Calibration and Uncertainty for multiRater Volume Assessment in multiorgan Segmentation: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Calibration and Uncertainty for multiRater Volume Assessment in multiorgan Segmentation

Challenge acronym

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

CURVAS

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.

In medical imaging, DL models are often tasked with delineating structures or abnormalities within complex anatomical structures, such as tumors, blood vessels, or organs. Uncertainty arises from the inherent complexity and variability of these structures, leading to challenges in precisely defining their boundaries. This uncertainty is further compounded by interrater variability, as different medical experts may have varying opinions on where the true boundaries lie. DL models must grapple with these discrepancies, leading to inconsistencies in segmentation results across different annotators and potentially impacting diagnosis and treatment decisions. Addressing interrater variability in DL for medical segmentation involves the development of robust algorithms capable of capturing and quantifying uncertainty, as well as standardizing annotation practices and promoting collaboration among medical experts to reduce variability and improve the reliability of DL-based medical image analysis. Interrater variability poses significant challenges in the field of DL for medical image segmentation.

Furthermore, achieving model calibration, a fundamental aspect of reliable predictions, becomes notably challenging when dealing with multiple classes and raters. Calibration is pivotal for ensuring that redicted probabilities align with the true likelihood of events, enhancing the model's reliability. It must be considered that, even if not clearly, having multiple classes account for uncertainties arising from their interactions. Moreover, incorporating annotations from multiple raters adds another layer of complexity, as differing expert opinions may contribute to a broader spectrum of variability and computational complexity.

Consequently, the development of robust algorithms capable of effectively capturing and quantifying variability and uncertainty, while also accommodating the nuances of multi-class and multi-rater scenarios, becomes imperative. Striking a balance between model calibration, accurate segmentation and handling variability in medical annotations is crucial for the success and reliability of DL-based medical image analysis.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Abdominal CT, multi-class, multi-rater, segmentation, calibration, uncertainty

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.

N/A

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.

A previously held challenge (QUBIQ 2021 - https://qubiq.grand-challenge.org/statistics/) with similar characteristics but more limited scope gathered 500 data downloads, and received up to 400 submissions from aproximately 40 participants. This challenge stands as a reference for us, as we have also requested their permission to employ their data. Therefore, it is safe to assume a similar or greater participation.

Publication and future plans

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

The first step will be to publish a paper together with the winning authors as well as making public all the code used by anyone involved in the publication. In the future, we plan to analyze the combination of algorithms through ensembling, inter-algorithm variability, common issues or biases present in the submitted methods, and the variability rankings. Our intention is to publish
this analysis in a leading journal within the field of medical image analysis.

Also, the winners of the challenge will be invited to present their methods and results in the challenge event hosted in MICCAI 2024. Furthermore, the data will remain open for researchers to use it freely, always referencing the challenge.

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.

The platform used will be grand-challenge.org.

Projectors, loud speakers and microphones.

TASK 1: Calibrated Segmentation and Volume Estimation of Abdominal Organs

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.

In medical imaging, DL models are often tasked with delineating structures or abnormalities within complex anatomical structures, such as tumors, blood vessels, or organs. Uncertainty arises from the inherent complexity and variability of these structures, leading to challenges in precisely defining their boundaries. This uncertainty is further compounded by interrater variability, as different medical experts may have varying opinions on where the true boundaries lie. DL models must grapple with these discrepancies, leading to inconsistencies in segmentation results across different annotators and potentially impacting diagnosis and treatment decisions. Addressing interrater variability in DL for medical segmentation involves the development of robust algorithms capable of capturing and quantifying uncertainty, as well as standardizing annotation practices and promoting collaboration among medical experts to reduce variability and improve the reliability of DL-based medical image analysis. Interrater variability poses significant challenges in the field of DL for medical image segmentation.

