# TrackRAD2025: Real-time tumor tracking for MRI-guided radiotherapy: Structured description of the challenge design

Remark: This challenge has been modified. All changes are highlighted in blue.

## CHALLENGE ORGANIZATION

### Title

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

TrackRAD2025: Real-time tumor tracking for MRI-guided radiotherapy

### Challenge acronym

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

TrackRAD2025

### 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.

The use of magnetic resonance imaging (MRI) to visualize and characterize motion is becoming increasingly important in treating cancer patients, especially with radiotherapy. In particular, motion management is crucial for tumors affected by respiratory motion such as liver, pancreatic or thoracic tumors, to ensure a high radiation dose to the tumor and the sparing of neighboring organs. The recent development of MRI-guided radiotherapy, based on hybrid MRI-linear accelerator (linac) systems (1), called MRI-linacs, offers the possibility to adapt to changes in tumor position during treatment. 2D cine-MRI allows real-time tumor motion visualization and allows closely following the tumor with the radiation beam, but requires tumor segmentation on all time-resolved frames. This needs to be done in real-time, with frame rates of up to 8 Hz or more, with high accuracy and robustness to ensure the sparing of critical organs. Currently, clinically available solutions rely on conventional deformable image registration (DIR) or template matching to propagate contours from a labeled frame and struggle with large non-rigid motion. This limits treatment to beam gating, where the beam is turned off for large motion. The fast inference of artificial intelligence (AI) methods, obtained by shifting computation time to the training phase, is promising for this task (2). TrackRAD2025 will impact the field of MRI-guided radiotherapy by providing cine-MRI data from multiple MRI-linac institutions to test competitive real-time tumor tracking methods based on a unified platform for comparison.

The objective of TrackRAD2025 will be real-time tumor tracking on time-resolved sagittal 2D cine-MRI sequences. Algorithms will be provided with a template tumor segmentation on the first frame, and the remaining 2D cine-MRI sequence requiring real-time segmentation. The submitted algorithms should produce a tumor segmentation mask on each frame.

TrackRAD2025 will provide the first public multi-institutional dataset and evaluation platform to compare the latest developments in cine-MRI-based tumor tracking methods competitively. Both unlabeled (477 patient cases) and labeled datasets (108 patient cases with 2D cine-MRI frame sequences where the tumor is manually segmented) will be provided for model development and testing. Six international centers will provide data (3 Dutch, 1 German, 1 Australian, and 1 Chinese). The data from 0.35 T and 1.5 T MRI-linacs will be divided into a public training set and a private test set to calculate evaluation metrics. The challenge will feature a preliminary testing (validation) phase with 8 cases and a final testing phase with 50 cases. Submitted algorithms will be rated for their ability to reproduce ground truth segmentation labels on the test set using the Dice similarity coefficient, Hausdorff and average surface distance, error of the tumor center of mass, estimated radiation dose delivery accuracy, and the inference speed.

TrackRAD2025 will allow determining the most promising methods to improve clinical tumor tracking on cine-MRI at MRI-linacs, which will benefit patients suffering from various motion-affected tumor entities with more accurate dose delivery. This will also lead the way to multi-leaf collimator tracking at MRI-linacs instead of gating, where the radiation beam continuously follows the movement of the tumor to deliver radiation more efficiently and shorten treatment times for an increased number of treatments per day.

1. Keall PJ, Brighi C, Glide-Hurst C, Liney G, Liu PZY, Lydiard S, et al. Integrated MRI-guided radiotherapy - opportunities and challenges. Nat Rev Clin Oncol. 2022;19(7):458-70

2. Lombardo E, Dhont J, Page D, Garibaldi C, Kunzel LA, Hurkmans C, et al. Real-time motion management in MRI-guided radiotherapy: Current status and AI-enabled prospects. RadiotherOncol. 2024;190:109970.

## Challenge keywords

List the primary keywords that characterize the challenge.

medical imaging, magnetic resonance imaging, MRI-linac, MRI-guided radiotherapy, tumor tracking

## Year

2025

## Novelty of the challenge

Briefly describe the novelty of the challenge.

TrackRAD2025 will provide the first publicly available 2D-cine MRI database of moving tumors featuring both unlabelled (477 patients) and manually labelled (108 patients) cases. This will be a high quality database for which all labels have been reviewed and corrected in a quality assurance process. This will allow developing object segmentation and tracking in video sequences in the biomedical field, which will push forward the performance of real-time algorithms. Algorithms will be ranked based on both their accuracy, including metrics relevant for radiotherapy, and their real-time capabilities. Moving tumors are usually treated with widened irradiation margins to cover their motion, but with better tumor tracking, margins can be reduced, higher doses delivered and better outcomes secured for patients. Furthermore, using tumor MLC-tracking instead of beam gating increases throughput and allows more patients to benefit from MRI-linac technology.

