Head and Neck Tumor Segmentation for MRI-Guided Applications Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Head and Neck Tumor Segmentation for MRI-Guided Applications Challenge

Challenge acronym

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

HNS-MRG 2023

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.

Radiation therapy (RT) is a cornerstone of cancer treatment for a wide variety of malignancies. Chief among the beneficiaries of RT as a treatment modality is head and neck cancer (HNC). Recent years have seen an increasing interest in MRI-guided RT planning. As opposed to more traditional CT-based RT planning, MRI-guided approaches afford superior soft tissue contrast and resolution, allow for functional imaging through special multiparametric sequences (e.g., diffusion-weighted imaging [DWI]), and permit daily adaptive RT through intra-therapy imaging using MRI-Linac devices (PMID: 28256898). Subsequently, improved treatment planning through MRI-guided adaptive RT approaches would maximize tumor destruction while minimizing side effects. Given the great potential for MRI-guided adaptive RT planning, it is anticipated that these technologies will transform clinical practice paradigms for HNC (PMID: 31632914).

The extensive data volume for MRI-guided HNC RT planning makes manual tumor segmentation by physicians — the current clinical standard — often impractical due to time constraints (PMID: 33763369). This is compounded by the fact that HNC tumors are among the most challenging structures for clinicians to segment (PMID: 27679540). Artificial intelligence (AI) approaches that leverage RT data to improve patient treatment have been an exceptional area of interest for the research community in recent years. The use of deep learning in particular has made significant strides in HNC tumor auto-segmentation (PMID: 36725406). These innovations have largely been driven by MICCAI public data challenges such as the HECKTOR Challenge (PMID: 35016077), and the SegRap Challenge (doi: https://doi.org/10.48550/arXiv.2312.09576). However, to-date, there exist no large publicly available AI-ready adaptive RT HNC datasets for public distribution. It stands to reason that community-driven AI innovations would be a remarkable asset to developing technologies for the clinical translation of MRI-guided RT.

In this proposal, we describe a data science challenge focused on the segmentation of HNC tumors for

MRI-guided adaptive RT applications. The challenge will be composed of 2 tasks focused on automated segmentation of tumor volumes on 1. pre-RT multiparametric MRI images and 2. mid-RT multiparametric MRI images. Our challenge is particularly unique as it seeks to ascertain whether incorporating prior timepoint data into auto-segmentation algorithms leads to enhanced performance for RT applications.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Segmentation, MRI, head and neck cancer, radiotherapy.

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.

None

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.

Based on conservative estimates from the first iteration of a similar data challenge (HECKTOR 2020, doi: https://doi.org/10.1016/j.media.2021.102336), we expect on the order of 64 teams registered and 18 final participants with individual submissions.

Publication and future plans

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

We plan to coordinate publications for each participant who is interested in publishing their results. Moreover, we plan to publish a subsequent comprehensive manuscript once the challenge is complete summarizing the challenge and applying meta analysis to the participants' data and results.

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 plan to use grand-challenge.org for hosting the challenge.

TASK 1: Segmentation on pre-RT imaging

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.

Radiation therapy (RT) is a cornerstone of cancer treatment for a wide variety of malignancies. Chief among the beneficiaries of RT as a treatment modality is head and neck cancer (HNC). Recent years have seen an increasing interest in MRI-guided RT planning. As opposed to more traditional CT-based RT planning, MRI-guided approaches afford superior soft tissue contrast and resolution, allow for functional imaging through special multiparametric sequences (e.g., diffusion-weighted imaging [DWI]), and permit daily adaptive RT through intra-therapy imaging using MRI-Linac devices (PMID: 28256898). Subsequently, improved treatment planning through MRI-guided adaptive RT approaches would maximize tumor destruction while minimizing side effects. Given the great potential for MRI-guided adaptive RT planning, it is anticipated that these technologies will transform clinical practice paradigms for HNC (PMID: 31632914).