Furthermore, achieving model calibration, a fundamental aspect of reliable predictions, becomes notably challenging when dealing with multiple classes and raters. Calibration is pivotal for ensuring that redicted probabilities align with the true likelihood of events, enhancing the model's reliability. It must be considered that, even if not clearly, having multiple classes account for uncertainties arising from their interactions. Moreover, incorporating annotations from multiple raters adds another layer of complexity, as differing expert opinions may contribute to a broader spectrum of variability and computational complexity.

Consequently, the development of robust algorithms capable of effectively capturing and quantifying variability and uncertainty, while also accommodating the nuances of multi-class and multi-rater scenarios, becomes imperative. Striking a balance between model calibration, accurate segmentation and handling variability in medical annotations is crucial for the success and reliability of DL-based medical image analysis.

Keywords

List the primary keywords that characterize the task.

Segmentation, Volume, Uncertainty, Calibration

ORGANIZATION

Organizers

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

Meritxell Riera i Marín
Affiliation: Sycai Medical / Universitat Pompeu Fabra (UPF)

Title: PhD Candidate

Javier García López
Affiliation: Sycai Medical
Title: CTO and co-founder PhD

Joy Kleiss
Affiliation: Universitätsklinikum Erlangen
Title: MD PhD candidate

Shika O K
Affiliation: Universitat Pompeu Fabra (UPF)
Title: Post-doc researcher, PhD

Adrian Galdrán
Affiliation: Universitat Pompeu Fabra (UPF)
Title: Post-doc researcher, PhD

Matthias May
Affiliation: Universitätsklinikum Erlangen
Title: MD

Miguel Angel González-Ballester
Affiliation: Universitat Pompeu Fabra (UPF)
Title: ICREA Professor, PhD

Júlia Rodríguez Comas
Affilitation: Sycai Medical
Title: CSO and co-founder PhD

Maximilan Schmidt
Affiliation: Universitätsklinikum Erlangen
Title: MD

Christopher Hessman
Affiliation: Universitätsklinikum Erlangen
Title: MD

Antón Aubanell
Affiliation: Hospital de Sant Pau
Title: MD

Andreu Antolín

Affiliation: Hospital Vall d'Hebron
Title: MD

b) Provide information on the primary contact person.

Meritxell Riera i Marín
m.riera@sycaitechnologies.com
+34 627182691

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

One-time 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 2024

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://www.sycaimedical.com/challenge

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.

Only the data provided by the challenge and/or publicly available with multiple annotators data may be used.

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.

Members of the organizers' institutes may participate but are not elegible for awards.

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

First prize: 1000€
Second prize: 500€
Third prize: 250€

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 five 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).

Two members of the participating team can be qualified as author (the person that submits the results). The participating team may publish their own results separately only after the organizer has published a challege paper and always mentioning the organizer's challenge paper.

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 challenge will have two phases, Validation and Testing. In the Validation phase participants will create docker containers and submit them wrapped as algorithms to Grand-Challenge. This process will allow participants to both measure performance on a small subset of the test set and debug their Docker containers. The Testing phase will follow a similar procedure: participants will submit their algorithms in the form of a Docker container with a short abstract highlighting the contributions. A baseline model and working docker container will be made available shortly after the start of the competition.

Participants will be asked to open-source their solutions after the end of the challenge to ensure reproducibility. Participants are required to share their source code under an open-source license, such as the MIT License. This license should grant the Organizer the right to distribute the code publicly for research and development purposes. While participants maintain copyright ownership of their submitted code, failure to specify a license will result in the assumption that the MIT License applies. By submitting without explicit licensing information, participants acknowledge and affirm that they possess all necessary rights to permit such submission under the MIT License.

Once the challenge is finished and the winners are announced, they will be asked to publicly share their code and model and, until the first dissemination paper is published, reproducibility checks will be performed in order to

evaluate the quality of the results. However, regardless of the quality of the code, the main focus will be set on whether the results are reproducible or not.

