Pulmonary CT registration through supervised learning with convolutional neural networks." TMI 2018

Further comments

Further comments from the organizers.

We will provide results of baseline algorithms (elastix, NiftyReg, deeds, ANTs) to be included in ranking but not eligible to win.