Towards the Automatization of Cranial Implant Design in Cranioplasty: 2nd MICCAI Challenge on Automatic Cranial Implant Design: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Towards the Automatization of Cranial Implant Design in Cranioplasty: 2nd MICCAI Challenge on Automatic Cranial Implant Design

Challenge acronym

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

AutoImplant 2021

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.

Cranioplasty is the surgical process where a skull defect resulting from previous surgery or injury is repaired using an implant that restores the original protective and aesthetic function of the skull. Implications range from decompressive craniectomies to performing brain surgery. Although the patient's autologous bone is routinely used as the implant material, it may not always be possible due to infection, fracture or bone tumour. Using autologous bone also poses significant risk of requiring a secondary surgery due to bone resorption [1]. Artificial patient-specific implants (PSI) created using computed-assisted design reduce overall patient risks as well as operating time [2]. Developing automatic skull reconstruction methods will increase the availability of PSIs as well as enable their design and subsequent manufacturing directly inside the operating room.
Prior to the first AutoImplant Challenge (AutoImplant 2020, held in conjunction with MICCAI 2020 https://autoimplant.grand-challenge.org/), automatic design of cranial implant has been a under-researched area, due to a lack of proper formulation of the problem from a technical perspective. In AutoImplant 2020, we formulated cranial implant design as a volumetric shape completion and 3D shape learning task. Based on the formulation, various data-driven approaches, such as deep learning and statistical shape model can be employed in solving the problem. AutoImplant 2021 is a substantial extension to the AutoImplant 2020 challenge, where only synthetic defects were used for training and evaluation. In AutoImplant 2021, real clinical defective skulls from craniotomy and skulls with traumatic defects will be used in the evaluation phase, each serving as a separate track (task). The original AutoImplant 2020 task will serve as a third track of AutoImplant 2021. Using real cases for evaluation will guarantee the clinical usability of the winning algorithms.

[1] Göttsche, J., Mende, K.C., Schram, A. et al. Cranial bone flap resorption—pathological features and their

implications for clinical treatment. Neurosurg Rev (2020). https://doi.org/10.1007/s10143-020-01417-w

[2] Gilardino, M. S., Karunanayake, M., Al-Humsi, T., Izadpanah, A., Al-Ajmi, H., Marcoux, J., Atkinson, J., & Farmer, J.-P. (2015). A Comparison and Cost Analysis of Cranioplasty Techniques. The Journal of Craniofacial Surgery, 26(1), 113–117. https://doi.org/10.1097/scs.0000000000001305

Challenge keywords

List the primary keywords that characterize the challenge.

Skull Reconstruction, Cranioplasty, Cranial implant design, Deep learning, Shape completion, 3D shape analysis.

Year

The challenge will take place in ...

2021

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

none

Duration

How long does the challenge take?

Full day.

Expected number of participants

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

We expect the number of participants for AutoImplant 2021 to be about 30. In last years' organization (AutoImplant 2020), more than 140 users have registered for our challenge to get access to the dataset prior to the challenge deadline. 11 teams managed to submit their results. During the conference, these teams have also expressed their interest in taking part again if the challenge is to take place in 2021. With more dissemination of the challenge (collaborations with other institutions) and more time for the participants to make preparations, we expect the challenge to attract about 30 participants this year.

Publication and future plans

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

We plan to coordinate a publication of the challenge results/algorithms as we did in AutoImplant 2020. The challenge papers will be peer-reviewed and the accepted papers will be published in Springer LNCS challenge proceedings, like in 2020: https://link.springer.com/book/10.1007%2F978-3-030-64327-0

Space and hardware requirements

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

We will set up the challenge website for AutoImplant 2021 on the grand-challenge platform (https://grand-challenge.org/). The MICCAI 2021 will be a venue for participants to present and discuss their algorithms on-site.

TASK: Cranial implant design for diverse synthetic defects on aligned skulls

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.

Skull defects can generally occur at various positions with different shapes and sizes. In this task, the skulls were aligned into standardized position, while the synthetic defects have random shapes and positions to simulate the variability. 570 cases are available for training and 100 for evaluation. Each case is comprised of a complete skull, a defective skull with a defect and the corresponding implant. The defects are further sorted by their general type such as unilateral, bilateral and fronto-orbital, which can be used to identify weak and strong points of implant design methods.

Keywords

List the primary keywords that characterize the task.

Cranioplasty, cranial implant design, 3D shape analysis

ORGANIZATION

Organizers

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

Oldich Kodym, Brno University of Technology, Czech Republic
Michal Španl, Brno University of Technology, Czech Republic

b) Provide information on the primary contact person.

oldrich.kodym@gmail.com (Oldich Kodym)
spanel@fit.vutbr.cz (Michal Španl)

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

Repeated event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

The grand-challenge platform (grand-challenge.org) will be used to set up the challenge website.

