

# DoseRAD2026: Real-time dose calculation in radiotherapy: Structured description of the challenge design

## CHALLENGE ORGANIZATION

### Title

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

DoseRAD2026: Real-time dose calculation in radiotherapy

### Challenge acronym

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

DoseRAD2026

### 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 modern clinical radiotherapy treatment planning systems, dose calculation times with Monte-Carlo simulations providing clinically acceptable dose accuracy can range from 13 to 70 seconds for GPU-based methods [1,2] and up to several minutes for CPU based approaches, which are still widespread in clinics [1-4]. The acceleration of radiotherapy dose calculation brings obvious benefits to treatment planning, online plan adaptation, quality assurance, and may ultimately foster real-time dose-guided radiotherapy, where radiation delivery is continuously tailored to the patient's moving anatomy via a real-time feedback loop [5]. Recent developments suggest that AI-based dose calculation directly on MRI may offer a pragmatic solution to close this loop by updating the delivered dose distribution in real-time [6,7]. It remains to be seen what the optimal strategy is to achieve (near-)real-time dose calculation on 3D anatomy to combine accuracy and calculation speed. The DoseRAD2026 Grand Challenge aims to benchmark state-of-the-art methods for real-time dose calculations directly on images (both CT and MRI), in photon and proton radiotherapy.

We propose to organize the DoseRAD2026 challenge, building on our previous experiences with the TrackRAD2025 [8] and SynthRAD2025 MICCAI challenges [9]. The challenge will feature four tasks for real-time photon and proton dose calculation using either CT or MRI images as inputs, and we plan to build on the existing SynthRAD2025 dataset which features paired MRI-CT images to generate ground truth Monte-Carlo based dose distributions for training and testing. We seek a broad dataset of photon beams generated from arbitrary multi-leaf collimator (MLC) apertures, and proton beams for pencil beam scanning. Ground truth data will be obtained from lengthy but accurate full physics modelling via Monte Carlo simulation of particle transport. Four tasks will address the latest technical developments in radiation therapy:

1) Photon therapy dose calculation on CT images is crucial for treatment planning of the delivery of intensity

modulated photon therapy using volumetric modulated arc therapy (VMAT), which is used for the vast majority of patients.

2) Photon therapy dose calculation on MRI images for cutting edge MRI-linacs which allow in-room online adaptation of radiation plans based on high soft tissue contrast MRIs.

3) Proton therapy dose calculation on CT images is required for treatment planning of high precision proton beams, which deposit dose more locally than photon beams and require advanced proton acceleration beamlines and gantries.

4) Proton therapy dose calculation on MRI images supports both MRI-only proton therapy radiation planning and prototype development of in-room MRI guidance for proton therapy, similarly to MRI-linacs.

Using the paired SynthRAD2025 dataset allows us to address these four tasks with a consistent dataset.

The objective of the challenge will be fast and accurate dose calculation of individual radiation beams on either CT or MRI images of patients. Beams will either be defined by a photon linear accelerator's MLC or a proton therapy system's pencil beam scanning parameters. The algorithms will be provided with the CT or MRI and beam parameters, and need to output a beam-specific radiation dose distribution in 3D.

The organization of this challenge will be supported by an international group of experts in photon therapy, proton therapy and MRI-linacs (the Netherlands, Switzerland, Germany, Sweden) and will be hosted on the open access grand-challenge.org platform. The top winning teams will earn prize money. The challenge is expected to launch in March 2026, with the testing phase in July 2026, and result announcement at MICCAI in Abu Dhabi in October 2026, depending on acceptance at MICCAI in early 2026.

## References

- [1] Aland T, et al. (2019) Accuracy and efficiency of graphics processing unit (GPU) based Acuros XB dose calculation within the Varian Eclipse treatment planning system. *Med Dosim*, 44(3):219-225.
- [2] Feygelman, V., et al. (2022). Maintaining dosimetric quality when switching to a Monte Carlo dose engine for head and neck volumetricmodulated arc therapy planning. *Journal of Applied Clinical Medical Physics*, 23(5), e13572.
- [3] Fracchiolla, F., et al. (2021). Clinical validation of a GPU-based Monte Carlo dose engine of a commercial treatment planning system for pencil beam scanning proton therapy. *Physica Medica*, 88, 226-234.
- [4] Cusumano, D. et al. (2024). Evaluation of clinical parallel workflow in online adaptive MR-guided radiotherapy: A detailed assessment of treatment session times. *Technical Innovations & Patient Support in Radiation Oncology*, 29, 100236.
- [5] Keall, Paul J., et al. "Critical Review: Real-Time Dose-Guided Radiation Therapy." *International Journal of Radiation Oncology\* Biology\* Physics* (2025).
- [6] Li, Muheng, et al. "A proof-of-concept study of direct magnetic resonance imaging-based proton dose calculation for brain tumors via neural networks with Monte Carlo-comparable accuracy." *Physics and Imaging in Radiation Oncology* (2025): 100806.
- [7] Xiao, Fan, et al. "Deep learningbased syntheticCTfree photon dose calculation in MRguided radiotherapy:

A proof-of-concept study." Medical Physics 52.11 (2025): e70106.

[8] Wang Y, et al. (2025) TrackRAD2025 challenge dataset: real-time tumor tracking for MRI-guided radiotherapy. Med Phys 52, e17964.

[9] Thummerer A, et al (2025) SynthRAD2025 Grand Challenge dataset: Generating synthetic CTs for radiotherapy from head to abdomen. Med Phys 52, e17981.

### Challenge keywords

List the primary keywords that characterize the challenge.

medical imaging, computed tomography, magnetic resonance imaging, MRI-linac, MRI-guided radiotherapy, real-time photon and proton dose calculation, real-time adaptive radiotherapy

### Year

2026

### Novelty of the challenge

Briefly describe the novelty of the challenge.

DoseRAD2026 will provide the first publicly available radiotherapy beam dose database with paired CT and MRI images. Carefully curated CT and MRI pairs will provide beam-wise dose labels directly on both the CT and MRI images. This will allow developing methods for real-time dose calculation at the radiotherapy beam level (the basis for any full plan optimization), which can be applied to either conventional CT images or real-time in-room images such as MRI. Furthermore, DoseRAD2026 will provide data for both photon-based radiotherapy and proton-based radiotherapy. For photon radiotherapy a beam refers to the radiation field shaped by a given MLC aperture, while for proton radiotherapy a beam refers to a proton pencil beam of a given energy and shape.

The challenge aims at covering the latest technological innovations in radiation therapy, starting with:

(i) the delivery of photon therapy using VMAT, where radiation is continuously delivered during linear accelerator gantry rotation. This requires sophisticated radiation treatment delivery planning using CT images (allowing electron density estimation) for radiation transport calculations.

(ii) The second area is MRI guided radiotherapy using hybrid MRI-linacs, where the acquisition of MRI images in-room allows superior soft tissue contrast for online radiation plan adaptation to ensure maximum personalization. Since MRI does not provide electron density directly, novel dose calculation approaches are required using either synthetic CTs or calculation directly on MRI.

(iii) Proton therapy is an alternative to photon therapy requiring highly specialized facilities bringing advantages due to the localized energy deposition in the proton Bragg peak, which allows sparing of critical structures but also requires very accurate dose calculation on CT images.

(iv) Finally, cutting edge research aims at developing prototype MRI-proton systems combining the superior soft tissue contrast of MRI with the precision of proton therapy, and where dose calculation on MRI is again required.

Algorithms submitted to the challenge will be ranked on both their accuracy, using metrics relevant for

radiotherapy dose distributions at individual beam level and full plan level (the sum of many beams forming the final radiotherapy dose distribution), and their real-time capabilities. The challenge will identify the next generation methods allowing fast dose calculation for real-time adaptive radiotherapy based on in-room imaging.

### Task description and application scenarios

Briefly describe the application scenarios for the tasks in the challenge.

The objective of DoseRAD2026 will be real-time dose calculation on both conventional CT images and in-room MRI images. Algorithms will be provided with an input volume (CT for tasks 1 and 3, and MRI for tasks 2 and 4) and will be required to compute photon doses (tasks 1 and 2) and proton doses (tasks 3 and 4). Algorithms will additionally have the MLC shape for photon beams (tasks 1 and 2) and the pencil beam shape and energy for protons (tasks 3 and 4).

DoseRAD2026's main application scenario is the real-time calculation of radiotherapy dose. This is a crucial component of real-time adaptive radiotherapy, allowing a continuous optimization of the dose as it accumulates.

Summarizing, we will have four tasks:

Task 1: Photon therapy dose calculation on CT images

Task 2: Photon therapy dose calculation on MRI images

Task 3: Proton therapy dose calculation on CT images

Task 4: Proton therapy dose calculation on MRI images

For future clinical implementation, expert approval will remain indispensable for any AI-based tool, especially in radiotherapy. This is consistent with current radiotherapy practice, where expert review of AI outputs of automatic planning (different than dose calculation) or organ segmentation is required before treatment delivery. AI-based methods from DoseRAD2026 are intended to accelerate this process, not to replace clinical oversight.

As the first challenge dedicated to AI-based beam-level dose calculation, DoseRAD2026 prioritizes establishing standardized benchmarks and fostering an open research ecosystem. To this end, the Monte Carlo dose calculation code and all planning tools based on the open-source matRad software will be made publicly available, lowering the barrier for researchers and clinics to reproduce and build upon the challenge outcomes. While detailed clinical deployment considerations such as system installation and staff training are beyond the current scope of a challenge proposal, we believe that open-source availability of both the reference dose engine and planning tools provides a solid foundation for future clinical translation efforts.

## FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

### Workshop

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

The challenge is currently not associated to a workshop, but if a suitable match is identified we would be happy to coordinate with a workshop.

### Duration

How long does the challenge take?

2 Hours

In case you selected half or full day, please explain why you need a long slot for your challenge.