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 be allowed to submit 3-5 times during validation and twice for the Testing phase - only the best submission will be considered.

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

Challenge website open: March 28th 2024
Registration date: April 1st 2024
Training set release: April 15th 2024
Validation submission open: May 15th 2024
Test submission open: June 15th 2024
Release date of the results: July 1st 2024
Challenge days (MICCAI): October 6-10th 2024

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 collected for the generation of the datasets involved in this challenge has been approved by an ethical commitee (number 23-243-B) held at the Universitätsklinikum Erlangen Hospital.

The data to be used during and after the challenge is pseudonymized and coded by the Hospital to assure that a re-identification of the data sample is not possible. Moreover, the patient information is only known by the IP of the Hospital so that the challenge collaborators do not have as well any means to identify patient's data at any point.

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

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 code to produce rankings will be released together with the training data.

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

The participants will have to sumbit their models in Docker containers during the test phase. The procedure to create the Docker containers will be given beforehand.

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.

There are no conflict of interest to be reported. The funding of the challenge is done by the organizers. The participants will have access to the test case labels during the challenge (release date of validation and test set: May 14th. 2024).

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

Segmentation, uncertainty, calibration

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.

N/A

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 90 CT images prospectively gathered at the University Hospital Erlangen between August 2023 and October 2023.

A technical requirement was the use of contrast-enhanced CTs in a portal venous phase and the acquisition of thin slices (0.6-1mm).

Inclusion criteria were a maximum of 10 cysts with a diameter of less than 2,0 cm. Furthermore, CT scans with major artefacts (e.g. breathing artefacts) or incomplete registrations were excluded.

Another inclusion criteria was that participants had to be older than 18 years and provided both verbal and written consent to use their CT images as part of the Challenge. Study specific and broad consent were gathered.

Among the 90 patients, there are 51 male and 39 female patients. Their age ranges from 37 to 94 years old. The average age is 65,7 years. All were treated at the University Hospital Erlangen, located in Bavaria, Germany. No other selection criteria were defined in order to obtain a representative sample of a typical patient collective.

Imaging modality(ies)

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

We used Thoracic-Abdominal CT images that were taken during the patients' stay at hopsital and because of different medical reasons. Further inclusion criteria were thin slices (0.6-1mm) and the use of contrast agent. Since the challenge is about abdominal organs we used Br40 soft kernel.

Context information

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

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

No other information about the image data will be released.

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

No other information about the patient is released since the CT images will be published online.
The CT images do not contain any further characteristics or information about the patient.

Target entity(ies)

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

The CT images are aqcuired at the University Hospital Erlangen in Germany.

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

The algorithm target are left and right kidney, pancreas and liver.

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)).

Dice Score for pancreas, liver and kidney segmentation; MSE for organ volume estimation and calibration error and brier score for uncertainty assessment.

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).

We used Siemens CT scanners. The models are Go.top, Xcite, Xceed, Alpha, Force, AS+.

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

CT examinations were performed with SIEMENS CT scanners at the university hospital Erlangen. The Rotation speed was 0.25 or 0.5 sec. Detector collimation ranged from 128x0.6mm single source to 98x0.6x2 and 144x0.4x2 dual source. Spiral pitch factors between 0.3 and 1.3 were used. Mean reference tube current was 200 mAs at 120 mAs. Automated tube voltage adaptation and tube current modulation was used in all cases.

Contrast agent was injected in all patients with an injection rate of 3-4 mL/s and body weight adapted dosage of 400 mg(iodine)/kg (= 1.14 ml/kg Iomeprol 350mg/ml).

All images were reconstructed using soft convolution kernels and iterative techniques. Slice thickness was 0.7-1.0 mm with overlapping increment.

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

The data was aqcuired in the department of radiology at the University Hospital Erlangen.

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).

The data selection was done by Joy-Marie Kleiß (MD PhD candidate) and then reviewed by radiologists with an experience of 2 and 4 years in abdominal CT imaging.

