Synthesizing computed tomography for radiotherapy: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Synthesizing computed tomography for radiotherapy

Challenge acronym

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

SynthRAD2023

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 impact of medical imaging on oncological patients' diagnosis and therapy, has grown significantly over the last decades. Especially in radiotherapy (RT), imaging plays a crucial role in the entire workflow, from treatment simulation to patient positioning and monitoring.

Traditionally, 3D computed tomography (CT) is considered the primary imaging modality in RT, providing accurate and high-resolution patient geometry and enabling direct electron density conversion needed for dose calculations and plan optimization [Chernak et al., 1975]. For patient positioning and monitoring before, during, and after dose delivery, 2D X-ray-based imaging has been widely adopted. In recent years, 3D cone-beam computed tomography (CBCT) - often integrated with the dose delivery machine - is increasingly playing a vital role in traditional and more advanced image-guided adaptive radiation therapy (IGART) workflows in both photon and proton therapy.

A key challenge in using CBCT is that due to the severe scatter noise and truncated projections, image reconstruction is affected by several artifacts, such as shading, streaking, and cupping.
As a result, CBCT is insufficient to perform accurate dose calculations or replanning. Consequently, patients need to be referred to a rescan CT when anatomical differences are noted between daily images and the planning CT [Ramella et al., 2017]. As an alternative, image synthesis has been proposed to improve the quality of CBCT to the CT level, producing the so-called “synthetic CT” (sCT) [Kida et al., 2018]. Additionally, conversions of CBCT-to-CT that enable accurate dose computation allow online adaptive CBCT-based RT workflows to improve the quality of IGART provided to the patients.

In parallel, over the last decades, magnetic resonance imaging (MRI) has also proved its added value for tumor and organs-at-risk delineation thanks to its superb soft-tissue contrast [Schmidt et al., 2015]. MRI can be acquired to match patient positioning to the planned one and monitor changes before, during, or after the dose delivery

[Lagendijk et al., 2004].

To benefit from the complementary advantages offered by different imaging modalities, MRI is generally registered to CT. Such a workflow requires obtaining both CT and MRI, increasing workload, and introducing additional radiation to the patient. Recently, MRI-only based RT has been proposed to simplify and speed up the workflow, decreasing patients' exposure to ionizing radiation. This is particularly relevant for repeated simulations or fragile populations like children. MRI-only RT may reduce overall treatment costs and workload and eliminate residual registration errors when using both imaging modalities. Additionally, MRI-only techniques can benefit MRI-guided RT [Edmund and Nyholm, 2017].

The main obstacle in introducing MRI-only RT is the lack of tissue attenuation information required for accurate dose calculations. Many methods have been proposed to convert MR to CT-equivalent images, yielding sCTs suitable for treatment planning and dose calculation.

Artificial intelligence algorithms such as machine learning or deep learning have become the best-performing methods for deriving sCT from MRI or CBCT. With so many algorithms available, all tested on different datasets; it is unclear which algorithm works better and which does not. Unfortunately, no public datasets or challenges have been designed to provide ground truth for this task and benchmark different approaches against each other. A recent review of deep learning-based sCT generation also advocated for public challenges to provide data and evaluation metrics for such open comparison [Spadea & Maspero et al., 2021].

We designed a challenge to provide the first platform offering public datasets and evaluation metrics comparing the latest developments in sCT generation approaches. Two tasks were defined: 1) MRI-to-sCT generation to facilitate MRI-only RT and 2) CBCT-to-sCT generation to facilitate IGART and online adaptive RT.

Two multi-center datasets of matched CBCTs (inputs) and CTs (targets) and MRIs (inputs) and CTs (targets) with heterogeneous acquisition protocols will each be divided into balanced training, validation, and test sets. We will initially share the inputs and targets from the training sets to design and evaluate algorithms generating sCT. In the first round, validation input images are shared for six weeks, and the teams can upload network dockers twice a week to validate their performance and see how they rank on the leaderboard (test phase). Then, for all teams, independent test cases
will be shared to generate sCT for our final evaluation (validation phase). Brain and pelvic datasets from three Dutch centers (UMC
Utrecht, UMC Groningen, and Radboud Nijmegen) have been collected, providing more than 500 patients for MRI-to-CT and 500 patients for CBCT-to-CT undergoing radiotherapy treatments in the corresponding departments.

The challenge will run on https://grand-challenge.org/, and the pre-processing and evaluation code used to rank the submissions based on image and dose evaluations will be shared.

Challenge participants may choose to participate only in task 1 (MRI-to-sCT), only in task 2 (CBCT-to-sCT), or in both, but for either task, sCT need to be generated for both anatomical regions. In the initial validation phase, training data will be made available and teams can upload a docker containing their networks to be evaluated. For validation, we only consider geometric evaluation metric, which will appear on an open leaderboard. After the validation phase, a final test phase, with new testing data, will be held for 4 weeks during which teams can only upload their algorithm twice. During testing, scores, containing a balance between geometric and dosimetric contributions, will appear on a separate open leaderboard.

Challenge keywords

List the primary keywords that characterize the challenge.

medical imaging, magnetic resonance imaging, computed tomography, cone beam computed tomography, image synthesis, generative model

Year

The challenge will take place in ...

2023

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

none.
The best-performing teams will present their algorithms in short talks at a dedicated event or workshop, according to MICCAI availability. Note that the event can also be held purely online if necessary. Alternatively, but not desired, we could try to align with one of the workshops, e,g. SASHIMI and present the results of the challenge

Duration

How long does the challenge take?

Half day.

Expected number of participants

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

is > 50. We expect that at least the ten best-ranked teams will participate at MICCAI 2023, other participants teams, and other interested attendees at the workshop. It is the first time that such a task has been considered for a challenge, but considering the research interest, we believe enough teams will participate. We think 20 teams could be a good estimate of the overall challenge participation. We have already started promoting the challenge, and, at the moment, we can confirm that five teams pre-subscribed on the challenge.org website.

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 arXiv if not yet fully published) along with the start of the challenge.

After the challenge, we plan to coordinate a publication (overview paper) describing the challenge results in a peer-reviewed journal paper in which we summarize the results and outcomes of the challenge. A new leaderboard will remain open after the challenge for new submissions for as long as there is budget to evaluate the metrics.

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.

None if the event is held online; otherwise, a room with PC and beamer + possibility to stream the event. Algorithms should be able to run on 16 GB GPU, with 8 cores and 32 GB RAM.

TASK: MRI to CT

SUMMARY

Keywords

List the primary keywords that characterize the task.

medical imaging, magnetic resonance imaging, computed tomography, image synthesis, generative model

ORGANIZATION

Organizers

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

Adrian Thummerer, UMC Groningen;

Arthur Jr. Galapon, UMC Groningen;

Peter Koopmans, Radboudumc (Nijmegen);

Evi Huijben, Eindhoven University of Technology;

Manya Afonso, Wageningen Research;
Maarten Terpstra, UMC Utrecht;

Matteo Maspero, UMC Utrecht;
Maureen, van Eijnatten, Eindhoven University of Technology;

Oliver J. Gurney-Champion, Amsterdam UMC;

Suraj Pai, Maastricht University;

Zoltan Perko, TU Delft.

b) Provide information on the primary contact person.

Matteo Maspero, Assistant professor, UMC Utrecht, Computational Imaging group, m.maspero@umcutrecht.nl / matteo.maspero.it@gmail.com

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 with an open leaderboard for submission after closing the challenge.

Challenge venue and platform

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

MICCAI.

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

https://synthrad2023.grand-challenge.org/

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

https://synthrad2023.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.

The data used to train algorithms are restricted to the data provided by the challenge. Pre-trained nets may NOT be used in the challenge.

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 if not listed among the organizers, contributors, or data providers and did not co-author any publication with the organizers in the last year; otherwise, they are not eligible for the prizes. Organizers, contributors, or data providers may not participate in the challenge.

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 team can be composed of 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 a condition to participating in the challenge, all teams must sign an greement, based on the restrictive TCIA policy form, stating that no attempt to de-anonymize the patients’ information will be made on behalf of the team.

As other conditions for being awarded a Prize, the best top ten performing teams must fulfil the following obligations:

1) Present their method at the final event of the challenge at MICCAI 2023;
2) Sign and return all prize acceptance documents as may be required by Competition Sponsor/Organizers.
3) The price eligibility is conditional on submitting a paper reporting the details of the methods in a short or long (up to the teams) LNCS format.
4) Commit to citing the data challenge paper and the data overview paper whenever submitting the developed method for scientific and non-scientific publications.

The participating teams are strongly encouraged to disclose or share their code, although not mandatory.

We are arranging prizes for 5000 eur per task.

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, and the top teams will be invited to present their approach during the final 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 so that their performance results will become part of presentations, publications, or subsequent analyzes derived from the challenge at the discretion of the organization. Specifically, all the performance results will be made public.
The SynthRAD2023 organizers will consolidate the results and submit a challenge paper (to IEEE TMI, MEDIA, LNCS issue, or similar).

Each team ranked among the top ten metrics 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 sCT 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 Section 8c. 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 final event. The embargo is counted from the final 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 overview 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.

We will organize a type 2 challenge, where algorithm submissions run through the website, as described in https://grand-challenge.org/documentation/challenges/. Training data (MRI or CBCT and corresponding CT) will be publicly available. Both for the test and validation phase, the teams must supply the algorithm for type 2 to the organizers following the submission link and instructions provided on https://synthrad2023.grand-challenge.org/. The same holds for the test phase: teams will submit their dockerized sCT algorithms to the challenge website without having the data at their disposal. Once the challenge is presented at MICCAI, the validation data and target will be made available but not the test data. After the challenge conclusion, the challenge will remain open as a type 2 and the leaderboard will be updated, until the budget is finished. this will be announced through the forum.

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 validation and test phase. The validation phase lasts six weeks and will allow up to 2 submissions per week for each team and let the teams get familiar with the submission system and compare their performance with the other teams. The results of the validation phase will be evaluated on the validation set with an open scoreboard.

The validation phase will be closed from a specific date (see below), and the test (final submission) phase will start. The validation phase start eight weeks after releasing the data, and will be open for six weeks. After this, the test phase starts, during which the algorithms can be submitted maximally twice to be applied to the test data. At the end of the test phase, top-performing
teams will be invited to present their methods.

The participating teams can submit up to two runs to evaluate their algorithms on this test 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 be identified with a description.

To ensure that no error occurs during the submission process, we plan to send warnings when the submission format is incorrect or when the performance is below a (very low) threshold, suggesting errors. We keep the test phase relatively short (four weeks) to avoid teams having too much additional time to refine their methods further but still allow teams to process all the input data (120 input cases per task).

The participating teams that want to be considered for an award are obliged to describe their methods by submitting a short or long paper in LNCS format (https://www.springer.com/gp/computerscience/lncs/conference-proceedings-guidelines).

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

Begin challenge: Release training cases: 1/04/2023

Training phase: 1/04/2023 - 31/05/2023 (8 weeks+ 6 weeks from validation + 4 weeks during the test)

Presentation of the challenge at ESTRO23: 13/05/2023

Validation phase: 1/06/2023-15/07/2023 (6 weeks)

Test phase: 16/07/2023-15/08/2023 (4 weeks)

Deadline for registration method paper: 15/08/2023

Dose evaluation for the top-ranked solutions in the leaderboard: 15/08/2023-20/09/2023

Submission deadline for the method paper: 19/09/2023

Announcements and invitation to present: 20/09/2023

Presentation of the challenge results in 8-12/10/2023, MICCAI, Vancouver
April/May 2024,
ESTRO (TBD)

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).

Each institution is providing data requesting ethics approval to the internal review board/Medical Ethical committee of its institute:
UMC Utrecht approved not-WMO on 4/03/2022 with number 22/474 entitled: "Synthetizing computed tomography for radiotherapy Grand Challenge (SynthRAD)." Contact persons: Department of Radiotherapy, University Medical Center Utrecht, m.maspero@umcutrecht.nl; trialburaucancercenter@umcutrecht.nl.
UMC Groningen approved non-WMO on 20/07/2022 with number 202200310 entitled: "Synthesizing computed tomography for radiotherapy - Grand Challenge" Contact persons: Department of Radiation Oncology, University Medical Center Groningen a.thummerer@umcg.nl / s.both@umcg.nl.
RadboudumcUMC declared the study non-WMO on 17/10/2022 with number 2022-15950 entitled "Synthetizing computed tomography for radiotherapy Grand Challenge" Contact persons: Department of Radiation Oncology, Radboud University Medical Center Nijmegen, erik.vanderbijl@radboudumc.nl.

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)

Additional comments: Data are released under CC BY-NC (Attribution-NonCommercial). This will be performed via Zenodo at 10.5281/zenodo.7260704.

