Unified Benchmarks for Imaging in Computational pathology, Radiology and Natural language: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Unified Benchmarks for Imaging in Computational pathology, Radiology and Natural language

Challenge acronym

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

UNICORN

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.

Challenges have undoubtedly shaped the field of medical image analysis and deep learning in the last decade, following a "many-to-one" approach, where numerous models competed on a specific single (narrow) task. But the emergence of (multimodal) foundation models powered by (vision) transformers comes with a paradigm change in benchmarking towards a "one-to-many" approach, where a single model is benchmarked across a selection of (multimodal) tasks.
Language models are benchmarked on different tests, including math, medicine and geometry, vision models are being benchmarked via zero-shot or few-shot learning for classification, detection and segmentation tasks, also in the field of medical imaging. One of the main promises of these large models pre-trained with very-large scale is to be "generalist" models, able to tackle multiple tasks without being specifically trained to solve that task with large-scale data. Instead, paradigms of fine-tuning and few-shot learning are increasingly being adopted and explored, envisioning a future where user can (re-)purpose foundation models to address different tasks with minimal interaction via prompts, similar to what is done via text with models such as ChatGPT. But for medical image analysis, the question is how far we are now with the development and usability of multi-modal foundation models, and to be able to answer this question, the field lacks a comprehensive publicly available benchmark that encompasses these new evaluation forms. We believe it is the right time to introduce such a benchmark, which we envision in the form of a challenge.

We propose UNICORN, a challenge to provide a unified set of benchmarks to assess the performance of multimodal foundation models. We focus on both image data in the fields of radiology and digital pathology, text data, using medical reports, and images + text, focusing on multi-modal approaches. We release multi-modal public data that can be used by participants to fine-tune existing (pre-trained) foundation models or to develop a strategy based on few-shot learning, and establish a battery of benchmarks based on sequestrated test data.

Considering both vision, and language and vision-language tasks, UNICORN is not just about the performance of models on individual tasks, but about understanding how well integrated vision-language models can adapt to the complexities of medical data. In UNICORN, we will put together multiple teams from one Dutch medical center (Radboudumc), one Dutch onology center (Maastro), a precision oncology group from Germany (TU Dresden), with multidisciplinary expertise and a strong track on challenge organization*. We aim at making UNICORN a unified evaluation platform for fair and uniform comparison of multimodal foundation models in medical imaging.

* We are currently discussing the inclusion of additional medical centers in the Netherlands, which have not fully confirmed their commitment yet due to time constraints, therefore not included in this proposal, but will likely join and contribute with data in existing tasks as well as new tasks, should this proposal being accepted.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

foundation models, radiology, pathology, vision, language and vision applications, multi-modality, classification, segmentation, detection, few-shot learning, textual prompting, visual prompting

Year

2025

Lighthouse challenge agreement

The organizers agree to all of the following points:

- The full labeling protocol will be sent to the challenge chairs in addition to the full proposal document.
- A set of a few representative data samples including annotations will be sent to the challenge chairs in addition to the full proposal document.
- The challenge will be open for at least 4 months.
- For the dataset review, the challenge chairs will get access to the data at least 3 months before challenge opening.

Challenge organizers have read and agree to all of the above terms and conditions.

Lighthouse challenge information

In two sentences or less, what sets your challenge apart from ordinary MICCAI challenges. In other words: What makes your challenge a lighthouse challenge?

Unlike traditional MICCAI challenges that benchmarks models for specific, single tasks, UNICORN sets itself apart by creating a comprehensive benchmark for evaluating multimodal foundation models across a wide range of medical imaging and language tasks, including radiology, digital pathology, and vision-language integration. This challenge aims at creating a platform to assess the potential of generalist foundation models in tackling complex, real-world clinical scenarios.

Previous challenge(s)

What is the closest challenge to your proposed lighthouse challenge? Are there previous versions of it? Specifically, if you applied for a 2024 challenge, what is the delta between the two iterations? (e.g., number of centers for new data, number of newly added data – This is not to be confused with details about the total data set)

The closest challenge to UNICORN is the MedFM challenge that focuses on leveraging early foundation models to tackle multiple downstream tasks. However, while the MedFM challenge includes three vision tasks, solely focuses on 2D data and on small regions of interests, UNICORN encompasses a more diverse set of at least 20 tasks that span across vision, language and vision-language applications, including 2D and 3D data problems directly derived from clinical practice, such as the analysis of entire digital pathology whole-slide images, or entire CT scans of the thorax, as well as analysis of clinical reports via NLP, and quantification of biomarkers in oncology.

Test set status

Was the test set (or parts of it) already used in previous challenges and/or previously made publicly available?

To offer a diverse range of tasks, UNICORN combines test sets from existing challenges with new test datasets. All test data in UNICORN, whether part of the sequestered test sets from existing challenges or from newly curated test sets, has never been made publicly available. This ensures the integrity and validity of our benchmarking process.

By incorporating test sets from established challenges in the medical field, UNICORN facilitates a direct comparison of foundation models' performance in a few-shot learning setting against that of the challenge-specific models that have been trained with extensive supervision.

Existing challenges whose (part of the) private test set is used in UNICORN: PI-CAI (https://pi-cai.grand-challenge.org), TIGER (https://tiger.grand-challenge.org), LEOPARD (https://leopard.grand-challenge.org), ULS (https://uls23.grand-challenge.org), SPIDER (https://spider.grand-challenge.org), DRAGON (https://dragon.grand-challenge.org), PANDA (https://www.kaggle.com/c/prostate-cancer-grade-assessment)

What major scientific advances or insights are expected from the challenge?

Please describe the major scientific advances ore insights you expect to be gained from the challenge. Please include references to the state of the art in your description and list open research questions to which the challenge seeks answers or solutions.

The UNICORN challenge aims at providing insights into the effectiveness of multimodal foundation models in handling diverse medical data types and performing across multiple clinically relevant tasks. Understanding how well these models can generalize to new tasks and application areas with minimal fine-tuning or few-shot learning is the main promise of foundation models, and understanding where we stand now, and where we will stand in the future in the developments of such models is crucial for advancements in this research field.

With UNICORN, we are addressing some fundamental questions that are of active research interest:
- Can a single foundation model be effectively adapted to various medical tasks, also across different data modalities?
- Can foundation models be fine-tuned for medical tasks using scarce labelled data via few-shot learning?
- What pre-training strategies and fine-tuning approaches are most effective for medical applications?
- What is the most effective way of using few-shot learning to address a clinical question with a foundation model?

To address these questions, UNICORN participants will develop "adapters" (see definition in the rest of the proposal) to adapt their single model to multiple tasks, and we will request them to make their solutions publicly available. This will allow the scientific community to gain insight on how models can be adapted to best address multiple tasks, for example via optimised techniques of few-shot detection, segmentation and classification, as well as optimised fine-tuning, to promote further research and development in the field.

Clinical body affiliation

Please describe your proposed challenge's affiliation with a clinical body, if any. How do you plan to engage the clinical community that your challenge is set to impact?

The proposed challenge is organized as a joint collaboration between the Department of Pathology and the Department of Medical Imaging, at the Radboud University Medical Center (Radboudumc) in Nijmegen, as well as the Maastro clinic in Maastricht (Netherlands), and the Technical University of Dresden (Germany, led by Prof. Jakob Kather, a physician with expertise in deep learning). The organising team natively involves pathologists, radiologists, computer scientists, research software engineers, medical professionals and researchers, thanks to the embedding of the organising teams.

During the preparation phase of UNICORN, we will connect to relevant societies and main events in the radiology and pathology space, via presence and advertisement of UNICORN at clinical international events such as RSNA, USCAP, ECDP, ECP and ECR. The UNICORN organizing team has strong presence (in organizing/steering committees, workshops/symposia organization), and collaborations within these events, where we plan to request for endorsement of UNICORN as an AI challenge and strengthen connections between the pathology/radiology and the machine learning societies. We will also maintain these connections after MICCAI 2025, in view of sustainability of UNICORN as a future reference benchmark.

Deadline for data acquisition and annotation

What's the deadline for data acquisition and annotation?

We already have the data and annotations required for the core tasks of the challenge. Data will be centralized at Radboudumc and then either uploaded to grand-challenge or shared publicly. To collect data and annotations from non-Radboudumc organizers, data transfer agreements are needed. We will start drafting those in case this proposal is accepted. For the optional tasks, we are working on curating the data and annotations. We anticipate completing these agreements and having all data ready for use by the end of 2024.

How much prize money has been secured?

Please state how much prize money has already been secured for the challenge.

No prize money has been secured yet. In case of acceptance, we will reach out to commercial partners within our network and start discussing sponsorship. We have successfully done this in the past for both challenges and academic events such as workshops and symposia, therefore we do not expect this to be a problem. The number of awards and the amount of prize money will depend on the number of sponsors.

Computing requirements per participant

Roughly estimate how much computing power would be required per challenge participant?

We considered three scenarios: a "best" case where participants use a relatively small foundation model, a "realistic" case where participants use a relatively large foundation model, and a "worst" case scenario where participants max out the compute constraint we aim to enforce (10 minutes per radiology vision case, 30 minutes per pathology vision case, and 30 seconds per language cases). Estimations are based on a T4 instance with 16GB RAM, which costs approximately 0.50 euros per hour.

For pathology vision tasks:

Best-case scenario: 440 seconds to process a case
Realistic-case scenario: 990 seconds to process a case
Worst-case scenario: 1800 seconds to process a case (30min in total)

For radiology vision tasks:

Best-case scenario: 178 seconds to process a case
Realistic-case scenario: 348 seconds to process a case
Worst-case scenario: 600 seconds to process a case (10min in total)

Assuming we allow up to 6 submissions per participant during the experimental test phase, and a single pass on the final test phase data, this would amount to:

Pathology (436 experimental test cases + 971 final test cases):

Best-case scenario: 436 * 440 sec * 6 + 971 * 440 sec = 438 hours
Realistic-case scenario: 436 * 990 sec * 6 + 971 * 990 sec = 987 hours
Worst-case scenario: 436 * 1800 sec * 6 + 971 * 1800 sec = 1794 hours

Radiology (335 experimental test cases + 2442 final test cases):

Best-case scenario: 335 * 178 sec * 6 + 2442 * 178 sec = 220 hours
Realistic-case scenario: 335 * 348 sec * 6 + 2442 * 348 sec = 430 hours
Worst-case scenario: 335 * 600 sec * 6 + 2442 * 600 sec = 742 hours

For language tasks:

Best-case scenario: 1 second to process a case
Realistic-case scenario: 10 seconds to process a case
Worst-case scenario: 30 seconds to process a case

Assuming we allow up to 6 submissions per participant during the experimental test phase, and a single pass during the final test phase, this would amount to:

Pathology (250 experimental test cases + 2128 final test cases):

Best-case scenario: 250 * 1 sec * 6 + 2128 * 1 sec = 1 hour
Realistic-case scenario: 250 * 10 sec * 6 + 2128 * 10 sec = 10 hours
Worst-case scenario: 250 * 30 sec * 6 + 2128 * 30 sec = 30 hours

Radiology (550 experimental test cases + 3526 final test cases):

Best-case scenario: 550 * 1 sec * 6 + 3526 * 1 sec = 2 hours
Realistic-case scenario: 550 * 10 sec * 6 + 3526 * 10 sec = 20 hours
Worst-case scenario: 550 * 30 sec * 6 + 3526 * 30 sec = 60 hours

For a participant going through all tasks with 6 submissions in the experimental phase and 1 submission in the final test phase, the cost would be:

Best-case scenario: 661 hours, i.e. 330 euros
Realistic-case scenario: 1447 hours, i.e. 723 euros
Worst-case scenario: 2626 hours, i.e. 1313 euros

This suggests that the realistic-case scenario could require around 1447 hours. To account for participants using more complex models, we propose planning around a maximum compute allocation of 1700 hours per participant as a safeguard. Note that this is an estimation solely based on the 20 core tasks, without considering any optional task.

