Learn2Reg: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Learn2Reg

Challenge acronym

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

Learn2Reg

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.

Biomedical image registration is nowadays ubiquitously used in science and hospitals yet research developments do not always directly translate into improved robustness and accuracy for clinical workflows. Learn2Reg has been striving to close this perceived gap by providing comprehensive metrics for fair evaluation and challenging tasks to address unmet clinical needs. This year we aim to make a further leap toward engaging both a wider range of researchers and opening up our established concepts for running a successful challenge to the community. First, Learn2Reg 2024 will be co-organised with the 11th Workshop on Biomedical Image Registration (WBIR) - this being the first time that WBIR happens at MICCAI - to bring together methodological and application oriented people from research and industry across the world. Second, Learn2Reg 2024 will be a meta-challenge that hosts three new sub-challenges proposed from within the biomedical image registration community. This will include ReMIND a multimodal intra-operative registration task, LUMIR a new large-scale unsupervised whole brain alignment task and COMULIS a biomedical application with very high resolution scans.
Throughout the last years the advent of deep learning has led to a shift of the research focus of MICCAI to segmentation and classification challenges, which does not fully reflect the clinical impact of image registration. Based on our previous Learn2Reg events we believe there is an inherent unsolved problem that introduces domain specific difficulties in adapting and advancing learning techniques for the ill-posed registration problem. There is also weaker connection across research groups that work on complementary registration problems or methodological concepts. Bringing WBIR to MICCAI, joining up WBIR and Learn2Reg and enabling community-driven challenges as part of our 2024 meta-challenge will be immensely beneficial for the registration and general MICCAI community.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

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

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

We plan to closely liaise with the 11th Workshop on Biomedical Image Registration (WBIR) and would ideally be assigned consecutive time slots in the same room (morning WBIR, afternoon Learn2Reg). That way researchers interested in both theory/methodological advancement and practical evaluation can benefit from the other event. The poster sessions of both events will be held jointly to further foster discussion. Participants of Learn2Reg are encouraged to submit a research paper or shorter abstract describing new methods used within their challenge submission to WBIR and WBIR presenters are welcome to join the evaluation session and contribute solutions to one of the subtasks. The panel discussion will also stimulate a gathering of researchers to exchange ideas and opinions about how to best bring the field forward both through technical innovation and by helping to mature methods into clinical practice.

The Learn2Reg challenge is planned to last from 1:30 to 6pm (4.5 hours including 15 minutes break). This includes oral and poster presentation of participants, validation focused paper presentations, details about metrics/evaluation and challenge results and a panel discussion. WBIR is scheduled from 8am-12:30 and includes several oral and poster presentation as well as a clinically focussed keynote speaker.
8:00 - 8:15 Welcome and Introduction
8:15 - 9:00 Keynote - (Radiological challenges in MSK) - Dr Amanda Isaac (TBC)
9:00 - 9:45 Oral Session 1
9:45 - 10:00 Break
10:00 - 11:00 Oral Session 2
11:00 - 11:30 WBIR poster 3 min presentations
11:30-12:30 WBIR and Learn2reg poster session
12:30-1:30 Lunch Break
1.30 - 2.00 Learn2Reg Task Introduction
2:00 - 3:00 Validation focused presentations
3:00 - 3:15 Break
3:15 - 4:00 Learn2reg and WBIR poster session.
4.00 - 5.00 Oral Learn2Reg participants and results.
5:00 - 6:00 Panel discussion - TBD

Duration

How long does the challenge take?

Half 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.

While previous Learn2Reg events had rather smaller number of participants (15-25, mainly the teams that submitted to the challenge) we expect a substantially larger number of 40-50 for 2024 due to the co-organisation with WBIR.
As reference: please see the 2023 Learn2Reg summary paper (https://doi.org/10.1109/TMI.2022.3213983) which had more than 50 contributors.

Publication and future plans

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

We encourage all participants to submit technical contributions involving the use of challenge data as full papers to WBIR. In addition we will enable a short paper challenge track that will also undergo peer-review and be published as joint proceedings with WBIR in LNCS. We welcome in particular papers that propose clinically useful evaluation criteria, new concepts for supervision of learning based registration and unsolved datasets or applications in the field of biomedical image registration.
Going forward we are confident that our concept of transferring Learn2Reg into a community-driven meta challenge will make it more inclusive and continue to establish relevant benchmarks for the registration community in future years.

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.

To ensure smooth transition between WBIR and Learn2Reg, despite the fact that both events can be attended separately, we would like to request the same room with a capacity of around 60 people. There is no online challenge part, all computations and rankings are evaluated offline (partly on grand-challenge.org) beforehand, so no specific hardware (apart from a projector, computer, loud speaker and microphones for presenters) is required.

TASK 1: ReMIND

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.

Surgical resection is the critical first step for treating most brain tumors, and the extent of resection is the major modifiable determinant of patient outcome. To optimize outcomes, surgeons must avoid causing neurologic deficits while leaving as little residual tumor as possible. To reach these competing objectives, intraoperative ultrasound (iUS) has raised attention as it is an affordable, real-time, and intraoperative imaging technology that can be easily integrated into existing surgical workflows, especially when compared with intraoperative Magnetic Resonance Imaging (MRI). In parallel, neuronavigation has helped considerably in providing intraoperative guidance to surgeons, allowing them to visualize the location of their surgical instruments relative to the tumor and critical brain structures visible in preoperative MRI. However, the utility of neuronavigation decreases as surgery progresses due to brain shift, which is caused by brain deformation and tissue resection during surgery, leaving surgeons without guidance.

The goal of the ReMIND-Learn2Reg challenge task is to register multi-parametric pre-operative MRI and intra-operative 3D ultrasound images. Specifically, we focus on the challenging problem of pre-operative to post-resection registration, requiring the estimation of large deformations and tissue resections. Preoperative MRI comprises two structural MRI sequences: contrast-enhanced T1-weighted (ceT1) and native T2-weighted (T2). However, not all sequences will be available for all cases. For this reason, developed methods must have the flexibility to leverage incomplete sets of data at inference time. To tackle this challenging registration task, we provide a large non-annotated training set (N=95). Evaluation will be performed using manual landmarks on a private test set.

Keywords

List the primary keywords that characterize the task.

Registration, deformable, brain, oncology, multimodal, realtime

ORGANIZATION

Organizers

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

Reuben Dorent (Harvard Medical School, Boston, USA)
Tina Kapur (Harvard Medical School, Boston, USA)
Sandy Wells (Harvard Medical School, Boston, USA)
Alexandra Golby (Harvard Medical School, Boston, USA)
Wiebke Heyer (University of Lübeck, Lübeck, Germany)

b) Provide information on the primary contact person.

Reuben Dorent (rdorent@bwh.harvard.edu)

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.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

We aim to provide continuous evaluation after the challenge ends.