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

Website will be set up on the grand-challenge platform, after acceptance.

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.

Publicly available data is allowed.

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 not participate.

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

We will award Best paper & highest ranking certificates to the winning teams. We will also look for company sponsorship for money related 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.

Top 3 performing methods will be announced publicly.

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).

Members that contribute to the development of the algorithms or the writing of papers qualify as authors. If a Springer LNCS proceeding is coordinated, participating teams should submit their papers to the organizing committee of the challenge for peer-review. Accepted papers will be included in the Springer challenge proceeding. Papers that are not accepted can be contributed to other venues.

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.

Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

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 can submit their results maximally 5 times. Participants can choose which submission is used for final evaluation and ranking.

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

The registration, release of training and test cases: April 15 2021
Results & paper submission opens: July 1 2021

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).

No ethics approval is required for this task.

Data usage agreement

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

Examples:

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

CC BY NC SA.

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 codes can be found at https://github.com/OldaKodym/evaluation_metrics.

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

Participants can choose to make public their codes.

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.

We will also look for company sponsorship for money related awards. The organizers have access to the ground truth of the test cases.

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

Surgery, CAD, Research.

Task category(ies)

State the task category(ies).

Examples:

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

Reconstruction.

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.

Synthetic defects are used for Task 1. The synthetic defects simulate variability of patients undergoing craniectomy.

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

The dataset for this task is derived from a public head CT collection. Synthetic surgical defects are used.

Imaging modality(ies)

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

The imaging technique applied in the challenge are Computed Tomography (CT) scans of the head.

Context information

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

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

not applicable

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

not applicable

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.

Synthetic skull defects are used for Task 1. The synthetic defects simulate variability of patients undergoing craniectomy.

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.

not applicable

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)).

Feasibility, Accuracy, Usability, Robustness.

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).

The dataset is adapted from a public head CT collection CQ500 (http://headctstudy.qure.ai/dataset). The training set and evaluation set are not overlapping with the datasets in Task 3.

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

The dataset is adapted from a public head CT collection CQ500 (http://headctstudy.qure.ai/dataset).

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 dataset is adapted from a public head CT collection CQ500 (http://headctstudy.qure.ai/dataset).

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.

In the training set, one case refers to a triple of complete skull, defective skull and the corresponding implant. In test set, only the defective skulls are released to the participants. The ground truth are kept by the organizers.

Therefore, one case refers to a defective skull for test set. All the cases are in Nrrd ("nearly raw raster data") format.

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

Task 1 offers 570 cases for training and 100 cases for evaluation.

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

The chosen cases for training and evaluation can reflect the shape variabilities if the skulls and defects.

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.

not applicable.

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.

No human annotators are involved.

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.

No human annotators are involved.

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.

No human annotators are involved.

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.

No human annotators are involved.

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.

Skull segmentation from head CT, denoising, alignment using rigid transformation.

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.

No human annotators are involved.

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)

Dice Similarity Coefficient (DSC) computed for the whole implant and for the implant border area, 95% Hausdorff Distance (HD95).

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

The use of DSC is supported by the following peer-reviewed publications:

Li, J., Pepe, A., Gsaxner, C., von Campe, G., & Egger, J. (2020). A baseline approach for autoimplant: the miccai 2020 cranial implant design challenge. In Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures (pp. 75-84). Springer, Cham.

Matzkin, F., Newcombe, V., Stevenson, S., Khetani, A., Newman, T., Digby, R., ... & Ferrante, E. (2020, October). Self-supervised Skull Reconstruction in Brain CT Images with Decompressive Craniectomy. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 390-399). Springer, Cham.

Computing DSC separately for the implant border area will allow us to evaluate smoothness of the fit between the predicted implant and the defective skull, which is a clinically significant property of implants.

HD95 is routinely used in segmentation challenge designs, such BraTS2020 (http://braintumorsegmentation.org/) or StructSeg2019 (https://structseg2019.grand-challenge.org/Evaluation/). We found that it correlates better with the perceived quality of reconstructed implant shape when compared to HD.

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.

Each metric will be aggregated using average over the whole test data set and ranked separately (HD in ascending order and DSC in descending order). The final ranking is the average of the two rankings.

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

Submissions with missing results on test cases will be excluded without consideration.

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

The same ranking method is used for AutoImplant 2020 as well as other MICCAI challenges such as StructSeg 2019

(https://structseg2019.grand-challenge.org/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.