### Task description and application scenarios

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

The objective of TrackRAD2025 will be real-time tumor tracking on time-resolved sagittal 2D cine-MRI sequences. Algorithms will be provided with a template tumor segmentation on the first frame, and the remaining 2D cine-MRI sequence requiring real-time segmentation. The submitted algorithms should produce a tumor segmentation mask on each frame. The task of the challenge will be the real-time segmentation of tumors on each frame of 2D cine-MRI sequences acquired at up to 8 Hz.

TrackRAD2025's main application scenario is the tracking of tumors during MRI-guided radiotherapy at novel MRI-linacs, to ensure increased precision when treating tumors in the abdominal, thoracic and pelvic region.

## FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

### Workshop

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

The best-performing teams will present their algorithms in short talks at a dedicated event or workshop, according to MICCAI availability. This is the first time TrackRAD is organized and we have contacted the organizers of a relevant workshop with a proposal to include TrackRAD2025 and are awaiting their feedback. If that workshop is not compatible we will contact other relevant workshops.

### 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.

If integrated with a workshop we expect half a day for the workshop and about 2 hours to present TrackRAD2025 and its results.

### 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.

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 a previous radiotherapy oriented challenge (SynthRAD2023) for which we have joined the 2025 organizers' group (SynthRAD2025).

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

At MICCAI 2025, we expect at least to draw participants from the five best-ranked teams, other participants teams, and other interested attendees at the workshop if TrackRAD2025 is integrated with it.

We will contact the teams that participated at SynthRAD2023 given their interest in radiotherapy image processing. TrackRAD2025 has been selected for presentation during ESTRO2025 which will draw additional participants from the radiotherapy community.

## 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 challenge SynthRAD2023 (https://doi.org/10.48550/arXiv.2403.08447). The leaderboard will remain open after the challenge
for new submissions.

## 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. 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 platform using newly available AWS g5 instances using a single GPU with 16 GB RAM, 8 cores CPU, and 32 GB RAM. At the MICCAI 2025 TrackRAD2025 workshop support would be needed for projectors, speakers, and microphones. We wish to hold a hybrid workshop with all the presenters available on-site.

# TASK 1: Real-time tumor tracking on cine-MRI

## 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 use of magnetic resonance imaging (MRI) to visualize and characterize motion is becoming increasingly important in treating cancer patients, especially with radiotherapy. In particular, motion management is crucial for tumors affected by respiratory motion such as liver, pancreatic or thoracic tumors, to ensure a high radiation dose to the tumor and the sparing of neighboring organs. The recent development of MRI-guided radiotherapy, based on hybrid MRI-linear accelerator (linac) systems (1), called MRI-linacs, offers the possibility to adapt to changes in tumor position during treatment. 2D cine-MRI allows real-time tumor motion visualization and allows closely following the tumor with the radiation beam, but requires tumor segmentation on all time-resolved frames. This needs to be done in real-time, with frame rates of up to 8 Hz or more, with high accuracy and robustness to ensure the sparing of critical organs. Currently, clinically available solutions rely on conventional deformable image registration (DIR) or template matching to propagate contours from a labeled frame and struggle with large non-rigid motion. This limits treatment to beam gating, where the beam is turned off for large motion. The fast inference of artificial intelligence (AI) methods, obtained by shifting computation time to the training phase, is promising for this task (2). TrackRAD2025 will impact the field of MRI-guided radiotherapy by providing cine-MRI data from multiple MRI-linac institutions to test competitive real-time tumor tracking methods based on a unified platform for comparison.

The objective of TrackRAD2025 will be real-time tumor tracking on time-resolved sagittal 2D cine-MRI sequences. Algorithms will be provided with a template tumor segmentation on the first frame, and the remaining 2D cine-MRI sequence requiring real-time segmentation. The submitted algorithms should produce a tumor segmentation mask on each frame.

TrackRAD2025 will provide the first public multi-institutional dataset and evaluation platform to compare the latest developments in cine-MRI-based tumor tracking methods competitively. Both unlabeled (477 patient cases) and labeled datasets (108 patient cases with 2D cine-MRI frame sequences where the tumor is manually segmented) will be provided for model development and testing. Six international centers will provide data (3 Dutch, 1 German, 1 Australian, and 1 Chinese). The data from 0.35 T and 1.5 T MRI-linacs will be divided into a public training set and a private test set to calculate evaluation metrics. The challenge will feature a preliminary testing (validation) phase with 8 cases and a final testing phase with 50 cases. Submitted algorithms will be rated for their ability to reproduce ground truth segmentation labels on the test set using the Dice similarity coefficient, Hausdorff and average surface distance, error of the tumor center of mass, estimated radiation dose delivery accuracy, and the inference speed.