N/A

### Expected number of participants

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

We base our estimate of the expected number of teams based on previous radiotherapy oriented challenges (SynthRAD2023/2025 and TrackRAD2025).

We are expecting about 30 teams of five participants, and thus about 150 participants to the online challenge.

At MICCAI 2026, we expect at least to draw participants from the four best-ranked teams per task (1 winner per task), other participating teams, and other interested attendees. Additionally two radiotherapy dose calculation companies sponsoring the prizes will participate in the event.

We will contact the teams that participated at SynthRAD2025 and TrackRAD2025 given their interest in radiotherapy image processing. We plan on announcing DoseRAD2026 at ESTRO 2026 and PTCOG 2026 to address interested participants in the photon and proton radiotherapy communities, respectively.

### Publication and future plans

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

A publication on the dataset will be released as a Medical Physics Dataset Article along with the start of the challenge. After the challenge is run, we plan to coordinate a publication (overview paper) describing the challenge results as a paper in a peer-reviewed journal that summarizes the results and outcomes. We will follow the template from our cousin radiotherapy challenges SynthRAD2023 (dataset paper: <https://aapm.onlinelibrary.wiley.com/doi/full/10.1002/mp.16529>, challenge report: <https://www.sciencedirect.com/science/article/pii/S1361841524002019>), SynthRAD2025 (dataset paper: <https://aapm.onlinelibrary.wiley.com/doi/full/10.1002/mp.17981>, challenge paper in preparation) and TrackRAD2025 (dataset paper: <https://aapm.onlinelibrary.wiley.com/doi/full/10.1002/mp.17964>, challenge paper submitted). The leaderboard will remain open after the challenge for new submissions.

### MICCAI LNCS proceedings

Indicate if you want to offer MICCAI Springer LNCS proceedings to the participants. Publishing a proceedings volume is optional and at the discretion of each challenge's organizers. At a minimum, organizers must ensure that a description of each participant's submission is publicly available. Organizers who wish to publish MICCAI Springer LNCS proceedings must adhere to the MICCAI Satellite events publication process.

No

### Collaboration with European Society of Radiology (ESR)

In collaboration with European Society of Radiology (ESR), we announce special clinical interest topics with associated clinicians who can help with the preparation of the proposals; the best 3 challenge proposals on these topics will get

the opportunity to present their challenges at the European Congress of Radiology (ECR) 2027 in a special session. If you want to organize a challenge in collaboration with ESR on one of these topics, please reach out to the MICCAI Challenges Team ([miccai-challenges-2026@dkfz-heidelberg.de](mailto:miccai-challenges-2026@dkfz-heidelberg.de)) and we will put you in contact with the corresponding clinician.

Challenge in collaboration with ESR. Ticking 'Yes' implies that the challenge has been prepared in collaboration with the clinical contact point.

No

### **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 challenge is online using the grand challenge platform. For training purposes, the computing environment is left to the participants. Example docker containers with evaluation scripts and baseline algorithms will be made public. The algorithms will run on the [grand-challenge.org](https://grand-challenge.org) platform using AWS g5 instances using a single GPU with 24 GB GPU RAM, 16 CPUs, and 64 GB RAM. At the MICCAI 2026 DoseRAD2026 workshop support would be needed for projectors, speakers, and microphones. We wish to hold a hybrid workshop with all the presenters ideally available on-site.

## TASK 1: Photon dose calculation on CT images

### 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.

Up to 50% of cancer patients will undergo radiation therapy, and the vast majority of these treatments will be delivered at medical electron linear accelerators producing photon beams. Radiation is delivered under continuous gantry rotation via beam modulation using a multi-leaf collimator (MLC), a modality called volumetric modulated arc therapy (VMAT). Prior to radiation delivery, patients undergo treatment planning, where the radiation shape and intensity from each direction (modelled as discrete beams) is optimized using the knowledge of each beam's dose distribution in the patient. This requires forward calculation of the radiation dose distribution in the patient, using a CT image to allow estimation of the electron density in each voxel, a quantity required to calculate photon transport and interactions. Once the dose of each beam is known, relative contributions can be optimized using a dedicated process aiming at maximizing dose in the tumor and minimizing dose in healthy tissue and organs. The speed at which this process can be executed is highly dependent on the time required to calculate the dose distribution of each beam. Currently the highest accuracy is achieved via Monte Carlo simulation based on full physics modelling of the transport of individual photons, which is highly time consuming.

The goal of this task is to foster fast photon therapy beam dose calculation on CT images for VMAT. We will provide CT images, MLC openings and corresponding ground truth dose distributions from VMAT beams from lengthy and accurate Monte Carlo simulations for training, and a hold out set for testing. Data will be from the challenging thoracic and abdominal sites where very heterogeneous density distributions are found due to the lungs and air pockets in the bowels.

#### Keywords

List the primary keywords that characterize the task.

medical imaging, computed tomography, real-time photon therapy dose calculation

### ORGANIZATION

#### Organizers

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

Group 1 - Data (collection, labeling, preparation)

Fan Xiao (Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany)

George Dedes (Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany)

Adrian Thummerer (Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany)

Miguel Palacios (Amsterdam UMC, Amsterdam, the Netherlands)

**Group 2 - Evaluation (metrics, manuscript)**

Muheng Li (PSI-CPT, Villigen, Switzerland)

Christopher Kurz (Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany)

Niklas Wahl (DKFZ, Heidelberg, Germany)

Viktor Rogowski (Skane University Hospital, Lund, Sweden)

Zoltan Perko (Delft University, Delft, the Netherlands)

Group 3: Technical organization (setup on the platform, sample code for public release and distribution of software)

Nikolaos Delopoulos (Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany)

Group 4: Project organization (MICCAI application, funding, platform)

Guillaume Landry (Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany)

Ye Zhang (PSI-CPT, Villigen, Switzerland)

Lennart Volz (GSI, Darmstadt, Germany)

Matteo Maspero (UMCU, Utrecht, the Netherlands)

b) Provide information on the primary contact person.

Guillaume Landry, W2 Professor, head of the LMU Adaptive Radiation Therapy Lab, Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany  
guillaume.landry@med.uni-muenchen.de

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

Clinical medical physicists are part of the organizing team. Their role is to guide the challenge design to be representative of the clinical task of radiotherapy dose calculation.

**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 deadline but with an open leaderboard for submission that will have a post-challenge phase to provide a platform for continuous evaluation of the algorithms.

**Challenge venue and platform**

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



We aim to present a report on the challenge principally to MICCAI 2026 (workshop with presentations from the top 4 teams) and additionally to the European Society for Radiotherapy and Oncology (ESTRO) 2027 annual conference (oral report on the outcomes of the challenge), reaching both the developer (MICCAI) and the end-user (ESTRO) communities. We also have submitted an abstract for presentation of the challenge design at ESTRO 2026 and have an invited lecture at the pre-meeting course, where we will advertise the challenge, which will then be in the early training phase.

b) Report the platform used to run the challenge.

[grand-challenge.org](https://grand-challenge.org)

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

Yes

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

<https://doserad2026.grand-challenge.org>

### Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

Only fully automatic methods are allowed. Methods should be submitted as specified on the submission page. Inference should run on an AWS g5 instance using a single GPU with 24 GB GPU RAM, 16 CPUs, and 64 GB RAM. The maximum inference time to produce a dose distribution for a beam setting should be compatible with the requirements of real-time dose calculation. This limit will be set at 1 second per beam. This will be measured as the time to process a given number of beams for practical reasons.

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 may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

DoseRAD2026 will provide training data. Participants are allowed to use additional data that is publicly available. Participants are also allowed to use publicly available pre-trained models. In either case, the data and/or models must be made publicly available before the start of the challenge on March 15th 2026.

Specifically:

Allowed: Using open-source codebases as a reference or for implementation.

Allowed: Training a model from scratch using only the permitted datasets.

Allowed: Initializing a model with pre-trained weights, as long as they were publicly available before March 15th, 2026.

Not Allowed: Fine-tuning a model trained on any private dataset or with private weights that were not publicly

available by March 15th, 2026.

The use of publicly available data and models must be reported in a document describing the submitted method and the corresponding submission form.

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 in the challenge and win prizes if not listed among the organizers, contributors, or data providers and if they did not co-author any publication (accepted publication date) with the organizers in the timeframe 2024-09/2026-09 (inclusive). If they do not meet these criteria, they may still participate but are not eligible for the prizes. Organizers, contributors, data providers, and sponsors may not participate in the challenge. The following five roles can be taken in the scope of the challenge: - Organizers: Take care of the challenge organization. They cannot participate in the challenge due to the possible access to data. - Contributors: These people support the challenge organization but are not actively involved in it. They cannot participate in the challenge due to possible access to data. - Data provider: Collect and provide data to the organizers, support the organizers in the organization without an active role, and not involved in the decision-making. They cannot participate in the challenge and will be listed in the dataset publication. - Prize sponsors: Industry partners sponsor prizes. Their employees can participate in the challenge, but may not win prizes. - Participant: Participate in the challenge organized in teams of up to five people. They cannot be listed among the organizers, contributors, or data providers.

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

Each participant/team can only use one account to participate in the competition. Participants who use multiple accounts will be disqualified from the competition. Each participant can only join a single team. Each team can comprise five participants, but the organizers reserve the right to reduce the number of co-authors of the top-performing teams to the challenge paper summarizing the results (see publication policy). Once a participant or a team submits, the submission or the team cannot withdraw from the challenge.