With 10 participants completing all tasks and 30 participants completing half the tasks, the total compute time would be roughly 42000 hours, leading to a total cost estimate of 21000 euros.

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

N/A

Duration

How long does the challenge take?

Full 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.

Given the diverse nature of the challenge, which encompasses Pathology and Radiology, Vision, Language, and Vision-Language, we anticipate a higher volume of submissions in each individual category, estimated at around 30-50 entries. Conversely, we expect fewer comprehensive submissions - approximately 10 - that address all categories. This relatively low number is attributable to the complexity involved in developing foundation models, despite their current prominence in the field.

Publication and future plans

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

Yes, we plan to coordinate a publication of the challenge results.

Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

grand-challenge.org

TASK 1: Vision - Classifying HE prostate biopsies into ISUP scores

SUMMARY

Abstract

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

Challenges have undoubtedly shaped the field of medical image analysis and deep learning in the last decade, following a "many-to-one" approach, where numerous models competed on a specific single (narrow) task. But the emergence of (multimodal) foundation models powered by (vision) transformers comes with a paradigm change in benchmarking towards a "one-to-many" approach, where a single model is benchmarked across a selection of (multimodal) tasks.
Language models are benchmarked on different tests, including math, medicine and geometry, vision models are being benchmarked via zero-shot or few-shot learning for classification, detection and segmentation tasks, also in the field of medical imaging. One of the main promises of these large models pre-trained with very-large scale is to be "generalist" models, able to tackle multiple tasks without being specifically trained to solve that task with large-scale data. Instead, paradigms of fine-tuning and few-shot learning are increasingly being adopted and explored, envisioning a future where user can (re-)purpose foundation models to address different tasks with minimal interaction via prompts, similar to what is done via text with models such as ChatGPT. But for medical image analysis, the question is how far we are now with the development and usability of multi-modal foundation models, and to be able to answer this question, the field lacks a comprehensive publicly available benchmark that encompasses these new evaluation forms. We believe it is the right time to introduce such a benchmark, which we envision in the form of a challenge.

We propose UNICORN, a challenge to provide a unified set of benchmarks to assess the performance of multimodal foundation models. We focus on both image data in the fields of radiology and digital pathology, text data, using medical reports, and images + text, focusing on multi-modal approaches. We release multi-modal public data that can be used by participants to fine-tune existing (pre-trained) foundation models or to develop a strategy based on few-shot learning, and establish a battery of benchmarks based on sequestrated test data. Considering both vision, and language and vision-language tasks, UNICORN is not just about the performance of models on individual tasks, but about understanding how well integrated vision-language models can adapt to the complexities of medical data. In UNICORN, we will put together multiple teams from one Dutch medical center (Radboudumc), one Dutch onology center (Maastro), a precision oncology group from Germany (TU Dresden), with multidisciplinary expertise and a strong track on challenge organization*. We aim at making UNICORN a unified evaluation platform for fair and uniform comparison of multimodal foundation models in medical imaging.

* We are currently discussing the inclusion of additional medical centers in the Netherlands, which have not fully confirmed their commitment yet due to time constraints, therefore not included in this proposal, but will likely join and contribute with data in existing tasks as well as new tasks, should this proposal being accepted.

Keywords

List the primary keywords that characterize the task.

foundation models, radiology, pathology, vision, language and vision applications, multi-modality, classification, segmentation, detection, few-shot learning, textual prompting, visual prompting

ORGANIZATION

Organizers

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

Francesco Ciompi (Radboudumc, Netherlands), Alessa Hering (Radboudumc, Netherlands), Clément Grisi (Radboudumc, Netherlands), Michelle Stegeman (Radboudumc, Netherlands), Judith Lefkes (Radboudumc, Netherlands), Marina D'Amato (Radboudumc, Netherlands), Luc Builtjes (Radboudumc, Netherlands), Lena Philipp (Radboudumc, Netherlands), Fennie van der Graaf (Radboudumc, Netherlands), Jakob Kather (TU Dresden, Germany), Marta Ligero Hernandez (TU Dresden, Germany), Leonard Wee (Maastro, Netherlands), Andre Dekker (Maastro, Netherlands), Bram van Ginneken (Radboudumc, Netherlands)

b) Provide information on the primary contact person.

Francesco Ciompi (Radboudumc, Netherlands), Alessa Hering (Radboudumc, Netherlands)

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

The challenge will initially be structured as a one-time event with a fixed submission deadline. Once the challenge has taken place and submissions have been evaluated, we plan to keep the challenge open for further submissions beyond the final deadline. After the initial event, depending on the outcomes and feedback received, the organizers will assess the feasibility and interest in transitioning the challenge into a repeated event with an annual fixed submission deadline. This would provide a regular opportunity for researchers to showcase advancements in multimodal foundation models and foster ongoing collaboration and innovation in the field.

Challenge venue and platform

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

MICCAI 2025

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

grand-challenge.org

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

https://unicorn.grand-challenge.org (we will create it if this proposal is accepted)

Participation policies

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

We will benchmark foundation models using a few-shot learning paradigm. For this, we will build a testing mechanism on the grand-challenge.org platform where participants will upload a docker container enclosing their algorithm to the grand-challenge platform, which will execute few-shot learning in the form of visual or textual prompts , and run the algorithm on sequestered test sets without exposing any of few-shot, or test data to the user. Once the algorithm is submitted, no user interaction is allowed, submitted algorithms will have to run in a fully-automatic fashion. However, the implementation of few-shot learning on grand-challenge de facto mimics some form of user interaction, but we will fully automate it.

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.

We set up UNICORN with the assumption that participants will join the challenge with their existing pre-trained foundation models. Therefore, we do not pose any restrictions to the type and source of training data used, any publicly available or private data including (open) pre-trained nets is allowed. Note that we do not release large-scale training data with UNICORN, but solely limited-size datasets meant for fine-tuning and few-shot learning.

Most data released publicly come from already existing publicly available datasets, such as existing challenges, which we re-purpose to the context of few-shot learning.

We expect participants to use these (limited-size) data to develop what we call an "prompt adapter", meaning an algorithmic solution that is "attached" to the foundation models, allowing it to address each task. The foundation model should always remain the same, at least within tasks of the same time (e.g. one vision model for vision tasks, one language model for language tasks, one vision-language model for vision-language tasks; but also a single vision-language model for all tasks is allowed, even preferred).

While we do not require participants to publicly release their foundation model, we will require them to make the code of their "adapter" open source with the final submission to UNICORN.

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.

Individuals affiliated with any of the sponsoring or organizing bodies (i.e. Radboud University Medical Center, Maastro and TU Dresden) are welcome to participate in the challenge. However, they will not be eligible for inclusion in the final ranking during the testing phase, nor for prizes.

Note that the group of TU Dresden is finalizing the development of their multi-modal foundation model, which will be used as the baseline model in UNICORN. While they are not asked to release their model publicly, they will open source the adapters that they will develop.

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

In case of acceptance, we will look for sponsorship to provide prizes for the best methods (e.g., top-3). Depending on the sponsorship availability, we will consider the following potential (but not limited to these) options:

- Prize for the best overall model
- Prize per type of downstream task (e.g. prize for segmentation, detection, classification)
- Prize per category (vision, vision-language, language)
- Prize per modality (radiology, pathology)

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 top N performing methods will be announced publicly during MICCAI 2025.
- We will use a public (experimental) and a final leaderboard. During the experimental phase, the performance and rankings of submitted algorithms on the experimental test set will be visible via the challenge's live leaderboard. The results of the final leaderboard will be formally presented at the MICCAI 2025 challenge session. At the same time, the final leaderboard for all participants will be visible at the grand-challenge platform.
- We will give top N performers the opportunity to present their results during the MICCAI 2025 challenge session.
- We will add top N algorithms and their authors to the challenge homepage.
- Invite x teams to present their method and results during the UNICORN workshop at MICCAI 2025. The teams are selected both based on the final leaderboard and because of high performance in individual categories (e.g. best algorithm in radiology). The final ranking will be announced during the UNICORN workshop.

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).

All team members who contributed to the submissions discussed in the paper will be invited to become co-authors of the UNICORN Journal article. This includes at least the x teams selected to present their methods at the workshop. Additionally, a larger number of submissions that have demonstrated scientific relevance may also be invited to co-author the journal paper.

The teams are permitted to publish their own papers. They may include the results of the UNICORN challenge after an embargo period, for the publication of the challenge journal paper.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Participants must submit their models as Docker containers, adhering to the format specified by the Grand Challenge platform. Each team is allowed to submit one foundation model for each task type, such as radiology classification or pathology detection. The Docker container must also include the postprocessing component (i.e., the "adapter") using few-shot examples such as KNN, linear probing, or a small network, specific to the related downstream task (e.g., classification, segmentation). The code for the adapter part must be made publicly available by the authors at the end of UNICORN. Restrictions on the complexity of the postprocessing, such as computational time or resource usage, will be defined in advance. Based on the challenge setup that we propose, fine-tuning the whole model on platform will not be allowed and also not possible due to the limited amount of few-shot cases and a time limitation to run the docker, which we will set on grand challenge, also to make the challenge computationally feasible.

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.

We will set up a two-phase challenge.
To check the functionality of the submission, participants will be allowed to submit multiple times during the experimental phase, and several validation leaderboards will be available for the different categories with N submission per team.
In the final phase, only one submission per team is allowed on the final test data set.