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 made publicly available on GitHub at the following location: https://github.com/SynthRAD2023

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

Openly sharing the teams' code is strongly encouraged but remains optional.
Conflicts of

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 endorsed by ESTRO and NVFK. A Seed Fund grant (https://ewuu.nl/en/collaboration/seed-fund/) has been granted to cover the challenge's computational costs on the website and organizational cost (and prices). Additional funding is under discussion with commercial parties, but no confirmed participation is ensured at the moment of submission of this template.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Research, Intervention planning.

Additional points: Radiotherapy;

Task category(ies)

State the task category(ies).

Examples:

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

Synthesis (image); Regression (image)

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 addressed patients undergoing radiotherapy (~50% of cancer patients). No gender restriction is considered, and a predominant adult population is collected. Inclusion criteria are the acquisition of CT and MRI during planning (task 1) to ensure accurate positioning during image-guided therapy. MRI and CBCT should be acquired within two months of the CT to reduce anatomical changes. Datasets for tasks 1 and 2 do not necessarily contain the same patients given the separated tasks.

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

Patients undergoing radiotherapy in the brain and pelvis will be considered for both tasks. For task 1, patients should have undergone MRI and CT.

Imaging modality(ies)

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

For task 1, patients should have undergone MRI and 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).

The following additional information is given along with the images:
1. Acquisition protocols for MRI, CBCT, and CT of the acquired images as extracted from Dicom header, and the image size and resolution of the images after pre-processing to provide the size of the finally provided images.

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

2. Gender, weight, and age (if available) of the patients.

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.

brain and pelvis oncological patients acquired in CT, MRI treated in three Dutch Radiotherapy departments.

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.

generated images similar to CT based on MRI (task 1) images. The whole patient volume in the FOV of the input image is considered.

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)).

Robustness, Accuracy.

Additional points: Properties of the algorithms to be optimized to perform well in

1. Generate synthetic CT from MRI (task 1).
2. Image similarity. Generate CT-like images (called synthetic CT (sCT)) according to image similarity metrics.
3. Dose evaluation. According to the dose evaluation metrics, the dose planned on the original CT and recalculated on the sCT should be similar.

The image similarity metrics rank the methods during the validation phase. In the test phase, also the effect on dose distribution is considered. The dose has the highest clinical relevance as it determines the treatment. Dose evaluation depends on matRad (https://e0404.github.io/matRad/) open source fully validated treatment planning system written in Matlab. We are investigating whether it is feasible to incorporate an automatic evaluation pipeline on the hosting site, making it possible to offer dose evaluation for all the tests. If this result is too cumbersome or computationally expensive, we will only apply dose metrics to the top-ranking teams in the image similarity metrics.

The image similarity is used as a surrogate to provide the participants quick feedback during the validation phase. The final metric combines image similarity and dose evaluation (see point 26).

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).

Images were acquired on the following devices:

MRI:
Center A (UMC Utrecht) - Brain/Pelvis: Philips Ingenia 1.5T/3.0T
Center B (UMC Groningen) - Brain: Siemens MAGNETOM Aera 1.5T, Siemens MAGNETOM Avanto_fit 1.5T; Pelvis: no data
Center C (Radboud Nijmegen) - Brain: Siemens Avanto fit 1.5T, Pelvis: Siemens MAGNETOM Vida fit 3.0T

CT:
Center A (UMC Utrecht) - Brain/Pelvis: Philips Brilliance Big Bore, Siemens Biograph20 PET-CT (5-10 % of the data)
Center B (UMC Groningen) - Brain: Siemens SOMATOM Definition AS; Pelvis: Siemens SOMATOM Definition AS, Siemens SOMATOM go.Open Pro, GE Medical Systems Optima CT580
Center C (Radboud Nijmegen) - Brain/Pelvis: Philips Brilliance Big Bore

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

Images were acquired with the clinically used imaging protocols of the respective centers for each anatomical site and reflect typical images found in daily routine. Due to different clinical routines, imaging protocols for CT vary within and between datasets, which reflects a realistic application scenario. A detailed table of imaging parameters will be distributed with the dataset.

For the MRI task (Task 1), a T1 weighted gradient echo sequence was selected for all datasets. Acquisition protocols varied between sites. The dataset of centers B and C only includes images acquired with a gadolinium contrast agent, while the selected MRIs for centers A were acquired without contrast.

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 radiotherapy treatments in the radiotherapy departments of UMC Utrecht, UMC Groningen, and Radboud Nijmegen and not provided in any previous challenge. The participant will not be able to recognize the origin center since the data are anonymized. Metadata is provided, and the institution's names are substituted with institutes A, B, and C.

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).

Data was acquired by clinical staff of the respective radiotherapy departments. Patients treated in the pelvis had either cervical, rectal, or prostate cancer. The clinically adopted delineation of target structures and organs at risk for the test set will be used for evaluation purposes (dose evaluation). Each center may have different routines concerning the patients' positioning in CT and MRI (e.g., different immobilization devices such as masks, or vacuum cushions, use of flat tops) and which guidelines were followed for delineations. The guidelines among the centers have been compared, finding that all centers followed the Dutch guidelines (https://richtlijnendatabase.nl/). Details about the delineations for each anatomical site will be reported in an upcoming publication focusing on the data. A pre-print (or, if time allows, a publication) will be released along with the challenge.

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 refers to a single patient consisting of a CT and MRI of the same patient acquired < 2 months apart. Radiotherapy treatment plans and structure sets (target volume and organs-at-risk) are also available for test cases. A train case has both input (MRI for task 1) image and reference (CT) case; while a test/validation case has only input image.

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

For the brain, all three institutes will provide 60/10/20 datasets for training, validation, and testing, respectively. For the pelvis, institute B does not have MRIs with a sufficiently large FOV. To compensate, institute A will provide twice the data for the pelvis: 120/20/40. Institute C will provide the normal amount of 60/10/20 cases.
Total cases task 1:

Pelvis: 180 training, 30 validation, 60 test cases
Brain: 180 training, 30 validation, 60 test cases

Total cases for both the tasks= 360 training, 60 validation, 120 test, for a total of 540 cases

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

The total number of cases and the separation into training, validation, and test sets are larger than previous studies in the field (see the review article by Spadea and Maspero et al., 2021) that considered on average between 30-50 patients. Considering the multicenter setting of the data, we have included -when available- at least 60 patients per center.

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.

For each center, the imaging protocol was included if the protocol was comparable to at least one-third of the population. This has been performed to ensure that class balance is preserved, helping the challenge participants develop methods to handle the multi-center variability. Case selection in the brain was blind to clinical information concerning primary tumor etiology, making the tumor characteristics a random sample of the clinical routine. In the pelvis, cervical, rectal, and prostate cases were considered equally distributed among training, validation, and test sets on an institute level.

Annotation characteristics

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

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

No annotators can be considered given that the task is regression, and no manual interaction can be considered on the provided data. The data has annotated whether the anatomy is pelvis or brain. This will facilitate running separate models on the two anatomy if desired

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.

N/A

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.

Image protocols on a per-patient basis will be provided in excel files reporting each imaging modality's relevant acquisition and reconstruction settings. However, this information will not be used for quantitative evaluation.

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.

N/A

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.

Data pre-processing includes the following steps:
1. Image conversion: All images are converted to the compressed nifty file format (.nii.gz, https://nifti.nimh.nih.gov/)
2. Image resampling: All images will be resampled to a uniform voxel spacing (Brain: 1x1x1 mm3, Pelvis: 1x1x2.5 mm3). Image size might vary from case to case and between datasets.
3. Image registration: A rigid image registration between CBCT (for task2) or MR (for task1) and resampled CT will be performed for all cases to align the images (using Elastix: https://elastix.lumc.nl/index.php). Participants can further improve the registration (rigid and deformable) if required. Elastix parameter files will be made available for participants for rigid registration and a suggestion for deformable registration. During the evaluation, rigid image registration between CBCT/MR and CT will reduce anatomical differences.

4. Image masking: The field-of-view of MR/CBCT and CT will be aligned by automatically segmenting based on the threshold and morphological operation of the patient outline on the MR/CBCT. The resulting mask will also be dilated to include surrounding air and applied to the CT image. The mask will be provided, and it can be used by the participant, if needed, for pre-processing. Also, image similarity metrics will be calculated in this mask whenever specified.
5. Image cropping: All images will be cropped to the bounding box of the patient outline (with a margin of 20 voxels in plane) to reduce the amount of data.
6. Defacing: For brain patients, face removal/de-identification will be performed. A parallelepiped mask for defacing has been obtained starting for the location of the eyes as available in the clinical delineation. If the eyes were not available, a surrogate structure has been selected or manually made. Defacing may vary between centers: UMC Utrecht: for task 1, the MRI has a FOV with no face included, but CT has a larger FOV. The MR mask from point 4 will be applied to CT to remove the face; for task 2, defacing is still under investigation.

The code used to perform pre-processing will be made publicly available at https://github.com/SynthRAD2023.

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.

Image registration may not perfectly match the two images (CT-MRI) due to slightly different positioning or anatomical changes. Even when using deformable registration, residual errors may still be present. The anatomy may change further between CBCT/MRI and CT. Especially in the pelvis, air will move around, and bladder/intestine filling can change. There is no fair way to correct this: deformable registration may be a viable option. However, MRI is affected by geometric distortion, and CBCT can be affected by cupping, which results in smaller body contours. Applying deformable registration will reduce (eliminate, not really, considering residual registration errors) the intrinsic difference between the input and reference image. However, such differences cannot be correct during inference without having at disposal the reference image. Therefore, avoiding deformable registration during evaluation is the fairest way to assess the quality of the sCT. Note that all challenge participants will deal with the same dataset, so the evaluation is deemed unbiased, even though some teams may develop methods that might be less sensitive to registration mismatches.

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

N/A

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 output sCT will be compared to CT undergoing an image similarity (validation and test phase) and a dose evaluation (test phase). For details and equations of the metrics, please, see the appendix (section 26) at https://drive.google.com/file/d/1jyGP8s99TfDGPMwIH05CAdmKuTxSCORI/view?usp=share_link.

Image evaluation between the sCT and CT using the metrics listed below.
a. Mean absolute error (MAE) on the mask;
b. Peak signal-to-noise ratio (PSNR) on the mask;
c. Structural similarity index (SSIM) on the mask.

Dose evaluation will be performed globally and locally by comparing photon and proton dose calculations between reference CT and sCT. Clinically, the most relevant question would be how a dose plan optimized on the CT (the "ground truth" where patients are currently treated) would perform on the sCT (the image at disposal in the new clinical workflow). Dose calculations will be performed with matRad (https://e0404.github.io/matRad/), an open-source treatment planning system where photon and proton intensity-modulated treatment plans will be optimized on CT. Dose prescriptions and plans will be chosen irrespectively of the original clinical goal for each anatomy, choosing the center of the planning target volume (PTV) as the isocenter. Specifically, we will plan to prescribe the target (PTV) of 30x2 Gy for the brain and 20x3 Gy for the pelvis at the 95% isodose level for both photons and protons. For simplicity, proton plans will be planned using the same PTV approach as photons without robust optimization. Dose delivery will be simulated via ten beams of 6MV for photons using the generic Linac model and 2-3 beams for proton plans using the generic proton system modeled in matRad. The number of beams may be optimized on a patient basis to comply with the dose prescription and limit the dose to the organs at risk following the international guidelines: for the pelvis [Hall et al., 2021] and the brain [Lambrecht et al., 2018]. To further reduce the dose to the healthy tissues and ensure plan uniformity between patients, we will use the same objective functions and constraints available in matRaD per treatment site, as reported in Table 1 (see the appendix). Organ-at-risk (OAR) dose limits will be handled as hard constraints whenever possible but might be turned into objectives on a patient-specific basis.

The following metrics are considered for this aspect of the evaluation:
d. Mean absolute dose differences relative to the prescribed dose;
e. A dose-volume histogram (DVH) provides information on the delivered dose to the volume of specific structures (see appendix);
f. Gamma index: The Gamma pass ratios will be calculated for sCTs using the CT doses as a reference. The calculation is performed in 3D, according to Low et al., 1998. The passing criteria are the dose-difference criterion DM = 2%

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

Image similarity: MAE, PSNR, and SSIM are image similarity metrics commonly used in medical image synthesis. For a detailed overview, see the review by Spadea and Maspero et al., 2021.

Relative dose differences: This metric gives insight into the dose distribution in the local regions that receive a high dose. Only including these regions ensures that the large regions receiving little to almost no dose do not unintentionally determine the outcome of this metric.

DVH parameters: In radiotherapy, DVH parameters are commonly used to assess dose distributions in target volumes and OARs [Drzymala et al., 1991]. Specifically, DVH parameters describe target coverage and OAR sparing. Considering differences in the DVH parameters is a way of verifying that clinically relevant objectives are maintained.

Gamma index: This is a commonly used metric in radiotherapy to compare two dose distributions. It combines dose difference criteria and distance difference criteria in a single metric. Low et al., 1998 describe the details of the theoretical background. The medical physicist association suggests using 3%,3mm distance-to-agreement criteria when comparing delivered and planned doses [Ezzel et al., 2009]. We will adopt an even more stringent criterion (2%, 2mm), considering that error in the planning phase leads to irradiating the patient with a systematically deviating dose from the planned one.