As further conditions for being awarded a prize, the teams must fulfill the following obligations:

- Present their method in person at the final event of the challenge at MICCAI 2026.
- Submit a paper reporting the details of the methods in a short or long LNCS format, following the checklist provided on the submission page. Organizers reserve the right to exclude submissions lacking any of these reporting elements.
- Submit a form reporting the details of the algorithm after the test submission has been completed, as the organizers will provide it.
- Sign and return all prize acceptance documents as may be required by the Competition Sponsor/Organizers.
- Commit to citing the challenge report and data overview paper whenever submitting the developed method for scientific and non-scientific publications.
- The top four teams must disclose and openly share their code and weights, as well as any additional data generated to set up their algorithms, to allow for future re-use of their algorithms. While all other teams are strongly encouraged to do so, it is not mandatory. The code should be provided within 14 days of the announcement of the winning participants.

The organizers will award cash prizes to the top four teams (1000 USD per winner, one winner per task, four tasks) sponsored by the companies Radformation (New York, USA) and RaySearch (Stockholm, Sweden). The DoseRAD2026 organizers will consolidate the results and submit a challenge report paper to Medical Image Analysis or similar. The first four teams (winners for each task, four tasks) will be invited to participate in this

publication, and they will be required to submit an algorithm summary in the requested form. The organizers reserve the right to reduce the number of co-authors among the team participants to at least two. The organizers reserve the right to invite additional teams with interesting submissions from a technical perspective.

e) Define the policy for result announcement.

Examples:

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

The results and winner will be announced publicly via the challenge website, and the top four teams (1 winner per task, four tasks) will be invited to present their approach during the MICCAI event.

Once participants submit their results on the test set to the challenge organizers via the challenge website, they will be considered fully vested in the challenge. Their performance results will become part of presentations, publications, or subsequent analyses derived from the challenge at the organization's discretion. Specifically, all the performance results will be made public.

The DoseRAD2026 organizers will consolidate the results and submit a challenge report paper to Medical Image Analysis (or similar).

Each winning team per task will be invited to participate in this publication, requiring that they submit an algorithm summary in the form of LNCS proceedings. The organizers will analyze their dose distributions as the challenge submission system will have automatically solicited them.

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).

To be eligible for the official ranking, the participants must submit a paper describing their method as described in point d above. The organizers, contributors, and data providers can independently publish methods based on the challenge data after an embargo of 6 months from the challenge's MICCAI event. The embargo is counted from the MICCAI event, considering the submission date of the work. Participants can submit their results elsewhere after an embargo of 6 months; however, if they cite the challenge report paper, no embargo will be applied.

## 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.

Submission instructions will be available on the challenge website.

We will organize a type 2 challenge with two phases (preliminary testing phase and final testing phase), where algorithm submissions run through the website during the 2 phases, as described in <https://grand-challenge.org/documentation/challenges/>.

Training data from 75 patients with CT (tasks 1 and 3)-MRI (tasks 2 and 4) pairs from the SynthRAD2025 training set having undergone quality assurance will be publicly available. Participants may split this data into training and local validation folds as they see fit.

During phase 1 (the preliminary testing phase), input pre-testing data (10 patient cases) will be available to provide predictions after type 2 model upload on the website (10 uploads per team per task). This phase will serve both as testing of the dockerized model and provide type 2 validation with a preliminary leaderboard, allowing participants to get an overview of their overall performance.

During phase 2 (the final testing phase, 37 patient cases with CT-MRI pairs), the teams must supply the algorithm for the type 2 challenge to the organizers following the submission link and instructions provided on <https://doserad2026.grand-challenge.org/>. Teams will have two tries during phase 2. Teams will submit their dockerized dose calculation algorithms to the challenge website without having the testing data at their disposal. Once the challenge is presented at MICCAI, a post-challenge phase will be opened, making the preliminary testing and final testing phases newly available with the testing data remaining private.

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.

The challenge is subdivided into a preliminary testing and a final testing phase. The preliminary testing phase allows each team to submit up to 10 submissions per task to familiarize themselves with the submission system. The results of this phase will be evaluated on a limited dataset (10 cases, separate from the final testing dataset) with an open dashboard. This will provide a preliminary overview of model performance. The preliminary testing phase will be with the “open logs” setting of the grand-challenge.org platform, meaning participants will have information on failed submissions. Additionally, this will serve as a type 2 validation and allow preliminary relative performance assessment by the teams.

The preliminary testing phase will remain open also during the final testing phase; in the last 4 weeks of the challenge, the final testing phase will start. The preliminary testing phase will last 7 plus 4 weeks and take place 11 weeks after the release of the training data to allow optimization of the algorithms.

After the preliminary testing phase, a new leaderboard will be created, and the type 2 final testing phase will start. During this final phase, all data and targets remain hidden, and logs are closed. The participating teams can submit up to two runs per task to evaluate their algorithms on the testing set. The second run is granted to accommodate possible errors during the submission process. Only the last run will be counted for the official ranking of the teams and the challenge results. We request that each run will be identified with a description.

## 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 online 01/01/2026

Start challenge: Release training cases 15/03/2026

Registration period 15/03/2026 - 01/06/2026

Training and validation phase (11+11 weeks) 15/03/2026 - 15/08/2026

Introduction of the challenge at ESTRO2026 15/05/2026 - 19/05/2026

Preliminary test phase (11 weeks, 10 submissions) 01/06/2026 - 15/08/2026

Test phase (4 weeks, max 2 submissions) 16/07/2026 - 15/08/2026

Announcements and invitation to present 15/09/2026

Presentation of the challenge results MICCAI26, Abu Dhabi 4 or 8/10/2026

Presentation of the challenge results ESTRO2027 April/May 2027

Post-challenge phase 15/09/2026-31/12/2026

## 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 LMU University Hospital Munich ethics committee approved a corresponding clinical study entitled 'Generating synthetic computed tomography scans for radiotherapy: SynthRAD2025 challenge' on 15/04/2024 (study number 24-0195).

Informed consent was obtained from all patients, and this study has been exempted by the VU University Medical Center Medical Ethics Review Committee (#2018.602, IRB00002991).

## 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)

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

CC BY-NC (Attribution-NonCommercial)

### 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.

Organizers' evaluation code and supporting data pre/post-processing code will be publicly available on GitHub at the following location:

<https://github.com/DoseRAD2026>

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

Openly sharing teams' code and weights when applicable is strongly encouraged but remains optional for all teams except for the four winning teams (winner of each task, four tasks). If additional data was generated for training it should also be made public. The code should be provided within 14 days of the announcement of the winning participants.

### 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.

The challenge is primarily funded by third party research funds at the LMU University Hospital Munich. Additional funding has been secured from Radformation (New York, USA) and RaySearch (Stockholm, Sweden), companies providing radiotherapy software, including dose calculation. The company funding will be used to sponsor the prizes for the winning teams. A Radformation employee is involved in the challenge organization. Sponsoring company employees are not eligible to win prizes. Endorsement from professional societies (ESTRO, European Radiation Oncology Society, DGMP, German Medical Physics Society, NVKF, Dutch Medical Physics Society, EFOMP, Italian Medical Physics Society, AIFM, European Medical Physics Society) will be sought, similarly as for TrackRAD2025 and SynthRAD2025.

Access to the test case labels is limited to the data providers and will be accessed by the organizers for data preparation. The test cases will not be made public.

## 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

Radiotherapy/radiation oncology,dose calculation,intervention assistance,Intervention planning,Research

### Task category(ies)

State the task category(ies)

Examples:

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

Real-time dose calculation,Regression

### Cohorts

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

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

The biomedical application addresses patients undergoing radiotherapy (about 50% of cancer patients). We will collect data randomly sampling cases of both sexes, ensuring a balanced representation of the sexes. An adult population will be collected. The vast majority of radiotherapy patients undergo CT-based treatment planning for photon therapy, and Task 1 will thus address dose calculation for photons on CT. Thus, for Task 1, only CT imaging is required. However, to provide homogeneous imaging data across the four challenge tasks, inclusion criteria for

Task 1 datasets would be treatment at a photon therapy MRI-linac with both CT and MRI acquisition during treatment planning (to allow using the same imaging dataset for all tasks).

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

Patients undergoing MRI-guided radiotherapy in the thorax and abdomen will be considered since their datasets provide both CT and MRI images needed to address tasks 1 to 4. Specifically, data from patients treated for lung tumors and abdominal tumors will be considered due to 1) the availability of paired CT and MRI images in the SynthRAD2025 dataset, 2) the high level of motion in these anatomies requiring real-time management such as fast dose recalculation, 3) the ability to generate curated CT-MRI pairs. In total, the challenge cohort will comprise about 122 0.35 T MRI-linac patients (62 abdomen cases, 60 lung cases). The cohort collected for the challenge is a subsample of the target cohort in the final biomedical application.

### Imaging modality(ies)

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

The challenge will focus on 3D CT and MRI as used as the basis of 3D radiotherapy dose calculation.

### Context information

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

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

In addition to the CT or MRI images, we will provide the necessary beam information for dose calculation.

For photon beams (tasks 1 and 2), this will include the multi leaf collimator aperture, the isocenter position of the simulated irradiation device, the source to isocenter distance and the beam gantry angle, as well as the magnetic field strength used for the MRI image acquisition. For proton beams (tasks 3 and 4) the aperture information will instead be the gaussian width of the pencil beam and its initial proton energy.

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

Age and sex will be provided for the centers from which sharing such demographic details has been granted.

### Target entity(ies)

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

The data will come from 3D CT and MRI scans acquired from patients treated at 0.35 T MRI-linacs for thoracic and abdominal lesions.

The data is representative of the target cohort since it was acquired retrospectively from clinical practice.

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.

For task 1, the algorithm should target photon therapy dose calculation on CT images of the thorax and abdomen regions.



## 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)).

Corresponding metrics are listed below (parameter 26). Given either a 3D CT or MRI image and the parameters of a radiation beam (either photons or protons), the algorithms should be able to output a 3D dose distribution as rapidly as possible. The accuracy should approach the accuracy which is feasible with slow accurate methods such as full Monte Carlo simulation of radiation transport on CT. Monte Carlo simulation will serve as ground truth. Since it is not possible to directly perform Monte Carlo dose calculation on MRI due to the lack of electron density information, dose calculation on the paired CT will be considered as the ground truth for the corresponding MRI in the pair.