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

October 2024 challenge announcement (during MICCAI 2024) and opening registration to challenge, challenge description on the grand-challenge website

Q4 2024: Releasing first public datasets with few-shots
Q1 2025: Release Github repo with example code for docker container implementing a full (baseline) solution for 1 radiology and 1 pathology task
Q1: Validation leaderboard available on GC.
July 2025: Final submission to test dataset

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

The local IRBs have approved the collection of data and their use for research purposes. Below we report the reference number of IRB approvals obtained so far. For new datasets that are being made for the purpose of UNICORN, IRB approval is not available yet. However, based on our experience with previous approvals, we do not expect major problems.

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)

All data used as part of the sequestered test set will not be released publicly, to prevent the community to (pre-)train their models on these data.
Data released publicly in the form of few-shots will be mostly derived from existing public datasets. In those cases, we will adhere to the original license of these datasets.
For new data that we will release publicly for the purpose of this challenge, we will adopt a CC BY-NC-SA license.

Code availability

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

The evaluation software will be made publicly accessible through a GitHub repository with a permissive license (Apache 2.0). This repository will contain the code necessary to produce rankings and evaluate the submissions received from participants.

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

Participants are required to make the postprocessing part of their code publicly available when they make their final submission. This includes the code for the postprocessing component, where they utilize few-shot examples and aggregate feature vectors for specific downstream tasks, such as classification or segmentation. The repository will be linked to their submission on GC. We will ask participants to release their code under a permissive license, at least for the top-N methods (N to be defined based on number of submissions and performance).

Additionally, while we encourage participants to release their pre-trained models, this is not mandatory. This decision aims to increase the number of industrial submissions and address privacy concerns.

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.

Neither the public nor final leaderboard data will be released. The organizing team will have full access to all the test case labels, the collaborating parties will only have access to test case labels of their own data.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

- Diagnosis
- Prognosis
- Screening

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction

- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Vision:
- Classification
- Segmentation
- Detection
- Prediction:

Language:
- Classification
- Extraction
- NER
- Generative task

Vision-language:
- Zero-shot classification
- Zero-shot segmentation
- Cross-model retrieval
- Generative task (Image captioning)

Cohorts

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

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

Patients included in prostate cancer screening programs, patients presenting for prostate cancer diagnosis at the hospital and undergoing a biopsy procedure to obtain tissue samples for further pathological examination.

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

The challenge cohort includes two sub cohorts, one being used for the experimental test phase, and the other for the final test phase. The first sub cohort includes subjects who underwent a prostate biopsy owing to a suspicion of prostate cancer at Radboud University Medical Center, between Jan 1, 2012, and Dec 31, 2017.The second sub cohort was gathered after making a call through social media platforms to gather a diverse and representative set of prostate biopsy slides from routine clinical practice. Both cohorts are subjects presenting for prostate cancer diagnosis and undergoing a biopsy procedure to obtain tissue samples for further pathological examination.

Imaging modality(ies)

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

Hematoxylin-eosin (HE) stained histopathology whole-slides.
Hematoxylin-eosin-saffron (HES) stained histopathology whole-slides.

Context information

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

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

No further information given.

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

No further information given.

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.

Prostate sample shown in histopathology data.

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.

Accurate classification of subjects into one of the International Society of Urological Pathology (ISUP) grade groups.

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)).

Runtime, Accuracy, Robustness, Complexity, Reliability.

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).

The data was acquired in 6 medical centers across Europe. Therefore, a variety of device types were used. Slides were scanned into whole-slide images using 4 different scanner vendors: 3DHistech, Hamamatsu, Leica and Phillips. The obtained slides had varying original minimum pixel resolutions: 0.23, 0.24, 0.25, 0.46, 0.48, 0.5 micron per pixel.

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

None.

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

The data was acquired in the following 6 medical centers:

1- Radboud University Medical Center (Netherlands)
2- Rennes University Hospital (France)
3- Institute of Pathology and Molecular Immunology of the University of Porto (Portugal)
4- Memorial Hospitals Istanbul (Turkey)
5- Foch Hospital (France)
6- Isala Zwolle (Netherlands)

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).

Not applicable.

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.

One case is one slide belonging to one subject.

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

Development phase: during this phase, participants can use the publicly available dataset from the PANDA challenge, containing 10616 cases. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases. This data consists of a total of 278 cases and is split into two non-overlapping sets.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset consists of 165 cases. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on 113 new cases.

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

By design, our challenge aims at evaluating foundation model capacities. Hence, we do not provide extensive training data, as it would still be relatively small compared to what these models are typically trained on. Instead, participants are encouraged to leverage publicly available datasets with similar labels for model development. In cases where such datasets are unavailable, we offer a set of few-shot labeled examples for participants to utilize. Participants will use the publicly released "shots" to fine-tune their model or to encode a task-specific "Adapter Tuning/Prompting" mechanism to repurpose their foundation model "on the fly" to address each specific task. Subsequently, participants will transition to the platform for the experimental phase. Here, they will receive a set of "shots" to further fine-tune their models or adapt them to perform prediction on the sequestered validation dataset.
The emphasis, therefore, shifts to curating a diverse set of private cases to evaluate the capacities of foundation models.

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.

Distribution of classes reflects real-world distribution.

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.

Experimental test phase cases:

The reference standard was determined in three rounds. In the first round, three uropathologists (C.H.v.d.K., R.V., H.v.B.) individually graded the cases digitally using the ISUP 2014 guidelines. For a number of cases, the majority vote was taken: cases with an agreement on ISUP grade group but a difference in Gleason pattern order, e.g., 5 + 4 versus 4 + 5; cases with an equal grade group but a disagreement on Gleason score; and cases for which two pathologists agreed while the third had a maximum deviation of one grade group. Cases with a disagreement on malignancy were always flagged for a second read in round two. In the second round, each biopsy without consensus was regraded by the uropathologist whose score differed from the other two. Biopsies without consensus after round two were discussed in a consensus meeting.

Final test phase cases:

Pathologists from data-contributing centers provided clinical Gleason grades for each case. A single representative biopsy was selected by a pathology resident (LT) under the supervision of a subspecialized uropathologist (JvI).

A reader study involving 10 pathologists was handled through https://grand-challenge.org/ platform to grade the slides. The final grade was assigned based on the majority vote among the 10 pathologists.

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.

Experimental test phase cases:

Additional to the pathologist's initial examination, the Gleason scores of the other pathologists were appended anonymously during round two.

Final test phase cases:

The 10 pathologists that took part in the reader study were presented with each of the 113 slides containing a single prostate biopsy. For each slide, they were asked two mandatory questions and one optional question. about each slide:

- is there a tumor on this slide, and if so, what Gleason grade would you assign it? (mandatory)

- considering the tumor area, what's the percentage of the most common Gleason pattern you assigned to this tumor? For example: If you said Gleason 3+4, with the Gleason 3 pattern representing 70% of the tumor area, enter 70. If you said 3+3 or 4+4, then enter 100. If you said no tumor: enter 0. Please enter only numbers from 0 to 100. (mandatory)
- do you have any comments you want to share on this specific slide? (optional)

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.

Experimental test phase cases:

The three expert uropathologists (C.H.v.d.K., R.V. and H.v.B.) had between 18 and 28 years of clinical experience after residency (mean of 22 years).

Final test phase cases:

The 10 pathologists that took part in the reader study had on average 17.5 years of experience in general pathology. Nine out of ten pathologists who participated in this study were subspecialized in uropathology, with on average 14 years of experience in uropathology.

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.

Experimental test phase cases:

See the full protocol description in 23a.

Final test phase cases:

Final grade was assigned based on the majority vote among the 10 pathologists. In case of ties, the higher Gleason grade is selected.

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.

Experimental test phase cases:

To reduce the overall size of the dataset, the images were downsampled and exported at a pixel spacing of 0.48 micron per pixel (mpp). The images were exported as resolution pyramids with three levels representing downsampling factors of 1, 4 and 16 relative to the full resolution. Images were converted to TIFF format with JPEG compression and a quality setting of 70.

Final test phase cases:

When multiple biopsies (or different levels of the same biopsy) are present in a slide, a pathology resident selected a single representative biopsy, which was then cropped from each slide and saved at the closest level to a resolution of 0.5 mpp. Then, new slides were generated with 0.5, 1.0, 2.0, 4.0, 8.0 mpp. All slides were saved as three-channel RGB multiresolution slides in a standard pyramidal TIFF format.

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.

Experimental test phase cases:

The pathologists who contributed to the reference standards showed high pairwise agreement (0.926 quadratic kappa).

Final test phase cases:

The average quadratic kappa of pathologists against the majority vote of the rest is 0.878 and it ranges from 0.847 to 0.914. The average pairwise agreement of pathologists is 0.858 and it ranges from 0.777 to 0.916.

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

None.

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 algorithm will be assessed by Cohen's quadratic weighted kappa. This metric measures the agreements between two outcomes, and typically ranges from 0 (random agreement to 1 (complete agreement).

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

Cohen's quadratic weighted kappa effectively handles ordinal data, penalizes disagreements appropriately, adjusts for random agreement, and aligns with standard practices in pathology.

Ranking method(s)

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

We will follow a procedure similar to what was done in past challenges such as the TIGER and the DECATHLON challenges. For each task, we will compute performance metrics as described in section 26a of each task. To compute the performance rank across all tasks, we will 1) rank algorithms per task to determine the position of each algorithm in the task's leaderboard, 2) compute their mean position by summing up the position across all tasks and divide it by the number of tasks, 3) rank each algorithm across tasks based on its mean position. This will give an overview of how each algorithm performs overall across all tasks in UNICORN. Next to that, we will also create category-specific leaderboards considering for example only vision tasks, only language tasks, only vision-language tasks, only pathology tasks, only radiology tasks, and within each category, possibly subdivide into the type of task, including classification, detection, segmentation, regression, etc. These sub-leaderboards will give an overview of how each algorithm performs within a certain data domain and task category.

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

Missing results will be awarded the lowest rank. With this approach, we want to keep the entry barrier for new participants as low as possible so that they can submit their methods to only some tasks. Still, we will also evaluate complete submissions for information.

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

Given the implicit multi-task nature of UNICORN, and the heterogeneity of tasks, ranking methods to build a leaderboard is compact yet informative approach to conveniently summarize the overall performance of algorithms across a variety of tasks. Next to a general leaderboard, where all tasks are considered, we will also create sub-leaderboard for each category and for each family of tasks, to account for methods that only compete in some categories, and to focus on category-dependent performance of methods without "diluting" them among the variety of all tasks.

Statistical analyses

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

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

In case of acceptance, we will start developing a statistical analysis plan (SAP) with statisticians from the institutes and the network of the organizing groups.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 2: Vision - Classifying nodule malignancy in CT

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

See Task 1.

Cohorts

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

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

Participants included in lung screening programs, typically high risk participants with at least 20 years smoking history, aged between 50-80 years old.