Challenge venue and platform

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

WBIR (Workshop on Biomedical Image Registration) @ MICCAI

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

grand-challenge.org

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

https://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.

Private data is allowed but must be reported.

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.

May participate but not eligible for awards and not listed in leaderboard. (all tasks)

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

Challenge organisers may submit algorithms as baselines but cannot win prizes. Up to three (different) teams can win awards.

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.

- The ranking of winning teams will be publicly available on the challenge website.
- Participating teams can ask to be removed from the leaderboard.
- 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. Organizers reserve the right to remove lower-scoring duplicate submissions from the same team of algorithms that are deemed too similar
- All results will be announced publicly unless an obvious error was made in data processing
- The challenge results will be summarized in a joint manuscript
(all tasks) to which usually up to two authors per contributing team will be invited as co-authors.

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).

There will be joint Learn2Reg and WBIR LNCS proceedings for challenge submissions and technical contributions. We aim to publish a joint summary journal paper on Learn2Reg 2024, but each subtask may also choose to publish an independent paper on specific aspects of the datasets and results. Usually up to two co-authors can qualify per contributing team. Participants are free to publish papers with their own findings somewhere else.
(all tasks)

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.

The validation and test submission processes are different:

- Validation leaderboard: The participating teams must submit their predictions (displacement fields) as a single compressed zip file via the grand-challenge.org interface. The tree directory structure will be described once the challenge is online.

- Test leaderboard: the participating teams are required to submit a short paper describing their method and a Docker container containing pre-trained models and training code. We encourage participants to only submit containers that can be run in a rootless mode to increase the usability of the submitted Docker containers on computational servers.

This choice is a trade-off between reducing the participant's workload and limiting the risk of cheating for the final ranking.

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.

For the avoidance of doubt pre-evaluation is limited to subsets of the data that is not used to compute test results, hence no strict limit is imposed on checking the evaluation on training data. (all tasks)
Participating teams may continuously upload displacement fields during the validation phase.
Trainings and validation split available.

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

- Registration period: March 2024
- Release date of the training set: March 2024
- Release date of the validation set: March 2024
- Start of the validation phase: April 2024
- Start of the evaluation phase: August 2024
- End of the evaluation phase: mid-August 2024 (short description of the proposed technique, results on the validation set)
- MICCAI 2024: presentation of results

(all tasks)

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).

Collection, analysis, and release of the ReMIND database have been performed in compliance with all relevant ethical regulations. The Institutional Review Board at the Brigham and Women's Hospital approved the protocol (2002-P-001238), and informed consent was obtained from all participants, including for public sharing of 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 4.0

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

Decision support, Intervention assistance, Diagnosis, Research.

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.

A variety of patient cohorts can benefit from improved medical image registration between pre-operative MR data and intraoperative ultrasound images. It especially includes brain, spine, prostate, and liver image-guided interventions.

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

The challenge cohort consists of subjects who are surgically treated with image-guided tumor resection and for whom pre-operative MRI and 3D intra-operative ultrasound data have been acquired.

Imaging modality(ies)

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

Intraoperative Ultrasound (iUS), 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).

The images will be released in compressed Nifti (.nii.gz) files. The image orientation and the voxel size can be found in these files.
As the registration task is unsupervised, no annotations are given.

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

No further information is given.

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 was acquired after substantial tumor resection was completed to the degree that either the surgeon was satisfied with the microscopically visible extent of resection or to identify the remaining portion of the tumor.

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 brain tumor 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).

MR data: These scans were acquired using 3T scanners from various manufacturers (Siemens, GE).

iUS data: All iUS series were acquired using a sterilizable 2D neuro-cranial curvilinear transducer on a cart-based ultrasound system (N13C5, BK5000, GE Healthcare, Peabody, MA, USA) in the AMIGO suite. The "Ultrasound Navigation Adapter Array" together with the "Ultrasound Navigation Adapter Base - BK N13C5" (Brainlab AG, Munich, Germany) were attached to the iUS probe to enable the Curve platform to track the probe relative to the patient. This enabled the reconstruction of a 3D volume from the tracked 2D sweeps using the "Ultrasound" module within the "Elements" software platform on the "Curve" hardware system.
More details can be found in our data paper [23]

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

The MR data include different 3D sequences (contrast-enhanced T1 and T2) scans. Not all sequences are available for all patients. Contrast-enhanced T1 images were acquired using MP-RAGE sequences from different manufacturers. T2 images were acquired using the SPACE protocol with Siemens scanners (3T).

The ultrasound probe had a contact area of 29mm x 10mm and a frequency range of 5-13 MHz. The imaging plane was chosen to be as parallel as possible to one of the three cardinal axes of the head (axial, sagittal, coronal). However, this was often limited by the size and shape of the craniotomy. The transducer was swept unidirectionally at a slow, consistent speed through the craniotomy. As mentioned above, 3D reconstruction was performed using the built-in proprietary reconstruction method in the Brainlab system.

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.

MR data: multi-institutional (USA)
Ultrasound data: Brigham and Women's Hospital, Boston, MA, USA

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 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) (parameter 21) and may include context information (parameter 18).

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

Training: 95 cases with post-resection 3D iUS and the following combinations of MR images:
- ceT1+T2: 55 cases
- ceT1: 35 cases
- T2: 5 cases

Validation: 5 cases with post-resection 3D iUS and both (ceT1 and T2) pre-operative MR images.

Consequently, a total of 15 predictions will be evaluated on the validation set: 5 predictions using both sequences as input, 5 using only ceT1, and 5 using T2 as input.

Test: we expect to be able to annotate around 30 cases with post-resection iUS. We will select to match the distribution of the MR data in the training and validation sets, i.e.:
- ceT1 + T2: 18 cases
- ceT1: 10 cases
- T2: 2 cases

Consequently, a total of 56 predictions will be evaluated on the test set: 18 predictions using both sequences as input, 28 predictions using only ceT1, and 10 predictions using only T2.

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

The training and validation sets correspond to a subset of the ReMIND dataset available on TCIA, where both pre-dura and post-resection ultrasound images were acquired. Moreover, only 3D MR data is used in this challenge.

The testing set corresponds to the consecutive cases that were surgically treated at the same institution.

Providing corresponding landmarks in multimodal scans is very time consuming and only possible for domain experts. We are confident that we can reach at least 5 validation cases and 30 test cases.

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.

We do not expect any difference between the training, validation and test cases. They correspond to consecutive 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.

First, salient features will be automatically detected in the ultrasound volumes before dural opening (data not included in this challenge) and after resection.
The features identified in the pre-dura images will then be transferred to the pre-operative MR space using affine image registration, which is a solved problem given the absence of deformation.