We will exclude the participants who fail to report on the whole testing set. Besides the statistical values such as mean, standard deviation of the DSCs and HDs, we use the p-value in t-test to assess whether the top performing/ranking algorithms are significantly better than the rest of algorithms. t test code is available at https://github.com/li-jianning/ttest. To measure the variability, we will also consider variance, squared deviation, average absolute deviation and the inter-quartile range.

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

The mean value of DSC and HD produced by the algorithms are indicators of their overall performance. The standard deviation measures the performance stability of the algorithms.

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.

TASK: Cranial implant design for real patient defects

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.

Craniotomy is the surgical procedure of removing and replacing part of the skull in order to conduct brain related surgeries, such as brain tumor removal. In case of craniectomy, the bone flap is not replaced during the primary surgery and the reconstruction is performed later, often including cranial implant design. In this task, 10 real craniectomy cases will be used for evaluation. For training, participants may make use of the complete skulls from Task 1 or/and Task 3 to create synthetic implants. It is up to the participants to decide how the synthetic implants should be created. This task will be used to see how the performance of the automatic implant design methods translate to different population and real defect shapes.

Keywords

List the primary keywords that characterize the task.

Cranioplasty, cranial implant design, skull reconstruction

ORGANIZATION

Organizers

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

David G. Ellis, Department of Neurosurgery, University of Nebraska Medical Center, Omaha, NE, USA
Michele R. Aizenberg, Department of Neurosurgery, University of Nebraska Medical Center, Omaha, NE, USA

b) Provide information on the primary contact person.

david.ellis@unmc.edu (David G Ellis)
maizenberg@unmc.edu (Michele R. Aizenberg)

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

Repeated event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

The grand-challenge platform (grand-challenge.org) will be used to set up the challenge website.

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

Website will be set up on the grand-challenge platform, after acceptance.

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.

Publicly available data is allowed.

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 not participate.

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

We will award Best paper & highest ranking certificates to the winning teams. We will also look for company sponsorship for money related 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.

Top 3 performing methods will be announced publicly.

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).

Members that contribute to the development of the algorithms or the writing of papers qualify as authors. If a Springer LNCS proceeding is coordinated, participating teams should submit their papers to the organizing committee of the challenge for peer-review. Accepted papers will be included in the Springer challenge proceeding. Papers that are not accepted can be contributed to other venues.

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.

Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

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 can submit their results maximally 5 times. Participants can choose which submission is used for final evaluation and ranking.

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

The registration, release of training and test cases: April 15 2021
Results & paper submission opens: July 1 2021

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).

No ethics approval is required for this task.

Data usage agreement

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

Examples:

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

CC BY NC SA.

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 codes can be found at https://github.com/OldaKodym/evaluation_metrics.

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

Participants can choose to make public their codes.

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.

We will also look for company sponsorship for money related awards. The organizers have access to the ground truth of the test cases.

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

Surgery, CAD, Research.

Task category(ies)

State the task category(ies).

Examples:

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

Reconstruction.

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.

The target cohort are patients with craniectomy defects undergoing cranioplasty.

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

For this task, the challenge cohort are patients with craniectomy defects undergoing cranioplasty.

Imaging modality(ies)

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

The imaging technique applied in the challenge are Computed Tomography (CT) scans of the head.

Context information

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

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

not applicable

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

not applicable

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 data origin of Task 2 is CT images of patients with craniectomy defects.

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.

not applicable

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)).

Feasibility, Accuracy, Usability, Robustness.

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).

Scans were acquired using GE Revolution CT scanners.

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

A standardized head CT imaging protocol was used to acquire the images.

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 dataset is acquired from clinical scans of patients at Nebraska Medicine/University of Nebraska Medical Center in Omaha, Nebraska.

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.

One case in Task 2 refers to a CT image of a defective skull following craniectomy and the segmented CT image of the implant following cranioplasty in Nrrd ("nearly raw raster data") format.

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

Task 2 offers 10 cases for evaluation. The complete skulls from Task 1 and Task 3 can be used to create training data by participants.

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

Craniectomy cases are considered to be difficult cases. In this task, we provide 10 craniectomy cases for evaluating the algorithms from participants. The performance on these cases can reflect the true generalization/modelling ability of the algorithms.

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.

not applicable.

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.

No human annotators are involved.

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.

No human annotators are involved.

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.

No human annotators are involved.

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.

No human annotators are involved.

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.

Anonymization via CT defacing, skull segmentation, alignment using rigid registration, and implant segmentation.

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.

No human annotators are involved.

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)

Dice Similarity Coefficient (DSC) computed for the whole implant and for the implant border area, 95% Hausdorff Distance (HD95).