## 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).

Data for Task 1 were acquired at either of the CT scanners listed below:

LMU University Hospital (LMU), Munich, Germany:  
Canon Aquillion

Amsterdam University Medical Center (AUMC), Amsterdam, the Netherlands:  
GE Optima CT 580

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

At LMU, all the lung and abdominal CT images were acquired at 120 kVp in breath hold with a slice thickness of 3.0 mm. The images were reconstructed with a  $512 \times 512$  pixel matrix, the FOV was 550 or 700 mm, and an in-plane pixel spacing of 1.07 mm or 1.37 mm, respectively. Images were reconstructed using the vendor software based on the filtered back-projection algorithm.

At AUMC, all the lung CT images were acquired at 120 kVp in breath hold with a slice thickness of 2.50 mm. The images were reconstructed with a  $512 \times 512$  pixel matrix, the FOV ranging from 500 to 700 mm, and an in-plane pixel spacing of 0.98 mm to 1.37 mm. CT images were reconstructed using the vendor's filtered back-projection algorithm with a Standard reconstruction filter.

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 CT data was acquired for MRI-guided radiotherapy treatments in the radiotherapy departments of LMU Munich DE and AUMC Amsterdam NL. The training set originates from LMU and was previously part of the training set for SynthRAD2025. The testing set will partially consist of data from the private SynthRAD2025 test set (from LMU), which was not released publicly, and data from AUMC which was never released publicly. We restrict ourselves to 0.35 T MRI-linac data due to the availability of CT-MRI pairs and for consistency of the CT to MRI intensity transformation.

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 clinical staff of the respective radiotherapy departments acquired the CT scans. All patients were treated with MRI-guided radiotherapy using repeated breath-hold conditions. Dedicated body coils were used for MRI-linac 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.

A case in this challenge refers to a patient's CT scan along with the photon dose distribution and photon beam parameters for a series of beams relevant for the treatment. For the training set, the photon beam shapes will be collected from clinical VMAT plans and will be randomly positioned in the patient CT dataset. This is to provide a wide variety of beam configurations to ensure that algorithms generalize. For the testing set, strictly beams optimized for the irradiation of the patient's tumor are used. In all cases, beams refer to those used with volumetric arc therapy (VMAT).

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

For task 1, we will have CT scans from 122 patients for thorax and abdomen.

Training and validation (publicly available):

Thorax: 39 patients

Abdomen: 36 patients

Preliminary testing:

Thorax: 5 patients

Abdomen: 5 patients

Final testing:

Thorax: 16 patients

Abdomen: 21 patients

For each patient, several hundred to thousand beams and corresponding photon dose distributions will be generated.

c) How much of the data are already annotated (stratified by train test in percentage)?

The lung and abdomen patient datasets have been reviewed for DIR quality and all abdomen patients datasets have undergone air cavity correction. Thus 100% of the training and testing sets have been reviewed and air cavity corrected.

Monte Carlo simulation on the GSI cluster has been set up and test, requiring 12 hours for 1800 photon beam calculations. The full simulation of the 40000 training beams should be done in 2 weeks and will be initiated in the coming days.

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

We chose all available SynthRAD training and testing cases for which CT and MRI pairs were available, and for which an accurate deformable image registration could be achieved, as assessed by visual inspection. While this is not strictly required for task 1 (photon dose calculation on CT), it is critical for tasks 2 and 4. We decided to keep a unified dataset for all tasks to ensure comparability of the results for the different tasks. While the number of cases may appear limited, one has to keep in mind that for a given CT image, several 100s of beams can be generated, each traversing different anatomy. Prior experience suggests that this allows accurate algorithm development.

For testing we followed the proportion of data available from SynthRAD2025, and added an external dataset to verify generalizability.

e) 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.

N/A

f) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

We have aimed at collecting data for abdomino-thoracic regions because their susceptibility to motion is high, and we expect the most benefit from dose adaptation by fast dose calculation methods.

### Annotation characteristics

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

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

Since the task of CT/MR dose prediction is formulated as regression, manual delineation for dose distributions was not required. However, manual data filtering and correction for matching CT and MRI before dose simulation was needed to ensure the anatomical consistency. A postdoctoral researcher with 7 years of experience and a PhD student with 2 years of experience in medical physics performed these manual preparations, which involved two phases:

1. Registration quality assurance was conducted on the synthRAD dataset. The annotators visually inspected the spatial alignment between the deformed CT and MRI slice-by-slice. Only cases exhibiting high registration accuracy were selected for the DoseRAD cohort.

2. An air cavity correction pipeline was implemented for abdominal cases, as deformable image registration (DIR) often fails to match these variable structures. In this context an air cavity refers to gas present in the patient's bowels, presenting as a low intensity region on CT and MRI. As on MRI air cavities share similar intensity as bony tissue, they were manually segmented on MRI for 20 patients to train an nnU-Net model. This model then segmented the remaining 42 patients, with results manually refined by the annotators to ensure accuracy. These MRI-derived masks were used to harmonize the CT anatomy: original air cavities on the CT (detected by simple thresholding) were filled with the median intensity of surrounding tissues, and the corrected cavities were subsequently assigned a value of -824 HU.

These efforts ensured well matched CT and MRI images. Subsequently, ground truth dose distributions were calculated using the Geant4 Monte Carlo simulation framework for a series of photon therapy VMAT beam apertures. All Monte Carlo simulations were implemented using the Geant4 toolkit (version 11.00-patch-03) with the predefined QGSP\_BIC\_EMV physics list to generate ground truth dose distributions.

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

The deformable image registration quality check annotators were instructed to visualize each case in the dataset after registration in an overlay and in a side by side view to determine whether gross anatomical mismatches remained. Cases where this was the case were removed from the dataset.

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

The DIR was verified by a postdoctoral researcher in medical physics with 7 years experience and a PhD student in medical physics with 2 years experience. The air cavities were delineated by the PhD student.

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

Not applicable.

### Data pre-processing method(s)

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

The total challenge cohort will comprise 122 0.35 T MRI-linac patients (62 abdomen cases, 60 lung cases, from two centers), which were divided into 75 patients for training, 10 patients for preliminary testing and 37 patients for final testing.

The 122 CT-MRI pairs were first processed following SynthRAD2025 preprocessing workflow, including rigid registration between CT and MRI, resampling with a consistent voxel spacing of  $2 \times 2 \times 2$  mm for both CT and MRI, and body outline segmentation using TotalSegmentator (version 2.3.0) (for more details, see <https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.17981>). Furthermore, instead of using the DIR from the Elastix framework in SynthRAD2025 preprocessing, the open-source ConvexAdam DIR method was adopted to achieve superior deformable registration accuracy between CT and MRI, as demonstrated by manual visual assessment of the anatomical alignment. Then, the body masks from the deformed CTs were used to remove the couch from CT images.

After the DIR quality assurance and air cavity correction (as described in the annotation characteristics section), the corrected CT images were subsequently converted into density and elemental composition maps using a scanner-specific calibration curve. All Monte Carlo simulations were implemented using the Geant4 toolkit (version 11.00-patch-03) with the predefined QGSP\_BIC\_EMV as physics list to generate ground truth segment dose distributions.

For the 75 training patients, 2D segments were extracted via control point sequences from 40 clinical VMAT plans (20 lung and 20 abdominal cases) designed for an Elekta Versa HD Linac with an Agility 160-leaf MLC. To augment the segment pool, each MLC leaf was randomly shifted by  $\pm 5$  mm along its direction of travel three times, expanding the training segments from approximately 10,000 to 40,000. For the 47 test patients (preliminary and final testing combined), 47 VMAT plans were generated based on the clinically contoured targets and organs using a scriptable research treatment planning system (TPS) configured for the Elekta Versa HD Linac. Segments from each testing plan were also extracted without augmentation.

After segment extraction, Geant4 dose simulations with a 6 MV photon energy spectrum derived from the ELEKTA\_PRECISE\_6MV phase space file were implemented. The source-to-axis distance was set to 100 cm, the binary segment mask served as a fluence filter on the isocenter plane with the full  $40 \text{ cm} \times 40 \text{ cm}$  field, and the dose grid matched the patient CT spacing ( $2 \times 2 \times 2$  mm). A uniform 0.35 T magnetic field was simulated in Geant4, aligned parallel to the superior-inferior patient direction and confined within a cylindrical region with a 50 cm diameter.

For each training patient, the isocenter was shifted along the superior–inferior axis by  $-2, 0, +2$  cm. At each isocenter, 180 gantry angles were simulated every 2 from 0 to 358; at each angle, one segment was randomly sampled from the 40,000-segment pool to generate a segment dose, yielding  $75 \times 3 \times 180 = 40,500$  training samples through randomized segments, gantry angles, and isocenters on 75 training patients. For 47 testing patients, segment doses were simulated according to each testing plan's control point sequence. Plan doses were then reconstructed by MU-weighted accumulation of segment doses. Besides, 470 segment doses with random segments (selected from the testing segment pool), gantry angle and iso centers were simulated on 47 test patients for the additional robustness test dataset. For both training and test segment dose simulation, photon histories were set to  $5e6/\text{cm}^2$ . The prescription was 60 Gy in 20 fractions. All MC testing plan doses were scaled so that  $D_{95\%}(\text{PTV}) \geq 0.95 \times 60 \text{ Gy} = 57 \text{ Gy}$ .

### 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 main source of error will be residual mismatches between the CT and MRI scans after DIR and air cavity matching. For photon therapy, these residual errors are considered to be relatively negligible since the dose calculation is not extremely sensitive to small localized density deviations.

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

Additional statistical uncertainty from Monte Carlo simulation of the photon segment dose is estimated to be at the level of 2% in the  $D > 10\% D_{\text{max}}$  dose region. This level was selected to offer a balance between precision and calculation time.