TrackRAD2025 will allow determining the most promising methods to improve clinical tumor tracking on cine-MRI at MRI-linacs, which will benefit patients suffering from various motion-affected tumor entities with more

accurate dose delivery. This will also lead the way to multi-leaf collimator tracking at MRI-linacs instead of gating, where the radiation beam continuously follows the movement of the tumor to deliver radiation more efficiently and shorten treatment times for an increased number of treatments per day.

1. Keall PJ, Brighi C, Glide-Hurst C, Liney G, Liu PZY, Lydiard S, et al. Integrated MRI-guided radiotherapy - opportunities and challenges. Nat Rev Clin Oncol. 2022;19(7):458-70

2. Lombardo E, Dhont J, Page D, Garibaldi C, Kunzel LA, Hurkmans C, et al. Real-time motion management in MRI-guided radiotherapy: Current status and AI-enabled prospects. RadiotherOncol. 2024;190:109970.

### Keywords

List the primary keywords that characterize the task.

medical imaging, magnetic resonance imaging, MRI-linac, MRI-guided radiotherapy, tumor tracking

## ORGANIZATION

### Organizers

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

Group 1 - Data (collection, labeling, preparation)
Elia Lombardo (Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany)
Yiling Wang (Sichuan Cancer Hospital, Chengdu, China)
Pim Borman (UMC Utrecht, NL)
Matteo Maspero (UMC Utrecht, NL)
Michael Jameson (GenesisCare Sydney, Australia)
Hilary Byrne (GenesisCare Sydney, Australia)
Rob Tijssen (Catharina Ziekenhuis, Eindhoven, NL)
Miguel Palacios (Amsterdam UMC, NL)

Group 2 - Evaluation (metrics, manuscript)
Christopher Kurz (Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany)
Marco Riboldi (Department of Medical Physics, Faculty of Physics, LMU Munich, Munich, Germany)
Denis Dudas (Czech Technical University in Prague, Czechia)

Group 3: Technical organization (containerization)
Adrian Thummerer (Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany)
Tom Blöcker (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)
Matteo Maspero (UMC Utrecht, NL)
Lorenzo Placidi (Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy)
Marco Fusella (Abano Terme Hospital, Italy)

Davide Cusumano (Mater Olbia Hospital, Italy)
Coen Hurkmans (Catharina Ziekenhuis, the Netherlands)
Paul Keall (University of Sydney, Australia)

b) Provide information on the primary contact person.

Guillaume Landry (Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany), guillaume.landry@med.uni-muenchen.de

## Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

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

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 to MICCAI 2025 and ESTRO, managing to reach both the developers' (MICCAI) and the end-users' (ESTRO) communities. We have an invited talk at ESTRO2025 (May 2025) to introduce the challenge to the radiotherapy community and attract participants.

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

grand-challenge.org
The challenge is already set up on grand challenge and requirements have been estimated.

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

https://trackrad2025.grand-challenge.org

## Participation policies

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

Fully automatic

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

TrackRAD2025 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 2025. 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, 2025.  Not Allowed: Fine-tuning a model trained on any private dataset or with private weights that were not publicly available by March 15th, 2025. The use of publicly available data and models must be reported in the 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 2023-09/2025-09 (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 2025.
- 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 five teams must disclose and openly share their code 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 five teams (2500 USD from sponsor Elekta (Stockholm, Sweden), 1st place 1000 USD, 2nd place 600 USD, 3rd place 400 USD, 4th place 300 USD, 5th place 200 USD). The TrackRAD2025 organizers will consolidate the results and submit a challenge report paper to Medical Image Analysis or similar. The first five teams 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.

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 five teams 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 TrackRAD2025 organizers will consolidate the results and submit a challenge report paper to Medical Image Analysis (or similar).
Each team ranked among the top five 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 frame labels 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 and validation data (from 477 patients with unlabelled cine-MRIs and 50 patients with labeled cine-MRIs with up to 248 frames each) will be publicly available. Participants may split this data into training and validation folds as they see fit.

During phase 1 (the preliminary testing phase), input pre-testing data (8 labeled patient cases with 47 to 100 frames each covering different MRI-linacs, different anatomy, and level of tracking difficulty) will be available to provide predictions after type 2 model upload on the website (10 uploads per team). 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, 50 labeled cine-MRIs with 47-100 frames each), the teams must supply the algorithm for the type 2 challenge to the organizers following the submission link and instructions provided on https://trackrad2025.grand-challenge.org/. Teams will have two tries during phase 2. Teams will submit their dockerized tumor tracking 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 to familiarize themselves with the submission system. The results of this phase will be evaluated on a limited dataset (8 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 13 weeks after the release of the 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 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 1/12/2024
Start challenge: Release training cases 15/03/2025
Registration period 15/03/2025-1/06/2025
Training and validation phase (11+11 weeks) 15/03/2025 - 15/08/2025
Introduction of the challenge at ESTRO2025 2/05/2025 - 6/05/2025
Preliminary test phase (11 weeks, 10 submissions) 1/06/2025 - 15/08/2025
Test phase (4 weeks, max 2 submissions) 16/07/2025 - 15/08/2025
Announcements and invitation to present 10/09/2025
Presentation of the challenge results MICCAI25, South Korea 23-27/09/2025
Presentation of the challenge results ESTRO2026 April/May 2026
Post-challenge phase (4.5 years, 2 submissions/60 days) 1/09/2025-1/03/2030