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

The use of DSC is supported by the following peer-reviewed publications:

Li, J., Pepe, A., Gsaxner, C., von Campe, G., & Egger, J. (2020). A baseline approach for autoimplant: the miccai 2020 cranial implant design challenge. In Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures (pp. 75-84). Springer, Cham.

Matzkin, F., Newcombe, V., Stevenson, S., Khetani, A., Newman, T., Digby, R., ... & Ferrante, E. (2020, October). Self-supervised Skull Reconstruction in Brain CT Images with Decompressive Craniectomy. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 390-399). Springer, Cham.

Computing DSC separately for the implant border area will allow us to evaluate smoothness of the fit between the predicted implant and the defective skull, which is a clinically significant property of implants.

HD95 is routinely used in segmentation challenge designs, such BraTS2020 (http://braintumorsegmentation.org/) or StructSeg2019 (https://structseg2019.grand-challenge.org/Evaluation/). We found that it correlates better with the perceived quality of reconstructed implant shape when compared to HD.

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.

Each metric will be aggregated using average over the whole test data set and ranked separately (HD in ascending order and DSC in descending order). The final ranking is the average of the two rankings.

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

Submissions with missing results on test cases will be excluded without consideration.

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

The same ranking method is used for AutoImplant 2020 as well as other MICCAI challenges such as StructSeg 2019 (https://structseg2019.grand-challenge.org/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.

We will exclude the participants who fail to report on the whole testing set. Besides the statistical values such as mean, standard deviation of the DSCs and HDs, we use the p-value in t-test to assess whether the top performing/ranking algorithms are significantly better than the rest of algorithms. t test code is available at https://github.com/li-jianning/ttest. To measure the variability, we will also consider variance, squared deviation, average absolute deviation and the inter-quartile range.

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

The mean value of DSC and HD produced by the algorithms are indicators of their overall performance. The standard deviation measures the performance stability of the algorithms.

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.

TASK: Cranial implant design for standardly positioned synthetic defects on unaligned skulls

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.

Skull defects are often created by cutting an artificial hole into the skull to access the brain tissue using cranial drills. In the original AutoImplant 2020 task (https://autoimplant.grand-challenge.org/), this type of defect was simulated at static positions at the back of skulls. The data will serve as Task 3 for the AutoImplant 2021 challenge. Synthetic defects will be used for both training (100 cases) and evaluation (110 cases).

Keywords

List the primary keywords that characterize the task.

Cranioplasty, cranial implant design, 3D shape analysis

ORGANIZATION

Organizers

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

Jianning Li, Graz University of Technology, Graz, Austria
Jan Egger, Graz University of Technology, Graz, Austria
Victor Alves, Center Algoritmi, University of Minho
Gord von Campe, Department of Neurosurgery, Medical University of Graz

b) Provide information on the primary contact person.

jianning.li@icg.tugraz.at (Jianning Li)
egger@tugraz.at (Jan Egger)

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

Repeated event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

The grand-challenge platform (grand-challenge.org) will be used to set up the challenge website.

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

Website will be set up on the grand-challenge platform, after acceptance.

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.

Publicly available data is allowed.

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 not participate.

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

We will award Best paper & highest ranking certificates to the winning teams. We will also look for company sponsorship for money related 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.

Top 3 performing methods will be announced publicly.

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).

Members that contribute to the development of the algorithms or the writing of papers qualify as authors. If a Springer LNCS proceeding is coordinated, participating teams should submit their papers to the organizing committee of the challenge for peer-review. Accepted papers will be included in the Springer challenge proceeding. Papers that are not accepted can be contributed to other venues.

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.

Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

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 can submit their results maximally 5 times. Participants can choose which submission is used for final evaluation and ranking.

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

The registration, release of training and test cases: April 15 2021
Results & paper submission opens: July 1 2021

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).

No ethics approval is required for this task.

Data usage agreement

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

Examples:

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

CC BY NC SA.

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 codes can be found at https://github.com/OldaKodym/evaluation_metrics.

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

Participants can choose to make public their codes.

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.

We will also look for company sponsorship for money related awards. The organizers have access to the ground truth of the test cases.

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

Surgery, CAD, Research.

Task category(ies)

State the task category(ies).

Examples:

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

Reconstruction.

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.

Synthetic defects are used for Task 3. The synthetic defects resembles the real defects from craniotomy.

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

The dataset for this task is derived from a public head CT collection. Synthetic surgical defects are used.

Imaging modality(ies)

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

The imaging technique applied in the challenge are Computed Tomography (CT) scans of the head.

Context information

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

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

not applicable

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

not applicable

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.