Subsequently, two raters will manually identify approximately 15 matches for each pair of data. Anatomical landmarks eligible for matching include deep grooves and corners of sulci, convex points of gyri, and vanishing points of sulci. In cases where raters disagree on matches (one accepts, and the other does not), a consensus will be reached through a joint investigation.

Finally, a clinical neurosurgery fellow will oversee and ensure the quality of all the matches.

Note that this protocol has been successfully tested on three cases at the time of the proposal submission. In total, 3 humans will be involved in the annotation process.

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.

First, salient features will be automatically detected in the ultrasound volumes before dural opening (data not included in this challenge) and after resection.

The features identified in the pre-dura images will then be transferred to the pre-operative MR space using affine image registration, which is a solved problem given the absence of deformation.

Subsequently, two raters will manually identify approximately 15 matches for each pair of data. Anatomical landmarks eligible for matching include deep grooves and corners of sulci, convex points of gyri, and vanishing points of sulci. In cases where raters disagree on matches (one accepts, and the other does not), a consensus will be reached through a joint investigation.

Finally, a clinical neurosurgery fellow will oversee and ensure the quality of all the matches.

Note that this protocol has been successfully tested on three cases at the time of the proposal submission.

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 who have +5 years of experience in brain anatomy. Final landmarks will be reviewed by a clinical neurosurgery fellow.

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.

A clinical fellow will control the quality of the annotations. If corrections are needed, the medical experts will either modify the landmarks if it is possible or discard the landmarks.

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 the same voxel resolution (0.5x0.5x0.5mm) and spatial dimension will be performed to ease the use of learning-based algorithms for participants with little prior experience in image registration. Pre-operative MR images will be co-registered using NiftyReg.

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.

The inter-rater variability between pre-dura and post-resection was found to be equal to 1.89 +/- 0.37 mm in a previous study using a similar annotating technique. As we now add an extra step (transferring to the MR data), this inter-rater variability may be higher.

We will estimate the inter-rater variability by presenting features in the post-resection volume. Each expert will then be asked to manually locate the correspondence for post-resection feature in the MR data.

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) DSC (Dice similarity coefficient) of segmentations
2) Robustness: 30% lowest DSC of all cases
3) Percentage of non-positive Jacobian determinant [20]
4) Percentage of non-diffeomorphic volume [20]
5) Runtime: computation time (only awarded when inference scripts provided)

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

(all tasks) DSC or TRE respectively measure accuracy; in addition to manual landmark alignment errors (TRE) it is common practice in image registration that Dice scores are used to evaluate the alignment manual segmentations using displacement fields. Surface distances and in particular the 95th percentile of the Haussdorff distance HD95 measures reliability; Outliers are penalized with the robustness score (30% of lowest DSC or 30% of highest TRE). That means the 30% of instances with lowest scores (highest TRE) **before** registration are selected as a subset to compute DSC and TRE of all participants solutions for these same instances. The smoothness of transformations (SD of log Jacobian determinant) are important in registration [21,22].
We will implement the Jacobian determinant computation using a new computation that is more robust following:

Liu, Yihao, et al. "On finite difference jacobian computation in deformable image registration." (2024). International Journal of Computer Vision (in press).
-DSC/HD95: The deformation fields generated by the methods of challenge participants will be applied to deform the label maps of the moving images. Subsequently, the DSC/HD95 metrics will be calculated by comparing the deformed label maps with the label maps of the fixed images, with the results being averaged over all anatomical structures and subjects.
-TRE: The deformation fields generated by the methods of challenge participants will be applied to displace the fiducial markers placed on the moving images. Subsequently, the TRE will be calculated by comparing the displaced fiducial markers with the fiducial markers in the fixed images, with the results being averaged over all markers and subjects.
-Deformation regularity: The smoothness and regularity of the deformation fields produced by the challenge participants' methods will be assessed and averaged over all subjects.
Run-time computation time is relevant for clinical applications.

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.

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. We will compute two separate rankings with and without including runtime, the one without will be considered for awards.

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.

All metrics but robustness-based metrics 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).

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.

The ranking scheme is in priniciple based on the ranking scheme of the Medical Decathlon. We rank methods using statistically significantly different results. For each metric applied in a task, methods are compared against each other (Wilcoxon signed rank test with $p<0.05$, see details in code [https://github.com/MDL-UzL/L2R/tree/main/ranking]), ranked based on the number of "won" comparisons and finally mapped to a numerical metric rank score between 0.1 and 1 (with possible score sharing).
A task rank score is then obtained as the geometric mean of individual metric rank scores. All methods for which no metric is available (not submitted to the task, no Docker container submitted) share the lowest possible metric rank score of 0.1.
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:
- NiftyReg/Deeds

TASK 2: LUMIR (Large-scale Unsupervised Whole Brain MRI Image Registration)

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.

Subcortical and deep brain structures are involved in many cognitive, memory, and emotional tasks. The hippocampus, e.g., is known to be involved in aging, memory, and spatial navigation. Regional analysis of whole brain MRI has also been identified as a key indicator for pathophysiology of Alzheimer's disease, schizophrenia, and epilepsy. The challenge in this task is the alignment of small structures of variable shape and size with high precision on mono-modal MRI images between different patients.

The previous Learn2Reg brain MRI challenge inspired our scientific community to address brain registration problems using learning-based methods. However, it revealed that weakly-supervised brain segmentation (training assisted by anatomical label maps) can lead to a bias toward the label maps. Specifically, a registration network trained solely on label maps, without incorporating image similarity measures and deformation regularizers, can achieve high Dice scores. Yet, such an approach often results in non-smooth and unrealistic deformations. In response to these findings, this year's challenge is pivoting towards unsupervised learning of image registration, omitting the use of label maps during training. This shift not only mitigates the aforementioned bias but also provides a basis for a more equitable comparison with traditional methods. Our aim is to establish the current status quo of deep learning's performance relative to optimization-based methods in the longstanding task of brain MRI registration.

To this end, a comprehensive dataset exceeding 4,000 neuroimaging datasets from the OpenBHB and AFIDs-OASIS datasets, which is a first of its scale in public registration challenges, has been employed. This dataset has been split into 3384, 110, and 520 images for training, validation, and testing. We computed segmentation maps involving 133 cortical and subcortical labels with an established segmentation method, SLANT. Additionally, 30 images from the AFIDs-OASIS dataset, featuring manually placed fiducial markers, are used for validation and testing to allow for more accurate registration accuracy quantification. This large-scale dataset marks a significant step towards developing a foundational model for brain image registration. For Learn2Reg, we will provide participants with the preprocessed images for 3,494 subjects (training and validation), keeping the label maps and fiducial markers confidential for evaluation purposes only.

Keywords

List the primary keywords that characterize the task.