## 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 assess algorithms along two main dimensions: dose accuracy and computational efficiency, using a two-level evaluation scheme that is identical for all four tasks (photon/proton, CT/MRI).

Both Level 1 and Level 2 evaluations use the same set of beams/segments from the treatment plan for each patient, but evaluate them differently:

- Level 1 assesses the accuracy of individual beams
- Level 2 assesses the accuracy of the complete plan (beams combined with their clinical weights)

All dose metrics are computed on a common reference dose grid. If a method internally uses a different

resolution, participants must resample their output to the reference grid before submission.

We distinguish:

- Level 1 – Single-beam / segment dose evaluation (per-beam dose accuracy)
- Level 2 – Full-plan evaluation (clinically oriented accuracy)
- Efficiency – Runtime for processing a fixed set of beams

#### Level 1: Single-beam / segment evaluation

For each patient, we use the set of beams/segments from the treatment plan (photon segments or proton pencil beams). These beams cover a wide range of incident angles, paths through heterogeneous anatomy, and energy settings (for protons).

For each individual beam in this set, we compare the predicted 3D dose distribution to the Monte Carlo ground truth using the following metrics.

##### 1. Masked Mean Absolute Error (MAE) per beam (all tasks)

- For each beam, we define a high-dose region as all voxels that receive at least 10% of the maximum dose for that beam in the ground truth.
- Within this high-dose region, we compute the mean absolute difference between predicted and ground-truth dose.
- The result is normalized by the beam-specific maximum dose (the same normalization will be used consistently for all tasks).

This yields a masked MAE per beam, which focuses the evaluation on the clinically relevant part of the beam path while ignoring low-dose background.

##### 2. Depth-dose metrics from integrated depth-dose (IDD) curves (all tasks)

To characterise longitudinal dose behaviour along the beam direction, we compute integrated depth-dose (IDD) curves for both photon and proton beams:

- For each beam, we compute the IDD curve along the nominal beam axis, both for ground truth and prediction. The IDD curve represents the accumulated dose as a function of depth along the beam, expressed in geometric depth from the patient surface along the beam axis.
- From these IDD curves, we derive:
  - A curve distance metric, summarising the overall difference between the predicted and ground-truth IDD curves across depth (a root-mean-square difference along the depth coordinate).

These IDD-based metrics are computed for each beam and later aggregated at the patient and task levels as described in Item 27 (Ranking Methods).

#### Level 2: Full-plan evaluation (clinical beam sets with weights)

For full-plan evaluation, we use the same set of beams/segments from the clinically inspired treatment plan that was evaluated in Level 1. Each beam/segment has an associated weight (for example, VMAT segments for photons or weighted pencil beams for protons).

The algorithms must output beam-wise dose distributions for all beams in the plan set. These are then combined using the stored weights to reconstruct a full 3D dose distribution for the plan, separately for each patient. We then compare the predicted and Monte Carlo reference plans using:

### 1. Stratified plan-level MAE

We compute the MAE in three dose strata for each plan:

- High-dose region: voxels receiving at least 80% of the prescription dose.
- Mid-dose region: voxels receiving between 30% and 80% of the prescription dose.
- Low-dose region: voxels receiving between 10% and 30% of the prescription dose.

For each of these regions, we compute the mean absolute difference between predicted and ground-truth dose, normalized by the prescription dose. We then compute a combined plan-level MAE as the average of these three stratified MAEs (low-dose, mid-dose, and high-dose), giving equal weight to each dose region.

### 2. 3D local gamma index

We compute a three-dimensional local gamma pass rate with strict criteria (1% dose difference and 1 mm distance-to-agreement), using the ground-truth plan as reference.

The gamma index is a composite metric widely used in clinical radiotherapy QA that simultaneously evaluates dose accuracy and spatial agreement. For each voxel in the evaluation volume, the gamma index quantifies the minimum distance (in a combined dose-distance space normalized by their respective criteria, here 1% dose and 1 mm distance) between that voxel's predicted dose and any voxel in the reference (ground truth) dose distribution.

More specifically, for our 1%/1mm criterion:

- The gamma index passes ( $\gamma \leq 1$ ) if, within a 1mm radius of the evaluated voxel, there exists at least one reference voxel whose dose differs from the predicted dose by no more than 1% of the evaluated voxel's dose.
- The gamma index fails ( $\gamma > 1$ ) if no such matching reference voxel exists. The magnitude of the index reflects the magnitude of the dose-distance disagreement, but in practice the index is usually used in a binary fashion (pass/fail).

The gamma pass rate is then defined as the percentage of evaluated voxels that pass this criterion. This metric is particularly valuable because it accounts for both:

1. Dose discrepancies: Small dose errors are tolerated if spatial agreement is good



## 2. Spatial misalignments: Small positional shifts are tolerated if dose values are accurate

In our evaluation:

- We apply this metric in 3D using local normalization (normalizing dose differences to the local reference dose rather than a global maximum), which is the current clinical standard for patient-specific QA.
- The evaluation is restricted to voxels receiving at least 10% of the prescription dose to focus on clinically relevant regions.
- This metric is sensitive to both spatial and dosimetric discrepancies and reflects established clinical quality assurance practice.

## 3. DVH-based clinical metric

We use a standardized DVH-based score for each plan:

- For each case, we consider:
  - One target structure (PTV, depending on the planning convention).
  - The three closest organs at risk (OARs) to this target.
- From the DVHs of these structures, we extract:
  - For the target:
    - A near-minimum dose (dose received by 98% of the target volume, D98).
    - A coverage measure (fraction of the target receiving at least 95% of the prescription dose, V95).
  - For each of the three OARs:
    - A near-maximum dose (dose received by 2% of the organ volume, D2).
    - The mean dose to the organ (Dmean).

For each DVH quantity, we compute the absolute relative difference between the predicted and ground-truth plans. The DVH score for a plan is the weighted average of these relative differences. The two target metrics are averaged together, and the six OAR metrics (two for each of the three OARs) are averaged and then divided by 3, so that target and OAR DVH parameters have equal weight in the final DVH score.

### Efficiency: Runtime metric

Real-time dose calculation is essential for the intended clinical application. We therefore define a runtime metric that measures the wall-clock time required to process a fixed, standardized set of beams.

- For each task, we will provide one canonical patient and use the beams from its treatment plan as the fixed set to be predicted. This set will be identical for all teams and clearly specified in advance.
- Participants are free to choose their batching strategy and internal implementation to minimize the total runtime without exceeding the GPU memory of the provided instance.

-On the grand-challenge.org platform, we record the wall-clock time between the start and successful completion of predictions for this scenario.

-The efficiency metric is the average runtime per beam, computed as total time divided by the number of beams in the runtime scenario.

We will impose an upper bound on the allowed average runtime per beam (one second per beam, including data loading and model initialization). Algorithms exceeding this limit on the provided hardware will be excluded from the official ranking for that task.

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

The masked beam-level MAE evaluates the dose accuracy of individual beams from the treatment plan, probing dose accuracy across a wide variety of paths through heterogeneous anatomy. This is useful to understand the intrinsic per-beam behaviour of each method.

The IDD-based depth-dose metrics quantify how well longitudinal dose build-up and fall-off are reproduced for both photons and protons. For protons, this directly reflects range accuracy; for photons, it provides sensitivity to build-up and fall-off regions and highly modulated segments.

The stratified plan-level MAEs on the clinical plans reveal whether methods fail primarily in high-, mid-, or low-dose regions, corresponding to different clinical priorities such as target coverage and normal tissue sparing. The local gamma index jointly captures spatial and dosimetric agreement at plan level and is aligned with routine clinical QA procedures.

The DVH-based score aggregates clinically interpretable endpoints (target coverage and OAR sparing) into a single measure, directly reflecting treatment quality.

The runtime per beam directly reflects the feasibility of near real-time adaptive workflows, ensuring that methods are not only accurate but also fast enough for online use.

By combining a single-beam evaluation with a plan-based evaluation and an explicit runtime metric, we obtain a comprehensive view of algorithm performance at the beam level, at the plan level, and in terms of computational efficiency.

### 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. Ideally, provide the ranking scheme as a concrete pseudo code.

Ranking is performed independently for each of the four tasks (photon/proton, CT/MRI).

#### 1. Aggregation of per-patient metrics

Level 1 and Level 2 metrics are both computed on the same beam set from the treatment plan, but evaluated

differently:

- For each patient and each metric that is defined per beam (masked MAE and the IDD-based metric in Level 1), we first average across all beams of that patient to obtain a single patient-level value.

- For each patient and each plan-level metric (stratified MAEs, DVH score, gamma pass rate in Level 2), we use the value directly obtained from the reconstructed plan (based on combining the beams with their clinical weights).

- The runtime metric is defined at the level of the standard runtime scenario and is therefore already a single value per submission.

For each submission and each metric, we then compute the average across all test patients (or, for runtime, across repeated measurements if applicable), yielding one scalar value per metric and submission.

## 2. Per-metric ranking

For each task and each metric, we rank all valid submissions:

- Metrics where lower values indicate better performance:

- Beam-level masked MAE.

- IDD curve distance.

- Combined plan-level MAE (averaging low-dose, mid-dose, and high-dose regions).

- DVH score.

- Runtime per beam (runtime scenario).

- Metrics where higher values indicate better performance:

- Local 1% / 1 mm gamma pass rate (plan beams).

This yields, for each submission, a list of metric-specific ranks (rank 1 being best).

## 3. Final ranking: RankThenMean with increased weight on efficiency

We use a RankThenMean scheme:

- For each submission, we compute the average of its ranks across all metrics.

- To ensure that the single runtime metric is not overshadowed by the larger number of accuracy metrics, we give double weight to the runtime rank when computing this average.