The extensive data volume for MRI-guided HNC RT planning makes manual tumor segmentation by physicians - the current clinical standard - often impractical due to time constraints (PMID: 33763369). This is compounded by the fact that HNC tumors are among the most challenging structures for clinicians to segment (PMID: 27679540). Artificial intelligence (AI) approaches that leverage RT data to improve patient treatment have been an exceptional area of interest for the research community in recent years. The use of deep learning in particular has made significant strides in HNC tumor auto-segmentation (PMID: 36725406). These innovations have largely been driven by MICCAI public data challenges such as the HECKTOR Challenge (PMID: 35016077), and the SegRap Challenge (doi: https://doi.org/10.48550/arXiv.2312.09576). However, to-date, there exist no large publicly available AI-ready adaptive RT HNC datasets for public distribution. It stands to reason that community-driven AI innovations would be a remarkable asset to developing technologies for the clinical translation of MRI-guided RT.

In this proposal, we describe a data science challenge focused on the segmentation of HNC tumors for MRI-guided adaptive RT applications. The challenge will be composed of 2 tasks focused on automated segmentation of tumor volumes on 1. pre-RT multiparametric MRI images and 2. mid-RT multiparametric MRI images. Our challenge is particularly unique as it seeks to ascertain whether incorporating prior timepoint data into auto-segmentation algorithms leads to enhanced performance for RT applications.

Keywords

List the primary keywords that characterize the task.

Segmentation, MRI, head and neck cancer, radiotherapy.

ORGANIZATION

Organizers

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

The organizers stem from the University of Texas MD Anderson Cancer Center (MDA) Department of Radiation Oncology. The organizing team is currently composed of Kareem Wahid, PhD, Mohamed Naser, PhD, Cem Dede, MD, Serageldin Attia, MD, Dina El-Habashy, MD, Michael Rooney, MD, Abdallah Mohamed, MD, PhD, and Clifton D. Fuller, MD, PhD.

b) Provide information on the primary contact person.

Kareem A. Wahid, kawahid@mdanderson.org

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 a fixed submission deadline. To ensure lasting accessibility and reference, the results of the challenge leaderboard will be permanently displayed online for public viewing.

Challenge venue and platform

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

MICCAI 2024 (current submission).

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).

We have successfully completed the initial onboarding process for the Grand Challenge and established a dedicated website URL for our challenge at https://hntsmrg23.grand-challenge.org/. In the coming weeks, we will fully update the site with comprehensive details about the challenge and officially open the website to the public.

Participation policies

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

Only fully-automated segmentation of the test cases will be assessed. We expect the majority of the participants to use deep learning-based approaches.

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.

Participants are granted full flexibility in utilizing the provided training data to develop their models. While the incorporation of extra data sources, whether public or private, is allowed during the training process, it will be mandatory to disclose such usage in the methodology description. It is important to note that participants who incorporate additional data sources (i.e., data not from the challenge) will be exempt from any prize/ranking eligibility. For assessment purposes, we will conduct a comprehensive evaluation, with all final numerical results being prominently displayed on our official website and detailed in our publications.

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 (The University of Texas MD Anderson Cancer Center) may participate in the challenge but are not eligible for awards and the final official ranking.

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

Task 1 (pre-RT segmentation): $500 award to 1st place winner. Top 3 places will be specifically acknowledged on the challenge website.

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.

For each task, the announcement of the results and the winner will be made public. In addition, the top 3 ranking participants will be acknowledged on the challenge website. When participants submit their results on the test set through the challenge website to the organizers, they commit fully to the challenge. Consequently, their performance results (anonymized unless explicit consent is provided) will be incorporated into various forms of dissemination, including promotional materials, presentations, publications, and further analyses (more details are provided in sections below). These inclusions are subject to the discretion of the organizing committee.

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).