Registration, deformable, brain, unsupervised, inter-subject, large-scale

ORGANIZATION

Organizers

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

Junyu Chen (Johns Hopkins,Maryland,USA)
Yihao Liu (Johns Hopkins,Maryland,USA)
Mattias Heinrich (University of Lübeck,Lübeck,Germany)

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.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

We aim to provide continuous evaluation after the challenge ends.

Challenge venue and platform

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

WBIR (Workshop on Biomedical Image Registration) @ MICCAI

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

grand-challenge.org

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

https://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.

Private data is allowed but must be reported.

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.

Challenge organisers may submit algorithms as baselines but cannot win prizes. Up to three (different) teams can win awards.

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).

There will be joint Learn2Reg and WBIR LNCS proceedings for challenge submissions and technical contributions. We aim to publish a joint summary journal paper on Learn2Reg 2024, but each subtask may also choose to publish an independent paper on specific aspects of the datasets and results. Usually up to two co-authors can qualify per contributing team. Participants are free to publish papers with their own findings somewhere else. (all tasks)

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.

see above, the validation phase asks for a displacement field, while the test phase asks for a algorithm submission + short paper that includes all necessary details on training/hyperparameter tuning and inference to reproduce the method.

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 algorithm output to be evaluated on the validation data of phase 1 in order to avoid any pitfalls (wrong orientation, etc.). We also provide our evaluation scripts to test them on the training data.

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

- Registration period: March 2024
- Release date of the training set: March 2024
- Release date of the validation set: March 2024
- Start of the validation phase: April 2024
- Start of the evaluation phase: August 2024
- End of the evaluation phase: mid-August 2024 (short description of the proposed technique, results on the validation set)
- MICCAI 2024: presentation of results

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

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

Research, Prognosis, Longitudinal Study, Diagnosis

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.

Neuroimaging registration is an important part of neuroscience and clinical treatment and planning for a wide range of diseases and procedures, such as neurodegenerative diseases (AD, schizophrenia, epilepsy) or tumor resections.

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

The dataset comprises T1-weighted MRIs sourced from the OpenBHB project [2] and the AFIDs-OASIS dataset project [3]. We utilized a specific portion of the OpenBHB dataset, encompassing a compilation of 10 publicly accessible datasets. This subset exclusively includes healthy control subjects, both male and female, with an average age of 24.9 +/- 14.3 years. In contrast, the AFIDs-OASIS dataset features MRI scans from 30 female subjects, averaging 58 +/- 17.9 years. The selected MRI scans from AFIDs-OASIS were intentionally chosen for their complexity (including regions with intricate anatomy and asymmetries), making them ideal for testing the robustness of deep neural networks.

Imaging modality(ies)

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

Magnetic Resonance Imaging (MRI) - T1 weighted Brain 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).

The OpenBHB dataset is composed of 3,984 T1-weighted MRI scans of healthy brains, collected from various centers (>60 centers worldwide). These brain MRIs underwent preprocessing using FreeSurfer for both pre-affine registration and skull-stripping. Additional preprocessing steps included intensity normalization and automatic segmentation to generate label maps for 133 anatomical structures. The dataset was divided into three parts:

3,384 for training, 100 for validation, and 500 for testing. For evaluation purposes, 10 images in the validation set and 40 images in the testing set will feature manual fiducial markers, annotated by an experienced radiologist and neurologist.
Additionally, we incorporated the AFIDs-OASIS brain dataset, primarily for validation and testing. It consists of 30 T1-weighted brain MRI scans, each annotated with manually placed fiducial markers. To align with the OpenBHB dataset, we performed similar intensity normalization, skull-stripping, and pre-affine registration on these scans. Additionally, we will conduct automatic segmentation on this dataset to generate label maps for 133 anatomical structures. The AFIDs-OASIS dataset is split into two groups: 10 scans for validation and 20 for testing.
For the Learn2Reg challenge, we will provide only the pre-processed images and exclude the label maps and fiducial markers. This aligns with our focus on unsupervised image registration.

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

The OpenBHB dataset comprises healthy control subjects with an average age of 24.9 +- 14.3 years.
The AFIDs-OASIS dataset includes cognitively intact subjects averaging 58 +- 17.9 years in age.

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.

Whole brain MRI showing important neuroanatomical regions: cortical, subcortical and deep brain structures.

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 subcortical and deep brain anatomy across patients within the same modality (MRI).

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, Complexity, 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).

Detailed information about the acquisition protocols and the specific scans utilized is available in the official documents of each contributing dataset:
- ABIDE 1 [6]
- ABIDE 2 [7]
- CoRR [8]
- GSP [9]
- IXI [10]
- Localizer [11]
- MPI-Leipzig [12]
- NAR [13]
- NPC [14]
- RBP [15, 16]

The AFIDs-OASIS dataset [3], which is a subset of the OASIS-1 dataset [17], can be accessed at https://github.com/afids/AFIDs-OASIS. Comprehensive details about this dataset are available in the OASIS-1 reference.

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

Data is either acquired or reshaped to 1x1x1 mm isotropic resolution of T1w-weighted MRI whole brain scans.

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 OpenBHB dataset [2], accessible at https://baobablab.github.io/bhb/dataset, features image data compiled from 10 publicly available datasets, originating from over 60 centers globally. Detailed information about the acquisition protocols and the specific scans utilized is available in the official documents of each contributing dataset:
- ABIDE 1 [6]
- ABIDE 2 [7]
- CoRR [8]
- GSP [9]
- IXI [10]
- Localizer [11]
- MPI-Leipzig [12]
- NAR [13]
- NPC [14]
- RBP [15, 16]

The AFIDs-OASIS dataset [3], which is a subset of the OASIS-1 dataset [17], can be accessed at https://github.com/afids/AFIDs-OASIS. Comprehensive details about this dataset are available in the OASIS-1 reference.

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 MR pairs from two different patients. The images in the OpenBHB dataset are annotated with segmentations according to the same protocol. Meanwhile, the images in the AFIDs-OASIS dataset feature fiducial markers that are placed manually.

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

Training: 3384 (OpenBHB - images only)
Validation: 90 (OpenBHB - segmentation only) + 10 (OpenBHB - segmentation & fiducial markers) + 10 (AFIDs-OASIS - segmentation & fiducial markers)
Testing: 460 (OpenBHB - segmentation only) + 40 (OpenBHB - segmentation & fiducial markers) + 20 (AFIDs-OASIS - segmentation & fiducial markers)

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

A substantial quantity of images from OpenBHB is used to effectively train and assess the deep neural networks designed for registration. Conversely, AFIDs-OASIS images, which are manually annotated with fiducial markers, are employed solely for validation and evaluation due to their limited availability. Since all images are subjected to the same preprocessing steps common in neuroimaging, we anticipate small dataset bias between OpenBHB and AFIDs-OASIS. However, to further reduce any potential dataset bias in testing, given that AFIDs-OASIS images are not included in the training phase, we will manually annotate a subset of approximately 50 images from the OpenBHB dataset with fiducial markers. This annotation will follow the same procedures used for the AFIDs-OASIS dataset, ensuring a consistent evaluation framework for the registration methods

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 processed, we will announce around 700 randomly selected pairs of test subjects that should be registered for evaluation.