- Submissions are then ordered from lowest to highest average rank. A lower final score corresponds to better overall performance.

Tie-breaking is defined as follows:

1. Prefer the method with lower runtime per beam.

2. If still tied, prefer the method with lower combined plan-level MAE.

3. If still tied, prefer the method with lower DVH score.

This ranking strategy preserves the relative strengths and weaknesses of each algorithm across metrics while explicitly incentivizing efficient methods.

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

-The challenge is organized as a type 2 grand-challenge.org challenge with dockerized algorithm submissions.

-If a submission completely fails for a given patient (for example, the container crashes or produces no outputs), that run is considered invalid and is not included in the final ranking. Teams still have a second allowed submission during the final testing phase.

-If a submission produces partial output for a patient (missing some beams from the treatment plan):

--For Level 1 evaluation, any missing beams are treated as if they delivered zero dose.

--For Level 2 evaluation, the missing beams are similarly treated as zero-dose beams when reconstructing the plan.

This leads to:

-Very large MAE at both single-beam and plan level.

-Poor gamma pass rates.

-Large DVH deviations.

In practice, this strongly penalizes incomplete predictions without requiring special-case metric definitions.

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

-General choices:

1. RankThenMean avoids the need to rescale heterogeneous metrics (MAE, DVH differences, gamma pass rate, IDD-based distances, runtime) to a common numerical range and has been used successfully in previous radiotherapy challenges.

2. Giving extra weight to runtime ensures that computational efficiency, which is critical for real-time applications, is appropriately reflected in the final ranking.

3. Evaluating both per-beam metrics (Level 1) and plan-level metrics (Level 2) on the same beam set from the treatment plan allows us to assess both the intrinsic per-beam quality and the clinically relevant plan reconstruction behaviour.

4. The tie-breaking rules make explicit that runtime and high-dose accuracy are primary priorities, followed by overall DVH agreement, in line with clinical relevance.

-Clinical rationale for tie-breaking priorities:

The tie-breaking sequence (runtime, plan-level MAE, DVH score) reflects the clinical priorities for real-time adaptive radiotherapy:

1. Runtime is prioritized first because real-time dose calculation requires methods to operate within strict time constraints. Without meeting the speed requirement, even the most accurate method cannot be deployed for online adaptive workflows. This aligns with our mission to enable real-time dose-guided radiotherapy.
2. Plan-level MAE is prioritized second because it represents the overall dose accuracy of the complete treatment plan. The plan-level dose distribution determines tumor coverage and organ-at-risk sparing, making it more clinically consequential than individual beam accuracy alone.
3. DVH score is prioritized third as it captures clinically interpretable endpoints (target coverage and OAR constraints). While DVH metrics are crucial for clinical evaluation, they are downstream consequences of the dose distribution quality captured by plan-level MAE.

This hierarchy ensures that we select methods that are both fast enough for clinical deployment and accurate enough to maintain treatment quality, with preference given to methods that excel in the most clinically decisive metrics.

-Justification for equal weighting in RankThenMean:

Our evaluation employs six accuracy metrics (beam-level MAE, IDD curve distance, plan-level MAE, gamma pass rate, DVH score, and runtime), with runtime receiving double weight. We chose equal weighting for the five accuracy metrics for the following reasons:

1. Complementary information: Each metric captures a distinct aspect of dose calculation quality:

- Beam-level MAE: Individual beam accuracy across diverse anatomical paths
- IDD curve distance: Longitudinal dose build-up and penetration (critical for protons)
- Plan-level MAE: Overall dose distribution quality in clinically relevant dose regions
- Gamma pass rate: Combined spatial and dosimetric agreement
- DVH score: Clinical endpoints for target and organs-at-risk

Since these metrics assess different properties that are all clinically relevant, imposing a predetermined weighting would unnecessarily favor certain aspects over others without clear clinical justification on the exact weight that should be applied per metric. 2. Avoiding implicit bias: Weighted averaging would require us to assign relative clinical importance values (e.g., "gamma is twice as important as DVH"). Such assignments would be inherently subjective and could bias the challenge toward methods optimized for the highest-weighted metrics, potentially overlooking methods with balanced performance across all clinically relevant dimensions. For example, a good

plan-level MAE may mask beam-level errors which cancel out when summed, which could become problematic in other beam configurations and decrease clinical confidence in the dose calculation.

3. Robustness across clinical scenarios: Equal weighting ensures that winning methods demonstrate broad competence rather than excelling in a single favored metric. Real-world clinical applications require algorithms that perform well across multiple evaluation criteria, as different clinical scenarios may emphasize different aspects (e.g., target coverage vs. OAR sparing, spatial accuracy vs. dose magnitude).

4. Precedent in radiotherapy challenges: This approach has been successfully employed in our previous radiotherapy challenges (SynthRAD2023, SynthRAD2025, TrackRAD2025), facilitating consistency and comparability with established benchmarks.

-Why double weight for runtime:

Runtime receives double weight to balance the fact that it is a single efficiency metric against five accuracy metrics. This prevents computational efficiency, which is essential for our real-time application, from being overshadowed by the larger number of accuracy metrics. The 2:5 weighting ratio (runtime:accuracy) reflects that both speed and accuracy are critical, with a slight emphasis on accuracy given the clinical consequences of dose miscalculation.

## Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

We will perform tests for estimating the precision of the performance estimates, variability across cases, ranking variability, statistical significance, missing data handling as described in the sections below. For clarity, the following 6 metrics are available per case:

Metrics where lower values indicate better performance:

1. Beam-level masked MAE.
2. IDD curve distance.
3. Combined plan-level MAE (averaging low-dose, mid-dose, and high-dose regions).
4. DVH score.
5. Runtime per beam (runtime scenario).

Metrics where higher values indicate better performance:

6. Local 1% / 1 mm gamma pass rate (plan beams).

For statistical tests below we will consider a single value per case for each metric (meaning we will test over cases,

not over the many individual beams making up a case).

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

For each metric and each submission, we derive 95% confidence intervals from the bootstrap distribution of the metric values (not ranks) across patient samples. For example, if the mean MAE for a submission is 2.5%, the confidence interval might be [2.3%, 2.7%]. These confidence intervals will be reported in the challenge overview to indicate the uncertainty associated with each metric value, especially when methods exhibit similar performance.

Provide a description of how variability of the performance of individual algorithms across test cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

For each metric we will compute the interquartile range (25th to 75th percentiles) across the test cases (patients) to characterize the spread of algorithm performance. We will define outliers as cases in the worst 5th percentile of performance for each metric.

-Outlier handling and impact on rankings:

Importantly, outliers are NOT excluded from the ranking computation. All test cases, including outliers, contribute equally to the final rankings through the following process:

1. Metric aggregation: For each algorithm and metric, we compute the mean across all test patients (including outliers). This mean value is then used for ranking.
2. Rationale for inclusion: Outliers often represent challenging anatomical scenarios (e.g., extreme heterogeneity, large air cavities, unusual beam geometries) that are clinically relevant. Excluding them would give an overly optimistic view of algorithm robustness and potentially reward methods that fail in edge cases.
3. Statistical robustness: Our bootstrap confidence intervals (see Item 28, Statistics - Precision of performance estimates) quantify the uncertainty in metric estimates, including the influence of outliers. This allows us to distinguish robust performance differences from variations driven by individual outlier cases.
4. Transparency: Outlier cases will be identified and reported in the challenge results to highlight which anatomical scenarios pose the greatest challenges. This information is valuable for understanding algorithm limitations and guiding future method development.

-Dataset noise characteristics:

The DoseRAD2026 dataset contains several sources of variability that contribute to performance variance across cases:

1. Anatomical heterogeneity: Test cases span thorax and abdomen regions with:
  - Variable lung density and air cavity distributions
  - Different body habitus and patient sizes

- Diverse tumor locations and shapes

This anatomical diversity is intentional and reflects real-world clinical variation.

2. Imaging noise: CT and MRI acquisitions include typical imaging artifacts:

- CT: Artifacts from residual beam hardening, motion artifacts from breathing
- MRI: Geometric distortions, signal intensity variations, motion artifacts

These are inherent to clinical imaging and are not artificially introduced.

3. Registration uncertainty: For MRI-based tasks (Tasks 2 and 4), residual misalignment between paired CT-MRI images (after deformable image registration and air cavity correction) introduces spatial uncertainty. We estimate this to be on the order of 2-3mm in challenging regions (e.g., near air-tissue interfaces with sliding motion).

4. Ground truth statistical uncertainty: Monte Carlo simulations used for ground truth have statistical uncertainty:

- Photon dose (Tasks 1-2): ~2% (1 standard deviation) in regions receiving >10% of maximum dose
- Proton dose (Tasks 3-4): ~1% (1 standard deviation) in regions receiving >10% of maximum dose

This uncertainty is small compared to the dose differences we aim to detect between algorithms.

-Expected noise impact on rankings:

Given the dataset size (37 final test cases) and the bootstrap-based confidence intervals, we expect rankings to be stable for algorithms with clearly distinct performance levels. Methods with very similar performance may show ranking uncertainty of  $\pm 1$ -2 ranks, which will be captured in our Kendall's tau rank correlation analysis (see Item 28, Statistics - Rankings variability). This uncertainty reflects genuine similarity in performance rather than a flaw in the evaluation design.

Provide a description of how variability of rankings is assessed.

To assess ranking stability:

- We repeatedly draw bootstrap samples of the test patients by resampling with replacement (for example, 1,000 bootstrap samples, each containing the full number of test patients but allowing repetitions).
- For each bootstrap sample, we recompute the patient-level metrics and the resulting RankThenMean rankings.

For each pair of submissions, we compute Kendall's tau rank correlation between the original ranking and the distribution of bootstrap rankings to quantify ranking consistency.