We will strongly encourage teams participating in the challenge to consider publishing their findings in the Lecture Notes in Computer Science (LNCS) proceedings, aligned with the MICCAI proceedings schedule and contingent upon peer-reviewed acceptance. Co-authorship on corresponding manuscripts will be up to the discretion of the teams but we strongly encourage any team members directly involved in algorithmic development or analysis to be listed as co-authors. Additionally, participants are free to present their findings in other publications (e.g., preprint servers), provided they reference the challenge's overview paper. In such cases, no embargo restrictions will be imposed.

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.

Teams taking part in the challenge are required to submit their segmentation results through the grand-challenge.org platform. They will be required to package their algorithms in a docker container that will be run through an evaluation container created by the organizers that will apply their method to the private test data. Participants will only ever have access to the training data; test data will not be made publicly available until after the challenge completion (see Further Analysis section). Comprehensive guidelines for submission will be thoroughly outlined on the challenge website and communicated directly to participants via email.

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 in the challenge must submit their segmentation algorithm in a docker container via the grand-challenge.org platform. A URL link with comprehensive instructions for submission will be provided to all teams. Each participating team is permitted to submit up to three attempts to assess their algorithms effectively. For the purpose of official rankings, only the highest-performing submission from each team will be considered. We require teams to clearly identify and elaborate on each attempt in their accompanying papers. Feedback prior to the submission deadline will be limited to notifications of upload errors such as empty segmentation masks. Additionally, an online evaluation tool will be available on a selection of training cases, assisting participants in validating their segmentation task outputs. The provision for multiple submissions is a standard practice in challenges. It serves to facilitate the comparison of various methodologies and to verify the absence of submission errors. To prevent format-related errors or unusually low performance (indicative of potential issues like empty segmentation masks), we will issue alerts. While multiple submissions can potentially lead to overfitting, limiting submissions to three per team strikes a balance between ample exploration and the risk of overfitting.

Challenge schedule

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

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

Registration period: May 1 - June 15, 2024
Release date of training cases: June 15, 2024
Release date of test cases: August 15, 2024
Submission dates: August 15 - September 15, 2024
Release date of results: September 20, 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).

Ethics approval was obtained from the University of Texas MD Anderson Cancer Center Institutional Review Board with protocol number RCR03-0800. This is a retrospective data collection protocol with a waiver of informed consent.

Data usage agreement

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

Examples:

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

CC BY (Attribution).

Code availability

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

We will create a publicly accessible GitHub repository containing Python scripts of our evaluation code and corresponding Jupyter Notebooks that will allow for users to easily understand how the challenge will evaluate their submissions. A link to the full code and corresponding documentation will be added to the challenge webpage. Using assistance from the Grand Challenge we have set up a GitHub page that we will soon populate with more challenge-specific information:
https://github.com/DIAGNijmegen/HNTSMRG23-challenge-pack/tree/main.

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

All participating teams will be strongly encouraged to share Github repositories of their algorithmic approach (e.g., model weights, network architectures) and any associated analysis code in their corresponding manuscripts.

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.

This challenge is funded by NIH administrative supplement grants 3R01DE028290-04S2 and 3R01CA257814-02S1. Only the challenge organizers (i.e., MDA Fuller Lab) will have access to the ground-truth segmentations (labels) for 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

Diagnosis, Intervention-planning, research.

Task category(ies)

State the task category(ies)

Examples:

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

Segmentation, detection.

Cohorts

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

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

Patients with HNC presenting to RT planning clinics. Outputs from the automated segmentation algorithms could be directly used for treatment planning interventions (e.g., RT dose intensification/deintensification).

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

Patients with histologically proven HNC who underwent RT.

Imaging modality(ies)

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

Magnetic resonance imaging. Specifically, T2-weighted (T2-w) anatomical sequences and apparent diffusion coefficient maps (ADC) derived from DWI scans.

Context information

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

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

DICOM meta-data including acquisition parameters (see Item 21) will be provided but are not expected to be used in algorithms.

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

As this is an expected end-to-end segmentation task, no specific patient-related clinical information will be provided directly to users.

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.