## 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 `Retrospective study on AI-based target tracking in magnetic resonance imaging-guided radiotherapy` on 23/01/2024 (study number 23-1043).
UMC Utrecht approved a non-WMO request on 12/08/2024 with number 24U-1509 entitled: `TrackRad – tracking tumors for real time MRI-guided radiotherapy`. Contact persons: Department of Radiotherapy, University Medical Center Utrecht, m.maspero@umcutrecht.nl; trialburaucancercenter@umcutrecht.nl.
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).
The Sichuan Cancer Hospital ethics committee approved a corresponding clinical study entitled 'AI assisted precise MR-guided radiotherapy' on 29.04.2024 (study number: SCCHEC-02-2024-076).
GenesisCare Research Governance granted approval on 16/12/2024 for a data access request with Project Title: "TrackRad Grand Challenge – tracking tumours for real-time MRI-guided radiotherapy" under the GenesisCare Oncology Outcomes protocol version 5 approved by St Vincent's Sydney HREC.
Patient data at the Catharina Hospital Eindhoven were retrospectively collected and fully anonymised. Therefore the requirement to obtain informed consent does not apply according to the Dutch Medical Research Involving Human Subjects Act.

## Data usage agreement

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

Examples:

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

CC BY-NC (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/LMUK-RADONC-PHYS-RES/trackrad2025

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

Openly sharing teams' code is strongly encouraged but remains optional for all teams except for the five winning teams. 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 Elekta (Stockholm, Sweden), a company that manufactures 1.5 T MRI-linacs. The Elekta funding will be used exclusively to sponsor the prizes for the five winning teams. Elekta has no participation in the challenge organization, and its 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) is secured.
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
-

Assistance,Research,Intervention planning

## Task category(ies)

State the task category(ies)

Examples:

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

real-time tumor segmentation

## 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 would be collected. Inclusion criteria would be treatment at an MRI-linac with cine-MRI acquisition during treatment delivery or treatment simulation. The cine-MRI must show the treated lesion or a clinical surrogate tracking target under movement.

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, abdomen, and pelvis will be considered. Specifically, data from patients treated for lung tumors, liver tumors, pancreas tumors, rectal tumors, gynecological tumors, prostate cancer, kidney cancer, and cancer of the abdominal lymph nodes will be considered due to the accompanying movement. In total, the challenge cohort will comprise about 253 0.35 T MRI-linac and 224 1.5 T MRI-linac unlabeled patients (roughly 15% abdominal nodes or kidney, 25% liver, 30% lung, 10% pancreas, 10% prostate, 5% gynecological and 5% rectum) and about 52 0.35 T MRI-linac and 56 1.5 T MRI-linac manually labeled patients (roughly 15% abdominal nodes or kidney, 20% liver, 30% lung, 15% pancreas, 10% prostate, 10% rectum).

Patients should have undergone 2D cine-MR imaging of a moving cancerous lesion. 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 2D cine-MRI in the sagittal plane acquired at up to 8 Hz, reflecting current clinical practice in MRI-guided radiotherapy at MRI-linacs.

## Context information

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

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

We provide the MRI field strength, the imaging frame rate, and the anatomical location (thorax, abdomen, or pelvis) for the training set for all the cine-MRIs. Due to the standardization of imaging at 0.35 T (Viewray MRIdian) and 1.5 T (Elekta Unity) MRI-linacs (the only two sources of data in this challenge), the field strength corresponds to the two types of cine-MRI sequences used clinically at each machine. Participants will be able to discriminate the tracked anatomical regions via the provided label but not able to discriminate between treatment center of origin.

We provide an unlabeled training set of cine-MRI and a labeled training set of cine-MRI. The labeled training set will contain 2D tumor segmentations on up to 248 cine-MRI frames. The private testing set allows the participants' algorithms to have access to the 2D tumor segmentation in the first frame.

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

We will provide this information at the patient level:
1) B-field strength of the MRI used to scan the patient
2) frame rate of the cine-MRI acquisition

3) anatomical area out of thorax, abdomen or pelvis

### 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 2D cine-MRIs acquired from patients treated at MRI-linacs for lesions such as lung tumors, liver tumors, pancreas tumors, gynecological tumors, prostate cancer, kidney cancer, rectal cancer, and cancer of abdominal lymph nodes where motion is observed. 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.