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

Subjects of lung cancer screening trials, such as the National Lung Screening Trial (NLST). NLST enrolled 53,454 current or former heavy smokers ages 55 to 74, will be sequestered challenge set for experimental and final phase use. This challenge cohort includes 16,077 nodules, including 1249 malignant and 14,828 benign nodules. These nodules are from 1,199 CT scans diagnosed with cancer and from 8,984 CT scans diagnosed as non-cancer, which were collected between 2002 and 2004. If permission is granted from another screening trial, this second screening trial will be used as the sequestered challenge dataset, while the NLST data will be used as the public training set for the development phase.

In the case that the NLST dataset cannot be used for training, then example training cases will be provided from the public dataset: the Lung Image Database Consortium and the Image Database Resource Initiative (LIDC-IDRI), which includes both diagnostic and lung cancer thoracic CT scans from 1,010 subjects.

Imaging modality(ies)

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

Computed Tomography (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).

No further information given.

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

No further information given.

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.

Lungs and thorax shown in CT data.

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

The lesion target is the nodule. Main goal of the algorithm is the output of a risk score between 0 and 100 for a nodule candidate, in which scores closer to 0 indicate that the participant has lower risk of malignancy while scores closer to 100 indicate a high risk of malignancy.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

A range of manufactures were used, including GE Medical Systems LightSpeed scanner models, Philips Brilliance scanner models, Siemens Definition, Emotion, and Sensation scanner models, Toshiba Aquilion scanners.

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

Baseline, typically low-dose CT scan of the thorax for the challenge cohort.

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

The sequestered challenge data includes the NLST data that is acquired at 33 U.S. medical centers. The public training data is from the Lung Image Database Consortium and the Image Database Resource Initiative (LIDC-IDRI) which was acquired from seven academic centers and eight medical imaging companies.

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).

Not applicable.

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.

One case is a pair of baseline LDCT, the bounding box coordinates of the candidate nodule, and the label malignant or benign. A baseline LDCT refers to the first scan made of the patient in the screening trial. While there were follow up scans performed in the trial, however these will not be included in this challenge.

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

Development phase: during this phase, participants can use the LIDC-IDRI dataset, containing scans from 1010 subjects. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases including 50% benign and 50% malignant, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases. This data consists of a total of 200 cases and is split into two non-overlapping sets, both derived from the same distribution.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset will consist of 160 cases, ensuring a balanced distribution of 50% benign and 50% malignant cases. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on 40 cases.

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

See Task 1.

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

An equal class distribution was chosen to help the model optimize performance on both classes.

Annotation characteristics

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

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

The method was for radiologists or human experts to classify each nodule as either benign or malignant.

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.

For the training cases, a two-phase image annotation process was performed by four experienced thoracic radiologists. In the initial blinded-read phase, each radiologist independently reviewed each CT scan and marked lesions belonging to one of three categories ("nodule > or =3 mm," "nodule <3 mm," and "non-nodule > or =3 mm"). The radiologists also provided the following data at the nodule level: unknown, benign or non-malignant disease, a malignancy that is a primary cancer, or a metastatic lesion that is associated with an extra-thoracic primary malignancy, the annotations provided only include the benign or the malignant classes.

For the validation cases, an experienced radiologist inspected all images in participants diagnosed with lung cancer within the study period to retrospectively locate the malignant nodules, with the results of a histopathological standard test. All nodules with malignant morphologic features located in the tumor-bearing lobe were registered as lung cancer, while other nodules were excluded. The CT images that were not diagnosed with cancer were annotated by two medical students.

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

The five radiologists had varying experience, for the training set, and for the validation set, an experienced radiologist and two medical students trained by that experienced radiologist were the annotators.

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

For the training set, in the following unblinded-read phase, each radiologist independently reviewed their own marks along with the anonymized marks of three other radiologists to form a final opinion. This process aimed to identify all lung nodules in each CT scan as comprehensively as possible, without requiring a forced consensus.

For the validation set, there was no merging.

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.

CT volumes will be provided as mha files, converted to HU already, no pre-processing will be provided.

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.

When annotating the training data, there were five different centers and so there may have been varying standards. When annotating the validation data there is reduced opportunity for error since a histopathological standard was used. However, if a participant leaves the trial before the diagnosis is made, then the nodule is recorded as benign, and the radiologists are unaware of the diagnosis, but there is high adherence so this was not a frequent occurrence. Other sources of error include missing small nodules, especially those less than 5 mm in diameter, since they can be overlooked due to their subtle appearance and the low resolution of LDCT scans.

Mis-identifying structures such as blood vessels, calcifications, and overlapping anatomical features, or other diseases as nodules.

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

None.

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)

AUC (Area Under the Receiver Operator Curve (ROC)) will be used to assess the performance of the models to distinguish between benign and malignant cases based on their predicted risk scores for the challenge dataset.

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

AUC evaluates the performance of the predictive model across all possible classification thresholds. This is particularly useful in this application since the optimal threshold for classifying a nodule as benign or malignant varies depending on clinical considerations and the need to balance sensitivity and specificity. Furthermore, AUC represents the probability that a randomly chosen malignant nodule is correctly rated with a higher risk score than a randomly chosen benign nodule. This provides a measure of the model's overall discriminative power, which is crucial for assessing its effectiveness in differentiating between benign and malignant nodules.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 3: Vision - Predicting the time to biochemical recurrence in HE prostatectomies

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to

compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Segmentation
- Tracking

See Task 1.

Cohorts

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

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

Patients undergoing prostatectomy, patients followed up for the concentration of prostate-specific antigen (PSA) in their blood.

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

Subjects who underwent prostatectomy at Radboud University Medical Center. The cohort includes patients treated between 1992 and 2012. Data was collected by the Urology department and involves monitoring prostate-specific antigen (PSA) levels through ordered lab tests. These tests were ordered by treating physicians in the hospital, or outside by general practitioners. Generally, after one year of follow-up, patients were transferred to their general practitioner for monitoring.

Imaging modality(ies)

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

Hematoxylin-eosin (HE) stained histopathology whole-slides.

Context information

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

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

Tissue segmentation mask.

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

No further information given.

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.

Prostate shown in histopathology data.

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.

Accurate biochemical recurrence risk estimation of subjects.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

Slides were scanned into whole-slide images on a 3DHistech P1000 scanner, at 0.25 micron per pixel resolution.

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

HE stained tissue slides corresponding to one prostate cross-section were picked from the hospital archive containing the highest-grade tumor, based on pathology reports.

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

The data was acquired at Radboud University Medical Center, Nijmegen, The Netherlands.

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).

Not applicable.

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.

One case is a set of slides belonging to one subject. The number of slides ranges from one to nine per subject.

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

Development phase: during this phase, participants can use the publicly available dataset from the LEOPARD challenge, containing 508 cases. Additionally, participants may choose to use other public datasets, such as PLCO, which includes over 700 cases, and TCGA, which has more than 300 cases. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases. This data consists of a total of 149 cases and is split into two non-overlapping sets, both derived from the same distribution.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset consists of 49 cases. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the remaining 100 cases.

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

See Task 1.

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

Distribution of classes reflect real-world distribution. The distribution of time to biochemical recurrence (event = 1) or time the last follow-up (event = 0) is similar between the different phases.

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 annotation involved as we use biochemical recurrence as primary endpoint.

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.

Not applicable.

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.

Not applicable.

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

Not applicable.

Data pre-processing method(s)

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

When multiple slides are available for one subject, we segment tissue in each slide to strip out white background, and concatenate the resulting tissue pieces into a single – bigger – slide. For tissue segmentation, we use a deep learning model developed internally. In addition to segmenting tissue, we additionally segment artefacts (mostly pen marks) and remove artefacts from the tissue mask. To segment artefacts, we used the open source model available at https://github.com/RTLucassen/slidesegmenter. The same pre-processing was applied to the training, validation and test cases.

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.

Not applicable as we use the patient outcome as label.

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

None.

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)

Performance is evaluated according to the censored concordance index. The concordance index is defined as the proportion of all comparable pairs in which the predictions and outcomes are concordant. Two subjects are comparable if:
- both experienced an event (at different times)
or
- the one with a shorter observed survival time experienced an event, in which case the event-free patient "outlived" the other

A pair is not comparable if the subject experienced events at the same time. Concordance intuitively means that two samples were ordered correctly by the model: two patients are concordant if the one with a higher estimated risk score has a shorter actual survival time.

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

The censored c-index was chosen as it effectively handles censored data, which is common in survival analysis, and provides a measure of the model's ability to correctly rank the predicted survival times.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 4: Vision - Predicting slide-level tumor proportion score in NSCLC IHC-stained WSI

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to

compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Segmentation
- Tracking

See Task 1.

Cohorts

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

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

Nowadays, immunotherapy in the form of immune checkpoint inhibitors (ICI) is standard of care to treat non-small cell lung cancer patients with metastatic disease (i.e., in a metastatic setting, meaning that patients will not undergo surgery and will only receive systemic treatment), and is increasingly being administered in patients that have undergone surgical resection of (early-stage) non-small cell lung cancer (i.e., in an "adjuvant" setting, meaning treating patients with ICI after surgery) and in patients that will receive surgery (i.e., in a "neoadjuvant" setting, meaning treating patients with ICI before surgery).

In all these cases, a patient selection process is used in the clinic, as it has been shown that ICI does not work for all patients, and only the ones that show a (sufficiently) high expression of the PD-L1 protein in tumor cells found in the tumor biopsy (or resection) are elegible for the treatment. The PD-L1 assessment is done on histopathology slides of biopsies or resections, which are cut and stained with a PD-L1-binding antibody (e.g., 22C3, SP263, E1L3N, etc.), staining PD-L1-positive cells. After that, pathologists have to assess the Tumor Proportion Score (TPS), which is computed as the total amount of PD-L1-positive tumor cells divided by the total number of tumor cells in the histology slide, resulting in a number between 0 and 100%. Pre-defined threshold are then applied to decide how to treat the patient. In the case of the ICI drug known as KEYTRUDA (Pemrolizumab), clinically relevant thresholds are 1% and 50%: patients with TPS < 1% are not eligible; patients with 1% < TPS < 49% get Pembrolizumab in combination with chemotherapy; patients with TPS > 50% get Pembrolizumab as monotherapy.

Among these aforementioned clinical scenarios (i.e., metastatic, adjuvant and neoadjuvant treatment settings), the reality in the clinic is that metastatic disease represents the vast majority of NSCLC patients, mostly due to the lack of the implementation of lung cancer screening programs in most countries world-wide. Therefore, most histopathology tissue consists of biopsies from metastatic sites, containing metastatic lung cancer but not taken from lung tissue. However, in some cases, tissue samples of surgical resections or pre-operative biopsies are also present.