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.

The final ranking will not be directly available in the leaderboards, but we will make it available at the end of phase given that ranking can be made only when all the submissions have been made.

---- (The ranking scheme will be revised) ----

To obtain one value per patient in the test set for every metric described in Section 26.2, the sum is taken over the photon and proton evaluations. Subsequently, the final result per metric is obtained by averaging all the test cases.

Phase 1 = validation: The first ranking phase includes only the image metrics on the validation set. The MAE, PSNR, and SSIM will be calculated and separately normalized between zero (worst result among participants) and one (the best result among participants). Subsequently, the normalized metrics will be summed with equal weights and converted to a standard ranking: 1 (best submission, highest summed normalized metrics) to n (worst submission, lowest summed normalized metrics).

Phase 2 = test: The second-ranking phase includes the image metrics & the dose evaluation on the test set. First, the automatic ranking based on the image similarity metrics will be performed, as in Phase 1. Secondly, dose evaluation will be run, and again the MAEtarget dose, DVHmetric and gamma index will be normalized between zero and one. Dose evaluation depends on matRad, an open source fully validated treatment planning system written in Matlab. We are investigating whether it is feasible to incorporate an automatic evaluation pipeline on the hosting site. If possible, dose evaluation will be offered for all the test submissions, along with a leaderboard. The normalized metrics for all image and dose metrics will be summed, where the dose evaluation metrics will weigh twice as much as the image similarity metrics. Suppose integration of this downstream task is too cumbersome. In that case, the best ten submissions based on the image similarity metrics will be considered for dose evaluation, and only the sum of the normalized metrics of the dose evaluation (equal weighting for MAEtarget dose, DVHmetric and gamma index) will contribute to the final ranking.

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

If a team lacks submitting sCTs for one or multiple patients, the team will be contacted and asked to resubmit within 24 hours. When a resubmission is still incomplete or corrupt, a black image (indicating an image that contains only air) will be used as an sCT of the specific patient.

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

- The approach of normalizing the metrics separately was chosen to overcome issues with metrics that do not have a predefined minimum and maximum value or metrics that always show a relatively high or low value.

- Aggregating and ranking were chosen to preserve large and small performance differences while combining metrics.
- The image similarity functions as a first threshold metric: we consider good similarity a prerequisite to access downstream dose evaluation.
- In the case of not having the dose evaluation integrated into the automatic pipeline. We consider only the best 10 participants based on the image similarity metrics because we consider good similarity a prerequisite to access downstream dose evaluation. Subsequently, having the dose evaluation determining the final winner makes sense in light of the clinical application. For the same reason, we double the weighting of the dose evaluation in case it can be included in the automated pipeline.

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.

--- (The ranking will be revised) ---

The variability of the ranking will be assessed with a Kandall’s tau analysis. We will investigate whether only including the image or the dose evaluation will lead to different rankings. Furthermore, the patterns of the different dose evaluations will also be analyzed.

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

Kendall’s tau quantifies differences between rankings, which can give insight into the quality of a metric.

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 generate the sCTs. This analysis will, for example, consider the difference between paired versus unpaired approaches and 2D versus 3D models. Furthermore, we will analyze the influence of the automated quantitative analysis on biases in our data and methods, considering, for example, the effect of registration and the difference in the quality of paired images for the brain and pelvis. Lastly, we will include qualitative analysis by expert observers on single cases, reporting their agreement on the quantitative analysis and their opinion on the sCT quality.

TASK: CBCT to CT

SUMMARY

Keywords

List the primary keywords that characterize the task.

medical imaging, computed tomography, cone beam computed tomography, image synthesis, generative model

ORGANIZATION

Organizers

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

Adrian Thummerer, UMC Groningen;

Arthur Jr. Galapon, UMC Groningen;

Peter Koopmans, Radboudumc (Nijmegen);

Evi Huijben, Eindhoven University of Technology;

Manya Afonso, Wageningen Research;
Maarten Terpstra, UMC Utrecht;

Matteo Maspero, UMC Utrecht;
Maureen, van Eijnatten, Eindhoven University of Technology;

Oliver J. Gurney-Champion, Amsterdam UMC;

Suraj Pai, Maastricht University;

Zoltan Perko, TU Delft.

b) Provide information on the primary contact person.

Matteo Maspero, Assistant professor, UMC Utrecht, Computational Imaging group, m.maspero@umcutrecht.nl / matteo.maspero.it@gmail.com

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 with an open leaderboard for submission after closing the challenge.

Challenge venue and platform

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

MICCAI.

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

https://synthrad2023.grand-challenge.org/

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

https://synthrad2023.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.

The data used to train algorithms are restricted to the data provided by the challenge. Pre-trained nets may NOT be used in the challenge.

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 if not listed among the organizers, contributors, or data providers and did not co-author any publication with the organizers in the last year; otherwise, they are not eligible for the prizes. Organizers, contributors, or data providers may not participate in the challenge.

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 team can be composed of 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 a condition to participating in the challenge, all teams must sign an greement, based on the restrictive TCIA policy form, stating that no attempt to de-anonymize the patients' information will be made on behalf of the team.

As other conditions for being awarded a Prize, the best top ten performing teams must fulfil the following obligations:

1) Present their method at the final event of the challenge at MICCAI 2023;
2) Sign and return all prize acceptance documents as may be required by Competition Sponsor/Organizers.
3) The price eligibility is conditional on submitting a paper reporting the details of the methods in a short or long (up to the teams) LNCS format.
4) Commit to citing the data challenge paper and the data overview paper whenever submitting the developed method for scientific and non-scientific publications.

The participating teams are strongly encouraged to disclose or share their code, although not mandatory.

We are arranging prizes for 5000 eur per task.

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, and the top teams will be invited to present their approach during the final 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 so that their performance results will become part of presentations, publications, or subsequent analyzes derived from the challenge at the discretion of the organization. Specifically, all the performance results will be made public.
The SynthRAD2023 organizers will consolidate the results and submit a challenge paper (to IEEE TMI, MEDIA, LNCS issue, or similar).

Each team ranked among the top ten metrics 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 sCT 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 Section 8c. 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 final event. The embargo is counted from the final 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 overview 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.

We will organize a type 2 challenge, where algorithm submissions run through the website, as described in https://grand-challenge.org/documentation/challenges/. Training data (MRI or CBCT and corresponding CT) will be publicly available. Both for the test and validation phase, the teams must supply the algorithm for type 2 to the organizers following the submission link and instructions provided on https://synthrad2023.grand-challenge.org/. The same holds for the test phase: teams will submit their dockerized sCT algorithms to the challenge website without having the data at their disposal. Once the challenge is presented at MICCAI, the validation data and target will be made available but not the test data. After the challenge conclusion, the challenge will remain open as a type 2 and the leaderboard will be updated, until the budget is finished. this will be announced through the forum.

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 validation and test phase. The validation phase lasts six weeks and will allow up to 2 submissions per week for each team and let the teams get familiar with the submission system and compare their performance with the other teams. The results of the validation phase will be evaluated on the validation set with an open scoreboard.

The validation phase will be closed from a specific date (see below), and the test (final submission) phase will start. The validation phase start eight weeks after releasing the data, and will be open for six weeks. After this, the test phase starts, during which the algorithms can be submitted maximally twice to be applied to the test data. At the end of the test phase, top-performing
teams will be invited to present their methods.

The participating teams can submit up to two runs to evaluate their algorithms on this test 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 be identified with a description.

To ensure that no error occurs during the submission process, we plan to send warnings when the submission format is incorrect or when the performance is below a (very low) threshold, suggesting errors. We keep the test phase relatively short (four weeks) to avoid teams having too much additional time to refine their methods further but still allow teams to process all the input data (120 input cases per task).

The participating teams that want to be considered for an award are obliged to describe their methods by submitting a short or long paper in LNCS format (https://www.springer.com/gp/computerscience/lncs/conference-proceedings-guidelines).

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

Begin challenge: Release training cases: 1/04/2023

Training phase: 1/04/2023 - 31/05/2023 (8 weeks+ 6 weeks from validation + 4 weeks during the test)

Presentation of the challenge at ESTRO23: 13/05/2023

Validation phase: 1/06/2023-15/07/2023 (6 weeks)

Test phase: 16/07/2023-15/08/2023 (4 weeks)

Deadline for registration method paper: 15/08/2023

Dose evaluation for the top-ranked solutions in the leaderboard: 15/08/2023-20/09/2023

Submission deadline for the method paper: 19/09/2023

Announcements and invitation to present: 20/09/2023

Presentation of the challenge results in 8-12/10/2023, MICCAI, Vancouver
April/May 2024,
ESTRO (TBD)

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).

Each institution is providing data requesting ethics approval to the internal review board/Medical Ethical committee of its institute:
UMC Utrecht approved not-WMO on 4/03/2022 with number 22/474 entitled: “Synthetizing computed tomography for radiotherapy Grand Challenge (SynthRAD).” Contact persons: Department of Radiotherapy, University Medical Center Utrecht, m.maspero@umcutrecht.nl; trialburaucancercenter@umcutrecht.nl.
UMC Groningen approved non-WMO on 20/07/2022 with number 202200310 entitled: “Synthesizing computed tomography for radiotherapy - Grand Challenge” Contact persons: Department of Radiation Oncology, University Medical Center Groningen a.thummerer@umcg.nl / s.both@umcg.nl.
RadboudumcUMC declared the study non-WMO on 17/10/2022 with number 2022-15950 entitled “Synthetizing computed tomography for radiotherapy Grand Challenge” Contact persons: Department of Radiation Oncology, Radboud University Medical Center Nijmegen, erik.vanderbijl@radboudumc.nl.

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)

Additional comments: Data are released under CC BY-NC (Attribution-NonCommercial). This will be performed via Zenodo at 10.5281/zenodo.7260704.

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 made publicly available on GitHub at the following location: https://github.com/SynthRAD2023

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

Openly sharing the teams' code is strongly encouraged but remains optional.
Conflicts of

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 endorsed by ESTRO and NVFK. A Seed Fund grant (https://ewuu.nl/en/collaboration/seed-fund/) has been granted to cover the challenge's computational costs on the website and organizational cost (and prices). Additional funding is under discussion with commercial parties, but no confirmed participation is ensured at the moment of submission of this template.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Research, Intervention planning.

Additional points: Radiotherapy;

Task category(ies)

State the task category(ies).

Examples:

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

Synthesis (image); Regression (image)

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 addressed patients undergoing radiotherapy (~50% of cancer patients). No gender restriction is considered, and a predominant adult population is collected. Inclusion criteria are the acquisition of CT and MRI during planning (task 1) to ensure accurate positioning during image-guided therapy. MRI and CBCT should be acquired within two months of the CT to reduce anatomical changes. Datasets for tasks 1 and 2 do not necessarily contain the same patients given the separated tasks.

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

Patients undergoing radiotherapy in the brain and pelvis will be considered for both tasks. For task 2, patients should have undergone CBCT and CT.

Imaging modality(ies)

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

For task 2, patients should have undergone CBCT and 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).

The following additional information is given along with the images:
1. Acquisition protocols for MRI, CBCT, and CT of the acquired images as extracted from Dicom header, and the image size and resolution of the images after pre-processing to provide the size of the finally provided images.

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

2. Gender, weight, and age (if available) of the patients.

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.

brain and pelvis oncological patients acquired in CT, MRI treated in three Dutch Radiotherapy departments.

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.

generated images similar to CT based on CBCT (task 2) images. The whole patient volume in the FOV of the input image is considered.

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)).

Robustness, Accuracy.

Additional points: Properties of the algorithms to be optimized to perform well in

1. Generate synthetic CT from CBCT (task 2).
2. Image similarity. Generate CT-like images (called synthetic CT (sCT)) according to image similarity metrics.
3. Dose evaluation. According to the dose evaluation metrics, the dose planned on the original CT and recalculated on the sCT should be similar.

The image similarity metrics rank the methods during the validation phase. In the test phase, also the effect on dose distribution is considered. The dose has the highest clinical relevance as it determines the treatment. Dose evaluation depends on matRad (https://e0404.github.io/matRad/) open source fully validated treatment planning system written in Matlab. We are investigating whether it is feasible to incorporate an automatic evaluation pipeline on the hosting site, making it possible to offer dose evaluation for all the tests. If this result is too cumbersome or computationally expensive, we will only apply dose metrics to the top-ranking teams in the image similarity metrics.

The image similarity is used as a surrogate to provide the participants quick feedback during the validation phase. The final metric combines image similarity and dose evaluation (see point 26).

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).