The algorithm should target lesions in the anatomical regions related to the thorax, the abdomen and the pelvis. The main target will be the tumor or lesion visible on 2D cine-MRI. In some cases, the target for radiation may be low contrast or minimal, and a surrogate structure with motion close to the tumor will be tracked, as done clinically, e.g tracking the liver instead of a low contrast liver lesion.

### 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)).

Accuracy,Runtime

## DATA SETS

### Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Data were acquired at either of the two MRI-linacs listed below:
-ViewRay MRIdian MRI-linac operating at 0.35 T
-Elekta Unity MRI-linac operating at 1.5 T

Specific centers:
LMU University Hospital (LMU), Munich, Germany: ViewRay MRIdian 0.35 T
Sichuan Cancer Center, Chengdu, China: Elekta Unity 1.5 T
University Medical Center Utrecht (UMC), Utrecht, the Netherlands: Elekta Unity 1.5 T
Amsterdam University Medical Center (AUMC), Amsterdam, the Netherlands: ViewRay MRIdian 0.35 T
Catharina Hospital (CZ), Eindhoven, the Netherlands: ViewRay MRIdian 0.35 T
GenesisCare, Sydney, Australia: Elekta Unity 1.5 T

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

All images were acquired using the clinically adopted imaging protocols of the respective centers for each anatomical site and reflect typical images found in daily clinical routines. The cine-MRI sequences used at the 0.35 T and 1.5 T MRI-linacs are standardized, which ensures uniformity of the data for a given field strength.

At the centers using the ViewRay MRIdian, the 2D cine-MRIs were acquired in sagittal orientation with the patient in treatment position in the MRI-linac bore. During treatment simulation or delivery, the patients performed breath holds to increase the duty cycle of the gated radiation delivery. The breath-holds are followed by periods of regular breathing. The sequence was a 2D-balanced steady-state free precession (bSSFP) at 4 Hz or 8 Hz with a slice thickness of 5, 7 or 10 mm and in-plane pixel size of 2.4x2.4 or 3.5x3.5 mm2.

At the centers using the Elekta Unity, the 2D cine-MRIs were acquired in interleaved sagittal, axial and coronal orientation at some centers, and interleaved sagittal and coronal at others, with the patient in treatment position in the MRI-linac. For the challenge, only the sagittal plane has been considered. During treatment simulation or delivery, some patients performed breath holds to increase the duty cycle of the gated radiation delivery, while others breathed freely. The breath-holds are followed by periods of regular breathing. The sequence was a balanced fast field echo (bFFE) sequence at 1.3 Hz to 3.5 Hz (for the sagittal slices) with a slice thickness of 5, 7 or 8 mm and in-place pixel size of 1.0x1.0 to 1.7x1.7 mm2.

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.

Data was acquired for MRI-guided radiotherapy treatments in the radiotherapy departments of LMU Munich DE, UMC Utrecht NL, AUMC Amsterdam NL, CZ Eindhoven NL, Sichuan Cancer Center Chengdu CN, GenesisCare Sydney AU and was not provided in any previous challenge. The challenge participants will not be able to recognize the center of origin since the data are anonymized and the institution's names are substituted with institute identifiers ‘A’ to ‘F’. One of the Elekta Unity centers also provided cine-MRI data acquired with the new Comprehensive Motion Management (CMM) protocol. As this protocol is available in very few centers worldwide, we introduce an additional center letter ‘X’ for this data to ensure proper anonymization of that center.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

The clinical staff of the respective radiotherapy departments acquired the 2D cine-MRIs. All patients were treated with MRI-guided radiotherapy in free breathing or repeated breath-hold conditions. The clinical staff chose the positioning of the sagittal plane used for the 2D cine-MRI to intersect the tumor or the surrogate structure used for tracking. 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 the sequence of 2D sagittal cine-MRIs of a given patient. For the training set, we distinguished between unlabeled and labeled cases. With a case is meant a cine-MRI series (including multiple acquisitions) of a single patient.
The unlabeled cases comprise up to 20543 frames per case acquired at 1 to 8 Hz. The labeled cases consist of at least 47 human-labeled cine MRI frames per case showing the tumor or surrogate tracking structure using a binary segmentation. The test cases are similar to the labeled training cases but will not be released. The unlabeled and labeled data will comprise an equal ratio of data from the MRIdian (0.35 T) and Unity (1.5 T) systems.

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

Overall, TrackRAD2025 will have over 2.8 million unlabelled sagittal cine-MRI frames from 477 individual patients, and over 10000 labelled sagittal cine-MRI frames (+8000 from frames with multiple observers) from 108 individual patients.