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

For this task, we included n=534 digital pathology whole-slide images of non-small cell lung cancer patients (one slide per patient) stained with PD-L1 immunohistochemical marker. Slides were retrieved by searching for the keywords "LUNG & PD-L1" in the pathology database of Radboud University Medical Center (Nijmegen, Netherlands) and included all available cases between 2016 and 2021. From the obtained list, we removed cases not containing non-small lung cancer. Tissue comprises both primary lung tumor and metastatic cases, therefore slides can contain both lung tissue and metastatic sites (e.g., liver, brain, bone, etc.). Most samples will contain biopsies, but several will contain surgical resections. In all cases, PD-L1 staining was done using either the 22C3 or the E1L3N monoclonal antibody. For each slide, experienced lung pathologists visually scored the "Tumor Proportion Score" (TPS), as the percentage of PD-L1 positive tumor cells over all tumor cells in the slide. These values have been extracted from pathology reports and will be used as targets in this task, resulting in a continuous value between 0 and 100%. Since relevant thresholds in clinical practice are 1% and 50%, we often found that TPS was reported as <1%, 1-49% or >50%.

From the total of 534 cases, we will include n=440 in the sequestered test set (to be split into n=220 for the experimental and n=220 for the final test set), n=64 (at least) as few-shots on platform, and release n=100 cases publicly as examples of "few shots", that can be used for fine-tuning. For the release of public few shots, we will use publicly available data used in the study of Vanguri et al. 2022 (https://www.nature.com/articles/s43018-022-00416-8); public dataset available at: https://www.synapse.org/{hashtag} !Synapse:syn26642505/wiki/615361. From the original dataset, containing n=234 slides, we will select at least n=100 slides (one slide per patient) with a distribution of biopsies/resections and primary/metastatic sites that is in line with the one of the final test sets.
In all cases, each slide will come together with a single number, the TPS at slide level.

Imaging modality(ies)

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

Digital Pathology Whole-Slide Image (WSI)

Context information

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

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

No further information given.

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

No further information given.

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.

Tissue samples of non-small cell lung cancer (NSCLC) taken from either the primary tumor (lung tissue) or the metastatic site (non-lung tissue). In all cases, NSCLC tumor cells are present in the histopathology tissue sample. Slides are scanned at 20X magnification, with a resolution of 0.5 micron/pixel.

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.

We focus on predicting the TPS from the entire slide. This is a clinically relevant target because TPS is the biomarker used in the clinic to select NSCLC patients to receive immunotherapy.

Main goal is the accurate prediction of TPS from the entire whole-slide image.

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)).

Predict TPS accurately with a high correlation with pathologists' scores, also in the presence of different antibodies used to stain the slide, which brings variability in the appearance of images. Therefore, key features are Accuracy, Robustness, Reliability.

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).

The data was acquired in a single center in the Netherlands (Radboud University Medical Center, Nijmegen) and scanned with a 3DHistech P1000 whole-slide image scanner at 40X magnification (0.25 micron/pixel resolution). Slides will be converted to standard TIFF format to maximize compatibility with whole-slide image libraries used by participants (some libraries might not support some versions of the MRXS files produced by the 3DHistech scanner) and saved at 20X magnification (0.5 micron/pixel).

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

Glass slides were scanned with a 3DHistech P1000 scanner (see details above).

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

The data used was generated at Radboud University Medical Center, Nijmegen, The Netherlands.

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).

Not applicable.

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.

One case is a pair of whole-slide image and slide-level TPS score.

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

Development phase: during this phase, participants can use the publicly available dataset from Vanguri et al. (https://www.nature.com/articles/s43018-022-00416-8, 2022), containing 234 cases. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases. This data consists of a total of 440 cases and is split into two non-overlapping sets, both derived from the same distribution.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset consists of 88 cases, representing 20% of the total 440 test cases. Participants can evaluate their models multiple times during this phase.

Final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the remaining 80% cases of the test data, i.e. 352 cases.

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

See Task 1.

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

Data comes from a real-world distribution as we extracted cases from clinical practice, therefore representing a real distribution of tissue samples and organs (primary and metastatic sites), and a real-world distribution of TPS values.

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.

Tumor Proportion Score was scored by board-certified lung pathologists with experience in PD-L1 scoring in NSCLC for clinical applications. No manual annotations are involved, and the TPS used in this task was derived from the clinical report.

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.

Not applicable.

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.

Not applicable.

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

Not applicable.

Data pre-processing method(s)

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

Slides scanned with the 3DHistech P1000 scanner produce files in the MIRAX format (.mrxs), which may not be compatible with some whole-slide image loaders. Therefore, we will convert all images to standard multiresolution TIFF files at 20X magnification (0.5 micron/pixel), which is the resolution typically used by state-of-the-art approaches to quantify TPS with computer algorithms, while reducing storage footprint compared to keeping the original 40X magnification.

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.

Visual estimation of the TPS is known to be subject to inter-observer variability. However, there is no additional clinical test that can be performed to obtain an objective quantification of the TPS, therefore the opinion of board-certified pathologists is the best we can obtain in this case. Note that the obtained TPS from pathology report is the result of a thorough diagnostic procedure where for example suboptimal PD-L1 stainings were rejected by pathologists and performed again in the pathology laboratories, also subject to quality checks using control tissue samples, minimizing the risk of poor image quality, while still keeping the implicit inter-observer variability.

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

None.

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)

Pearson correlation coefficient between predicted TPS and visual TPS from pathologists computed on the validation set(s) (n=220 samples).

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

The Pearson correlation coefficient was chosen because it is a standard metric for regression tasks, measuring the linear relationship between the algorithm's predicted Tumor Proportion Score (TPS) and the pathologists' visual TPS. This metric is crucial for ensuring that the algorithm's predictions align closely with expert assessments, which is vital for clinical decision-making in cancer treatment. By using Pearson correlation, we can effectively evaluate the algorithm's performance and its reliability in a clinical context.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 5: Vision - Detecting signet ring cells in HE-stained WSI of gastric cancer

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to

compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Segmentation
- Tracking

See Task 1.

Cohorts

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

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

Patients with hereditary diffuse gastric cancer (HDGC) who underwent tissue sampling for further pathological analysis.

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

Patients with hereditary diffuse gastric cancer (HDGC) who underwent tissue sampling for further pathological analysis.

Imaging modality(ies)

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

Hematoxylin-eosin (HE) stained histopathology whole-slides.

Context information

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

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

No further information given.

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

No further information given.

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.

Stomach sample shown in histopathology data.

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.

Accurate detection of cancer cells into one of 3 classes: signet ring cells, atypical signet ring cells, poorly differentiated cells.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

Slides were scanned into whole-slide images at a pixel resolution of 0.25 micron per pixel using a 3DHISTECH Panoramic 1000 scanner.

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

None.

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

The data was acquired at Radboud University Medical Center (Netherlands).

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).

Not applicable.

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.

One case is one region of interest belonging to one subject.

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

Development phase: during this phase, participants can use the publicly available dataset from the DigestPath 2019 challenge, containing 455 cases. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases. This data consists of a total of 458 cases and is split into two non-overlapping sets, both derived from the same distribution.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset consists of 92 cases, representing 20% of the total 458 test cases. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the remaining 80% cases of the test data, i.e. 366 cases

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

See Task 1.

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

Distribution of classes reflects real-world distribution.

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.

Regions of interest (ROIs) are selected in tumor areas within whole-slide images. Cells within the ROIs are manually annotated by placing a point in the cell and assigning this point one of three labels.

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.

A resident pathologist was asked to draw polygonal annotation boxes (ROIs) around all diffuse gastric cancer (DGC) tumor lesions, including pre-invasive lesions.

Student annotators were asked to annotate all tumor cells present in those ROIs by placing point annotations near the center of each tumor cell and labelling the cell according to the three target cell types (signet ring cell, atypical signet ring cell, poorly differentiated cell).

The pathologist was asked to check the student's cell-level annotations and make corrections / additions where necessary.

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.

ROI were selected by a pathologist resident. Individual cells were annotated by students. The resulting annotations were checked by a pathologist.

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

Not applicable.

Data pre-processing method(s)

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

All slides from which ROIs are extracted have all been saved as three-channel RGB multiresolution slides in a standard pyramidal TIFF format with starting resolution of 0.25 microns per pixel (mpp), and subsequent resolution levels increasing with a factor 2 (0.5, 1, 2, 4, 8, 16 ... mpp).

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.

Tumor cells can sometimes be difficult to distinguish from healthy tissue. The largest possible source of error is confusing the 3 different cell types. We have not estimated the magnitude of these errors.

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

None.

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)

Detection performance will be evaluated using the mean average precision (mAP) metric.
We will compute mAP across multiple IoU thresholds and classes:

- for each class, we compute the average AP across multiple IoU thresholds
- we average the AP scores across all classes to get the final mAP

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

The area under the precision-recall curve (AP) provides a balanced measure of both precision and recall, which is crucial in medical imaging tasks where both false positives (misidentifying non-cancerous cells) and false negatives (missing cancerous cells) can have significant clinical consequences. Averaging the AP scores per class ensures a comprehensive evaluation of the model's performance across all classes.

Using the mAP, we ensure that we evaluate both detection accuracy (via IoU thresholds) and classification accuracy (via the AP for each class).

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 6: Vision - Detecting clinically significant cancer in prostate MRI exams

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to

compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Segmentation
- Tracking

See Task 1.

Cohorts

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

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

Patients included in prostate cancer screening programs.

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

Subjects suspected of harboring clinically significant prostate cancer (ISUP greater or equal to 2 2), e.g. due to elevated levels of PSA, abnormal DRE findings. Patients are included only if they do not have a history of treatment or prior ISUP greater or equal to 2 findings.

Imaging modality(ies)

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

Biparametric magnetic resonance imaging (bpMRI) scans.

Context information

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

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

No further information given.

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

No further information given.

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.

Prostate shown in bpMRI scans.

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.

Accurate 3D detection of clinically significant cancerous (ISUP greater or equal to 2) lesions.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

The data was acquired using Siemens Healthineers or Philips Medical Systems-based scanners with surface coils.

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

Imaging consists of the following sequences:
- axial, sagittal and coronal T2-weighted imaging (T2W)
- axial high b-value (greater or equal to 1000 s/squared mm) diffusion-weighted imaging (DWI)
- axial apparent diffusion coefficient maps (ADC)

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

The data was acquired at three Dutch medical centers (Radboud University Medical Center, Ziekenhuis Groep Twente, Prostaat Centrum Noord-Nederland) and one Norwegian medical center (St. Olav's Hospital, Trondheim University Hospital).

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).

Not applicable.

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.

For the training data, one case corresponds to a patient with at least three imaging sequences: axial T2W, axial DWI and axial ADC scans. Additionally, they can also have either, both or none of these optional imaging sequences: sagittal and coronal T2W scans.