Images were acquired on the following devices:

CBCT:
Center A (UMC Utrecht) - Brain/Pelvis: Elekta XVI
Center B (UMC Groningen) - Brain: IBA Proteus Plus; Pelvis: Elekta XVI
Center C (Radboud Nijmegen) - Brain/Pelvis: Elekta XVI

CT:
Center A (UMC Utrecht) - Brain/Pelvis: Philips Brilliance Big Bore, Siemens Biograph20 PET-CT (5-10 % of the data)
Center B (UMC Groningen) - Brain: Siemens SOMATOM Definition AS; Pelvis: Siemens SOMATOM Definition AS, Siemens SOMATOM go.Open Pro, GE Medical Systems Optima CT580
Center C (Radboud Nijmegen) - Brain/Pelvis: Philips Brilliance Big Bore

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

Images were acquired with the clinically used imaging protocols of the respective centers for each anatomical site and reflect typical images found in daily routine. Due to different clinical routines, imaging protocols for CBCT and CT vary within and between datasets, which reflects a realistic application scenario. A detailed table of imaging parameters will be distributed with the dataset.

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

Data was acquired for radiotherapy treatments in the radiotherapy departments of UMC Utrecht, UMC Groningen, and Radboud Nijmegen and not provided in any previous challenge. The participant will not be able to recognize the origin center since the data are anonymized. Metadata is provided, and the institution's names are substituted with institutes A, B, and C.

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).

Data was acquired by clinical staff of the respective radiotherapy departments. Patients treated in the pelvis had either cervical, rectal, or prostate cancer. The clinically adopted delineation of target structures and organs at risk for the test set will be used for evaluation purposes (dose evaluation). Each center may have different routines concerning the patients' positioning in CT, and CBCT (e.g., different immobilization devices such as masks, or vacuum cushions, use of flat tops) and which guidelines were followed for delineations. The guidelines among the centers have been compared, finding that all centers followed the Dutch guidelines (https://richtlijnendatabase.nl/). Details about the delineations for each anatomical site will be reported in an upcoming publication focusing on the data. A pre-print (or, if time allows, a publication) will be released along with the challenge.

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 refers to a single patient consisting of a CT and CBCT of the same patient acquired < 2 months apart. Radiotherapy treatment plans and structure sets (target volume and organs-at-risk) are also available for test cases. A train case has both input (CBCT for task 2) image and reference (CT) case; while a test/validation case has only input image.

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

For the brain, all three institutes will provide 60/10/20 datasets for training, validation, and testing, respectively. For the pelvis, institute B does not have MRIs with a sufficiently large FOV. To compensate, institute A will provide twice the data for the pelvis: 120/20/40. Institute C will provide the normal amount of 60/10/20 cases.
Total cases task 2:

Pelvis: 180 training, 30 validation, 60 test cases
Brain: 180 training, 30 validation, 60 test cases

Total cases for both the tasks= 360 training, 60 validation, 120 test, for a total of 540 cases

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

The total number of cases and the separation into training, validation, and test sets are larger than previous studies in the field (see the review article by Spadea and Maspero et al., 2021) that considered on average between 30-50 patients. Considering the multicenter setting of the data, we have included -when available- at least 60 patients per center.

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.

For each center, the imaging protocol was included if the protocol was comparable to at least one-third of the population. This has been performed to ensure that class balance is preserved, helping the challenge participants develop methods to handle the multi-center variability. Case selection in the brain was blind to clinical information concerning primary tumor etiology, making the tumor characteristics a random sample of the clinical routine. In the pelvis, cervical, rectal, and prostate cases were considered equally distributed among training, validation, and test sets on an institute level.

Annotation characteristics

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

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

No annotators can be considered given that the task is regression, and no manual interaction can be considered on the provided data. The data has annotated whether the anatomy is pelvis or brain. This will facilitate running separate models on the two anatomy if desired

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.

N/A

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.

Image protocols on a per-patient basis will be provided in excel files reporting each imaging modality's relevant acquisition and reconstruction settings. However, this information will not be used for quantitative evaluation.

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.

N/A

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.

Data pre-processing includes the following steps:
1. Image conversion: All images are converted to the compressed nifty file format (.nii.gz, https://nifti.nimh.nih.gov/)
2. Image resampling: All images will be resampled to a uniform voxel spacing (Brain: 1x1x1 mm3, Pelvis: 1x1x2.5 mm3). Image size might vary from case to case and between datasets.
3. Image registration: A rigid image registration between CBCT (for task2) or MR (for task1) and resampled CT will be performed for all cases to align the images (using Elastix: https://elastix.lumc.nl/index.php). Participants can further improve the registration (rigid and deformable) if required. Elastix parameter files will be made available for participants for rigid registration and a suggestion for deformable registration. During the evaluation, rigid image registration between CBCT/MR and CT will reduce anatomical differences.
4. Image masking: The field-of-view of MR/CBCT and CT will be aligned by automatically segmenting based on the threshold and morphological operation of the patient outline on the MR/CBCT. The resulting mask will also be dilated to include surrounding air and applied to the CT image. The mask will be provided, and it can be used by the participant, if needed, for pre-processing. Also, image similarity metrics will be calculated in this mask

whenever specified.
5. Image cropping: All images will be cropped to the bounding box of the patient outline (with a margin of 20 voxels in plane) to reduce the amount of data.
6. Defacing: For brain patients, face removal/de-identification will be performed. A parallelepiped mask for defacing has been obtained starting for the location of the eyes as available in the clinical delineation. If the eyes were not available, a surrogate structure has been selected or manually made. Defacing may vary between centers: UMC Utrecht: for task 1, the MRI has a FOV with no face included, but CT has a larger FOV. The MR mask from point 4 will be applied to CT to remove the face; for task 2, defacing is still under investigation.

The code used to perform pre-processing will be made publicly available at https://github.com/SynthRAD2023.

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.

Image registration may not perfectly match the two images (CT-MRI) due to slightly different positioning or anatomical changes. Even when using deformable registration, residual errors may still be present. The anatomy may change further between CBCT/MRI and CT. Especially in the pelvis, air will move around, and bladder/intestine filling can change. There is no fair way to correct this: deformable registration may be a viable option. However, MRI is affected by geometric distortion, and CBCT can be affected by cupping, which results in smaller body contours. Applying deformable registration will reduce (eliminate, not really, considering residual registration errors) the intrinsic difference between the input and reference image. However, such differences cannot be correct during inference without having at disposal the reference image. Therefore, avoiding deformable registration during evaluation is the fairest way to assess the quality of the sCT. Note that all challenge participants will deal with the same dataset, so the evaluation is deemed unbiased, even though some teams may develop methods that might be less sensitive to registration mismatches.

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

N/A

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 output sCT will be compared to CT undergoing an image similarity (validation and test phase) and a dose evaluation (test phase). For details and equations of the metrics, please, see the appendix (section 26) at https://drive.google.com/file/d/1jyGP8s99TfDGPMwIH05CAdmKuTxSCORI/view?usp=share_link.

Image evaluation between the sCT and CT using the metrics listed below.
a. Mean absolute error (MAE) on the mask;
b. Peak signal-to-noise ratio (PSNR) on the mask;

c. Structural similarity index (SSIM) on the mask.

Dose evaluation will be performed globally and locally by comparing photon and proton dose calculations between reference CT and sCT. Clinically, the most relevant question would be how a dose plan optimized on the CT (the "ground truth" where patients are currently treated) would perform on the sCT (the image at disposal in the new clinical workflow). Dose calculations will be performed with matRad (https://e0404.github.io/matRad/), an open-source treatment planning system where photon and proton intensity-modulated treatment plans will be optimized on CT. Dose prescriptions and plans will be chosen irrespectively of the original clinical goal for each anatomy, choosing the center of the planning target volume (PTV) as the isocenter. Specifically, we will plan to prescribe the target (PTV) of 30x2 Gy for the brain and 20x3 Gy for the pelvis at the 95% isodose level for both photons and protons. For simplicity, proton plans will be planned using the same PTV approach as photons without robust optimization. Dose delivery will be simulated via ten beams of 6MV for photons using the generic Linac model and 2-3 beams for proton plans using the generic proton system modeled in matRad. The number of beams may be optimized on a patient basis to comply with the dose prescription and limit the dose to the organs at risk following the international guidelines: for the pelvis [Hall et al., 2021] and the brain [Lambrecht et al., 2018]. To further reduce the dose to the healthy tissues and ensure plan uniformity between patients, we will use the same objective functions and constraints available in matRaD per treatment site, as reported in Table 1 (see the appendix). Organ-at-risk (OAR) dose limits will be handled as hard constraints whenever possible but might be turned into objectives on a patient-specific basis.

The following metrics are considered for this aspect of the evaluation:
d. Mean absolute dose differences relative to the prescribed dose;
e. A dose-volume histogram (DVH) provides information on the delivered dose to the volume of specific structures (see appendix);
f. Gamma index: The Gamma pass ratios will be calculated for sCTs using the CT doses as a reference. The calculation is performed in 3D, according to Low et al., 1998. The passing criteria are the dose-difference criterion DM = 2%

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

Image similarity: MAE, PSNR, and SSIM are image similarity metrics commonly used in medical image synthesis. For a detailed overview, see the review by Spadea and Maspero et al., 2021.

Relative dose differences: This metric gives insight into the dose distribution in the local regions that receive a high dose. Only including these regions ensures that the large regions receiving little to almost no dose do not unintentionally determine the outcome of this metric.

DVH parameters: In radiotherapy, DVH parameters are commonly used to assess dose distributions in target volumes and OARs [Drzymala et al., 1991]. Specifically, DVH parameters describe target coverage and OAR sparing. Considering differences in the DVH parameters is a way of verifying that clinically relevant objectives are maintained.

Gamma index: This is a commonly used metric in radiotherapy to compare two dose distributions. It combines dose difference criteria and distance difference criteria in a single metric. Low et al., 1998 describe the details of the theoretical background. The medical physicist association suggests using 3%,3mm distance-to-agreement criteria when comparing delivered and planned doses [Ezzel et al., 2009]. We will adopt an even more stringent criterion (2%, 2mm), considering that error in the planning phase leads to irradiating the patient with a systematically deviating dose from the planned one.

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.

The final ranking will not be directly available in the leaderboards, but we will make it available at the end of phase given that ranking can be made only when all the submissions have been made.

---- (The ranking scheme will be revised) ----

To obtain one value per patient in the test set for every metric described in Section 26.2, the sum is taken over the photon and proton evaluations. Subsequently, the final result per metric is obtained by averaging all the test cases.

Phase 1 = validation: The first ranking phase includes only the image metrics on the validation set. The MAE, PSNR, and SSIM will be calculated and separately normalized between zero (worst result among participants) and one (the best result among participants). Subsequently, the normalized metrics will be summed with equal weights and converted to a standard ranking: 1 (best submission, highest summed normalized metrics) to n (worst submission, lowest summed normalized metrics).

Phase 2 = test: The second-ranking phase includes the image metrics & the dose evaluation on the test set. First, the automatic ranking based on the image similarity metrics will be performed, as in Phase 1. Secondly, dose evaluation will be run, and again the MAEtarget dose, DVHmetric and gamma index will be normalized between zero and one. Dose evaluation depends on matRad, an open source fully validated treatment planning system written in Matlab. We are investigating whether it is feasible to incorporate an automatic evaluation pipeline on the hosting site. If possible, dose evaluation will be offered for all the test submissions, along with a leaderboard. The normalized metrics for all image and dose metrics will be summed, where the dose evaluation metrics will weigh twice as much as the image similarity metrics. Suppose integration of this downstream task is too cumbersome. In that case, the best ten submissions based on the image similarity metrics will be considered for dose evaluation, and only the sum of the normalized metrics of the dose evaluation (equal weighting for MAEtarget dose, DVHmetric and gamma index) will contribute to the final ranking.

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

If a team lacks submitting sCTs for one or multiple patients, the team will be contacted and asked to resubmit within 24 hours. When a resubmission is still incomplete or corrupt, a black image (indicating an image that contains only air) will be used as an sCT of the specific patient.

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

- The approach of normalizing the metrics separately was chosen to overcome issues with metrics that do not have a predefined minimum and maximum value or metrics that always show a relatively high or low value.
- Aggregating and ranking were chosen to preserve large and small performance differences while combining metrics.
- The image similarity functions as a first threshold metric: we consider good similarity a prerequisite to access downstream dose evaluation.

- In the case of not having the dose evaluation integrated into the automatic pipeline. We consider only the best 10 participants based on the image similarity metrics because we consider good similarity a prerequisite to access downstream dose evaluation. Subsequently, having the dose evaluation determining the final winner makes sense in light of the clinical application. For the same reason, we double the weighting of the dose evaluation in case it can be included in the automated pipeline.

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.

--- (The ranking will be revised) ---

The variability of the ranking will be assessed with a Kandall's tau analysis. We will investigate whether only including the image or the dose evaluation will lead to different rankings. Furthermore, the patterns of the different dose evaluations will also be analyzed.

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

Kendall's tau quantifies differences between rankings, which can give insight into the quality of a metric.