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.

All the data provided in this challenge are CT images of a human abdomen. Each CT will have three different annotations of three different structures: pancreas, kidney and liver. However, the different CTs are classified in three different groups depending on the amount of lesions and pathologies.

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

Our overall data consists on 90 CTs splitted in three different groups:
- Group A: cases with 2 cysts or less with no contour altering pathologies - 45 CTs
- Group B: cases with 3-5 cysts with no contour altering pathologies - 22 CTs
- Group C: cases with 6-10 cysts with some pathologies included (liver metastases, hydronephrosis, adrenal gland metastases, missing kidney) - 23 CTs

All cysts included have a maximum of 2 cm of diameter.

For the training set, we will be using 20 CTs from the Group A. This training set is small since it is a high interest of us to study whether this task can be learned with a reduced number of of CTs or using other public data bases. Participants will be encouraged to leverage publicly available external data annotated by multiple raters. QUIBQ21 organisers have already been contacted and have given consent (with proper attribution) on using their multi-annotator data.

For the validation set, only images from Group A and Group B will be given:
- Group A: 5 CTs
- Group B: 5 CTs

For the test set, the amount of images is the following:
- Group A: 20 CTs
- Group B: 17 CTs
- Group C: 23 CTs

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

The idea of giving a small amount of data for the training set and giving the opportunity of using a public dataset for training is to make the challenge more inclusive, giving the opportunity to develop a mehtod by using data that is in anyones hands. Also, by using this data to train and using other data to evaluate, makes it more robust to shifts and other sources of variability between datasets.

Then, for the validation set, we found it enough use 10 CTs from the two first groups so the algorithm and the docker can be tested, since they are close enough to the 20 CTs given for the training set.

Finally, for the evaluation of the algorithm, we considered it relevant to split it in subgroups in order to study how the methods perfom in different scenarios: first, healthier easy cases; second, cases that have with a few more cysts than the first group; then, cases in which big cysts and pathologies appear.

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.

Among the 90 patients whose data were uploaded, there are 51 male and 39 female patients. Their age ranges from 37 to 94 years old. The average age is 65,7 years. All were treated at the University Hospital Erlangen, located in Bavaria, Germany.

No other selection criteria were defined in order to obtain a representative sample of a typical patient collective.

Annotation characteristics

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

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

The first step for obtaining de labels was using the TotalSegmentator (https://github.com/wasserth/TotalSegmentator) to get rough annotations. Then, the labels were sent to three radiologists, to both correct the automatic annotations and add possible missing organs. One of the three labeling radiologists, the MD PhD candidate Joy-Marie Kleiß, previously defined both the dataset cohort and the criteria of what belongs to the parenchyma and what does not and it was given to the other two labeling radiologists to follow the same criteria to be coherent with each other. Separately, two other radiologists, with an experience of 2 and 4 years in abdominal CT imaging, supervised the criteria of the cohort of the MD PhD student, but they did not have anything to do with the labeling itself, hence, there is no bias between the annotations of the different radiologists.

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.

Each labeled for this challenge has specific instructions. Below are listed per organ.
- Liver:
Generally speaking, we define the liver 'as the entire liver tissue including all internal structures like vessel systems, tumors etc.' (1) Thus, the portal vein itself is excluded from contouring. The two main branches of the portal vein are excluded from the segmentation. Any branch of the following generations is included. 'In case of partial enclosure (occurring where large vessels as Vena Cava and portal vein enter or leave the liver), the parts enclosed by liver tissue are included in the segmentation, thus forming the convex hull of the liver shape.' (1) Any fatty tissue that pulls into the liver is excluded. The gallbladder should not be marked. Wide and especially pathologically widened bile ducts are included in the segmentation of the liver.
- Kidney:
The right and left kidney will be segmented. Included in the segmentation will be the kidney parenchyma including the renal medulla. Excluded is the renal pelvis (4) and the ureter as a urinary stasis could alter the original volume.
- Pancreas:

When segmenting the pancreas, we will not differentiate between head, body and tail. Moreover neither the splenic vein nor the mesenterial vein will be included in segmentation. (3) However, it is important the whole pancreas in its course is tracked and marked.