Training and validation (publicly available):
Unlabeled:
253 patients with at least 20 frames from multiple radiotherapy fractions at 0.35 T
224 patients with at least 20 frames at 1.5 T (in practice most unlabelled cases have much more than 20 frames)

Labeled:
25 patients with up to 248 frames at 0.35 T
25 patients with up to 100 frames at 1.5 T

Preliminary testing (also type 2 validation):
labeled:
2 patients with up to 100 frames at 0.35 T
6 patients with up to 100 frames at 1.5 T

Final testing:
Labeled:
25 patients with 100 frames at 0.35 T
25 patients with 47-97 frames at 1.5 T

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

Labeling 2D cine-MRIs is time consuming due to the high frame rate and corresponding high number of frames per MRI. We propose to share a relatively large database of unlabeled data (477 patients with at least 20 frames each) that may allow unsupervised training strategies.

An additional 50 cases with at least 47 frames with ground truth labels is provided as a public training set. The participants will be free to subdivide that set into training and validation as they see fit.

Clinically meaningful evaluation of the proposed methods requires labeled ground truth data. We have aimed at having a representative private test set of 50 cases, with up to 100 frames per case labeled. In some cases (n=34), we have multiple observers, allowing us to compare to inter-observer variability.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

We have aimed to have a broadly representative set of data. Firstly, the data is closely split 50/50 between the two main types of MRI-linacs operating clinically (0.35 T and 1.5 T systems). Secondly, we have aimed at distributing the types of tumors according to the likelihood of motion. We have sampled more thoracic+abdominal tumors than pelvic lesions. Nonetheless, we have aimed to include most of the tumor types treated clinically. All cases were selected from clinical practice.

e) 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.

The test set will feature unseen, unpublished data of 50 labelled patient cases. These cases will be labelled.

## 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.

Human observers have performed the reference annotations. For dataset D, 4 radiation oncologists and one medical physicist have independently segmented the tumor on the cine-MRI frames using software provided by the 0.35 T MRI-linac vendor (Viewray Technologies).

For dataset C, two radiation oncologists independently segmented the tumor on the cine-MRI frames using itk-snap.

For dataset A, two observers (a medical student and a dentistry student) labeled the cine-MRI frames using an in-house labeling tool developed for the challenge. A medical physics doctoral student with 4 years experience in tumor tracking then reviewed and corrected the labels.

For dataset B a medical physics researcher (assistant professor) with more than 10 years experience in radiotherapy used the labeling tool developed for the challenge (same as in A) to delineate the cine-MRI frames.

All datasets were reviewed by the cohort A medical physics doctoral student with 4 years of experience in tumor tracking who performed corrections for errors such as erroneous clicks generating separated islands of segmentation and inconsistent inclusion of structures over the course of a cine-MRI (for example, adding a vessel to the label in the later part of a cine-MRI, but not in the first part).

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 annotators had access to a delineated static 3D MRI showing the tumor segmentation. When available, they also had access to clinical videos showing the cine-MRI and the target used clinically. This target could either correspond to the tumor segmented on the 3D MRI, or an alternative surrogate tracking structure. If a surrogate structure was used, the annotators were instructed to segment that on the cine-MRIs. Otherwise, they were instructed to segment the tumor on the cine-MRIs. During annotation with the LMU labeling tool they were instructed to visualize the contour on the previous frame to reduce frame-to-frame intra-observer variations.

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.

For the dataset D:
Expert 1: 4 years of professional experience as Radiation Oncologist.
Expert 2: 3 years of professional experience as Radiation Oncologist.
Expert 3: 5 years of professional experience as Radiation Oncologist.
Expert 4: 14 years of professional experience as Radiation Oncologist.
Expert 5: 9 years of professional experience as Medical Physics Expert.

For dataset C, the 1st annotator has 2 years of professional experience as an Associate Chief Physician in radiation therapy. The second annotator has 6 years of professional experience as a Chief Attending Physician in radiation therapy.

For dataset A, both observers are medically trained (one medical student and one dentistry student). The medical physics doctoral student who reviewed the labels has two years of experience as a Qualified Medical Physicist in radiotherapy and four years of experience in tumor tracking research.

For dataset B, one observer performed all the delineations. The observer has more than ten years of experience in radiotherapy, with ten years of experience as an MRI physicist, and is undergoing a trainee as a Qualified Medical Physicist.

All datasets were reviewed by the cohort A medical physics doctoral student with 4 years of experience in tumor tracking who performed corrections for errors such as erroneous clicks generating separated islands of segmentation and inconsistent inclusion of structures over the course of a cine-MRI (for example, adding a vessel to the label in the later part of a cine-MRI, but not in the first part).

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.

For cases with multiple annotators, we will use the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm to generate a single ground truth [*]. These cases will also be used to estimate the baseline inter-observer variations.

[*] Warfield, Simon K., Kelly H. Zou, and William M. Wells. "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation." IEEE transactions on medical imaging 23.7 (2004): 903-921.