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 generate the sCTs. This analysis will, for example, consider the difference between paired versus unpaired approaches and 2D versus 3D models. Furthermore, we will analyze the influence of the automated quantitative analysis on biases in our data and methods, considering, for example, the effect of registration and the difference in the quality of paired images for the brain and pelvis. Lastly, we will include qualitative analysis by expert observers on single cases, reporting their agreement on the quantitative analysis and their opinion on the sCT quality.

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.

Chernak E.S., Rodriguez-Antunez A., Jelden G.L., Dhaliwal R.S., Lavik P.S. The use of computed tomography for radiation therapy treatment planning. Radiology. 1975 Dec;117(3):613-4. https://doi.org/10.1148/117.3.613
Ramella, S., Fiore, M., Silipigni, S., Zappa, M. C., Jaus, M., Alberti, A. M., ... & D'Angelillo, R. M. (2017). Local control and toxicity of adaptive radiotherapy using weekly CT imaging: results from the LARTIA trial in stage III NSCLC.

Journal of Thoracic Oncology, 12(7), 1122-1130. https://doi.org/10.1016/j.jtho.2017.03.025
Kida, S., Nakamoto, T., Nakano, M., Nawa, K., Haga, A., Kotoku, J. I., ... & Nakagawa, K. (2018). Cone beam computed tomography image quality improvement using a deep convolutional neural network. Cureus, 10(4). https://doi.org/10.7759%2Fcureus.2548
Schmidt M. A., Payne G. S. Radiotherapy planning using MRI. Phys Med Biol. 2015;60:R323
Lagendijk, J. J., Raaymakers, B. W., Van den Berg, C. A., Moerland, M. A., Philippens, M. E., & Van Vulpen, M. (2014). MR guidance in radiotherapy. Physics in Medicine & Biology, 59(21), R349. https://doi.org/10.1088/0031-9155/59/21/r349
Edmund, J. M., & Nyholm, T. (2017). A review of substitute CT generation for MRI-only radiation therapy. Radiation Oncology, 12(1), 1-15. https://doi.org/10.1186/s13014-016-0747-y
Spadea, M. F. & Maspero, M., Zaffino, P., & Seco, J. (2021). Deep learning-based synthetic-CT generation in radiotherapy and PET: A review. Medical Physics, 48(11), 6537–6566. https://doi.org/10.1002/mp.15150
Schwarz, C. G., Kremers, W. K., Wiste, H. J., Gunter, J. L., Vemuri, P., Spychalla, A. J., ... & Alzheimer's Disease Neuroimaging Initiative. (2021). Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives. NeuroImage, 231, 117845. https://doi.org/10.1016/j.neuroimage.2021.117845
Low, D.A., Harms, W.B., Mutic, S., and Purdy, J.A. (1998), A technique for the quantitative evaluation of dose distributions. Med. Phys., 25: 656-661. https://doi.org/10.1118/1.598248
Hall, W. A., Paulson, E., Davis, B. J., Spratt, D. E., Morgan, T. M., Dearnaley, D., ... & Lawton, C. A. (2021). NRG oncology updated international consensus atlas on pelvic lymph node volumes for intact and postoperative prostate cancer. International Journal of Radiation Oncology* Biology* Physics, 109(1), 174-185. https://doi.org/10.1016/j.ijrobp.2020.08.034
Lambrecht, M., Eekers, D. B., Alapetite, C., Burnet, N. G., Calugaru, V., Coremans, I. E., ... & Troost, E. G. (2018). Radiation dose constraints for organs at risk in neuro-oncology; the European Particle Therapy Network consensus. Radiotherapy and Oncology, 128(1), 26-36 https://doi.org/10.1016/j.radonc.2018.05.001
Ezzell, G. A., Burmeister, J. W., Dogan, N., LoSasso, T. J., Mechalakos, J. G., Mihailidis, D., ... & Xiao, Y. (2009). IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119. Medical physics, 36(11), 5359-5373. https://doi.org/10.1118/1.3238104
Drzymala, R. E., Mohan, R., Brewster, L., Chu, J., Goitein, M., Harms, W., & Urie, M. (1991). Dose-volume histograms. International Journal of Radiation Oncology* Biology* Physics, 21(1), 71-78. https://doi.org/10.1016/0360-3016(91)90168-4
International Commission on Radiation Units and Measurements, ICRU Report 83, Prescribing, recording, and reporting intensity-modulated photon-beam therapy (IMRT)(ICRU Report 83), Bethesda, MD (2010) https://www.fnkv.cz/soubory/216/icru-83.pdf

Further comments

Further comments from the organizers.

Challenge origin

The idea of a challenge organization arose within the "Image synthesis & reconstruction" subgroup of the Dutch deep learning in radiotherapy initiative. Such subgroups arose from an initiative independently organized by professors affiliated with all the Dutch University medical centers, focusing on deep learning and working in radiotherapy departments. The organizer group encompasses early-stage researchers, PhDs, postdocs, three assistant professors, and one associate professor.

Additional information and details on the metrics can be found at:

https://drive.google.com/file/d/1jyGP8s99TfDGPMwIH05CAdmKuTxSCORI/view?usp=share_link. Such appendix will be attached to the pdf submitted to the submission form.

Duplicate submission

After submitting the challenge, we received confirmation of its funding, so we repeated the submission stating that a type 2 challenge will be performed.

SynthRAD2023

Synthesizing computed tomography for radiotherapy

CHALLENGE ORGANIZATION

This is the challenge organization approved by MICCAI in February 2023.

1a Title: Synthesizing computed tomography for radiotherapy
1b Acronym: SynthRAD2023

2 Abstract
The impact of medical imaging on oncological patients' diagnosis and therapy, has grown significantly over the last decades. Especially in radiotherapy (RT), imaging plays a crucial role in the entire workflow, from treatment simulation to patient positioning and monitoring.

Traditionally, 3D computed tomography (CT) is considered the primary imaging modality in RT, providing accurate and high-resolution patient geometry and enabling direct electron density conversion needed for dose calculations and plan optimization [Chernak et al., 1975]. X-ray-based imaging has been widely adopted for patient positioning and monitoring before, during, and after dose delivery. Recently, 3D cone-beam computed tomography (CBCT) - often integrated with the dose delivery machine - has been widely adopted, playing a vital role in traditional and more advanced image-guided adaptive radiation therapy (IGART) workflows in both photon and proton therapy.

A key challenge in using CBCT is that several artifacts, such as shading, streaking, and cupping, affect image reconstruction. As a result, CBCT quality is insufficient to perform accurate dose calculations or replanning. Consequently, patients must be referred to a rescan CT when anatomical differences are noted between daily images and the planning CT [Ramella et al., 2017]. As an alternative, image synthesis has been proposed to improve the quality of CBCT to the CT level, producing the so-called "synthetic CT" (sCT) [Kida et al., 2018]. Additionally, CBCT-based sCT allows online adaptive CBCT-based RT workflows to improve the quality of IGART provided to the patients.

In parallel, over the last decades, magnetic resonance imaging (MRI) has also proved its added value for tumor and organs-at-risk delineation thanks to its superb soft-tissue contrast [Schmidt et al., 2015]. MRI can be acquired to match patient positioning to the planned one and monitor changes before, during, or after the dose delivery [Lagendijk et al., 2004].

To benefit from the complementary advantages offered by different imaging modalities, MRI is generally registered to CT. Such a workflow requires obtaining both CT and MRI, increasing workload, and introducing additional radiation to the patient. Recently, MRI-only based RT has been proposed to simplify and speed up the workflow, decreasing patients' exposure to ionizing radiation. This is particularly relevant for repeated simulations or fragile populations like children. MRI-only RT may reduce treatment costs and workload and eliminate residual registration errors using both imaging modalities. Additionally, MRI-only techniques can benefit MRI-guided RT [Edmund and Nyholm, 2017].

The main obstacle in introducing MRI-only RT is the lack of tissue attenuation information required for accurate dose calculations. Many methods have been proposed to convert MR to CT-equivalent images, yielding sCTs suitable for treatment planning and dose calculation.

Artificial intelligence algorithms such as machine learning or deep learning have become the best-performing methods for deriving sCT from MRI or CBCT. With many algorithms available, all tested on different datasets, it is unclear which algorithms are better than others. Unfortunately, no public datasets or challenges have been designed to benchmark and compare different approaches. A recent review of deep learning-based sCT generation also advocated for public challenges to provide data and evaluation metrics for such open comparison [Spadea & Maspero et al., 2021].

We now designed a challenge to provide the first platform offering public datasets and evaluation metrics to benchmark and compare the latest algorithms in sCT generation. For this purpose, two tasks were defined: 1) MRI-to-sCT generation to facilitate MRI-only RT and 2) CBCT-to-sCT generation to facilitate IGART and online adaptive RT.

A multi-center dataset of matched input (CBCT or MRI) and target (CT) image pairs with heterogeneous acquisition protocols will be divided into balanced training, validation, and test sets. We will share the inputs and targets from the training set to design and evaluate algorithms generating sCT. In the first evaluation round, called validation, input images will be shared. The participating team can upload their outputs to validate their performance and see how they rank on the leaderboard on image similarity metrics. Finally, independent test cases will be shared to generate sCT for our final evaluation. Brain and pelvic datasets from three Dutch centers (UMC Utrecht, UMC Groningen, and Radboud Nijmegen) have been collected, providing more than 500 patients for MRI-to-CT (Task 1) and 500 patients for CBCT-to-CT (Task 2) undergoing radiotherapy treatments in the corresponding departments.

The challenge runs on https://synthrad2023.grand-challenge.org/, and the pre-processing and evaluation code used to rank the submissions based on image and dose evaluations will be shared.

Challenge participants may choose to participate only in task 1 (MRI-to-sCT), only in task 2 (CBCT-to-sCT), or in both, for both anatomical regions. In the initial training/validation phases, training and validation data will be made available to submit to an open leaderboard to enable the development of algorithms. The test input MRI/CBCT and ground truth CT will not be shared with the participants to avoid optimistic biases.

We envision that this challenge will enable a fair and open assessment of different approaches.

3 Keywords: medical imaging, magnetic resonance imaging, computed tomography, cone beam computed tomography, image synthesis, generative model

Year: 2023

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop The best-performing teams will present their algorithms in short talks at a dedicated event or workshop, according to MICCAI availability. Note that the event can also be held purely online if necessary. Alternatively, but not desired, we could try to align with one of the workshops, e,g. SASHIMI and present the results of the challenge.

Duration Half day.

The expected number of participants is > 50. We expect participation at least from the ten best-ranked teams at MICCAI 2023, other participants teams, and other interested attendees at the workshop. It is the first time such a task has been considered a challenge, but considering the research interest, we believe enough teams will participate. We think more than 50 teams could estimate the overall challenge participation well.

Space and hardware requirements

None if the event is held online; otherwise, a room with PC and beamer + possibility to stream the event.

Publication and plans

A publication on the dataset will be released (as arXiv if not yet fully published) 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. The leaderboard will remain open after the challenge for new submissions. When the challenge is closed, we will release the code to perform the downstream task offline.

CHALLENGE ORGANIZATION

4 Organizers & Stakeholders

We define the following four categories with their tasks and requirements

Role	Objective	Requisite/Do-don't
Organizer	Take care of challenge organization	1) Person only 2) Active participation to the subgroups 3) NOT participate in the challenge (possible access to data)
Contributor	Support the challenge organization	1) Person or institution 2) Support the organizer in the organization but without an active role, not involved in the decision-making 3) NOT participate in the challenge
Data provider	Collect and provide data to the organizers	1) Person or institution 2) Active role in collecting the data/organizing the legal business related to collection 3) If a person is listed, canNOT participate in the challenge (possible access to part of the data)
Participant	Take part in the challenge	1) Teams of up to 5 people 2) No organizer, contributor, or data provider 3) Need to provide a description of the method in the form of a (short) paper Members of the organizers' or data provider's institutes may participate in the challenge if not listed among the organizers, contributors, or data providers and did not co-author any publication with the organizers in the last year; otherwise, they can participate, but they are not eligible for the prizes.

a) Organizing team, contributors, and data providers

Organizers

Adrian	Thummerer	UMC Groningen
Arthur Jr.	Galapon	UMC Groningen
Peter	Koopmans	Radboudumc (Nijmegen)
Evi	Huijben	Eindhoven University of Technology

Manya	Afonso	Wageningen Research
Maarten	Terpstra	UMC Utrecht
Matteo	Maspero	UMC Utrecht
Maureen	van Eijnatten	Eindhoven University of Technology
Oliver	Gurney-Champion	Amsterdam UMC
Suraj	Pai	Maastricht University
Zoltan	Perko	TU Delft

List: Adrian Thummerer, Arthur Jr. Galapon, Peter Koopmans, Evi Huijben, Manya Afonso, Maarten Terpstra, Maureen van Eijnatten, Oliver Gurney-Champion, Suraj Pai, Zoltan Perko, Matteo Maspero

Contributors

Cornelis (Nico) AT	van den Berg	UMC Utrecht
Joost JC	Verhoeff	UMC Utrecht
Stefan	Both	UMC Groningen
ESTRO	https://www.estro.org/	Event is endorsed
NVFK	https://www.nvkf.nl/	Event is endorsed

Data Providers

Adrian	Thummerer	UMC Groningen
Stefan	Both	UMC Groningen
Johannes A	Langendijk	UMC Groningen
		UMC Groningen
Joost	JC Verhoeff	UMC Utrecht
Matteo	Maspero	UMC Utrecht
		UMC Utrecht
Erik	van der Bijl	Radboudumc(Nijmegen)
		Radboud UMC (Nijmegen)

b) Primary contact person: Matteo Maspero, Assistant professor, UMC Utrecht, Computational Imaging group, m.maspero@umcutrecht.nl / matteo.maspero.it@gmail.com.