Data originates from multiparametric MRI of the head and neck region. All patients are immobilized using a thermoplastic mask and acquisitions are taken under a consistent imaging protocol so that images (i.e., T2-w and DWI) are standardized and implicitly co-registered.

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.

Primary gross tumor volumes (abbreviated GTVp) - at most 1 per patient (can be 0), and metastatic lymph nodes (abbreviated GTVn) - variable number per patient (can be 0).

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)).

Find tumor segmentation algorithm that processes multiparametric MRI images of a certain size with high accuracy as per standard geometric evaluation measures.

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).

Multiparametric MRI imaging was performed on a Siemens Aera scanner with a magnetic field strength of 1.5 T and standardized acquisition parameters (see below).

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

T2-w images had a repetition time of 4800 ms, echo time of 80ms, echo train length of 15, flip angle of 180 degrees, slice thickness of 2mm, and in-plane resolution of 0.5. DWI images had a repetition time of 5400 ms, echo time of 50 ms, echo train length of 15, flip angle of 120 degrees, slice thickness of 4mm, and in-plane resolution of 2mm. ADC parametric maps were derived from DWI sequences through a proprietary Siemens algorithm (Munich, Germany) using a monoexponential model, as per a previous publication (PMID: 34765748). All images at a given timepoint were collected with patients immobilized in a thermoplastic mask. Therefore, images at a given timepoint are implicitly coregistered.

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 University of Texas MD Anderson Cancer Center.

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).

Raw images (T2-weighted, DWI) are automatically extracted from a centralized institutional imaging repository (Evercore). Annotation (i.e., segmentation) characteristics are provided below.

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.

Task 1 (pre-RT): Training and test cases both represent one pre-therapy 3D T2-weighted MRI volume with a co-registered 3D ADC map. Training cases will contain segmentation of the annotated ground truth tumors. Test cases will not contain any annotations. The labels will have 3 values: background = 0, GTVp = 1, GTVn = 2 (in the case of multiple lymph nodes they will be concatenated into one label). The goal is to successfully predict the tumor segmentation on the pre-RT images. Participants will be free to use any combination of images (pre-RT T2, pre-RT ADC) to develop auto-segmentation algorithms. Subsequently, participants will submit algorithms for blinded evaluation.

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

The total number of cases is approximately 200 (final case number TBD based on final rigorous QA process). The total number of training cases will be approximately 140. No specific validation cases will be provided but the training set can be split for any manner with cross-validation. The total number of test cases will be approximately 60.

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

In determining the total number of cases and the specific proportions for training and testing, we closely aligned our approach with the data distribution framework used in a recent, relevant MICCAI challenge - SegRap 2023 (doi: https://doi.org/10.48550/arXiv.2312.09576). This prior challenge focused on head and neck target structure segmentation, closely paralleling the scope of our current endeavor. The decision to mirror the 140/60 train/test split from SegRap 2023 was strategic. This specific proportion was selected based on its demonstrated effectiveness in providing a balanced yet comprehensive dataset for both training and testing. By employing a similar distribution, we aim to ensure that our model is trained on a robust and representative sample, but allowing for reliability in our testing phase.

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.

Data from the training and test sets are representative of real-world cases from a large cancer institute treating HNC. Since it is well known that segmentation algorithm performance is often correlated to structure volume/size, training and test sets will be partitioned such as to contain similar distributions based on TNM 8th edition staging criteria to ensure a representative split. Moreover, cases will also be split to ensure a similar distribution of patients experiencing mid-RT response (relevant for Task 2).

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.

Multiple physician expert observers (n = 3 to 4) have independently segmented GTVp and GTVn structures for all cases (pre-RT and mid-RT) based on MRI images provided (T2-weighted, ADC). Based on recent literature from our group (PMID: 36761036), a minimum of 3 annotators is suggested to yield acceptable segmentations when combined via the simultaneous truth and performance level estimation algorithm (STAPLE) in these structures, therefore we have collected independent segmentations from 3 annotators for each structure.

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.