For the validation and testing data, one case corresponds to a patient with exactly five imaging sequences: axial, sagittal and coronal T2W; axial DWI and axial ADC scans.

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

Development phase: during this phase, participants can use the publicly available dataset from the PI-CAI challenge, containing 1500 cases. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases. Replicating the setup of the PI-CAI challenge, 100 cases will be used for the experimental test phase, and 1000 cases for the final test phase.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset will consist of 100 cases. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the final test set, consisting of 1000 cases.

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

See Task 1.

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

Distribution of classes reflects real-world distribution.

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.

Reference standard for training cases: cases are annotated with the same reference standard as used for the ProstateX challenge (Armato et al., 2018), i.e. histopathology (ISUP greater or equal to 2) positives, and histopathology (ISUP less or equal to 1) or MRI (PI-RADS less or equal to 2) negatives, without follow-up.

Reference standard for validation & test cases: the reference standard aims to utilize the best possible evidence to define the ground-truth for every case, i.e. histologically confirmed (ISUP greater or equal to 2) positives, and histopathology (ISUP less or equal to 1) or MRI (PI-RADS less or equal to 2) negatives, with follow-up (greater or equal to 3 years), as detailed below:

Cases with negative MRI (i.e. benign or carrying PI-RADS 1-2 lesions) generally do not undergo biopsies or radical prostatectomy and lack histologically-confirmed evidence for the absence of csPCa. It is likely that they do not harbor csPCa, but a small percentage (less than 1% at Radboud University Medical Center; Venderink et al., 2019) can still be missed. To alleviate this, up to 40% of the validation and testing cohorts is composed of multi-center patient data from the 4M cohort (van der Leest et al., 2019), where all patients with negative MRI had received systematic biopsies and subsequent grading was supervised by an expert uropathologist (greater than 25 years of experience). In other words, by using data from the 4M cohort, we are able to acquire histopathology evidence for a large fraction of the patient population, that is encountered, but typically not histologically confirmed during clinical routine.

Biopsies alone can still be prone to undersampling csPCa, especially in the case of smaller lesions (Srivastava et al., 2019). Hence, all negative cases (negative MRI and/or histopathology) in the validation and testing cohorts are confirmed with follow-up data (e.g. using the national Dutch Pathology Registry (PALGA) for centers based in The Netherlands). Negative patient exams found to be positive (via MRI or histopathology) in greater or equal to 3 years of follow-up, were inspected with an expert radiologist for retrospective signs of potentially missed csPCa. If the presence of csPCa can be definitively confirmed, they are included as positive cases; otherwise, they are excluded. Negative patient exams with 100% csPCa diagnosis-free survival (DFS) after at least 3 years, are included.

Annotations: voxel-level annotations for significant cancer lesions were delineated in NifTI format and resampled to the resolution of the T2-weighted imaging for validation and test datasets, as well as for a subset of the training dataset.

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.

Lesion delineations were created using the ITK-SNAP v3.80 software.

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.

Cases were annotated by one of 11 trained investigators, under the supervision of expert radiologists at each participating center (9-11 years of experience in reading prostate MRI). Annotations for every case were derived using biparametric MRI examinations, diagnostic reports (radiology, pathology) and whole-mount prostatectomy specimen (if applicable). All annotations underwent independent cross-examination and quality control at the central coordinating center of this study (Radboud University Medical Center), and were deferred back to site investigators when revisions were deemed necessary.

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

Not applicable.

Data pre-processing method(s)

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

No preprocessing carried out.

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.

Not applicable.

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

None.

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)

Detection performance will be evaluated using the area under the precision-recall curve or average precision (AP) metric.

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

The area under the precision-recall curve (AP) is particularly effective in evaluating performance on imbalanced datasets, where the number of cancerous instances is much smaller than non-cancerous ones. This metric focuses on the model's ability to maintain high precision and recall, which is critical in medical diagnostics to minimize false positives and ensure that significant cancer cases are correctly identified.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 7: Vision - Detecting nodules in thoracic CT

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

See Task 1.

Cohorts

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

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

Patients that receive a thorax CT part of clinical care and participants that receive a thorax CT as part of an organized screening program.

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

The challenge cohort includes a sequestered dataset from RadboudUMC: 100 scans from 100 patients during clinical routine that had no prior scans for cancer.

Example training cases will be provided from the public dataset from the Lung Image Database Consortium and the Image Database Resource Initiative (LIDC-IDRI), which includes both diagnostic and lung cancer thoracic CT scans from 1,010 subjects.

Imaging modality(ies)

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

Computed Tomography (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).

No further information given.

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

No further information given.

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.

Lungs and thorax shown in CT data

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

The lesion target is the nodule. The main goal of the algorithms is the accurate detection of pulmonary nodules in clinical routine chest CT scans and in screening.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

A range of manufactures were used, including GE Medical Systems LightSpeed scanner models, Philips Brilliance scanner models, Siemens Definition, Emotion, and Sensation scanner models, Toshiba Aquilion scanners.

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

For both datasets, non-contrast and contrast-enhanced CT were included. For the private set from RadboudUMC, a mean slice thickness of 0.67, and the axial plane resolution of 0.64 mm was used. From the LIDC-IDRI dataset for training, scans with slice thickness less than 2 mm are provided.

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

The validation cohort includes scans collected at RadboudUMC, and the training cohort Lung Image Database Consortium and the Image Database Resource Initiative (LIDC-IDRI), which was acquired from seven academic centers and eight medical imaging companies.

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).

Not applicable.

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.

One case is CT volume (mha file) with corresponding 3D bounding box coordinates of 0 to N nodules, this refers to one series, which could be either a contrast CT or a non-contrast CT. One case always refers to one series.

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

Development phase: during this phase, participants can use the LIDC-IDRI dataset, containing scans from 1010 subjects. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases including 50% benign and 50% malignant, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases. This data consists of a total of 100 cases and is split into two non-overlapping sets, both derived from the same distribution.

- experimental-test phase: After fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset consists of 20 cases, representing 20% of the total 100 test cases. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the remaining 80% cases of the test dataset, i.e. 80 cases.

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

See Task 1.

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

Distribution of classes reflects real-world distribution.

Annotation characteristics

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

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

The method was for radiologists to identify nodules.

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.

For the training cases, a two-phase image annotation process was performed by four experienced thoracic radiologists. In the initial blinded-read phase, each radiologist independently reviewed each CT scan and marked lesions belonging to one of three categories ("nodule > or =3 mm," "nodule <3 mm," and "non-nodule > or =3 mm"). The radiologists also provided the following data at the nodule level: unknown, benign or non-malignant disease, a malignancy that is a primary cancer, or a metastatic lesion that is associated with an extra-thoracic primary malignancy, the annotations provided only include the benign or the malignant classes.

For the validation dataset, a panel of five thoracic radiologists with 2 to 21 years of experience independently annotated and measured all intrapulmonary nodules in the test datasets using in-house software (CIRRUS Lung Screening, version 19.9.2, DIAG, Radboudumc, Nijmegen, The Netherlands). They manually identified nodules and used a semi-automatic segmentation algorithm for volumetric measurement, with manual corrections allowed. Radiologists also recorded the lobe location and nodule type (solid, part-solid, non-solid, perifissural, calcified). The task included annotating all intrapulmonary nodules, defined as any round or irregular density within the lung parenchyma with a diameter between 3 and 30 mm.

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

For the training set, the five radiologists had varying experience, and for the validation dataset, a panel of five thoracic radiologists with 2 to 21 years of experience were the annotators.

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

For the training set, in the following unblinded-read phase, each radiologist independently reviewed their own marks along with the anonymized marks of three other radiologists to form a final opinion. This process aimed to identify all lung nodules in each CT scan as comprehensively as possible, without requiring a forced consensus.

For the validation dataset, the nodule annotations from different radiologists and used a majority vote system, including only nodules detected by at least three radiologists as reference standards. Annotations of lesions smaller than 3 mm, larger than 30 mm, or identified by fewer than three radiologists were considered indeterminate and moved to an exclusion list.

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.

CT volumes will be provided as mha files, HU converted, no pre-processing will be provided.

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.

Common sources of error include missing small nodules, especially those less than 5 mm in diameter, since they can be overlooked due to their subtle appearance.

Mis-identifying structures such as blood vessels, calcifications, and overlapping anatomical features, or other diseases as nodules.

There is a large interobserver variability amongst radiologists about what constitutes a nodule, so a majority vote is used for the reference standard, possible error source is a nodule in which some think is it a nodule, but majority do not.

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

None.

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)

Sensitivity and False Positive Rate per scan.

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

For detecting pulmonary nodules, false positive and false negative rates refer to a lack of sensitivity best translate to clinical decision making. False positives can lead to unnecessary further testing, anxiety for patients, and additional healthcare costs. False negatives, on the other hand, can result in missed diagnoses of potentially malignant nodules, delaying treatment and reducing patient survival rates. By assessing FPR and FNR per scan, healthcare providers can better manage and mitigate risks associated with misdiagnoses. High FPRs can be addressed by refining the criteria for nodule detection to reduce unnecessary follow-ups, whereas high FNRs can prompt improvements in sensitivity to ensure critical nodules are not overlooked.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 8: Vision - Detecting mitotic figures in breast cancer HE-stained WSIs

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to

compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Segmentation
- Tracking

See Task 1.

Cohorts

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

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

Breast cancer patients from routine diagnostics that have undergone surgical resection of (early stage) tumor. Most tumor show the presence of proliferation of tumor cells, i.e., mitosis. The number of mitotic figures depend on the aggressiveness of the tumor, with more mitoses usually indicating more aggressive tumors, and therefore usually found in higher grade cancers, which typically need more aggressive treatment.

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

To benchmark models on this task, we will use data from the training dataset of Tellez et al., 2018 (doi:10.1109/TMI.2018.2820199), consisting of n=18 whole-slide images of surgical resection of triple-negative breast cancer (TNBC) patients (detailed as TNBC-H&E; see Table I in the paper). Since the objective of this task is to detect the location of mitotic figures in entire WSIs, we built a ground truth of mitoses in H&E; based on immunohistochemistry using the PHH3 marker, which stains for mitotic figures. In brief, for each tissue slide, we first stained the slide with H&E;, then we scan it, then we de-stained it and restained it with PHH3, scanned it again, and then ran WSI registration to align the H&E; and the PHH3 digital slides. Using the process described in Tellez et al., 2018, we obtained mitoses detections in PHH3 and transferred those to H&E; slides. In total, we obtained > 1,000 mitoses per slide, for a total of at least 22,000 mitoses. These will be used as ground truth in the sequestered test sets of this task.

For the public data that we release in the form of few-shots, we will use WSIs from the publicly available TCGA-BRCA archive, and visual prompts (i.e., patches containing mitoses and non-mitoses) from the recent MIDOG22 challenge.