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

We will generate an array containing the values of the 6 metrics per case (size of array will be number of cases x 6). The gamma pass rate will be inverted by 100%-gamma pass rate so that decreasing values are better like the other metrics (100% is the maximum value for perfect dose calculation).



We will perform a non-parametric Friedman test with post hoc Nemenyi test to determine whether algorithms are significantly different than others.

Provide a description of the missing data handling.

See above

Indicate any software product that is used for all data analysis methods.

python open source libraries such as scipy will be used. All code will be release open source along with the metrics calculation code.

### Further analyses

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

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

We will categorize performances based on the participants' methods to calculate the dose distributions. This analysis will, for example, consider the difference between fast GPU-based Monte Carlo simulation or data driven deep learning models, or whether MRI-base dose calculation requires synthetic CT generation as a first step.

Our challenge report will additionally evaluate the feasibility of integrating the outcomes into clinical practice, with a focus on hurdles to translate the resulting open source dose calculation and prediction framework into a clinical tool. The involvement of company representatives among the challenge organizers will be beneficial for this.

## TASK 2: Photon dose calculation on MRI images

### SUMMARY

#### Abstract

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

The most advanced form of image-guided photon therapy is currently delivered at hybrid MRI-linear accelerator devices (MRI-linacs). The combination of MRI and linac gantry allows high soft tissue contrast imaging of patients just before they receive their radiation treatment. This way, the treatment can be personalized just before radiation delivery, a process called online adaptive radiotherapy. This differs from conventional photon therapy in Task 1, where a single treatment plan calculated and optimized on CT is used for the delivery of several radiation fractions. Thus, at MRI-linacs a new daily treatment plan is made at each radiation fraction. This requires radiation beam dose calculation on the MRI images, which is currently achieved via the generation of a synthetic CT (using deformable image registration of a CT acquired as part of the treatment workflow to match the daily MRI). The goal of this task is to foster the development of methods for direct dose calculation on MRI images, bypassing the need for synthetic CT generation.

To this end, we will provide a training set of MRI images and corresponding ground truth dose distributions. Since direct Monte Carlo simulation on MRI is not feasible due to the lack of electron density information, we will make use of the SynthRAD2025 paired CT-MRI dataset (originating from clinical practice at an MRI-linac) to calculate the dose with Monte Carlo on CT (essentially the same doses as in Task 1) so that they also map onto the MRI. A dedicated registration and quality assurance pipeline for data curation has been developed.

The challenge will not only foster faster dose calculation for the online adaptive workflow, but also opens the possibility of real-time dose reconstruction on real-time MRI images during radiation delivery. For this reason, it will be crucial that dose calculation speed is as high as possible.

#### Keywords

List the primary keywords that characterize the task.

medical imaging, magnetic resonance imaging, MRI-linac, MRI-guided radiotherapy, real-time photon dose calculation, real-time adaptive radiotherapy

### ORGANIZATION

#### Organizers

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

See Task 1

b) Provide information on the primary contact person.

See Task 1

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

See 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

See Task 1

### Challenge venue and platform

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

See Task 1

b) Report the platform used to run the challenge.

See Task 1

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

Yes

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

See Task 1

### Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

See 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 may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

**See 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.

**See Task 1**

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

**See Task 1**

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.

**See 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).

**See 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.

**See 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.

**See 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

See 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).

See 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)

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

**CC BY-NC (Attribution-NonCommercial)**

### **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.

See Task 1

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

See 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.

See 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

See Task 1

## Task category(ies)

State the task category(ies)

Examples:

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

See 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.

The biomedical application addresses patients undergoing radiotherapy (about 50% of cancer patients). We will collect data randomly sampling cases of both sexes, ensuring a balanced representation of the sexes. An adult population will be collected. Cutting edge MRI-linacs afford the highest precision in photon radiotherapy via excellent soft tissue contrast and require photon dose calculation on MRI images (Task 2). Thus, for Task 2 (photon dose calculation on MRI), both CT and MRI are required so that CT imaging can be used to compute ground truth doses, which can be mapped onto the MRI via the CT to MRI deformable image registration performed by the challenge organizers. Inclusion criteria for Task 2 datasets would thus be treatment at an MRI-linac with both CT and MRI acquisition during treatment planning.

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

See Task 1

### Imaging modality(ies)

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

MRI

### Context information

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

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

See Task 1

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

See 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.

See 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.

For task 2, the algorithm should target photon therapy dose calculation on MRI images of the thorax and abdomen regions.

### 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)).

See 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).

MRI for Task 2 were acquired at either of the MRI-linacs listed below:

LMU University Hospital (LMU), Munich, Germany:

ViewRay MRIdian 0.35 T

Amsterdam University Medical Center (AUMC), Amsterdam, the Netherlands:

ViewRay MRIdian 0.35 T

For the CT data see Task 1.

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

All MRI images were acquired using a 3D balanced steady state free-precession sequence (bSSFP) for thorax and abdomen in breath hold, yielding a T2/T1- weighted contrast.

At LMU, abdominal MR images were primarily acquired with a matrix size of  $266 \times 266$ , a slice thickness of 3.0 mm, and an in-plane pixel spacing of 1.50 mm, with some variations including  $276 \times 276$  or  $234 \times 234$  or  $300 \times 334$  matrix and 1.63 mm pixel spacing. Lung MR images were primarily acquired with a matrix size of  $310 \times 360$ , a slice thickness of 3.0 mm, and an in-plane pixel spacing of 1.50 mm, with variations including  $300 \times 334$  and  $266 \times 266$  matrix.

At AUMC, all lung MR images were acquired with a matrix size of  $276 \times 276$ , a slice thickness of 3.0 mm, and an in-plane pixel spacing of 1.63 mm.

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



The MRI data was acquired for MRI-guided radiotherapy treatments in the radiotherapy departments of LMU Munich DE and AUMC Amsterdam NL. The training set originates from LMU and was previously part of the training set for SynthRAD2025. The testing set will partially consist of data from the private SynthRAD2025 test set (from LMU), which was not released publicly, and data from AUMC which was never released publicly. We restrict ourselves to 0.35 T MRI-linac data due to the availability of CT-MRI pairs and for consistency of the CT to MRI intensity transformation.

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 clinical staff of the respective radiotherapy departments acquired the MRI scans. All patients were treated with MRI-guided radiotherapy using repeated breath-hold conditions. Dedicated body coils were used for MRI-linac 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.

A case in this challenge task refers to a patient's MRI scan along with the photon dose distribution and photon beam parameters for a series of beams relevant for the treatment. For the training set, the photon beams are generated in a random fashion to provide a wide variety of beam configurations to ensure that algorithms generalize. For the testing set, strictly beams used for the irradiation of the patient's tumor are used. In all cases, beams refer to those used with volumetric arc therapy (VMAT).

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

For task 2, we will have MRI scans from 122 patients for thorax and abdomen.

Training and validation (publicly available):

Thorax: 39 patients

Abdomen: 36 patients

Preliminary testing:

Thorax: 5 patients

Abdomen: 5 patients

Final testing:

Thorax: 16 patients

Abdomen: 21 patients

For each patient, several hundred to thousand beams and corresponding photon dose distributions will be generated.

c) How much of the data are already annotated (stratified by train test in percentage)?

See Task 1

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

See Task 1

e) 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.

N/A

f) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

See 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.

See 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.

See 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.

See 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.

See 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.

See 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.

See Task 1

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

See 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)

See Task 1

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

See 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. Ideally, provide the ranking scheme as a concrete pseudo code.

See Task 1

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

See Task 1

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

See Task 1

### Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

**See Task 1**

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

**See Task 1**

Provide a description of how variability of the performance of individual algorithms across tests cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

**See Task 1**

Provide a description of how variability of rankings is assessed.

**See Task 1**

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

**See Task 1**

Provide a description of the missing data handling.

**See Task 1**

Indicate any software product that is used for all data analysis methods.

**See 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.

**See Task 1**

## TASK 3: Proton dose calculation on CT images

### 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.

Proton therapy offers an alternative to photon therapy which brings advantageous dose distributions due to the localized energy deposition found in the Bragg peak, where no dose is delivered downstream from the peak. This allows better sparing of organs at risk distal to tumors compared to photon therapy. In state of the art proton therapy, dose is delivered by pencil beam scanning, where a narrow pencil beam of Gaussian profile (full width half max approximately a centimeter) is scanned in a raster pattern to deliver a 3D dose to the tumor. Different depths are reached by changing the pencil beam's energy.

In proton therapy, dose calculation is particularly sensitive to the electron density of tissues (or more accurately, to their proton stopping power), making this task one of the most challenging. The goal of this challenge task is to foster the development of fast and accurate proton therapy dose calculation algorithms on CT images.

The challenge will provide a set of proton therapy pencil beam ground truth doses calculated on CT images via Monte Carlo simulation, covering a range of anatomical locations in the lung and abdomen and of pencil beam energies and angles.

#### Keywords

List the primary keywords that characterize the task.

medical imaging, computed tomography, real-time proton therapy dose calculation

### ORGANIZATION

#### Organizers

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

See Task 1

b) Provide information on the primary contact person.

See Task 1

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

See 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

See Task 1

### Challenge venue and platform

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

See Task 1

b) Report the platform used to run the challenge.

See Task 1

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

Yes

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

See Task 1

### Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

See 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 may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

See 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.

See Task 1

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

See Task 1

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.

See 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).

See 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.

See 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.

See 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

See 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).

See 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)

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

CC BY-NC (Attribution-NonCommercial)

### 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.

See Task 1

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

See 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.

See 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

See Task 1

### Task category(ies)

State the task category(ies)

Examples:

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

See 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.