5 Lifecycle type
One-time event with a fixed deadline with an open leaderboard for submission after closing the challenge. The test leaderboard will be available for evaluation after the challenge is finished.

6 Challenge venue and platform
a) Event MICCAI 2023
b) Platform: grand-challenge.org
c) URL: https://synthrad2023.grand-challenge.org/

7 Participation policies

a) **Allowed user interaction** Only fully automatic methods are allowed. Methods should be submitted as specified on the submission page. Inference should run on an AWS g4dn.2xlarge instance using a single GPU with 16 GB RAM, 8 cores CPU and 32 GB RAM. The maximum inference time to produce sCT for a single case (one patient) is 15 minutes.

b) **Usage of training data/pre-trained models** The data used to train algorithms are restricted to the data provided by the challenge. Pre-trained nets may NOT be used in the challenge.

c) **Participation policy for organizers' institutes** Members of the organizers', contributors', and data providers' institutes may participate in the challenge if not listed among the organizers, contributors, or data providers and did not co-author any publication with the organizers, contributors, or data providers in the last year; otherwise, they are not eligible for the prizes. Organizers, contributors, and data providers may not participate in the challenge.

d) **Participation and award policy** 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 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 other conditions for being awarded a prize, the best performing teams must fulfil the following obligations:

1) Present their method at the final event of the challenge at MICCAI 2023;
2) Sign and return all prize acceptance documents as may be required by Competition Sponsor/Organizers.
3) The price eligibility is conditional on submitting a paper reporting the details of the methods in a short or long (up to the teams) LNCS format.
4) Commit to citing the data challenge paper and the data overview paper whenever submitting the developed method for scientific and non-scientific publications.

The participating teams are strongly encouraged to disclose or share their code, although not mandatory.

Depending on the available funding, organizers reserve the possibility to award prizes to the top teams in both tasks.

e) **Result announcement**
The results and winner will be announced publicly, and the top teams will be invited to present their approach during the final 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 so that their performance results will become part of presentations, publications, or subsequent analyzes derived from the challenge at the discretion of the organization. Specifically, all the performance results will be made public.
The SynthRAD2023 organizers will consolidate the results and submit a challenge paper (to IEEE TMI, MEDIA, LNCS issue, or similar).
Each team ranked among the top ten metrics 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 sCT as the challenge submission system will have automatically solicited them.

f) **Publication policy**
To be eligible for the official ranking, the participants must submit a paper describing their method as described in Section 8c. 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 final event. The embargo is counted from the final 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 overview paper, no embargo will be applied.

8 Submission method

a) **Submission instructions** will be available on the challenge website.

We will organize a type 2 challenge, where algorithm submissions run through the website during the test phase, as described in https://grand-challenge.org/documentation/challenges/. Training data (MRI or CBCT and corresponding CT) will be publicly available. Input validation data (only MRI or CBCT) will be available to provide prediction to upload on the site (so type 1 is used for validation). During test phase, the teams must supply the algorithm for the type 2 challenge to the organizers following the submission link and instructions provided on https://synthrad2023.grand-challenge.org/. Teams will submit their dockerized sCT algorithms to the challenge website without having the test data at their disposal. Once the challenge is presented at MICCAI, the target validation data will be made available but not the test data and its target. It is possible that after the challenge conclusion, the challenge will be reverted to type 1 to allow future submissions of the output images to avoid computational costs.

b) **Evaluate algorithms**

The challenge is subdivided into a validation and test phase. The validation phase lasts six weeks and will allow up to 2 submissions per week for each team to allow the teams to get familiar with the submission system and compare their performance with the intermediate results of other teams. The results of the validation phase will be evaluated on the validation set with an open dashboard.

The validation phase will be closed from a specific date (see below), and the test (final submission) phase will start. The validation phase will last six weeks and take place after eight weeks after the release of the data to allow developing and training of the algorithms. After the validation phase, a new leaderboard will be created and the type 2 test phase starts. During this final test phase, all data and targets remain hidden, and it requires teams to upload their dockerized algorithms. At the end of the test phase, top-performing teams will be invited to present their methods.

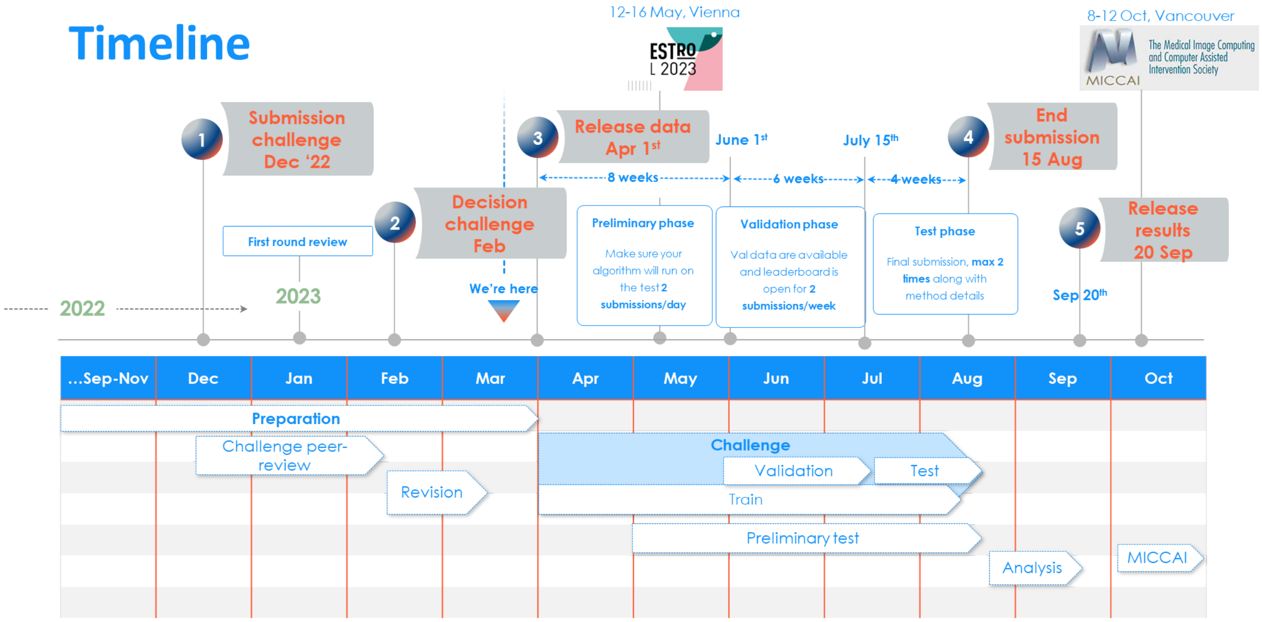
The participating teams can submit up to two runs to evaluate their algorithms on this test set. The second run is granted to accommodate possible errors during the submission process. Only the last run will be counted for the official ranking of the teams and the challenge results. We request that each run be identified with a description.

To ensure no error occurs during the submission process, we will also set up a preliminary test phase where participants can familiarize themselves with the submission system and test whether their dockerized algorithms run on a subset of the validation patients. Furthermore, we plan to send warnings when the submission format is incorrect or performance is below a (very low) threshold, suggesting errors. We keep the test phase relatively short (four weeks) to avoid teams having too much additional time to refine their methods further.

c) Submission method description The participating teams that want to be considered for an award must describe their methods by submitting a short or long paper in LNCS format (https://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines).

9 Timetable

Begin challenge: Release training cases	1/04/2023
Training phase	1/04/2023 - 31/05/2023 (8 weeks + 6 weeks for validation + 4 weeks for test)
Presentation of the challenge at ESTRO 23	13/05/2023
Validation phase	1/06/2023-15/07/2023 (6 weeks)
Preliminary phase for testing algorithms	1/06/2023-15/08/2023 (10 weeks)
Test phase	16/07/2023-15/08/2023 (4 weeks)
Deadline for submission method paper	15/08/2023
Analysis results	15/08/2023-20/09/2023
Announcements and invitation to present	20/09/2023
Presentation of the challenge results at	8-12/10/2023, MICCAI, Vancouver April/May 2024, ESTRO (TBD)



10 Ethics approval

Each institution is providing data requesting **ethics approval** to the internal review board/Medical Ethical committee of its institute:

- UMC Utrecht approved not-WMO on 4/03/2022 with number 22/474 entitled: "Synthetizing computed tomography for radiotherapy Grand Challenge (SynthRAD)." Contact persons: Department of Radiotherapy, University Medical Center Utrecht, m.maspero@umcutrecht.nl; trialburaucancercenter@umcutrecht.nl.
- UMC Groningen approved non-WMO on 20/07/2022 with number 202200310 entitled: "Synthesizing computed tomography for radiotherapy - Grand Challenge" Contact persons: Department of Radiation Oncology, University Medical Center Groningen a.thummerer@umcg.nl / s.both@umcg.nl.
- RadboudumcUMC declared the study non-WMO on 17/10/2022 with number 2022-15950 entitled "Synthetizing computed tomography for radiotherapy Grand Challenge" Contact persons: Department of Radiation Oncology, Radboud University Medical Center Nijmegen, erik.vanderbijl@radboudumc.nl.

11 Data usage agreement

Data are released under CC BY-NC (Attribution-NonCommercial). This will be performed via Zenodo at https://doi.org/10.5281/zenodo.7260705.

12 Code availability

a) **Accessibility of the organizers' evaluation software** Organizers' evaluation code and supporting data pre/post-processing code will be publicly available on GitHub at the following location: https://github.com/SynthRAD2023

b) **Accessibility of the participating teams' code.** Openly sharing teams' code is strongly encouraged but remains optional.

13 Conflicts of interest

Sponsoring/ funding the challenge: The challenge is endorsed by ESTRO and NVFK. A Seed Fund grant (https://ewuu.nl/en/collaboration/seed-fund/) has been granted to cover the challenge's computational costs on the website and organizational costs (and prices). Additional funding is discussed with commercial parties, but no confirmed participation is ensured at the moment of submission of this file. Further funding will be openly disclosed.

Access to the test case labels: data providers have access to the test case target and have provided the data to the organizers. The test cases will be accessed during the preparation of the test phase by the organizers.

MISSION OF THE CHALLENGE

14 Fields of application

- Radiotherapy/ radiation oncology
- Intervention planning
- Research

15 Task categories

- Synthesis (image)
- Regression (image)

16-17 Cohorts and imaging modalities

a) **Target cohort**: The biomedical application addressed patients undergoing radiotherapy (~50% of cancer patients). No gender restriction is considered, and a predominant adult population is collected. Inclusion criteria are the acquisition of CT and MRI during planning (task 1) or CT and CBCT (task 2) to ensure accurate positioning during image-guided therapy. MRI and CBCT should be acquired within two months of the CT to reduce anatomical changes. Datasets for tasks 1 and 2 do not necessarily contain the same patients given the separated tasks.

b) **Challenge cohort:** Patients undergoing radiotherapy in the brain and pelvis will be considered for both tasks. For task 1, patients should have undergone MRI and CT. For task 2, patients should have undergone CBCT and CT.

18 Context information

The following additional **information is given along with the images**:

1. Acquisition protocols for MRI, CBCT, and CT of the acquired images as extracted from Dicom header, and the image size and resolution of the images after pre-processing to provide the size of the finally provided images.
2. Gender, weight, and age (if available) of the **patients**.

19 Target entities

a) **Data origin:** brain and pelvis oncological patients acquired with CT, MRI, and CBCT treated in three Dutch Radiotherapy departments.
b) **Algorithm target:** generated images similar to CT based on MRI (task 1) or CBCT (task 2) images. The whole patient volume in the FOV of the input image is considered.

20 Assessment aim(s)

Properties of the algorithms to be optimized to perform well in

- *Generate synthetic CT* (sCT) from MRI (task 1) or CBCT (task 2).
- *Image similarity*. Generate CT-like images (sCT), which are assessed by image similarity metrics.
- *Dose evaluation.* According to the dose evaluation metrics, the dose planned on the original CT and recalculated on the sCT should be similar.