### Data pre-processing method(s)

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

All frames were resampled to 1x1 mm using linear interpolation. The annotation of labels then took place using the resampled 1x1 mm images.

### 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 the inter-observer variation. We have used the D dataset with 5 observers to obtain the STAPLE ground truth and calculated geometric accuracy of the five observers against this ground truth.

We have estimated the inter-observer agreement in terms of Dice similarity coefficient of 0.94 +- 0.07, 95th percentile Hausdorff distance of 4.2 +- 2.8 mm, mean average surface distance of 1.0 +- 1.7 mm and mean center distance of 1.2 +- 0.9 mm. For a frame, these metrics were calculated as the median of the five observers vs the STAPLE ground truth. The results of all frames from a case were averaged, and finally the median over cases was taken. We used the median when the data was not normally distributed (observers and cases) and the mean when it was (over the frames of one scan).

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

Geometric distortion could impact the target segmentation accuracy. This can be mitigated by positioning the clinical tracking target at the center of the field of view. Additionally, such an error would affect both the ground truth and model predictions similarly, and would thus not impact the ranking of the models themselves.

# 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)

The model predictions will be evaluated with geometric metrics [A] for each single frame in comparison against the ground truth expert label :
The Dice similarity coefficient (DSC) will be computed [B]
The surface distance distribution will be computed, and the 95th percentile of the distribution will be reported. This is also often called the 95th percentile Hausdorff distance (HD95). [C]
The mean average surface distance (MASD). [D]
The center of mass of the ground truth and prediction will be computed and the Cartesian distance in 2D will be computed (center distance, CD) [E].

[A] Maier-Hein, Lena, et al. "Metrics reloaded: recommendations for image analysis validation." Nature methods 21.2 (2024): 195-212.
[B] https://metrics-reloaded.dkfz.de/metric?id=dsc
[C] https://metrics-reloaded.dkfz.de/metric?id=hd95
[D] https://metrics-reloaded.dkfz.de/metric?id=masd
[E] https://metrics-reloaded.dkfz.de/metric?id=center_distance

We will include a radiotherapy-specific dose metric per cine-MRI:

5) We will estimate in 2D the accuracy of the radiotherapy dose in a multileaf collimator tracking scenario based on the model predictions. For this, the ground truth label of the first frame will be converted to an approximated radiation dose by applying a 3 mm expansion of the gross tumor volume (GTV) indicated by the label to the clinical target volume (CTV). Subsequently, this expanded mask will be smoothed by a Gaussian of 6 mm standard deviation for targets in the lung (simulates a dose fall-off similar to those observed for lung patients) and of 4 mm for all other targets (dose fall-off for targets in higher density tissue). This dose distribution will be shifted by the difference between the ground truth centroid position of the tracking target and the centroid position obtained by the investigated model for each frame. These shifted distributions will be averaged to get a centroid-error shifted dose. The relative difference between the GTV (or tracking target) D98% (from the cumulative dose volume histogram) for the ground truth distribution and the final shifted distribution will be calculated for each patient.

Additionally, the average model runtime will be considered:

6) Models must run sufficiently fast to be compatible with few Hz cine-MRI imaging, as done clinically. We will calculate the average model runtime per frame over the test set.

The execution time for each case is measured using the started_at and completed_at timestamps provided by Grand Challenge. This total runtime consists of four distinct components: the time required to load the container, the time needed to load the frames, the time required to load the submission code, and the time spent processing the frames within the submission.

To analyze the runtime more effectively, we perform a linear regression of the total execution time as a function of the number of frames. This approach allows us to calculate the variable runtime per frame, which accounts for both the per-frame loading time and the per-frame processing time. Our analysis shows that the loading time per frame is negligible and consistent across participants. As a result, we use the variable runtime per frame as the basis for ranking the submissions.

We note that this approach can be exploited by artificially increasing the fixed (loading) components of the runtime. However, the requirement for publication of the algorithm will allow us to spot such and exclude such submissions or evaluate them without any malicious components.

We also impose a maximum allowable runtime per case. This limit was established by benchmarking a relatively large transformer model that just met the real-time requirement of 8 Hz on an RTX A6000 GPU from 2020. We measured this model's performance on the Grand Challenge platform and set the runtime limit accordingly with an additional margin.

Algorithms exceeding this maximum runtime of 1 sec per frame (plus model and data loading time) on the provided hardware will be excluded from the challenge due to concerns for real-time applicability of the algorithm.

Missing output on single frames:

In cases where an algorithm produces no output on a given frame with ground truth label available, the following default metric values will be used: (1) DSC=0, (2) HD95/(3) MASD/(4) CD=image size along the largest dimension in mm, (5) dose set to zero for that frame, (6) calculate as in normal cases.