The biomedical application addresses patients undergoing radiotherapy (about 50% of cancer patients). We will collect data randomly sampling cases of both sexes, ensuring a balanced representation of the sexes. An adult

population will be collected. Proton therapy offers ballistic advantages over photon therapy due to the localized energy deposition of the Bragg peak, and dose calculation is typically done on CT images (Task 3). Thus, for Task 3 (proton dose calculation on CT), only CT imaging is required. However, to provide homogeneous data across the four challenge tasks, inclusion criteria for Task 3 datasets would be treatment at an MRI-linac with both CT and MRI acquisition during treatment planning (to allow using the same imaging dataset for all tasks). While MRI-linacs are used for photon therapy and not proton therapy, the CT datasets obtained for these machines are representative of CT data which would be acquired for proton therapy.

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

See Task 1

### Imaging modality(ies)

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

CT

### Context information

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

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

See Task 1

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

See 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.

See 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.

For task 3, the algorithm should target proton therapy dose calculation on CT images of the thorax and abdomen regions.

### 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)).

See 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).

See Task 1

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

See 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.

See 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).

See 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.

A case in this challenge task refers to a patient's CT scan along with the proton dose distribution and proton pencil beam's parameters (lateral size, energy, location, direction) for a series of pencil beams relevant for the treatment. For the training set, the pencil beams are generated in a random fashion to provide a wide variety of beam configurations to ensure that algorithms generalize. For the testing set, strictly pencil beams used for the irradiation of the patient's tumor are used.

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

For task 3, we will have CT scans from 122 patients for thorax and abdomen.

Training and validation (publicly available):

Thorax: 39 patients

Abdomen: 36 patients

Preliminary testing:

Thorax: 5 patients

Abdomen: 5 patients

Final testing:

Thorax: 16 patients

Abdomen: 21 patients

For each patient, several hundred to thousand beams and corresponding proton pencil beam dose distributions will be generated.

c) How much of the data are already annotated (stratified by train test in percentage)?

See Task 1

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

See Task 1

e) 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.

N/A

f) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

See 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.

See 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.

See 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.

See 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.

See 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.

Similar to the pre-processing in Task 1, the 122 CT-MRI pairs were first processed following SynthRAD2025 preprocessing workflow, including rigid registration between CT and MRI, resampling with a consistent voxel spacing of  $1 \times 1 \times 3$  mm for both CT and MRI, and body outline segmentation using TotalSegmentator. Furthermore, instead of using the DIR from the Elastix framework in SynthRAD2025 preprocessing, the open-source ConvexAdam DIR method was adopted to achieve superior deformable registration accuracy between CT and MRI, as demonstrated by manual visual assessment of the anatomical alignment. Then, the body masks from the deformed CTs were used to remove the couch from CT images.

After the DIR quality assurance and air cavity correction (as described in the annotation characteristics section), the corrected CT images were subsequently converted into density and elemental composition maps using a scanner-specific calibration curve, allowing Geant4 to inherently calculate the material-specific proton stopping powers. All Monte Carlo simulations were performed using the Geant4 toolkit (version 11.00-patch-03) with the predefined QGSP\_BIC\_EMV as physics list to generate ground truth proton beamlet dose distributions. The initial proton energies ranged from 31.72 to 236.11 MeV (114 float energy levels). A hypothetical generic beam model representing a cyclotron with pencil beam scanning delivery was employed to define the energy-dependent beamlet energy spread (sigma ranging from 0.27 to 19.27 MeV) and spot size (sigma from 3.72 to 9.63 mm), under a single-Gaussian spatial approximation. The proton source-to-axis distance was set to 1000 cm and the dose grid matched the patient CT spacing ( $1 \times 1 \times 3$  mm). No magnetic field was simulated for the proton dose in Geant4.

For the 75 training patients, in each training patient, the isocenter was shifted along the superior–inferior axis by -4, -2, 0, +2, +4 cm. At each isocenter, 36 gantry angles were simulated every 10 from 0 to 350; at each angle, 3 energies were randomly sampled from the 114 energies to generate 3 proton beamlet dose, yielding  $75 \times 5 \times 36 \times 3 = 40,500$  training samples through randomized energies, gantry angles, and isocenters on 75 training patients.

For 47 testing patients, 47 intensity modulated proton therapy (IMPT) plans were generated based on the clinically contoured targets and organs using a scriptable research TPS. Plan doses were then calculated by MU-weighted accumulation of beamlet doses. Besides, 470 proton beamlet doses with random energies, gantry angle and iso centers were simulated on 470 test patients for the additional robustness test dataset. For both training and test proton beamlet dose simulation, proton histories were set to  $1e6$ . The prescription was 60 Gy in 20 fractions. All MC testing plan doses were scaled so that  $D_{95\%}(PTV) \geq 0.95 \times 60 \text{ Gy} = 57 \text{ Gy}$ .

### 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 primary source of uncertainty stems from residual geometric mismatches between CT and MRI scans following DIR and air cavity correction. Given the high sensitivity of proton therapy to anatomical variations, minimizing these residual errors is critical. To address this, we employed an improved DIR method and a rigorous air cavity correction workflow to mitigate discrepancies between the MRI and CT datasets.

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

Additional statistical uncertainty from Monte Carlo simulation of the proton beamlet dose is estimated to be at the level of 1% in the  $D > 10\% D_{max}$  dose region. This level was selected to offer a balance between precision and calculation time.

## 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)

See Task 1

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

See 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. Ideally, provide the ranking scheme as a concrete pseudo code.

See Task 1

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

See Task 1

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

See Task 1

### Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

**See Task 1**

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

**See Task 1**

Provide a description of how variability of the performance of individual algorithms across tests cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

**See Task 1**

Provide a description of how variability of rankings is assessed.

**See Task 1**

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

**See Task 1**

Provide a description of the missing data handling.

**See Task 1**

Indicate any software product that is used for all data analysis methods.

**See 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.

**See Task 1**

## TASK 4: Proton dose calculation on MRI images

### SUMMARY

#### Abstract

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

The very latest development in proton therapy has seen the introduction of prototype systems combining a proton therapy delivery system with an in-room MRI imaging device to allow workflows similar to those achievable at MRI-linacs for photons (see task 2). Similarly to task 2, such workflows will require proton therapy dose calculation directly on MRI images.

Similarly to task 2, the challenge will provide a set of proton therapy pencil beam ground truth doses calculated on CT images via Monte Carlo simulation, which will be mapped to MRI via the same DIR used in task 2. Like in task 3, the proton beams will cover a range of anatomical locations in the lung and abdomen and of pencil beam energies.

#### Keywords

List the primary keywords that characterize the task.

medical imaging, magnetic resonance imaging, MRI-proton prototype, MRI-guided proton therapy, real-time proton therapy dose calculation, real-time adaptive proton therapy

### ORGANIZATION

#### Organizers

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

See Task 1

b) Provide information on the primary contact person.

See Task 1

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

See 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

See Task 1

### Challenge venue and platform

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

See Task 1

b) Report the platform used to run the challenge.

See Task 1

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

Yes

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

See Task 1

### Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

See 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 may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

See 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.

See Task 1

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

See Task 1

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.

See 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).

See 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.

See 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.

See 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

See 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).

See 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)

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

**CC BY-NC (Attribution-NonCommercial)**

### **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.

See Task 1

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

See 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.

See 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

See Task 1

### Task category(ies)

State the task category(ies)

Examples:

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

See 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.

Proton therapy offers ballistic advantages over photon therapy due to the localized energy deposition of the Bragg peak, and dose calculation is typically done on CT images, but efforts to improve soft tissue contrast with MRI require dose calculation methods on that modality (task 4). Thus, for Task 4 (proton dose calculation on MRI), both CT and MRI are required so that CT imaging can be used to compute ground truth proton doses, which can be mapped onto the MRI via the CT to MRI deformable image registration performed by the challenge organizers. Inclusion criteria for Task 4 datasets would be treatment at an MRI-linac with both CT and MRI acquisition during

treatment planning (to allow using the same imaging dataset for all tasks). While MRI-linacs are used for photon therapy and not proton therapy, the CT and MRI datasets obtained for these machines are representative of data which would be acquired for proton therapy.

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

See Task 1

### Imaging modality(ies)

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

MRI

### Context information

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

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

See Task 1

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

See 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.

See 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.

For task 4, the algorithm should target proton therapy dose calculation on MRI images of the thorax and abdomen regions.

### 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)).

See 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).

See Task 2

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

See Task 2

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.

See Task 2

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).

See Task 2

### Training and test case characteristics

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

Examples:

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

A case in this challenge task refers to a patient's MRI scan along with the proton dose distribution and proton pencil beam parameters (size, energy, location, direction) for a series of pencil beams relevant for the treatment. For the training set, the pencil beams are generated in a random fashion to provide a wide variety of beam configurations to ensure that algorithms generalize. For the testing set, strictly pencil beams used for the irradiation of the patient's tumor are used.

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

For task 4, we will have MRI scans from 122 patients for thorax and abdomen.

Training and validation (publicly available):

Thorax: 39 patients

Abdomen: 36 patients

Preliminary testing:

Thorax: 5 patients

Abdomen: 5 patients

Final testing:

Thorax: 16 patients

Abdomen: 21 patients

For each patient, several hundred to thousand beams and corresponding proton pencil beam dose distributions will be generated.

c) How much of the data are already annotated (stratified by train test in percentage)?

See Task 1

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

See Task 1

e) 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.

N/A

f) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

See 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.

See 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.

See 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.

See 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.

See 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.

See Task 3

### **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.

See Task 3

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

See Task 3

## **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)

See Task 1

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

See 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. Ideally, provide the ranking scheme as a concrete pseudo code.

See Task 1

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

See Task 1

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

See Task 1



## Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

See Task 1

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

See Task 1

Provide a description of how variability of the performance of individual algorithms across tests cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

See Task 1

Provide a description of how variability of rankings is assessed.

See Task 1

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

See Task 1

Provide a description of the missing data handling.

See Task 1

Indicate any software product that is used for all data analysis methods.

See 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.

See 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.

N/A

### Further comments

Further comments from the organizers.

N/A