All observers used the same segmentation research software commonly used by our institution (Velocity AI, version 3.0) to visualize images and generate segmentations. Observers were instructed to segment the GTVp and GTVn structures on the T2-weighted image (primary image) with the ADC image made available to the observer during the segmentation process. Observers had access to the patient's chart and clinical history if needed via our electronic health system (EPIC EMR).

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 annotators were medical doctors with at least 2 years of experience in head and neck cancer segmentation. Final verification of segmentation quality was performed by experienced radiation oncology faculty members with greater than 10 years of experience.

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.

Segmentations will be combined via the STAPLE algorithm to yield the final ground truth segmentation for each case. Of note, we have previously shown that a small number of observers can approximate expert consensus benchmark values using the STAPLE algorithm (PMID: 36761036). Final segmentations will be checked by an experienced independent physician observer and modified if needed.

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.

Anonymized DICOM files (images and structure files) will be converted to NIfTI format for ease of use by participants. Since head and neck images contain potentially discernible facial features when using 3D reconstructions, facial tissue will be anonymized via the mri_reface tool which has recently been shown to work exceptionally well for mpMRI data (PMID: 37269958) before data sharing. All images will be quality assessed to ensure segmentations are not affected by the anonymization process.

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 largest source of error naturally emerges from differences in observer segmentations. We hope to mitigate this by combining segmentations via the STAPLE algorithm which we have shown can yield acceptable segmentations given a minimal number of observer inputs (PMID: 37269958). From preliminary calculations (note: annotations have not been completed yet) average inter-observer DSC variation is comparable to what has been observed for literature in head and neck tumor segmentation based on MRI (GTVp DSC = 0.77, GTVn DSC = 0.67, PMID: 35124134). The final interobserver variation will be calculated across all observers in our dataset and reported in our summary manuscript.

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

Only segmentation data will be collected as annotations so we do not anticipate any other sources of errors as our image acquisitions have all been standardized with rigorous quality assurance.

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)

We will use a metric termed aggregated Dice Similarity Coefficient (DSCagg), which was initially employed by Andrearczyk et al. for the segmentation task of the 2022 edition of the HECKTOR Challenge (doi: https://doi.org/10.1007/978-3-031-27420-6_1). Specifically, this metric is defined as 2iAi BiiAi +Bi, where Ai and Bi are the ground truth and predicted segmentation for image i, where i spans the entire test set. Conceptually, the 2022 edition of the HECKTOR Challenge had similar segmentation outputs (GTVp and GTVn for head and neck cancer patients) as our proposed challenge, so we deem this an appropriate metric.

For both GTVp and GTVn, we will accumulate the intersections and unions between GTVs and the respective predicted volumes across all images. Note that the intersection and union in an image can be zero for both GTVp and GTVn, as some cases may not contain GTVn or GTVp. Ultimately, we will divide the aggregated intersection by the aggregated union, both for GTVp and GTVn, and will compute the average of these two aggregated indices. In other words, DSCagg will be computed separately for GTVp and GTVn on the test set, and the average of the two will be used for the final ranking. This choice was made to give equal importance to the two GTV types since both primary tumors and metastatic lymph nodes serve as target structures for RT and should be given the full treatment dose.

In addition to DSCagg, we will also incorporate the mean average precision (mAP) as defined by Maier-Hein et al. (doi: https://doi.org/10.48550/arXiv.2206.01653), providing a balanced assessment of both segmentation accuracy and the ability to detect relevant structures, irrespective of their size. The average precision will be computed for each class separately (GTVp and GTVn) and the average between these two values will yield the final mAP.

For the final ranking, participants will be ranked on an equally weighted score of DSCagg and mAP.

The metrics will be calculated for task 1 (pre-RT segmentations) and task 2 (mid-RT segmentations) separately.