Synthetic skull defects are used for Task 3. The synthetic defects resemble the real defects from craniotomy.

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.

not applicable

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)).

Feasibility, Accuracy, Usability, Robustness.

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).

The dataset is adapted from a public head CT collection CQ500 (http://headctstudy.qure.ai/dataset). The training set and evaluation set are not overlapping with the datasets in Task 1.

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

The dataset is adapted from a public head CT collection CQ500 (http://headctstudy.qure.ai/dataset).

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 dataset is adapted from a public head CT collection CQ500 (http://headctstudy.qure.ai/dataset) and was used in AutoImplant 2020.

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.

In the training set, one case refers to a triple of complete skull, defective skull and the corresponding implant. In test set, only the defective skulls are released to the participants. The ground truth are kept by the organizers. Therefore, one case refers to a defective skull for test set. All the cases are in Nrrd ("nearly raw raster data") format.

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

Task 3 offers 100 cases for training and 110 cases for evaluation.

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

The chosen cases for training and evaluation can reflect the shape variabilities if the skulls and defects.

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.

not applicable.

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.

No human annotators are involved.

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.

No human annotators are involved.

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.

No human annotators are involved.

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.

No human annotators are involved.

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.

Skull segmentation from head CT and denoising.

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.

No human annotators are involved.

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)

Dice Similarity Coefficient (DSC) computed for the whole implant and for the implant border area, 95% Hausdorff Distance (HD95).

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

The use of DSC is supported by the following peer-reviewed publications:

Li, J., Pepe, A., Gsaxner, C., von Campe, G., & Egger, J. (2020). A baseline approach for autoimplant: the miccai 2020 cranial implant design challenge. In Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures (pp. 75-84). Springer, Cham.

Matzkin, F., Newcombe, V., Stevenson, S., Khetani, A., Newman, T., Digby, R., ... & Ferrante, E. (2020, October). Self-supervised Skull Reconstruction in Brain CT Images with Decompressive Craniectomy. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 390-399). Springer, Cham.

Computing DSC separately for the implant border area will allow us to evaluate smoothness of the fit between the predicted implant and the defective skull, which is a clinically significant property of implants.

HD95 is routinely used in segmentation challenge designs, such BraTS2020 (http://braintumorsegmentation.org/) or StructSeg2019 (https://structseg2019.grand-challenge.org/Evaluation/). We found that it correlates better with the perceived quality of reconstructed implant shape when compared to HD.

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.

Each metric will be aggregated using average over the whole test data set and ranked separately (HD in ascending order and DSC in descending order). The final ranking is the average of the two rankings.

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

Submissions with missing results on test cases will be excluded without consideration.

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

The same ranking method is used for AutoImplant 2020 as well as other MICCAI challenges such as StructSeg 2019 (https://structseg2019.grand-challenge.org/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.

We will exclude the participants who fail to report on the whole testing set. Besides the statistical values such as mean, standard deviation of the DSCs and HDs, we use the p-value in t-test to assess whether the top performing/ranking algorithms are significantly better than the rest of algorithms. t test code is available at https://github.com/li-jianning/ttest. To measure the variability, we will also consider variance, squared deviation, average absolute deviation and the inter-quartile range.

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

The mean value of DSC and HD produced by the algorithms are indicators of their overall performance. The standard deviation measures the performance stability of the algorithms.

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.

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.

[1] Li, J., Pepe, A., Gsaxner, C., & Egger, J. (2020). An Online Platform for Automatic Skull Defect Restoration and Cranial Implant Design. arXiv preprint arXiv:2006.00980.

[2] Kodym, O., Španl, M., & Herout, A. (2020). Skull shape reconstruction using cascaded convolutional networks. Computers in Biology and Medicine, 123, 103886.

[3] Li, J., Pepe, A., Gsaxner, C., von Campe, G., & Egger, J. (2020). A baseline approach for autoimplant: the miccai 2020 cranial implant design challenge. In Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures (pp. 75-84). Springer, Cham.

[4] Li, J., & Egger, J. Towards the Automatization of Cranial Implant Design in Cranioplasty: First Challenge, AutoImplant 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings.

[5] Li, J. Gsaxner, C., Pepe, A., Morais, A., Alves, V., von Campe, G., Wallner, J.,& Egger, J. Synthetic skull bone defects for automatic patient-specific craniofacial implant design. Scientific Data, Nature Publishing Group, accepted.

[6] Li, J. Gsaxner, C., Pepe, A., Morais, A., Alves, V., von Campe, G., Wallner, J.,& Egger, J. Head ct collection for patient-specific craniofacial implant (psi) design. Figshare, doi.org/10.6084/m9.figshare.12423872