The image similarity metrics rank the methods during the validation phase. In the test phase, also the effect on dose distribution is considered. The dose has the highest clinical relevance as it determines the treatment. Dose evaluation depends on matRad open-source, fully validated treatment planning system written in Matlab. We are investigating whether it is feasible to incorporate an automatic evaluation pipeline on the hosting site, making it possible to offer dose evaluation for all the tests. If this result is too cumbersome or computationally expensive, we will only apply dose metrics to the top-ranking teams in the image similarity metrics.
The image similarity is used as a surrogate to provide the participants quick feedback during the validation phase. The final metric combines image similarity and dose evaluation (see point 26).

CHALLENGE DATASETS

21 Data source(s):
a) **Devices**

Images were acquired on the following devices:

	Center A (UMC Utrecht)	**Center B (UMC Groningen)**	**Center C (Radboud Nijmegen)**
CBCT	Brain/Pelvis: ● Elekta XVI	Brain: ● IBA Proteus Plus Pelvis: ● Elekta XVI	Brain/Pelvis: ● Elekta XVI
MRI	Brain/Pelvis ● Philips Ingenia 1.5T/3.0T	Brain: ● Siemens MAGNETOM Aera 1.5T ● Siemens MAGNETOM Avanto_fit 1.5T	Brain: ● Siemens Avanto fit 1.5T Pelvis: ● Siemens MAGNETOM Vida fit 3.0T
CT	Brain/Pelvis ● Philips Brilliance Big Bore ● Siemens Biograph20 PET-CT (5-10 % of the data)	Brain: ● Siemens SOMATOM Definition AS Pelvis: ● Siemens SOMATOM Definition AS ● Siemens SOMATOM go.Open Pro ● GE Medical Systems Optima CT580	Brain/Pelvis: ● Philips Brilliance Big Bore

b) **Data acquisition protocols:**

Images were acquired with the clinically used imaging protocols of the respective centers for each anatomical site and reflect typical images found in daily routine. Due to different clinical routines, imaging protocols for CBCT and CT vary within and between datasets, which reflects a realistic application scenario. A detailed table of imaging parameters will be distributed with the dataset.

For the MRI task (Task 1), a T1 weighted gradient echo sequence was selected for all datasets. Acquisition protocols varied between sites. The dataset of centers B and C only includes images acquired with a gadolinium contrast agent, while the selected MRIs for centers A were acquired without contrast.

c) **Center of origin and platform** Data was acquired for radiotherapy treatments in the radiotherapy departments of UMC Utrecht, UMC Groningen, and Radboud Nijmegen and not provided in any previous challenge. The participant will not be able to recognize the origin center since the data are anonymized. Metadata is provided, and the institution's names are substituted with institutes A, B, and C.

d) **Data characteristics**: Data was acquired by the clinical staff of the respective radiotherapy departments. Patients treated in the pelvis had either cervical, rectal, or prostate cancer. The clinically adopted delineation of target structures and organs at risk for the test set will be used for evaluation purposes (dose evaluation). Each center may have different routines concerning the patients’ positioning in CT, CBCT, and MRI (e.g., different immobilization devices such as masks, or vacuum cushions, use of flat tops) and which guidelines were followed for delineations. The guidelines among the centers have been compared, finding that all centers followed the Dutch guidelines (https://richtlijnendatabase.nl/). Details about the delineations for each anatomical site will be reported in an upcoming publication focusing on the data. A pre-print (or, if time allows, a publication) will be released along with the challenge.

22 Training and test case characteristics

a) **Definition of a case:** A case refers to a single patient consisting of a CT and, depending on the task, CBCT or MR of the same patient. Radiotherapy treatment plans and structure sets (target volume and organs-at-risk) are also available for test cases. A train case contains an input (MRI for task 1, CBCT for task 2) image and a reference (CT); while a validation case has only an input image. Test input will not be made available until the full closure of the challenge.

b) **Total** training/validation/test cases:

Task 1 MR-to-CT:
For the brain, all three institutes will provide datasets with a 60/10/20 split for training, validation, and testing, respectively. Only two institutes have MRI acquired with a sufficiently large FOV for the pelvis for use in MRI-only RT. To compensate for the lack of data, institute A will provide twice the data for the pelvis: 120/20/40. Institute C will provide the normal amount of 60/10/20 cases.
Total cases task 1:
Pelvis: 180 training, 30 validation, 60 test cases
Brain: 180 training, 30 validation, 60 test cases
Total cases for both the tasks= 360 training, 60 validation, 120 test, for a total of 540 cases

Task 2 CBCT-to-CT:
Institute A, B, and C provide each 60/10/20 datasets (train, val, test) for the brain and 60/10/20 datasets for the pelvis.
Total cases task 2:
Pelvis: 180 training, 30 validation, 60 test cases
Brain: 180 training, 30 validation, 60 test cases
Total cases for both the tasks= 360 training, 60 validation, 120 test, for a total of 540 cases

Total cases for both the tasks= 720 training, 120 validation, 240 test, for a total of 1080 cases

c) **Justification total number/splitting:** The total number of cases and the separation into training, validation, and test sets are larger than previous studies in the field (see the review article by Spadea and Maspero et al., 2021) that considered on average between 30-50 patients. Considering the multicenter setting of the data, we have included -when available- at least 60 patients per center.

d) **Further important characteristics** of the training, validation, and test cases: the imaging protocol was included for each center if the protocol was comparable to at least one-third of the population. This has been performed to preserve class balance, helping the challenge participants develop methods to handle the multi-center variability. Case selection in the brain was blind to clinical information concerning primary tumor etiology, making the tumor characteristics a random sample of the clinical routine. In the pelvis, cervical, rectal, and prostate cases were considered equally distributed among training, validation, and test sets on an institute level.

23 Annotation characteristics
a,b,c,d) **method for determining the reference annotation**: No annotators can be considered given that the task is regression, and no manual interaction can be considered on the provided data.

Other annotations used for the evaluation
Image protocols on a per-patient basis will be provided in excel files reporting each imaging modality's relevant acquisition and reconstruction settings.

24 Data pre-processing method(s)

Data pre-processing includes the following steps:

1) **Image conversion:** All images are converted to the compressed nifty file format (.nii.gz, https://nifti.nimh.nih.gov/)
2) **Image resampling:** All images will be resampled to a uniform voxel spacing (Brain: 1x1x1 mm3, Pelvis: 1x1x2.5 mm3). Image size might vary from case to case and between datasets.
3) **Image registration:** A rigid image registration between CBCT (for task2) or MR (for task1) and resampled CT will be performed for all cases to align the images (using Elastix: https://elastix.lumc.nl/index.php). Participants can further improve the registration (rigid and deformable) if required. Elastix parameter files will be made available for participants for rigid registration and a suggestion for deformable registration. During the evaluation, rigid image registration between CBCT/MR and CT will reduce anatomical differences.
4) **Image masking:** The field-of-view of MR/CBCT and CT will be aligned by automatically segmenting based on the threshold and morphological operation of the patient outline on the MR/CBCT. The resulting mask will also be dilated to include surrounding air and applied to the CT image. The mask will be provided and can be used by the participant, if needed, for pre-processing. Also, image similarity metrics will be calculated in this mask whenever specified.
5) **Image cropping:** All images will be cropped to the bounding box of the patient outline (with a margin of 20 voxels in plane) to reduce the data.
6) **Defacing:** Face removal/de-identification will be performed for brain patients. Open-source software is currently being investigated for face removal, as presented by [Schwarz e. Al 2021: https://doi.org/10.1016/j.neuroimage.2021.117845]. Defacing may vary between centers:
    a) UMC Utrecht: for task 1, the MRI has a FOV with no face included, but CT has a larger FOV. The MR mask from point 4 will be applied to CT to remove the face; for task 2, defacing is still under investigation.
    b) Groningen: for task 1, one of the open-source methods reported by Schwarz et al. will be adopted; for task 2, the same patients are present the patients also have an MRI, the chosen open-source method will be applied to MRI, and the face mask will be used for face removal on CBCT and CT.
    c) Nijmegen: for task 1, one of the open-source methods reported by Schwarz et al. will be adopted; for task 2, defacing is still under investigation.

The code used to perform pre-processing will be made publicly available at https://github.com/SynthRAD2023.

25 Sources of error

a) **Possible error source:** Image registration may not perfectly match the two images (CT-MR or CT-CBCT) due to slightly different positioning or anatomical changes. Even when using deformable registration, residual errors may still be present. The anatomy may change further between CBCT/MRI and CT. Especially in the pelvis, air will move around, and bladder/intestine filling can change. There is no fair way to correct this: deformable registration may be viable. However, MRI is affected by geometric distortion, and CBCT can be affected by cupping, which results in smaller body contours. Applying deformable registration will reduce but not completely eliminate (considering residual registration errors) the intrinsic difference between the input and reference image. However, such differences cannot be corrected during inference without having access to the reference image. Therefore, avoiding deformable registration during evaluation is the fairest way to assess the quality of the sCT. Note that all challenge participants will deal with the same dataset, so the evaluation is deemed unbiased, even though some teams may develop methods less sensitive to registration mismatches.

ASSESSMENT METHODS

26 Metric(s)

a) **metric(s) to assess a property of an algorithm**.

The output sCT will be compared to CT undergoing an *image similarity* (validation and test phase) and a *dose evaluation (test phase)*.

1. *Image evaluation* between the sCT and CT using the metrics listed below.

a. Mean absolute error (MAE) on the mask, defined as $MAE = \frac{1}{n}\sum_{i=1}^{n}\left|CT_i - sCT_i\right|$,
Where n is the number of voxels in this dilated mask as described in Section 24.

b. Peak signal-to-noise ratio (PSNR) on the mask, defined as

$$PSNR = 10\, log_{10}\left(\frac{Q^2}{\frac{1}{n}\sum_{i=1}^{n}\left(CT_i - sCT_i\right)^2}\right),$$

where n again the number of voxels in this dilated mask as described in Section 24, and Q is the population-wide dynamic range of voxel intensities in the CTs (global maximum - minimum over the patient population).

c. Structural similarity index (SSIM) on the mask as defined by:

$$SSIM = \frac{\left(2\mu_{CT}\mu_{sCT}+C_1\right)\left(2\delta+C_2\right)}{\left(\mu_{CT}^2+\mu_{sCT}^2+C_1\right)\left(\delta_{CT}^2+\delta_{sCT}^2+C_2\right)}\text{d}$$

where $\mu_{(s)CT}$ is the mean pixel value of the (s)CT, $\delta_{(s)CT}$ is the variance of the (s)CT, and δ is the covariance of CT and sCT. $C_1 = (0.01 \cdot Q)^2$ and $C_2 = (0.03 \cdot Q)^2$ are two variables to stabilize the division with weak denominators, where Q is again the population-wide dynamic range of voxel intensities of the CTs.

2. *Dose evaluation* will be performed globally and locally by comparing photon and proton dose calculations between reference CT and sCT. Clinically, the most relevant question would be how a dose plan optimized on the CT (the "ground truth" where patients are currently treated) would perform on the sCT (the image at the disposal in the new clinical workflow). Dose calculations will be performed with matRad (https://e0404.github.io/matRad/), an open-source treatment planning system where photon and proton intensity-modulated treatment plans will be optimized on CT. Dose prescriptions and plans will be chosen irrespectively of the original clinical goal for each anatomy, choosing the center of the planning target volume (PTV) as the isocenter. Specifically, we will plan to prescribe the target (PTV) of 30x2 Gy for the brain and 20x3 Gy for the pelvis at the 95% isodose level for both photons and protons. For simplicity, proton plans will be planned using the same PTV approach as photons without robust optimization. Dose delivery will be simulated via ten beams of 6MV for photons using the generic Linac model and 2-3 beams for proton plans using the generic proton system modeled in matRad. The number of beams may be optimized on a patient basis to comply with the dose prescription and limit the dose to the organs at risk following the international guidelines: for the pelvis [Hall et al., 2021] and the brain [Lambrecht et al., 2018]. To further reduce the dose to the healthy tissues and ensure plan uniformity between patients, we will use the same objective functions and constraints available in matRaD per treatment site, as reported in Table 1. Organ-at-risk (OAR) dose limits will be handled as hard constraints whenever possible but might be turned into objectives on a patient-specific basis.

Table 1. Dose constraints and planning objectives used in matRaD for the brain and pelvis cases, respectively. Values are taken from [Lambrecht et al., 2018] and [Hall et al., 2021].