(1) Heimann T, van Ginneken B, Styner MA, Arzhaeva Y, Aurich V, Bauer C, Beck A, Becker C, Beichel R, Bekes G, Bello F, Binnig G, Bischof H, Bornik A, Cashman PM, Chi Y, Cordova A, Dawant BM, Fidrich M, Furst JD, Furukawa D, Grenacher L, Hornegger J, Kainmüller D, Kitney RI, Kobatake H, Lamecker H, Lange T, Lee J, Lennon B, Li R, Li S, Meinzer HP, Nemeth G, Raicu DS, Rau AM, van Rikxoort EM, Rousson M, Rusko L, Saddi KA, Schmidt G, Seghers D, Shimizu A, Slagmolen P, Sorantin E, Soza G, Susomboon R, Waite JM, Wimmer A, Wolf I. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging. 2009 Aug;28(8):1251-65. doi: 10.1109/TMI.2009.2013851. Epub 2009 Feb 10. PMID: 19211338.
(2) Rädsch, T., Reinke, A., Weru, V. et al. Labelling instructions matter in biomedical image analysis. Nat Mach Intell 5, 273-283 (2023). https://doi.org/10.1038/s42256-023-00625-5 (3) Westenberger, Jasmin Barbara. Automatische Gewebesegmentierung der Nieren und des Pankreas im Ganzkörper-MRT mittels Deep Learning. PhD thesis, Eberhard Karls Universität Tübingen, 2021. http://dx.doi.org/10.15496/publikation-63135
(4) Brachmann, Franz Xaver. Evaluation einer semi-automatischen Segmentierungsmethode für die Volumetrie des renalen Kortex, der Medulla und des gesamten Nierenparenchyms in nativen T1-gewichteten MR-Bildern. PhD thesis, Medizinischen Fakultät def Friedrich-Alexander-Universität Eralngen-Nürnberg, 2021. https://open.fau.de/items/cb6cd403-3178-4884-98c4-5098178658ba

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.

All CTs were labeled by three radiologists. The MD PhD student (and then reviewed by other radiologists with an experience of 2 and 4 years in abdominal CT imaging that had nothing to do with the labeling itself) and two more radiologists.

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.

In this challenge we are working with multiple classes and multiple lables. It is a critical decision on how it is decided to merge the different annotations and it can change greatly the final results of the challenge (1). Because of that, we ended up by using one of the simplest methods, averaging the three different annotators, obtaining a soft label as a golden ground truth. This soft label is used for the uncertainty and calibration volume assessment.

(1) Maier-Hein, L., Eisenmann, M., Reinke, A. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat Commun 9, 5217 (2018). https://doi.org/10.1038/s41467-018-07619-7

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.

The studies included in the training, validation and testing datasets were downloaded from the Hospital's PACS and the pseudonimized to remove any metadata that would allow their identification. This pseudonimization process did not affect the image quality.

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 image annotation process it has been defined really thoroughly. Firstly, high quality images were considered for this challenged by an MD PhD candidate and reviewed by two experienced radiologists, that had nothing to do with the labeling. Furthemore, when given to the labeling radiologists, they followed a clearly defined methodology on what to consider kidney, pancreas and liver. Hence, the three annotations are coherent but not biased, since neither of the three radiologists had contact with the others. However, there is the inherent ambiguity of the image that cannot be fought; in the edges of the structures there is always going to be different considerations on what belongs or not to the structure due to the blurriness and maybe low quality of the image we could be working with.