Imaging modality(ies)

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

Digital Pathology Whole-Slide Image (WSI)

Context information

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

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

No further information given.

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

No further information given.

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.

Tissue samples of breast cancer taken via surgical resection. In all cases, breast cancer cells are present in the histopathology tissue sample. Slides are scanned at 40X magnification, with a resolution of 0.25 micron/pixel.

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.

We focus on detecting the mitotic figures in the entire slide. The number of mitosis and their location is a clinically relevant target because breast cancer grading is based on the mitotic count performed in the 2 mm^2 area with the highest density of mitoses (i.e., the "hot spot"). Knowing the location of each mitotic figure is therefore essential to later identify the hot spot.

Main goal is the accurate predict the (x,y) coordinate of each mitotic figure in the entire whole-slide image.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

The data was acquired from several medical centers in the central-west area of the Netherlands (data described in Tellez et al., 2018, PMID: 29994086). However, all slides were produced in a single center in the Netherlands (Radboud University Medical Center, Nijmegen). All slides were first stained with H&E; and scanned, then de-stained and re-stained using a PHH3 immunohistochemical marker, which highlights the presence of mitoses, and then scanned again, Therefore, two whole-slide images of the same tissue were obtained, co-registered, and used to make annotations of mitoses in H&E; slides at full-slide level. Note that this process has led to obtaining around 1,000 mitoses/slide. Each H&E; slide was then scanned with a 3DHistech P1000 whole-slide image scanner at 40X magnification (0.25 micron/pixel resolution). Slides will be converted to standard TIFF format to maximize compatibility with whole-slide image libraries used by participants (some libraries might not support some versions of the MRXS files produced by the 3DHistech scanner).

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

Glass slides were scanned with a 3DHistech P1000 scanner (see details above).

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

The data used was generated at Radboud University Medical Center, Nijmegen, The Netherlands.

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).

Not applicable.

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.

We make a distinction between what is considered as a "shot" and what is a considered as a "case" here.

A "shot" is a small portion of a digital pathology image, a "patch" that contains a cell at its center; the cell can be a mitosis or a non-mitosis and the patch will be labeled accordingly as containing a mitosis or not.

A "case" is what belongs to the test set and consists in a whole-slide paired with a list of (x,y) coordinates of mitotic figures present in the slide.

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

Development phase: during this phase, participants can use the publicly available datasets from the MIDOG21 and MIDOG22 challenges. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases. This data consists of a total of 18 cases and is split into two non-overlapping sets, both derived from the same distribution.

- experimental-test phase: After fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset consists of 4 cases, representing 20% of the total 18 test cases. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the remaining 80% cases of the test dataset, i.e. 14 cases.

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

See Task 1.

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

Data come from cases with TNBC, which is the most aggressive form of breast cancer, usually showing a high mitotic activity. Therefore, the test sets will contain a high number of mitotic figures. Publicly released data contains breast cancers belonging to different molecular subtypes, including TNBC.

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.

In the test set, mitotic figures were annotated using the immunohistochemical marker PHH3 as gold standard, to overcome the subjectivity of pathologists when annotating mitoses. In MIDOG21, MIDOG22 and TUPAC16, a group of pathologists was involved in making manual annotations, but no immunohistochemistry was used to confirm annotations.

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.

Not applicable.

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.

Not applicable.

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

Not applicable.

Data pre-processing method(s)

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

Slides scanned with the 3DHistech P1000 scanner produce files in the MIRAX format (.mrxs), which may not be compatible with some whole-slide image loaders. Therefore, we will convert all images to standard multiresolution TIFF files at 40X magnification (0.25 micron/pixel), which is the resolution typically used by state-of-the-art approaches as well as pathologists to detect mitotic figures in WSIs.

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.

Although PHH3 was used to generate the ground truth in the test set, and cure was taken in excluding staining artifacts in the immune staining, some detections derived from PHH3 might not have their exact correspondence of mitosis in H&E.; However, using PHH3 allows to obtain a much more reliable ground truth compared to annotations generated based on the opinion of pathologists, who are known to show inter-observer variability for this task. This is the case in data that will be publicly released from MIDOG21+22 and TUPAC16 challenges. Therefore, we should consider a regime of potentially noisy tuning (or few-shot) data, and a more objective ground truth in the test set.

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

None.

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)

F1-score computed as the harmonic mean of precision and recall. Precision and recall are computed based on true positive, false positive and false negative obtained as result of the detection output. We will define a "hit criterion" to measure whether an annotated mitosis is "hit" or not by a prediction. Following previous work, we will define a "hit" as a detection within a distance of 25 micron from the manually annotated (x,y) location of a mitosis.

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

F1-score is the standard performance metric used in mitosis detection since the AMIDA challenge in 2013 and has been adopted by all studies and challenges on this topic in the past decade. Using this same metric will also allow to compare methods in UNICORN with previous work in the literature.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 9: Vision - Segmenting ROIs in breast cancer HE-stained WSIs

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

See Task 1.

Cohorts

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

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

Patients with [Her2 positive (Her2+) and Triple Negative (TNBC)] breast cancer who underwent a core needle biopsy or resection.

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

Patients with Her2+ and TNBC breast cancer that underwent a core needle biopsy or resection of the breast tumor.

Imaging modality(ies)

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

Hematoxylin & Eosin-stained histopathology whole-slide images

Context information

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

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

For test data, a mask will be provided to guide the algorithm to the region of interest for assessment

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

No further information given.

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.

Breast tumor mass shown in a digital histopathology image.

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.

Accurate segmentation of three tissue types in breast slides: tumor, stroma, and rest.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Find breast segmentation algorithm for histopathological images with high accuracy on tumor and stroma segmentation. Therefore key features are Runtime, Accuracy, Reliability.

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).

The data was acquired in two medical centers in Europe.

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

None.

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.

Radboud University Medical Center (RUMC) (Nijmegen, Netherlands)

Jules Bordet Institut (JB) (Bruxelles, Belgium)

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).

Not applicable.

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.

One case is a single annotated region of interest within a Whole slide image

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

Development phase: during this phase, participants can use the publicly available datasets from the TIGER challenge, which consists of 151 whole-slide images with 453 annotated examples. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases. This data consists of a total of 279 regions of interest and is split into two non-overlapping sets.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset consists of 38 whole-slide images with 149 regions of interest manually annotated. A subset will be selected based on computational constraints. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the test set of the TIGER challenge, containing 26 whole-slide images with 130 regions of interest manually annotated.

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

See Task 1.

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

Slides were selected to have a variety of characteristics, including source (two European clinical centers); molecular subtypes, tissue type (11 resections, 8 biopsies); morphological subtype. In all slides, we performed manual annotations in pre-selected regions of interest. Furthermore, because we solely focused on regions of interest, the distribution of tissue types (58% resections, 42% biopsies) does not affect the evaluation.

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.

Manual annotations were made by five board-vertified breast pathologists from the tumor infiltrating lympchoytes working group, among the international experts in quantification of TILs

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.

Annotators were given instructions on how to use the annotation software (CIRRUS Pathology, via the grand-challenge web platform) with written instructions and with recorded videos. After that, we had meetings with the annotators, both all together and individually, to go over specific questions or problems with the interface, which were then solved and communicated to all the others.

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

All annotators are board certified breast pathologists with many years of experience

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

Annotations were not merged.

Data pre-processing method(s)

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

All slides have been converted to JPEG compressed TIF multiresolution images with maximum resolution set to 0.5 microns/px. This makes data in a homogenous format, coming from different scanners and different file formats.

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.

Inter- or intra-observer variability was not analyzed.

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

None.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

We will evaluate performance at segmentation of tumor and stroma, which are large regions of tissue in selected regions of interest. We will compute the Dice Similarity Coefficient for each class, and then average them to provide the contribution to the performance in the leaderboard.

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

Dice is a largely used metric for segmentation. Issues are known for this metric, but we believe that given the current setting, where large regions are evaluated, such as tumor and stroma, which often represent a large part of the region of interest, those issues have a minor effect.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 10: Vision - Segmenting lesions within ROIs in CT

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

See Task 1.

Cohorts

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

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

Patients with measurable lesions.

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

Patients who had undergone imaging exams where target lesions were measured.

Imaging modality(ies)

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

Computed Tomography (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).

Fully annotated 3D segmentation masks for lesions are provided.

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

No further information given.

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.

Lesion shown in chest-abdomen-pelvis CT.

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.

Segmentation of different types of lesions.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

The data was acquired using imaging scanners from Siemens, Philips, Toshiba, GE, and Canon, with varying distributions across the datasets.

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

None.

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

The dataset includes data from Radboudumc and Jeroen Bosch Ziekenhuis in the Netherlands.

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).

Not applicable.

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 CT sub-image showing a lesion and a segmentation mask.

The segmentation masks are fully-annotated in 3D.

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

Development phase: during this phase, participants can use the publicly available datasets from the ULS challenge (https://zenodo.org/records/10035161), consisting of 38693 cases. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases. This data consists of a total of 775 cases and is split into two non-overlapping sets.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset consists of 50 cases from the original ULS challenge validation set.
Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the 725 test cases of the ULS challenge.

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

See Task 1.

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

The dataset comprises 46.5% female patients.

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.

Preselection: Natural Language Processing tools analyzed radiological reports to identify patients with measured target lesions, detailing each lesion's location and size as recorded by radiologists.

Annotation process: Six biomedical annotators were tasked with locating and remeasuring lesions from the selected reports, which were then reviewed and adjusted by an experienced radiologist. The annotators used these verified measurements to segment the lesions in 3D. The test segmentations were again checked and corrected by the radiologists after the initial segmentation by the annotators

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.

Both the biomedical annotators and the radiologists were trained/familiarized with the software for the test set by segmenting a subset of the DeepLesion dataset in 3D first, with multiple review rounds to improve segmentation quality.

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

The radiologist that reviews and corrects the work of the annotators has over 10 years of professional experience.

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

Three annotators independently segmented each lesion, and the segmentation mask was determined by majority vote. This mask was reviewed, and corrected where necessary, by a radiologist.

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.

Imaging data is provided as volumes-of-interest (VOI's) of 256x, 256y, 128z voxels in the original scan spacing. Lesions with an axial diameter larger than this VOI were excluded from the test set. Where necessary, scans were padded with the minimum intensity value of the VOI minus one, allowing participants to establish where padding was added. The VOI's were sampled such that there always is a lesion voxel in the middle of the volume, simulating a click by a radiologist on that lesion. This voxel in the center of the volume was selected randomly from within the lesion mask.

Within each VOI, there is only one annotated lesion. Any masks representing nearby lesions not connected to the central lesion were removed.

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.

Not applicable.

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

None.