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

Since the presence of GTVp and GTVn will not be consistent across all cases, the proposed DSCagg metric is well-suited for this task. Unlike the conventional volumetric DSC, which may be disproportionately affected by a single false negative result (yielding a DSC of 0), this metric is designed to accommodate such occurrences more effectively.

mAP was selected because it effectively balances the need for high sensitivity in detecting small or subtle features, crucial in early disease diagnosis, with the specificity required to avoid over-treatment or unnecessary interventions. By accounting for varying levels of prediction confidence, mAP offers a nuanced assessment of model performance, aligning with clinical decision-making processes where both the detection of conditions and the minimization of false alarms are critical.

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.

DSCagg will be computed individually for GTVp and GTVn and the average of the two will be used for the final challenge ranking (similar to HECKTOR 2022). Additionally, a similar process will be applied for average precision for GTVp and GTVn leading to calculation of mAP. For the final ranking, participants will be ranked on an equally weighted score of DSCagg and mAP.

The metrics will be calculated for Task 1 (pre-RT segmentations) and Task 2 (mid-RT segmentations) separately.

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

If we receive missing results for a test case, we will consider the model predicted no structures for that case (i.e., no GTVp and GTVn). The imputed prediction will be used for calculating the average metric. In the scenario that the majority of the submitted cases contain empty segmentations, we will send a warning to users before submission (in this case flagging the possibility of an upload error).

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

Our proposed evaluation metric, incorporating both DSCagg and mAP is designed to offer a comprehensive assessment of segmentation and detection capabilities. Geometric volume overlap, a standard in segmentation studies, ensures accuracy in quantifying how well the segmented structures match with the target areas crucial for radiotherapy. The addition of mAP addresses the critical aspect of detecting relevant structures with varying sizes and appearances, ensuring models are not only precise in segmentation but also effective in identifying key anatomical features. This balanced approach aligns with clinical needs, prioritizing both the accurate delineation and reliable detection of structures, which are essential for informed decision-making in treatment planning.

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.

For each task, bootstrap resampling will be performed to calculate test set DSCagg and mAP distributions for use in statistical analyses, and also allow for determining ranking robustness. To statistically compare algorithm results, we will employ paired non-parametric tests (e.g., Wilcoxon signed rank test) and apply a Bonferroni correction to correct for multiple comparisons. The primary objective of these tests is to evaluate if there is a significant disparity between the submitted algorithms. The statistical tests will be performed in a stepwise fashion by rank to determine when performance scores start to significantly diverge. Additionally, comparisons will be made against a control group (distribution of interobserver variability for a subset of cases). Subanalyses will be performed to determine if there were any significant differences between test cases from different anatomic sites (i.e., base of tongue vs. tonsil) and stage (early stage vs. late stage). All analysis will be conducted in Python with relevant computational/statistical libraries such as numpy, scipy, and statsmodels.

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

DSCagg and mAP data from participants will likely be non-normally distributed (will be formally verified with the Shapiro-Wilk test) hence the selection of non-parametric approaches.

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.

For each task, we will conduct an ensemble analysis of submitted algorithms using a consensus segmentation method (i.e., super-ensemble). For comparison purposes, STAPLE, majority voting, and SIMPLE (PMID: 20667809) will be employed as consensus segmentation methods. The consensus segmentations will incorporate predictions from 1. all participants and 2. from the top three performers. We will determine how the performance of this consensus method compares to the top teams using previously described non-parametric tests. We will also examine the relationship between tumor size and predictive performance by using the Spearman correlation coefficient. Furthermore, we plan to qualitatively analyze the locations of segmentation false positives to determine whether these inaccuracies predominantly occur near the primary tumor or in other areas of the head and neck region. In addition, we will generate cumulative probability and entropy maps by combining the binary mask outputs of each segmentation algorithm to determine where algorithms generally disagreed and agreed in a qualitative manner. These maps could be used in future studies for voxel-wise uncertainty experiments, and compared to ground truth probability maps of human interobserver variability. Finally, should the challenge attract a substantial number of participating teams, we plan to conduct an ordinary least squares regression analysis. This analysis will utilize demographic factors collected from the participant intake forms - such as the team's background (academic, industry, clinical), the number of team members, and the type of algorithm implemented (e.g., U-net, others) - as fixed input variables. These will be evaluated against the DSCagg and mAP from the participant test sets. The objective of this analysis is to identify any significant factors that contribute to a team's success in this challenge, thereby providing valuable insights into the elements that drive effective performance.