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

As previously said, one of the relevant sources of error is the image quality. However, this is being studied in the challenge. On the other hand, we tried to reduce as possible the interrater variability by setting a really thorough and clear process to label the different images and structures. Nonetheless, there is a human variability that cannot and should not be ignored. This is why, in this challenge, we highlight the importance of studying this areas in which there is ambiguity that cannot be eliminated. In order to keep improving the dataset, we will establish a GitHub repository to solicit possible errors from data users.

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)

As indicated in (1) it is not a trivial to define a an evaluation criteria that reflects the biomedical need. It is highlighted that the Dice Similarity Score (DSC) is often used without considering the fact that there are small structures that should be analysed otherwise. As the QUBIQ challenge states: "Dice scores have many shortcomings, in particular in the novel application domain of uncertainty-aware image quantification." Consequently, additional metrics will be considered to have a proper and thorough assessment.

In our study, we will performe a DSC assessment but under some specific considerations. First, we found ourselves in a case in which there are no small structures that should be analysed otherwise; since we are working with big enough organs, we are able to assess the quality of the segmentation with the DSC. Secondly, since we are working with interrater variability, we will asses the DSC only in the consensus areas defined previously. This consensus areas are the areas in which there is no ambigüity: all three raters have labeled this area as belonging to the structure. The discensus areas (areas in which at least one rater did not consider them as structure) will not be assessed by the DSC; this would not make sense, because these areas need to be assessed

otherwise due to the fact that there is no a clear criteria in which value should we binarize or what rater do we consider as right or wrong.

In order to add a clinically relevant metric, we will be studying highly relevant biomarkers, such as volumes. We will evaluate the different volumes corresponding to each part of the sctructure, depending on whether it is a consensus area or an discensus area. Furthermore, this consensus volume will have an uncertainty assessment; since this will be the intersection of the three labelers, the uncertainty in this volume will need to be as close as possible to 0.

Finally, there will be a second part of the challenge in which calibration of the model will be evaluated. Calibration for multiclass and multi rater is not something trivial to define. In this challenge, we tried to be as clear as possible and as intuitive as possible when defining this metric. For this calibration study, the areas will be split in three: the consensus background area, the discensus area that can be both background or structure and the consensus structure area. Each area will be studied separately with the Brier Score metric (1).

(1) Maier-Hein, Lena, et al. "Metrics Reloaded: Recommendations for Image Analysis Validation" (Version 7). arXiv:2206.01653 [cs.CV]. 2022. Web. https://doi.org/10.48550/arXiv.2206.01653

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

In this challenge, we are looking to develop a way to study not only the performance of a Neural Network, but how calibrated it is as well as define a metric to assess the clinical value. Since we are trying to work in a scope as close to the real world as possible, we are dealing with multi rater variability; there will be three annotations per each organ defined by three different radiologists. We tried to eliminate as much as possible the ambiguity among them by stating clear criteria on what to segment and how. However, it is something that is not going to disappear due to, for example, the inherent variability of the image itself, hence, we need to learn to live with it. For this reason, the different regions of study will be splitted in three: the consensus background area, the discensus area which can belong both to the background or the structure and, the consensus foreground area, which will be defined by the region that can only belong to the organ without any ambiguity.

Furthermore, generally there is a lack of connection between the technical part and the clinical relevance of the different algorithms that are developed. Because of that, the first part of evaluating the algorithms will be focusing on developing a relevant way of giving information to the radiologists or oncologists. This is why biomarkers such as volume are considered relevant for the assessment of this challenge. It consists of studying the total volume of the organ. However, as already said, the interrater variability is something that cannot be eliminated. Hence, the different volume assessments will be carried separately. In addition, we find it interesting to have a classical technical assessment as well. Consequently, we consider that both DSC and volume are complementary and relevant since it is a way to study how the model is working in a classical way as well as in a clinical context.