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.

For the AFIDs-OASIS dataset, each image underwent annotation by one expert and two novice raters, assigned at random [2]. The average locations of fiducial markers placed by all raters were established as the ground truth. Additionally, we used automatic segmentation on the AFIDs-OASIS dataset to delineate 133 anatomical structures (including subcortical and deep brain areas) using SLANT software [1]. Similarly, the OpenBHB dataset underwent automatic segmentation to identify the same 133 anatomical structures using SLANT [1]. In addition, Dr. Harrison Bai, an experienced radiologist from Johns Hopkins and one of the challenge organisers, along with a neurologist will manually annotate fiducial markers on approximately 50 images from the OpenBHB dataset for validation and testing purposes. This annotation process will adhere to the methodology previously described in [2]. For the Learn2Reg challenge, the pre-processed images for the 3464 training and validation subjects will be distributed, but the segmentation labels and fiducial markers will not be shared.

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.

The AFIDs-OASIS dataset was annotated as part of an pre-existing, independent study, and the comprehensive protocol can be found at this link: https://www.nature.com/articles/s41597-023-02330-9.
The manual annotation process for the OpenBHB dataset will similarly adhere to the methodology outlined in this document.

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.

SLANT [1] is a widely recognized and established method for automatic image segmentation in neuroimaging. The AFIDs-OASIS dataset [2] was annotated by medical experts with extensive experience. Dr. Harrison Bai, an experienced radiologist with over nine years of experience in the field of radiology and more than ten years in medicine, currently serving as an Associate Professor at Johns Hopkins, will undertake the manual annotation for the subset of the OpenBHB dataset in collaboration with an experienced neurologist.

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.

To merge multiple annotations, an average was calculated from the fiducial markers placed by different raters.

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.

Standard neuroimaging pre-processing pipeline was used, including procedures like resampling to uniform voxel resolution, intensity normalization, skull-stripping, ensuring consistent spatial dimensions, and providing affine pre-registration. These steps are intended to simplify the application of learning-based algorithms for participants who have limited 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.

Previous research has demonstrated that the automatic segmentation tool SLANT [13] attains over 75% Dice overlap for all 133 structures, though this varies considerably among them. Note that SLANT significantly surpasses many of the traditional state-of-the-art methods of its time in terms of both speed and accuracy. However, it faces challenges with certain small structures, such as the inferior horn of the lateral ventricle and the cortex, where there is often inadequate resolution or contrast.

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) DSC (Dice similarity coefficient) of segmentations
2) HD95 (95% percentile of Haussdorff distance) of segmentations
3) Robustness: 30% lowest DSC/TRE of all cases
4) Percentage of non-positive Jacobian determinant [20]
5) Percentage of non-diffeomorphic volume [20]
6) TRE of manually fiducial markers

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 penalized with the robustness score (30% of lowest DSC or 30% of highest TRE). Furthermore, the smoothness of transformations, indicated by the percentage of non-positive Jacobian determinant and non-diffeomorphic volume, plays a crucial role in registration. Recent advancements [20] have improved the handling of finite approximations in existing Jacobian determinant computations, mitigating related issues.

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.

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. We will compute two separate rankings with and without including runtime, the one without will be considered for awards.

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.

All metrics but robustness-based metrics 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).

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.

The ranking scheme is in priniciple based on the ranking scheme of the Medical Decathlon. We rank methods using statistically significantly different results. For each metric applied in a task, methods are compared against each other (Wilcoxon signed rank test with $p<0.05$, see details in code [https://github.com/MDL-UzL/L2R/tree/main/ranking]), ranked based on the number of "won" comparisons and finally mapped to a numerical metric rank score between 0.1 and 1 (with possible score sharing).
A task rank score is then obtained as the geometric mean of individual metric rank scores. All methods for which no metric is available (not submitted to the task, no Docker container submitted) share the lowest possible metric rank score of 0.1.
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 (IEEE TMI'19)
- TransMorph (MedIA'22)
- Im2Grid (MMML'22)
- ANTs, NiftyReg, Deeds

TASK 3: COMULISglobe SHG-BF

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.

Second-harmonic generation (SHG) microscopy is a non-invasive imaging technique that does not require the use of exogenous labels. This is particularly beneficial for studying live tissues, allowing for real-time observations without introducing artefacts or potential toxicity associated with staining agents. In addition, due to its nonlinear excitation (multi-photons), SHG signals exhibit minimal scattering and absorption, enabling deep tissue penetration. This is crucial for imaging thick biological specimens and makes SHG microscopy suitable for visualising collagen in various tissues, including skin, cartilage, and blood vessels.

The use of second-harmonic generation (SHG) microscopy in biomedical research is rapidly increasing. This is largely due to the interest of using this imaging technique as a means to examine the role of fibrillar collagen organisation in diseases such as cancer. Collagen is the major component of the tumour microenvironment and participates in cancer fibrosis and targeting tumour collagen may be a way to provide therapeutic efficacy.

However, SHG imaging only gives partial information, whereby co-examination of SHG images and traditional bright-field (BF) images of hematoxylin and eosin (H&E;) stained tissue is usually required. H&E; staining enables differentiation of tissue components while SHG imaging is particularly sensitive to the collagen fibres. The combination of the two facilitates interpretation of the role of collagen fibres in the tissue microenvironment, of high importance for .

Due to the large information difference between BF and SHG images, existing registration methods fail to readily register the images from these two modalities, while manual image registration is difficult, labour intensive and error prone. Reliable and fast automated image registration would greatly facilitate correlative analysis of these two modalities.

Several virtual staining methods have also demonstrated their potential in research use. Reliable automated registration would allow to construct wider databases of paired SHG-BF staining images in order to be able to learn to generate (predict) equivalent stained image (H&E;) from the stain free image (SHG), thereby avoiding several negative consequences of staining (including labour, toxicity, stain variations). The combination and evaluation against conventional (tedious and often subjective) H&E; histology can pave the way for SHG as a fast label-free optical diagnostic tool to assess a variety of biomedical diseases in cancer, fibrosis, or wound healing (which all involve collagen remodelling or degradation).

Keywords

List the primary keywords that characterize the task.

Registration, cancer, histopathology, multimodal, tissue microarray, bright-field, second-harmonic generation

ORGANIZATION

Organizers

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

Andreas Walter (Aalen University, Aalen, Germany)
Joakim Lindblad (Uppsala University, Uppsala, Sweden)
Natasa Slodaje (Uppsala University, Uppsala, Sweden)
Perrine Paul-Gilloteaux (Nantes Université, Nantes, France)
Lasse Hansen (echoscout.ai, Lübeck, Germany)

b) Provide information on the primary contact person.