Brain = 30 x 2 Gy to 95% of the PTV		Pelvis = 20x3 Gy to 95% of the PTV	
Structure	Objective/Constraint	Structure	Objective/Constraint
Brain stem	$D0.03\,cc < 60$ Gy $D0.03\,cc < 54$ Gy	Rectum	$V60$ Gy $< 1\%$ $V50$ Gy $< 22\%$ $V40$ Gy $< 38\%$

			$V30$ Gy $<$ 57% $V20$ Gy $<$ 85%
Chiasm	$D0.03\ cc < 55$ Gy	Bladder	$V60$ Gy $<$ 3% $V56.8$ Gy $<$ 5% $V48$ Gy $<$ 25% $V40$ Gy $<$ 50%
Optical Nerve	$D0.03\ cc < 55$ Gy	Femur heads	$Dmax < 37$Gy
Cochlea	$Dmean < 45$ Gy	Colon	$Dmax < 50$Gy
Brain	$V60$ Gy $< 3\ cc$	Small bowel	$Dmax < 40$Gy

For each sCT, the plan will be recalculated without replanning to avoid possible differences due to plan optimization.

The following metrics are considered for this aspect of the evaluation.

a. Mean absolute dose differences relative to the prescribed dose $D_{prescribed}$ as described by

$$MAE_{target\ dose} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|D_{\geq 90\%,CT,i} - D_{\geq 90\%,\, sCT,i}\right|}{D_{prescribed}},$$

with $D_{\geq 90\%,\,(s)CT}$ the dose distribution in the (synthetic) CT within the regions that receive at last 90% of $D_{prescribed}$ in de CT, and n the number of voxels within this region.

b. A dose-volume histogram (DVH) provides information on the delivered dose to the volume of specific structures. Small differences in DVH parameters between the CT and the sCT indicate a good radiotherapy treatment based on the sCT. We consider four DVH parameters: near-minimum dose $D98\%_{target}$, which is the dose that at least 98% of the target volume received, $V95\%_{target}$, which is the target volume that received at least 95% of the prescribed dose, near-maximum dose $D2\%_{OAR}$, which is the dose 2% of the volume of a specific organ-at-risk (OAR) received, and $Dmean_{OAR}$, which is the mean dose a specific OAR received. Specifically, using the near-minimum and near-maximum was suggested by ICRU83 (ttps://www.fnkv.cz/soubory/216/icru-83.pdf). For the evaluation, we recognize three organs at risk per region. The bladder, rectum, and pelvic bones are considered in the pelvic region, and the brain stem, brain, and optic chiasm are considered in the brain. To obtain one metric for all DVH parameters, we sum the relative absolute differences of the abovementioned parameters between the CT and sCT. The final metric is defined as

$$DVH_{metric} = \frac{\left|D98\%_{target,CT} - D98\%_{target,sCT}\right|}{D98\%_{target,CT}} + \frac{\left|V95\%_{target,CT} - V95\%_{target,\,sCT}\right|}{V95\%_{target,CT}}$$
$$+ \frac{1}{n_{OARs}}\sum_{OARs}\frac{\left|D2\%_{OAR,\,CT} - D2\%_{OAR,\,sCT}\right|}{D2\%_{OAR,\,CT}} + \frac{1}{n_{OARs}}\sum_{OARs}\frac{\left|Dmean_{OAR,\,CT} - Dmean_{OAR,\,sCT}\right|}{Dmean_{OAR,\,CT}},$$

where n_{OARs} is the number of OARs.

c. Gamma index: The Gamma pass ratios will be calculated for sCTs using the CT doses as a reference. The calculation is performed in 3D, according to Low et al., 1998. The

passing criteria are the dose-difference criterion $\Delta D_M = 2\%$ and the distance-to-agreement criterion $\Delta d_M = 2mm$. The gamma index at each position vector r in the sCT is:
$\gamma\left(r_{sCT}\right) = min\left\{\Gamma\left(r_{sCT}, r_{CT}\right)\right\} \forall \left\{r_{CT}\right\}$, where r_{sCT} indicates a position vector in the CT and

$$\Gamma(r_{sCT}, r_{CT}) = \sqrt{\left(\frac{\left|r_{CT}-r_{sCT}\right|}{\Delta d_M}\right)^2 + \left(\frac{D_{CT}\left(r_{CT}\right)-D_{sCT}\left(r_{sCT}\right)}{\Delta D_M}\right)^2}.$$

Gamma pass rates will be calculated within the regions that receive at dose higher than 10% of $D_{prescribed}$, based on the dose calculation of the CT, as suggested by Ezzel et al., 2009.

b) **Metric justification**

Image similarity: MAE, PSNR, and SSIM are image similarity metrics commonly used in medical image synthesis. For a detailed overview, see the review by Spadea and Maspero et al., 2021.

Relative dose differences: This metric gives insight into the dose distribution in the local regions that receive a high dose. Only including these regions ensures that the large regions receiving little to almost no dose do not unintentionally determine the outcome of this metric.

DVH parameters: In radiotherapy, DVH parameters are commonly used to assess dose distributions in target volumes and OARs [Drzymala et al., 1991]. Specifically, DVH parameters describe target coverage and OAR sparing. Considering differences in the DVH parameters is a way of verifying that clinically relevant objectives are maintained.

Gamma index: This is a commonly used metric in radiotherapy to compare two dose distributions. It combines dose difference criteria and distance difference criteria in a single metric. Low et al., 1998 describe the details of the theoretical background. The medical physicist association suggests using 3%,3mm distance-to-agreement criteria when comparing delivered and planned doses [Ezzel et al., 2009]. We will adopt an even more stringent criterion (2%, 2mm), considering that error in the planning phase leads to irradiating the patient with a systematically deviating dose from the planned one.

27 Ranking method(s)

a) **Method used to compute a performance rank**

To obtain one value per patient in the test set for every metric described in Section 26.2, the sum is taken over the photon and proton evaluations. Subsequently, the final result per metric is obtained by averaging all the test cases.

Phase 1 = validation: The first ranking phase includes only the image metrics on the validation set. The MAE, PSNR, and SSIM will be calculated and separately normalized between zero (worst result among participants) and one (the best result among participants). Subsequently, the normalized metrics will be summed with equal weights and converted to a standard ranking: 1 (best submission, highest summed normalized metrics) to *n* (worst submission, lowest summed normalized metrics).

Phase 2 = test: The second-ranking phase includes the image metrics & the dose evaluation on the test set. First, the automatic ranking based on the image similarity metrics will be performed, as in Phase 1. Secondly, dose evaluation will be run, and again the $MAE_{target\ dose}$, DVH_{metric} and gamma index will be normalized between zero and one. Dose evaluation depends on matRad, an open source fully validated treatment planning system written in Matlab. We are investigating whether it is feasible to incorporate an automatic evaluation pipeline on the hosting site. If possible, dose evaluation will be offered for all the test submissions, along with a leaderboard. The normalized metrics for all image and dose metrics will be summed, where the dose evaluation metrics will weigh twice as much as the image similarity metrics. If the integration of this downstream task turns out to be too cumbersome, in that case, only the best ten submissions based on the image similarity metrics will be considered for dose evaluation, and only the sum of the normalized metrics of the dose evaluation (equal weighting for $MAE_{target\ dose}$, DVH_{metric} and gamma index) will contribute to the final ranking.

b) **Submissions with missing results** on test cases.
If a team lacks submitting sCTs for one or multiple patients in the validation phase, the team will be contacted and asked to resubmit within 24 hours. When a resubmission is still incomplete or corrupt, a black image (indicating an image that contains only air) will be used as an sCT of the specific patient. During the test, the dockerized algorithms need to be submitted, so if the preliminary phase works, there cannot be missing result cases.

c) **Justification of ranking scheme**

- Normalizing the metrics separately was chosen to overcome issues with metrics that do not have a predefined minimum and maximum value or metrics that always show a relatively high or low value.
- Aggregating and ranking were chosen to preserve large and small performance differences while combining metrics.
- The image similarity functions as a first threshold metric: we consider good similarity a prerequisite to access downstream dose evaluation.

28 Statistical analyses

a) **Details of the statistical methods**
The variability of the ranking will be assessed with a Kandall's tau analysis. We will investigate whether including only the image or the dose evaluation will lead to different rankings. Furthermore, the patterns of the different dose evaluations will also be analyzed.

b) **Justification of statistical method**
Kendall's tau quantifies differences between rankings, which can provide insight into the quality of a metric.

29 Further analyses

We will categorize performances based on the participants' methods to generate the sCTs. This analysis will, for example, consider the difference between paired versus unpaired approaches and 2D versus 3D models. Furthermore, we will analyze the influence of the automated quantitative analysis on biases in our data and methods, considering, for example, the effect of registration and the difference in the quality of paired images for the brain and pelvis. Lastly, we will include qualitative analysis by expert observers on single cases, reporting their agreement on the quantitative analysis and their opinion on the sCT quality.

REFERENCES

- Chernak E.S., Rodriguez-Antunez A., Jelden G.L., Dhaliwal R.S., Lavik P.S. The use of computed tomography for radiation therapy treatment planning. Radiology. 1975 Dec;117(3):613-4. https://doi.org/10.1148/117.3.613
- Ramella, S., Fiore, M., Silipigni, S., Zappa, M. C., Jaus, M., Alberti, A. M., ... & D'Angelillo, R. M. (2017). Local control and toxicity of adaptive radiotherapy using weekly CT imaging: results from the LARTIA trial in stage III NSCLC. Journal of Thoracic Oncology, 12(7), 1122-1130. https://doi.org/10.1016/j.jtho.2017.03.025

- Kida, S., Nakamoto, T., Nakano, M., Nawa, K., Haga, A., Kotoku, J. I., ... & Nakagawa, K. (2018). Cone beam computed tomography image quality improvement using a deep convolutional neural network. *Cureus*, *10*(4). https://doi.org/10.7759%2Fcureus.2548
- Schmidt M. A., Payne G. S. Radiotherapy planning using MRI. Phys Med Biol. 2015;60:R323
- Lagendijk, J. J., Raaymakers, B. W., Van den Berg, C. A., Moerland, M. A., Philippens, M. E., & Van Vulpen, M. (2014). MR guidance in radiotherapy. Physics in Medicine & Biology, 59(21), R349. https://doi.org/10.1088/0031-9155/59/21/r349
- Edmund, J. M., & Nyholm, T. (2017). A review of substitute CT generation for MRI-only radiation therapy. *Radiation Oncology*, *12*(1), 1-15. https://doi.org/10.1186/s13014-016-0747-y
- Spadea, M. F. & Maspero, M., Zaffino, P., & Seco, J. (2021). Deep learning-based synthetic-CT generation in radiotherapy and PET: A review. Medical Physics, 48(11), 6537–6566. https://doi.org/10.1002/mp.15150
- Schwarz, C. G., Kremers, W. K., Wiste, H. J., Gunter, J. L., Vemuri, P., Spychalla, A. J., ... & Alzheimer's Disease Neuroimaging Initiative. (2021). Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives. *NeuroImage*, *231*, 117845. https://doi.org/10.1016/j.neuroimage.2021.117845
- Low, D.A., Harms, W.B., Mutic, S., and Purdy, J.A. (1998), A technique for the quantitative evaluation of dose distributions. Med. Phys., 25: 656-661. https://doi.org/10.1118/1.598248
- Hall, W. A., Paulson, E., Davis, B. J., Spratt, D. E., Morgan, T. M., Dearnaley, D., ... & Lawton, C. A. (2021). NRG oncology updated international consensus atlas on pelvic lymph node volumes for intact and postoperative prostate cancer. International Journal of Radiation Oncology* Biology* Physics, 109(1), 174-185. https://doi.org/10.1016/j.ijrobp.2020.08.034
- Lambrecht, M., Eekers, D. B., Alapetite, C., Burnet, N. G., Calugaru, V., Coremans, I. E., ... & Troost, E. G. (2018). Radiation dose constraints for organs at risk in neuro-oncology; the European Particle Therapy Network consensus. *Radiotherapy and Oncology*, *128*(1), 26-36 https://doi.org/10.1016/j.radonc.2018.05.001
- Ezzell, G. A., Burmeister, J. W., Dogan, N., LoSasso, T. J., Mechalakos, J. G., Mihailidis, D., ... & Xiao, Y. (2009). IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119. *Medical physics*, *36*(11), 5359-5373. https://doi.org/10.1118/1.3238104
- Drzymala, R. E., Mohan, R., Brewster, L., Chu, J., Goitein, M., Harms, W., & Urie, M. (1991). Dose-volume histograms. *International Journal of Radiation Oncology* Biology* Physics*, *21*(1), 71-78. https://doi.org/10.1016/0360-3016(91)90168-4
- International Commission on Radiation Units and Measurements, ICRU Report 83, Prescribing, recording, and reporting intensity-modulated photon-beam therapy (IMRT)(ICRU Report 83), Bethesda, MD (2010) https://www.fnkv.cz/soubory/216/icru-83.pdf

OTHER COMMENTS: Challenge origin

The idea of a challenge organization arose within the “Image synthesis & reconstruction” subgroup of the Dutch deep learning in radiotherapy initiative (www.DLinRT.org). Such subgroups arose from an initiative independently organized by professors affiliated with all the Dutch University medical centers, focusing on deep learning and working in radiotherapy departments. The organizer group encompasses early-stage researchers, PhDs, postdocs, three assistant professors, and one associate professor.