Regarding the calibration and uncertainty study, it should be essential and mandatory when providing new model results. It should be a must when considering whether a model is good, thorough and reliable. The justification for considering both calibration and uncertainty assessment lies in their complementary roles in providing a holistic evaluation of a predictive model. The Brier score, focusing on calibration assessment and following the criteria stated in (1), measures the discrepancy between predicted probabilities and observed outcomes, providing

insight into model calibration. Meanwhile, uncertainty assessment evaluates the model's confidence in predictions. Together, they offer a comprehensive view, ensuring not only accurate calibration but also a robust understanding of the model's uncertainty, enhancing the overall reliability and effectiveness of the predictive system.

(1) Maier-Hein, Lena, et al. "Metrics Reloaded: Recommendations for Image Analysis Validation" (Version 7). arXiv:2206.01653 [cs.CV]. 2022. Web. https://doi.org/10.48550/arXiv.2206.01653

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.

We will have multiple validation metrics, resulting in a merged ranking (1).

Ranking for segmentation:
- For the evaluation of the segmentations, the DSC will be used as a primary metric, since it yields to more robust tankings than the HD (2). Furthermore, since the structures are not too small it is not a missuse of the DSC.
- Another metric that will be studied, to make the challenge as clinically relevant as possible, will be the volume of the parenchyma. However, since we have different splits of the segmentations, depending on whether it is a consensus or an discensus area, the volumes will be studied separately.
- This volume assessment will be paired with an uncertainty metric, that will give how certain is the consensus volume.

Ranking for calibration:
- The Brier Score will be used in each volume separately to study the calibration.

Mean and metric-based aggregation will define the approach to follow in terms of considering the final ranking, since it will consider both segmentation and calibration performance. Afterwards, bootstraping will be used in order to study the robustness of the ranking as well as the Wilcoxon test will be used to confirm the final ranking.

(1) Maier-Hein, Lena, et al. "Metrics Reloaded: Recommendations for Image Analysis Validation" (Version 7). arXiv:2206.01653 [cs.CV]. 2022. Web. https://doi.org/10.48550/arXiv.2206.01653
(2) Maier-Hein, L., Eisenmann, M., Reinke, A. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat Commun 9, 5217 (2018). https://doi.org/10.1038/s41467-018-07619-7

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

Missing results on the submissions will be scored as 0.

Even though assigning a score of 0 to the missing results may be unfair in close competition scenarios, we do not find a better way to assess these results. Participants should take time on studying if they have all the results needed. Furthermore, for the validation phase they will have several attempts to submit their results so they can check and understand the procedure. In addition, for the testing phase they will have two attempts and only the best submission will be considered, giving them an advantage if any of the submissions have a missing result.

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

This ranking system is designed to consider essential metrics individually in the processes of segmentation, calibration, and uncertainty assessment for distinct organs such as the liver, kidney, and pancreas. Additionally, it aims to assess the overall performance of participant teams. To ensure a comprehensive evaluation for the different capabilities, we recommend the use of multiple metrics to avoid bias and discourage teams from optimizing their methods solely for a particular performance metric.

Statistical analysis will be conducted to validate the ranking outcomes. The selection of statistical methods, including bootstrap and the Wilcoxon test, for ranking submissions is underpinned by their suitability for the challenge's data characteristics and objectives. Bootstrap resampling allows for robust estimation of uncertainty in metrics, crucial in evaluating diverse submissions across various metrics. The Wilcoxon test offers a non-parametric approach to compare performances, ensuring fair evaluation regardless of metric variations. Together, these methods provide a rigorous and unbiased framework for ranking submissions, aligning with the challenge's objective of impartially assessing diverse model performances.

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.

In instances of missing data or reporting errors, organizers are tasked with assisting participants in rectifying and completing the missing results initially. For incomplete test cases, scores will be designated as 0. As already said, participants should take time on studying if they have all the results needed. Furthermore, for the validation phase they will have several attempts to submit their results so they can check and understand the procedure. In addition, for the testing phase they will have two attempts and only the best submission will be considered, giving them an advantage if any of the submissions have a missing result.