Lasse Hansen (lasse@echoscout.ai, Lübeck, Germany)

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.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

We aim to provide continuous evaluation after the challenge ends.

Challenge venue and platform

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

WBIR (Workshop on Biomedical Image Registration) @ MICCAI

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

grand-challenge.org

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

https://learn2reg.grand-challenge.org/

Participation policies

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

Only (semi-) automatic methods allowed.

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.

Private data is allowed but must be reported.

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.

Challenge organisers may submit algorithms as baselines but cannot win prizes. Up to three (different) teams can win awards.

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).

There will be joint Learn2Reg and WBIR LNCS proceedings for challenge submissions and technical contributions. We aim to publish a joint summary journal paper on Learn2Reg 2024, but each subtask may also choose to publish an independent paper on specific aspects of the datasets and results. Usually up to two co-authors can qualify per contributing team. Participants are free to publish papers with their own findings somewhere else.
A joint evaluation publication for Task 3 and 4 is planned after the challenge ends.

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.

see above, the validation phase asks for a displacement field, while the test phase asks for a algorithm submission + short paper that includes all necessary details on training/hyperparameter tuning and inference to reproduce the method.

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.

Participating teams may continuously upload displacement fields during the validation phase. We also provide our evaluation scripts to test them on the training data.

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

- Registration period: March 2024
- Release date of the training set: March 2024
- Release date of the validation set: March 2024
- Start of the validation phase: April 2024
- Start of the evaluation phase: August 2024
- End of the evaluation phase: mid-August 2024 (short description of the proposed technique, results on the validation set)
- MICCAI 2024: presentation of results

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).

The data is collected in a study approved by the University of Wisconsin Institutional Review Board (https://irb.wisc.edu/).

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.

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. Only teams that made their code publicly available will be associated to the joint publication.

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

Diagnosis, Prognosis, (Cancer) Research, Treatment planning, Education

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 suspected of having diseases such as cancer, fibrosis, and connective tissue disorders, which all typically cause collagen structure alterations. Fibrillar collagen topology and organization are important indicators of disease progression and prognostication. The co-examination of SHG images and traditional bright-field (BF) images of hematoxylin and eosin (H&E;) stained tissue, as a gold standard clinical validation, is usually required.

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

Women at high risk for the development of breast cancer, as well as patients with pancreatic tumors who underwent resection surgery with curative intent.

Imaging modality(ies)

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

Second Harmonic Generation (SHG) and H&E; stained Bright Field (BF) Microscopy imaging of human breast and pancreatic cancer tissue.

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 Histological dataset comprises 206 aligned Second Harmonic Generation (SHG) and Bright-Field (BF) tissue microarray (TMA) image pairs of size 2048 × 2048 px.

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

The Histological dataset comprises 206 aligned Second Harmonic Generation (SHG) and Bright-Field (BF) tissue microarray (TMA) image pairs of size 2048 × 2048 px.

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 target is any tissue where collagen structure alterations may occur, shown in BF and SHG data.

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.

Given a BF image of a H&E; stained tissue and an SHG image of the collagen in the same tissue section, the goal is to find a transformation that maps the source image to the target image, which allows to harness the complementary information from both modalities.

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, 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).

Bright-field imaging of the pancreatic samples: Aperio CS2 Digital Pathology Scanner (Leica Biosystems) at 40x magnification. All SHG imaging and bright-field imaging of the breast samples was done with a custom built integrated SHG/bright field imaging system [5].

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

Tissues were formalin-fixed and paraffin-embedded, then cut into 5 micrometer thin slices, affixed to a slide and stained with hematoxylin and eosin (H&E;) before mounting with a coverslip. To alleviate out of focal plane issues due to the unevenness of the tissue slice, 3 z-planes were captured per SHG image and then maximum-intensity projected to capture the entire axial field of view.

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.

Images acquired at the Laboratory for Optical and Computational Instrumentation, Department of Biomedical Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA. The data providing platform is Zenodo: https://zenodo.org/records/4550300

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 represents a pre-aligned image pair of SHG and BD modality.

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

156+10(training+validation), 40 (test)

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

The dataset with the total number of 206 cases was chosen as the number is high enough to develop machine learning methods, but still manageable to attract as many participating teams as possible. A commonly used data split of 80:20 was chosen.

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.

We do not expect any difference between the training, validation and test cases as the data splits are chosen at random from a homogenous challenge cohort.

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.

Since imaging the same fixed slides, the two views acquired from the same specimen differ mainly by a rigid transformation (and small deformable deformations e.g. due to temperature changes). For each image, at least 3 landmark pairs distributed across the image are selected by an observer.

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.

Landmark annotation is following in principle https://imagej.net/plugins/name-landmarks-and-register.

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.

The observer is an expert in the field with several years of experience in biomedical imaging.

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.

Not applicable.

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 image pairs are pre-aligned and have the same size of 2048x2048 pixels.

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 3-5 pixels are expected for manual landmarks, which is planned to be confirmed by annotating a small subset of cases by a second observer.

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) Mean Target Registration Error (mTRE) of manually annotated landmarks
2) Robustness: 30% highest TRE of all cases
3) Smoothness: SD of log Jacobian determinant
4) Runtime: computation time (only awarded when inference scripts provided)

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

TRE measures accuracy; Outliers are penalised with the robustness score (30% of highest mean TRE)
The smoothness of transformations (SD of log Jacobian determinant) are important in registration [21,22].
Run-time computation time is relevant for clinical applications.

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.

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. We will compute two separate rankings with and without including runtime, the one without will be considered for awards.

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.

All metrics but robustness-based metrics 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.

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.

The ranking scheme is in priniciple based on the ranking scheme of the Medical Decathlon. We rank methods using statistically significantly different results. For each metric applied in a task, methods are compared against each other (Wilcoxon signed rank test with $p<0.05$, see details in code [https://github.com/MDL-UzL/L2R/tree/main/ranking]), ranked based on the number of "won" comparisons and

finally mapped to a numerical metric rank score between 0.1 and 1 (with possible score sharing).
A task rank score is then obtained as the geometric mean of individual metric rank scores. All methods for which no metric is available (not submitted to the task, no Docker container submitted) share the lowest possible metric rank score of 0.1.
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:
- NiftyReg/Elastix/greedy
- Corrfield
- VoxelMorph

TASK 4: COMULISglobe 3D-CLEM

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.