All meta-analytic data (e.g., participant STAPLE consensus and probability maps), along with training and test data, will be made available for public reuse through The Cancer Imaging Archive. This will be accompanied by a detailed data descriptor, which we plan to submit to Nature Scientific Data or another appropriate publisher.

TASK 2: Segmentation on mid-RT imaging

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.

Same as task 1.

Keywords

List the primary keywords that characterize the task.

Same as task 1.

ORGANIZATION

Organizers

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

Same as task 1.

b) Provide information on the primary contact person.

Same as task 1.

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

Same as task 1.

Challenge venue and platform

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

Same as task 1.

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

Same as task 1.

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

Same as task 1.

Participation policies

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

Same as task 1.

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.

Same as task 1.

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.

Same as task 1.

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

Task 2 (mid-RT segmentation): $500 award to 1st place winner. Top 3 places will be specifically acknowledged on the challenge website.

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.

Same as task 1.

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).

Same as task 1.

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.

Same as task 1.

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.

Same as task 1.

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

Same as task 1.

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).

Same as task 1.

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)

Same as task 1.

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.

Same as task 1.

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

Same as task 1.

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.

Same as task 1.

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

Same as task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Segmentation
- Tracking

Same as task 1.

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.

Same as task 1.

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

Same as task 1.

Imaging modality(ies)

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

Same as task 1.

Context information

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

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

Same as task 1.

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

Same as task 1.

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.

Same as task 1.

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.

Same as task 1.

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)).

Same as task 1.

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).

Same as task 1.

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

Same as task 1.

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.

Same as task 1.

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).

Same as task 1.

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.

Task 2 (mid-RT): Essentially, the patients will be the same as Task 1 but with an additional mid-RT scan. Training and test cases both represent one pre-therapy 3D T2-weighted MRI volume with a co-registered 3D ADC map AND one mid-therapy 3D T2-weighted MRI volume with a co-registered 3D ADC map. Training cases will contain segmentation of the annotated ground truth tumors. Test cases will not contain any annotations. The labels will have 3 values: background = 0, GTVp = 1, GTVn = 2 (in the case of multiple lymph nodes they will be concatenated into one label). The goal is to successfully predict the tumor segmentation on the mid-RT images. Participants will be free to use any combination of images (pre-RT T2, pre-RT ADC, mid-RT T2, mid-RT ADC) to develop auto-segmentation algorithms. Subsequently, participants will submit algorithms for blinded evaluation.

Note: Participants who decide to partake in both tasks will be asked to not use any mid-RT data from Task 2 for Task 1 algorithms. Should they use any mid-RT data for Task 1, they will be ineligible for prizes/ranking.

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

Same as task 1.

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

Same as task 1.

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.

Same as task 1.

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.

Same as task 1.

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.

Same as task 1.

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.

Same as task 1.

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.

Same as task 1.

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.

Same as task 1.

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.

Same as task 1.

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

Same as task 1.

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)

Same as task 1.

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

Same as task 1.

Ranking method(s)

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

Same as task 1.

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

Same as task 1.

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

Same as task 1.

Statistical analyses

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

- description of the missing data handling,
- details about the assessment of variability of rankings,

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

Same as task 1.

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

Same as task 1.

Further analyses

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

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

Same as task 1.

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.

Cardenas, Carlos E., Sanne E. Blinde, Abdallah S. R. Mohamed, Sweet Ping Ng, Cornelis Raaijmakers, Marielle Philippens, Alexis Kotte, et al. 2022. Comprehensive Quantitative Evaluation of Variability in Magnetic Resonance-Guided Delineation of Oropharyngeal Gross Tumor Volumes and High-Risk Clinical Target Volumes: An R-IDEAL Stage 0 Prospective Study. International Journal of Radiation Oncology, Biology, Physics 113 (2): 426 to 36.