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

The devised ranking scheme deliberately steers clear of favoring a single metric that might outperform other methods. Instead, it prioritizes metrics aligned with medical requirements and substantial clinical significance. As a result, various evaluation metrics from different perspectives are incorporated. The statistical analysis focuses on methods consistently ranked higher across multiple metrics, providing a deeper scrutiny of the ranking outcomes.

The selection of statistical methods, including bootstrap and the Wilcoxon test, for ranking submissions is underpinned by their suitability for the challenge's data characteristics and objectives. Bootstrap resampling allows for robust estimation of uncertainty in metrics, crucial in evaluating diverse submissions across various metrics. The Wilcoxon test offers a non-parametric approach to compare performances, ensuring fair evaluation regardless of metric variations. Together, these methods provide a rigorous and unbiased framework for ranking submissions, aligning with the challenge's objective of impartially assessing diverse model performances.

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.

The first step will be to publish a paper together with the winning authors as well as making public all the code used by anyone involved in the publication. In the future, we plan to analyze the combination of algorithms through ensembling, inter-algorithm variability, common issues or biases present in the submitted methods, and the variability rankings. Our intention is to publish this analysis in a leading journal within the field of medical image analysis.

Also, the winners of the challenge will be invited to present their methods and results in the challenge event hosted in MICCAI 2024. Furthermore, the data will remain open for researchers to use it freely, always referencing the challenge.

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) Heimann T, van Ginneken B, Styner MA, Arzhaeva Y, Aurich V, Bauer C, Beck A, Becker C, Beichel R, Bekes G, Bello F, Binnig G, Bischof H, Bornik A, Cashman PM, Chi Y, Cordova A, Dawant BM, Fidrich M, Furst JD, Furukawa D, Grenacher L, Hornegger J, Kainmüller D, Kitney RI, Kobatake H, Lamecker H, Lange T, Lee J, Lennon B, Li R, Li S, Meinzer HP, Nemeth G, Raicu DS, Rau AM, van Rikxoort EM, Rousson M, Rusko L, Saddi KA, Schmidt G, Seghers D, Shimizu A, Slagmolen P, Sorantin E, Soza G, Susomboon R, Waite JM, Wimmer A, Wolf I. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging. 2009 Aug;28(8):1251-65. doi: 10.1109/TMI.2009.2013851. Epub 2009 Feb 10. PMID: 19211338.
(2) Rädsch, T., Reinke, A., Weru, V. et al. Labelling instructions matter in biomedical image analysis. Nat Mach Intell 5, 273-283 (2023). https://doi.org/10.1038/s42256-023-00625-5
(3) Westenberger, Jasmin Barbara. Automatische Gewebesegmentierung der Nieren und des Pankreas im Ganzkörper-MRT mittels Deep Learning. PhD thesis, Eberhard Karls Universität Tübingen, 2021. http://dx.doi.org/10.15496/publikation-63135
(4) Brachmann, Franz Xaver. Evaluation einer semi-automatischen Segmentierungsmethode für die Volumetrie des renalen Kortex, der Medulla und des gesamten Nierenparenchyms in nativen T1-gewichteten MR-Bildern. PhD thesis, Medizinischen Fakultät def Friedrich-Alexander-Universität Eralngen-Nürnberg, 2021. https://open.fau.de/items/cb6cd403-3178-4884-98c4-5098178658ba
(5) Maier-Hein, Lena, et al. "Metrics Reloaded: Recommendations for Image Analysis Validation" (Version 7). arXiv:2206.01653 [cs.CV]. 2022. Web. https://doi.org/10.48550/arXiv.2206.01653
(6) Maier-Hein, L., Eisenmann, M., Reinke, A. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat Commun 9, 5217 (2018). https://doi.org/10.1038/s41467-018-07619-7

Further comments

Further comments from the organizers.

N/A