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

Tumor tracking is ultimately a geometric problem, and the four geometric metrics selected are complementary and state-of-the-art to evaluate the geometric accuracy of segmentations in many fields. The DSC correlates well with ground truth-prediction overlap. The DSC may saturate towards unity for large structures [A], but the HD95 remains sensitive to contour mismatches even for large volumes. The HD95 relates to the worst mismatches on top of the average mismatch (MASD). We chose the 95th percentile instead of the 100th percentile (which is the Hausdorff distance) to be robust to outliers. The CD is important for MLC tracking in radiotherapy and for tumor tracking. The dosimetric accuracy is the closest to the application of MLC tracking for better dose delivery. It reflects how well we would cover the target if using such a model clinically [B]. Finally, faster model run-time limits latencies between tumor motion and beam adaptation and should be reduced.

[A] Reinke, Annika, et al. "Understanding metric-related pitfalls in image analysis validation." Nature methods 21.2 (2024): 182-194.

[B] Lombardo, Elia, et al. “Patient-specific deep learning tracking framework for real-time 2D target localization in MRI-guided radiotherapy.” International Journal of Radiation Oncology, Biology, Physics. (2024)

## Ranking method(s)

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

We will consider no-registration (copying the ground truth label of the first frame to all other frames) as the baseline model, which will serve as a submission threshold.
Teams will not be considered in the ranking if their method does not outperform the no-registration baseline in at least one of DSC, HD95, MASD and CD.

For each of the 4 geometric metrics, the obtained results per cine frame will be averaged over all frames. The average over all test cases will be determined to obtain a single value per metric and model. The dosimetric and runtime metrics are already averaged over all frames by definition.

For each submission, the rank for the 6 average metrics will be calculated compared to all submissions (DSC, higher better, HD95/MASD, lower better, CD error, lower better, relative D98%, higherbetter, runtime, lower better). The final rank for a submission is obtained by computing the average rank over the 6 metrics, ranging from 1 (best submission) to n (worst submission).

We will incorporate an automatic evaluation pipeline on the hosting site, along with a leaderboard. The ranking among all metrics will be determined according to the RankThenMean procedure described above. In case of a tie, the team with the fastest model will rank higher than teams with an equal average rank.

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

In the preliminary testing phase, if the algorithm does not produce an output, the participants will have access to the log file and output result to debug their submission. During the final testing phase, the dockerized algorithms must be submitted, so if the preliminary test phase works, all the result cases must be present. If single frames of a given submission are empty, penalties as outlined at the end of section ASSESSMENT METHODS Metric(s) a) will be applied.

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

- Ranking per metric is a transparent way to obtain the final ranking, preserving the strengths and weaknesses of a method while placing equal weight on the importance of each metric.
- Aggregating and ranking were chosen to preserve large and small performance differences while combining metrics.
- After a previous investigation from another challenge (SynthRAD2023 (see https://doi.org/10.1016/j.media.2024.103276 sections 2.6 and 4.5)), RankThenMean could be used instead of MeanThenRank without having to resort to normalization of metrics, something which requires arbitrary parameters to balance the weight of each metric when calculating the mean.

## Statistical analyses

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

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

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

Missing data will be handled as described at the end of section ASSESSMENT METHODS
Metric(s) a).

Details about the assessment of variability of rankings:
We will use bootstrapping to generate 1000 test sets consisting of 50 randomly selected patient cases from the test dataset, with patients potentially being selected more than once. The variability of the ranking will be assessed with a Kandall`s tau analysis using the original ranking and the 1000 bootstrapping test sets.

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

The combination of bootstrapping and Kendall`s tau provides a good estimation of the robustness of a ranking.

## 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 investigate whether including only the geometric accuracy, the dose evaluation, or inference time will lead to different rankings. Additionally, we will investigate the impact of adding object detection metrics such as the J&F; scores (Jaccard index and Contour accuracy as defined here https://openaccess.thecvf.com/content_cvpr_2016/papers/Perazzi_A_Benchmark_Dataset_CVPR_2016_paper.pdf) and the failure rate (CD larger than a threshold of 3 mm as defined here https://www.sciencedirect.com/science/article/pii/S0360301624035089).

We will categorize performances based on the participants' methods to generate the labels on cine-MRI frames. This analysis will, for example, consider the difference between DIR-based vs instance segmentation approaches, the use of general purpose pre-trained foundation models using either non-domain specific open data (natural images or videos), domain-specific open data (medical images in general), or using the labeled and unlabeled cine-MRI training data provided by this challenge (supervised, self-supervised or unsupervised approaches). Furthermore, we will analyze the influence of the automated quantitative analysis on biases in our data and methods, considering, for example, the level of inter- and intra-observer variability in generating the ground truth and the amount of motion exhibited in the cine-MRI (by evaluating the performance of the no-registration baseline).

# 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