Correlative light and electron microscopy (CLEM) is a workflow that enables researchers to add functional information about the identity of a cell or organelle to the ultrastructure of that cell or organelle revealed by electron microscopy. There are many different CLEM workflows that combine different light and electron microscopy modalities, each custom-designed for the specific research question.
In this challenge, we focus on a pre-embedding volume CLEM workflow, in which the fluorescence microscopy is performed after primary fixation with formaldehyde and before further processing into resin for volume electron microscopy (vEM). This workflow delivers two datasets from the same sample, one from fluorescence microscopy and one from electron microscopy. The benchmark datasets provided for the challenge were acquired from a monolayer of HeLa cells grown on glass coverslips. The first dataset was acquired using fluorescence microscopy (Zeiss Airyscan LSM900 with fluorescent markers highlighting nucleus, mitochondria, lysosomes, endosomes/golgi and plasma membrane, depending on the experiment). The second dataset was acquired using vEM, specifically focused ion beam scanning electron microscopy (FIB-SEM; Zeiss Crossbeam 540).
This challenge addresses the problem of aligning the two datasets, which is difficult because light and electron microscopy have different resolutions and contrast mechanisms, and thus contain different information about the sample. It is also a difficult problem because of the shrinkage and warping of the sample that happens during processing of the sample into resin, so that the raw light and electron images will not perfectly align.
The aim of the challenge is to automate the alignment of the two datasets so that the fluorescent signals marking the lysosomes and golgi (the target organelles) can be assigned to membranous structures in the EM data. The mitochondria and plasma membrane fluorescent markers can be used as landmarks to aid this alignment. The alignment must not be performed by first selecting structures in the EM data that look like lysosomes or golgi and then warping the light microscopy data onto these structures, since this introduces bias into the process and makes the pipeline unusable for real-world research where the target organelle structure is unknown.
The benefit of an automated alignment pipeline is that it simplifies and speeds up the process of CLEM alignment for non-experts, and it removes the bias inherent in a manual pipeline where the user manually selects matching landmarks in each imaging modality.
The impact for bioscience and clinical research is the ability to assign identity to an organelle, cell or structure and to find the subcellular localisation of fluorescently-tagged proteins in EM images. The results can localise parasites to organelles to follow membrane damage and host responses, visualise processes of cell migration across blood vessel walls during inflammation or metastasis, and track pathogenic protein aggregates through the cell in neurodegenerative diseases.

Keywords

List the primary keywords that characterize the task.

Registration, cellular analysis, multimodal, correlative light microscopy, volume electron microscopy

ORGANIZATION

Organizers

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

Andreas Walter (Aalen University, Aalen, Germany)
Joakim Lindblad (Uppsala University, Uppsala, Sweden)
Natasa Slodaje (Uppsala University, Uppsala, Sweden)
Perrine Paul-Gilloteaux (Nantes Université, Nantes, France)
Marie-Charlotte Domart (The Francis Crick Institute, London, UK)
Lucy Collinson (The Francis Crick Institute, London, UK)
Martin Jones (The Francis Crick Institute, London, UK)
Lasse Hansen (echoscout.ai, Lübeck, Germany)

b) Provide information on the primary contact person.

Lasse Hansen (lasse@echoscout.ai, Lübeck, Germany)

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.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

We aim to provide continuous evaluation after the challenge ends.

Challenge venue and platform

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

WBIR (Workshop on Biomedical Image Registration) @ MICCAI

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

grand-challenge.org

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

https://learn2reg.grand-challenge.org/

Participation policies

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

Only (semi-) automatic methods allowed.

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.

Private data is allowed but must be reported.

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.

Challenge organisers may submit algorithms as baselines but cannot win prizes. Up to three (different) teams can win awards.

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).

There will be joint Learn2Reg and WBIR LNCS proceedings for challenge submissions and technical contributions. We aim to publish a joint summary journal paper on Learn2Reg 2024, but each subtask may also choose to publish an independent paper on specific aspects of the datasets and results. Usually up to two co-authors can qualify per contributing team. Participants are free to publish papers with their own findings somewhere else.
A joint evaluation publication for Task 3 and 4 is planned after the challenge ends.

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.

see above, the validation phase asks for a displacement field, while the test phase asks for a algorithm submission + short paper that includes all necessary details on training/hyperparameter tuning and inference to reproduce the method.

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.

Participating teams may continuously upload displacement fields during the validation phase. We also provide our evaluation scripts to test them on the training data.

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

- Registration period: March 2024
- Release date of the training set: March 2024
- Release date of the validation set: March 2024
- Start of the validation phase: April 2024
- Start of the evaluation phase: August 2024
- End of the evaluation phase: mid-August 2024 (short description of the proposed technique, results on the validation set)
- MICCAI 2024: presentation of results

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).

Human cervical cancer epithelial cells (HeLa) were obtained from Cell Services at the Francis Crick Institute, whose code of conduct and ethical regulations apply.

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 (Attribution)

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. Only teams that made their code publicly available will be associated to the joint publication.

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

Research

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.

vCLEM is a general tool in bioimage analysis, used in various disciplines in the life sciences, including neuroscience, tissue research, and protein research. One of the most common applications of vCLEM is the analysis of the cellular structure and function, helping researchers to understand the cellular mechanisms underlying various biological processes.

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

Human cells grown in culture.

Imaging modality(ies)

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

Correlative light (CL) and volume electron microscopy (EL) (vCLEM).

Context information

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

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

CLEM datasets (EMPIAR-10819 and EMPIAR-11537) of human cervical cancer epithelial cells (HeLa), originating from the American Type Culture Collection (ATCC; CCL-2).

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

The line is derived from cervical cancer cells taken on 8 February 1951 from Henrietta Lacks, a 31-year-old African American mother of five. Lacks died of cancer on 4 October 1951. For further background and ethical implications see e.g. [4].

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.

Human cells grown in culture.

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.

Given a CL volume and an EM volume of the same human cell, the goal is to find a transformation that maps the source volume to the target volume, which allows to harness the complementary information from both modalities.

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, Memory Requirements.

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).

LM imaging: Zeiss Airyscan LSM900. Cells were washed twice and imaged in 0.1 M PB on an AxioObserver 7 LSM900 with Airyscan 2 microscope with Zen 3.1 software (Carl Zeiss Ltd). Cells were first mapped with a 10x objective (NA 0.3) using brightfield microscopy to determine their position on the grid and tile scans were generated. The cells of interest were then imaged at high resolution in Airyscan mode with a 63x oil objective (NA

1.4). Smart setup was used to set up the imaging conditions. A sequential scan of each channel was used to limit crosstalk and z-stacks were acquired throughout the whole volume of the cells.
EM imaging: Focused ion beam scanning electron microscope (FIB-SEM). Focused ion beam scanning electron microscopy (FIB-SEM) data was collected using a Crossbeam 540 FIB-SEM with Atlas 5 for 3D tomography acquisition (Zeiss, Cambridge). A segment of the cell monolayer containing the cell of interest was trimmed out, and the coverslip removed using liquid nitrogen prior to mounting on a standard 12.7 mm SEM stub using silver paint, and coating with a 10 nm layer of platinum. The region of interest was relocated by briefly imaging through the platinum coating at an accelerating voltage of 10 kV and correlating to previously acquired fluorescence microscopy images. On completion of preparation of the target ROI for Atlas-based milling and tracking, images were acquired at 5 nm isotropic resolution throughout the cell of interest, using a 6 microsecond dwell time. During acquisition, the SEM was operated at an accelerating voltage of 1.5 kV with 1.5 nA current. The EsB detector was used with a grid voltage of 1.2 kV. Ion beam milling was performed at an accelerating voltage of 30 kV and current of 700 pA.