Hindocha, S., K. Zucker, R. Jena, K. Banfill, K. Mackay, G. Price, D. Pudney, J. Wang, and A. Taylor. 2023. Artificial Intelligence for Radiotherapy Auto-Contouring: Current Use, Perceptions of and Barriers to Implementation. Clinical Oncology 35 (4): 219 to 26.

Kiser, Kendall J., Benjamin D. Smith, Jihong Wang, and Clifton D. Fuller. 2019. Après Mois, Le Déluge: Preparing for the Coming Data Flood in the MRI-Guided Radiotherapy Era. Frontiers in Oncology 9 (September): 983.

Langerak, Thomas Robin, Uulke A. van der Heide, Alexis N. T. J. Kotte, Max A. Viergever, Marco van Vulpen, and Josien P. W. Pluim. 2010. Label Fusion in Atlas-Based Segmentation Using a Selective and Iterative Method for Performance Level Estimation (SIMPLE). IEEE Transactions on Medical Imaging 29 (12): 2000 to 2008.

Lin, Diana, Kareem A. Wahid, Benjamin E. Nelms, Renjie He, Mohammed A. Naser, Simon Duke, Michael V. Sherer, et al. 2023. E Pluribus Unum: Prospective Acceptability Benchmarking from the Contouring Collaborative for Consensus in Radiation Oncology Crowdsourced Initiative for Multiobserver Segmentation. Journal of Medical Imaging (Bellingham, Wash.) 10 (Suppl 1): S11903.

Luo, Xiangde, Jia Fu, Yunxin Zhong, Shuolin Liu, Bing Han, Mehdi Astaraki, Simone Bendazzoli, et al. 2023. SegRap2023: A Benchmark of Organs-at-Risk and Gross Tumor Volume Segmentation for Radiotherapy Planning of Nasopharyngeal Carcinoma. arXiv [eess.IV]. arXiv. http://arxiv.org/abs/2312.09576.

Maier-Hein, Lena, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, et al. 2022. Metrics Reloaded: Recommendations for Image Analysis Validation. arXiv [cs.CV]. arXiv. http://arxiv.org/abs/2206.01653.

Pollard, Julianne M., Zhifei Wen, Ramaswamy Sadagopan, Jihong Wang, and Geoffrey S. Ibbott. 2017. The Future of Image-Guided Radiotherapy Will Be MR Guided. The British Journal of Radiology 90 (1073): 20160667.

Schwarz, Christopher G., Walter K. Kremers, Arvin Arani, Marios Savvides, Robert I. Reid, Jeffrey L. Gunter, Matthew L. Senjem, et al. 2023. A Face-off of MRI Research Sequences by Their Need for de-Facing. NeuroImage 276 (August): 120199.

Segedin, Barbara, and Primoz Petric. 2016. Uncertainties in Target Volume Delineation in Radiotherapy - Are They Relevant and What Can We Do about Them? Radiology and Oncology 50 (3): 254 to 62.

Thorwarth, Daniela, and Daniel A. Low. 2021. Technical Challenges of Real-Time Adaptive MR-Guided Radiotherapy. Frontiers in Oncology 11 (March): 634507.

Wahid, Kareem A., Sara Ahmed, Renjie He, Lisanne V. van Dijk, Jonas Teuwen, Brigid A. McDonald, Vivian Salama, et al. 2022. Evaluation of Deep Learning-Based Multiparametric MRI Oropharyngeal Primary Tumor Auto-Segmentation and Investigation of Input Channel Effects: Results from a Prospective Imaging Registry. Clinical and Translational Radiation Oncology 32 (January): 6 to 14.

Further comments

Further comments from the organizers.

We have prepared a response to reviewers' document where we indicate the changes made to our uploaded design document to address comments, and where the changes can be found.