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

Mitochondria (MitoTracker Deep Red), nucleus (Hoechst 33342), Golgi apparatus protein TGN46 (GFP-TGN46) and lysosomes (LysoTracker Red) in EMPIAR-10819 and cell membrane (WGA) in EMPIAR-11537 were tagged so that the unbiased registration performance could be assessed using target structures. The acquired images were preprocessed following standard EM protocols, see [Krentzel et al. 2023] for details. The sample is milled by a focused ion beam and imaged in a resin-block which provides stability.
The Airyscan data was processed in Zen software using the z-stack alignment tool to correct z-shift. The settings used were highest quality, translation and linear interpolation, with the mitotracker channel as reference.

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.

Institute: Electron Microscopy Science Technology Platform at The Francis Crick Institute.
Data platforms: EMPIAR and Bioimage Archive.
EMPIAR-10819: EM (https://www.ebi.ac.uk/empiar/EMPIAR-10819/), FM (https://www.ebi.ac.uk/biostudies/bioimages/studies/S-BSST707);
EMPIAR-11537: EM (https://www.ebi.ac.uk/empiar/EMPIAR-11537/) FM (https://www.ebi.ac.uk/biostudies/bioimages/studies/S-BSST1075).

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 represents a single patch of a volume pair of CL and EM modality.

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

We aim for a 50:50 data split with 100 (90+10 for training+validation) cases each.

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

Due to the extremely large electron microscopy volumes we will provide appropriate volume patches to the participating teams that cover the whole area of the cells. With 100 sampled patches for each cell we assume to cover enough important regions of interest. However, we may decide for a larger number of sampled patches, if first benchmarks (see item 29) show insufficient coverage.

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.

We do not expect any difference between the training, validation and test cases as the data splits are chosen at random from a homogenous challenge cohort.

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.

Target structures (lysosomes in EMPIAR-10819 and endosomes in EMPIAR-11537) were manually segmented in 3D in the EM volume by one annotator using TrakEM2 plugin in fiji.. These lysosomes and endosomes were then cropped with a bounding box from the corresponding warped FM volumes and segmented using Otsu thresholding to generate the correlating FM segmentation.
2-3 manual selected corresponding landmark pairs per validation/test case (20-30 and 200-300 landmarks in total, respectively) will be annotated by at least 2 experts in the field with at least 5 years of experience in biomedical imaging (following in principle [19])

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.

For manual segmentation, annotators can refer to TrakEM2 tutorials (https://imagej.net/plugins/trakem2/tutorials) in particular the Video tutorial on the basic recipe for TrakEM2

and Video tutorial on segmenting/outlining objects over multiple sections in 3D. The landmark annotation process follows in principle [19].

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.

Segmentations labelled by electron microscopist with >10 years of experience in vCLEM. The landmarks will be annotated by up to 2 experts in the field with at least 5 years of experience in biomedical imaging.

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.

Not applicable.

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.

LM imaging: Zeiss Airyscan LSM900. The Airyscan data was first processed in Zen software using the z-stack alignment tool to correct z-shift. The settings used were highest quality, translation and linear interpolation, with the mitotracker channel as reference. The .czi file was then opened in Fiji and saved as .tif.
EM imaging: Focused ion beam scanning electron microscope (FIB-SEM) After initial registration with template matching by normalised cross correlation in Fiji (https://sites.google.com/site/qingzongtseng/template-matching-ij-plugin), the FIB-SEM images were contrast normalised as required across the entire stack and converted to 8-bit grayscale. To fine-tune image registration, the alignment to median smoothed template method was applied. The aligned XY 5 nm FIB-SEM stack was opened in Fiji and resliced from top to bottom (i.e., to XZ orientation), and rotated to match the previously acquired Airyscan data. Finally, the EM volume was binned by a factor of 4 in X, Y and Z using Fiji resulting in an isotropic voxel-size of 20 nm.
In addition to the cell volumes, we will prepare patches for training/valiation and testing that cover all relevant regions of interest.

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.

Manual segmentation is prone to errors and is known to result in inter- and intra-annotator variability. The lengthiness of the process means that estimating the magnitude of the error is difficult in general due to the small number of annotators. But a rough estimate based on the chosen structure and the fact that an expert annotator was involved would be <10% error, mostly at the edge of the structures. For the light microscopy data, the diffraction limit and poor z resolution will be a source of error.
Based on experience inter-rater variabilities of 20 voxels are expected for manual landmarks, which is planned to be confirmed by annotating a small subset of cases by a second observer.

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) DSC (Dice similarity coefficient) of segmentations
2) HD95 (95% percentile of Haussdorff distance) of segmentations
3) Mean Target Registration Error (mTRE) of manually annotated landmarks
4) Robustness: 30% lowest DSC of all cases/30% highest TRE of all cases
5) Smoothness: SD of log Jacobian determinant
6) Runtime: computation time (only awarded when inference scripts provided)
7) Memory: GPU/RAM usage (only awarded when inference scripts provided)

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 [21,22].
Run-time computation time and memory requirement is relevant for clinical applications.

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.

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. We will compute two separate rankings with and without including runtime, the one without will be considered for awards.

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.

All metrics but robustness-based metrics 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 and memory ranks are only considered with 50% weight (since not all participants are able to use docker containers for evaluation).

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.

The ranking scheme is in priniciple based on the ranking scheme of the Medical Decathlon. We rank methods using statistically significantly different results. For each metric applied in a task, methods are compared against each other (Wilcoxon signed rank test with $p<0.05$, see details in code [https://github.com/MDL-UzL/L2R/tree/main/ranking]), ranked based on the number of "won" comparisons and finally mapped to a numerical metric rank score between 0.1 and 1 (with possible score sharing).
A task rank score is then obtained as the geometric mean of individual metric rank scores. All methods for which no metric is available (not submitted to the task, no Docker container submitted) share the lowest possible metric rank score of 0.1.
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:
- NiftyReg/Elastix/greedy
- Corrfield
- VoxelMorph

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.

As reference for challenge design and scope of previous editions of Learn2Reg: please see the 2023 Learn2Reg summary paper (https://doi.org/10.1109/TMI.2022.3213983) which had more than 50 contributors.

Further comments

Further comments from the organizers.

N/A