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)

For the average segmentation performance (SP), the Sørensen-Dice coefficient is used to evaluate the predicted 3D masks against the reference segmentation. Moreover, the symmetric mean absolute percentage error for the long- and short-axis measurements (LAE and SAE) is calculated.

A subset of lesions is included multiple times in the evaluation of the validation and test set, using randomly sampled lesion foreground voxels as the center locations. This introduces slight variations in the scan context for each cropped Volume of Interest (VOI). The consistency of model outputs is checked across these different click locations by comparing the Sørensen-Dice coefficient of the re-aligned segmentation masks. A model robust to click location variation should demonstrate a high average segmentation consistency score (SCS).

The final score (CS) is defined as:

CS = 0.8 * SP + 0.05 * LAE + 0.05 * SAE + 0.1 * SCS

where LAE and SAE are determined using the Symmetric Mean Absolute Percentage Error.

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

LAE and SAE were chosen since in the clinic 2D measurements are used to determine lesion response. SCS was measured to get an idea of the "intra-observer variability" of the model and due to different click points being inevitable when a radiologist would use such a model.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 11: Vision - Segmenting three anatomical structures in lumbar spine MRI

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to

compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Segmentation
- Tracking

See Task 1.

Cohorts

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

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

Patients with a history of low back pain

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

Patients with a history of low back pain were retrospectively collected from four different hospitals in the Netherlands acquired between January 2019 and March 2022.

Imaging modality(ies)

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

MRI with at least a sagittal T1 or a sagittal T2 sequence

Context information

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

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

Radiological gradings per intervertebral disc level are available for all patients. Graded was either the presence or the severity of the degenerative changes (Modic changes (type I, II or III), Upper and lower endplate changes/Schmorl nodes (binary), Spondylolisthesis (binary), Disc herniation (binary), Disc narrowing (binary), Disc bulging (binary), Pfirrman grade (grade 1 to 5)) that can be observed on MRI images and that have a known or suspected relation to low back pain.

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

39 patient studies (62% female) with (chronic) low back pain

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.

Lumbar spine shown in MRI data

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.

Segmentation of vertebrae, intervertebral discs, and spinal canal in lumbar MRI.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

Details for the test cases are not provided, but they come from the same distribution as the training and validation data. See overview file for further details on acquisition parameters: https://doi.org/10.5281/zenodo.10159290.

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

Sagittal lumbar MRI, T1 or T2 weighted (regular resolution, or high resolution generated using a SPACE sequence)

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.

Studies were collected from four different hospitals in the Netherlands, including one university medical center (UMC), two regional hospitals and one orthopedic hospital (data acquired between January 2019 and March 2022).

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).

All imaging acquisitions were performed by trained MRI radiographers.

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.

One case corresponds to one MRI sequence. For each series sequence, segmentation masks with vertebrae, intervertebral discs, and spinal canal are provided.

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

Development phase: during this phase, participants can use the publicly available datasets from the SPIDER challenge (https://zenodo.org/records/10159290), consisting of 218 cases. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases. This data consists of a total of 44 cases and is split into two non-overlapping sets.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset consists of 5 cases. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the remaining 39 test cases of the SPIDER challenge.

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

See Task 1.

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

The dataset comprises 62% female patients.

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.

All visible vertebrae (excluding the sacrum), intervertebral discs, and the spinal canal were manually segmented. The segmentation was performed using 3D Slicer version 5.0.3.

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.

Not applicable.

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

The segmentations were performed by a medical trainee who was trained and supervised by both a medical imaging expert and an experienced musculoskeletal radiologist.

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

Not applicable each case was annotated by one/same rater

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.

No preprocessing carried out.

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.

Unknown.

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

None.

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 segmentation performance will be evaluated using two metrics:

(1) The Dice coefficient (measured in 3D) and
(2) the average absolute surface distance (ASD)

Both metrics will be calculated separately for all individual structures and were averaged per anatomical structure (vertebrae, IVDs, or spinal canal). Additionally, the average Dice coefficient and average ASD per MRI sequence (T1 vs. T2) will be calculated for each anatomical structure. To ensure the Dice score and ASD are not influenced by labelling differences, the individual structures of the reference segmentation are matched to the structured in the predicted segmentation based on the largest found overlap.

In addition, a completeness classification error will be used. It is defined as the binary cross-entropy between the true label and the predicted probability. The completeness classification performance will be determined by the percentage of accurate predictions, as well as the average number of false positives and false negatives.

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

The Dice coefficient measures the volume overlap, and the average absolute surface distance (ASD) provides an indication of the segmentation accuracy along the surface of all structures.

The completeness classification error provides insight into the classification aspect of the segmentation task, where the goal is to determine the presence or absence of a structure.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 12: Language - Predicting histopathology sample origin

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

See Task 1.

Cohorts

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

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

The target cohort for this language task consists of pathology reports obtained from patients undergoing diagnostic or therapeutic procedures in clinical settings. These reports contain unstructured text descriptions of pathology slides, including details about the cancer type or tissue type present in each sample. The final biomedical application aims to accurately predict the origin of each pathology slide solely based on the information provided in the pathology reports. The motivation for this application lies in the need to automate the prediction of sample origin from unstructured pathology reports, facilitating more efficient and accurate analysis of pathology data.

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

The challenge cohort includes patients suspected of having non-small cell lung cancer between 1 January 2016 and 31 December 2022 at Radboud University Medical Center, Nijmegen.

Imaging modality(ies)

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

Reports refer to histopathology whole-slides.

Context information

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

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

No further information given.

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

No further information given.

Target entity(ies)

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

The pathology reports derived from patients undergoing diagnostic or therapeutic procedures in clinical settings.

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

The algorithm target is to extract relevant diagnostic information from the pathology reports. Specifically, for each report, the algorithm needs to determine the sample origin (i.e. lung, liver, brain, etc.).

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

The data was acquired in a single center in the Netherlands (Radboud University Medical Center, Nijmegen) and the images to which the pathology reports refer were acquired using a 3DHISTECH P1000 scanner at 40X magnification (0.25 micron/pixel resolution).

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

None.

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

The images to which the pathology reports refer were acquired at Radboud University Medical Center, Nijmegen, The Netherlands.

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).

Not applicable.

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.

One case refers to a pathology report.

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

Development phase: during this phase, at least 32 cases from the challenge cohort will be released. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset will consist of up to 100 cases, with the final number depending on available computational resources. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the 297 test cases of the DRAGON challenge.

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

See Task 1.

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

Distribution of classes reflects real-world distribution.

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.

Annotations were manually made by trained student assistants.

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.

Not applicable.

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.

Not applicable.

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

Not applicable.

Data pre-processing method(s)

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

The reports are all anonymized during preprocessing by an in-house developed algorithm called (HIPS) in which personally identifiable information such as dates, person names, report identifiers, places, telephone numbers, hospitals, patient ages, and full names of medical practitioners are replaced by sensible but fake surrogates. This is to protect the patient and doctor's privacy. The reports provided to the participating teams therefore do not include any sensitive information.

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.

Incorrect annotations are possible however since annotations were performed by trained student assistants we estimate the magnitude of these errors low.

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

None.

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 labels for this task are categorical. The frequency of the labels is unbalanced. The labels are not ordinal. Therefore, we use the Unweighted Kappa to evaluate model performance.

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

The Unweighted Kappa coefficient was selected as the evaluation metric due to its suitability for assessing model performance in tasks with categorical labels. The choice of the Unweighted Kappa coefficient as the evaluation metric for this task is justified by its suitability for assessing model performance in scenarios with categorical labels, particularly when dealing with unbalanced label frequencies and non-ordinal categories.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 13: Language - Classifying pulmonary nodule presence

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

See Task 1.

Cohorts

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

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

Patients with a clinically ordered chest CT.

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

The cohort includes 1000 chest CT reports from patients from two Dutch hospitals within the period of 1 January 2008 to 31 December 2019 [1].

1. Hendrix W, Rutten M, Hendrix N, van Ginneken B, Schaefer-Prokop C, Scholten ET, et al. Trends in the incidence of pulmonary nodules in chest computed tomography: 10-year results from two Dutch hospitals. Eur Radiol [Internet]. 2023 Jun 20 [cited 2023 Aug 17]; Available from: https://doi.org/10.1007/s00330-023-09826-3

Imaging modality(ies)

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

Not applicable.

Context information

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

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

No further information given.

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

No further information given.

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.

Unstructured radiology reports from lung cancer (screening) patients undergoing a CT.

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

The target of the algorithm is to predict whether a pulmonary nodule is described in the report.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

Not applicable

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

The radiology reports are manually written by radiologists.

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

The data was acquired at Radboud University Medical Center, Nijmegen, The Netherlands and Jeroen Bosch Ziekenhuis, s-Hertogenbosch, The Netherlands. 1000 randomly sampled reports from 1000 unique patients from hospital A (n=500) and hospital B (n=500).

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).

Not applicable.

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.

One case is single report containing information about a lung CT

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

Development phase: during this phase, at least 32 cases from the challenge cohort will be released. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset will consist of up to 100 cases, with the final number depending on available computational resources. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the 200 test cases of the DRAGON challenge.

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

See Task 1.

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

Distribution of classes reflects real-world distribution.

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.

These reports were annotated by a medical student under supervision of an experienced radiologist (26 years of experience). Each report was given a binary label indicating whether a pulmonary nodule was reported or not.

The testset is an independent dataset of 200 randomly sampled radiology reports from 200 unique patients from hospital A (n = 100) and hospital B (n = 100). There was no overlap between patients from the development and test set. The test set was annotated by an experienced radiologist (32 years of experience) according to the same annotation procedure as applied for the development set.

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.

Not applicable.

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.

Subjects annotated the reports in a previous study (see https://doi.org/10.1007/s00330-023-09826-3)

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

Not applicable each case was annotated by one/same rater

Data pre-processing method(s)

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

The reports are all anonymized during preprocessing by an in-house developed algorithm called (HIPS) in which personally identifiable information such as dates, person names, report identifiers, places, telephone numbers, hospitals, patient ages, and full names of medical practitioners are replaced by sensible but fake surrogates. This is to protect the patient and doctor's privacy. The reports provided to the participating teams therefore do not include any sensitive information.

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.

Since they are annotated by a single person it is possible that reports are assigned incorrect labels, however since annotations were under the supervision of an experienced radiologist, we estimate the magnitude of these errors low.

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

Unknown.

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 labels for this task are binary. The frequency of the labels is unbalanced. We use the Area Under the Receiver Operating Characteristic Curve (AUC) to evaluate model performance.

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

The Area Under the Receiver Operating Characteristic Curve (AUC) was chosen because it effectively handles imbalanced datasets common in biomedical applications, evaluating model performance across all classification thresholds.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

- combining algorithms via ensembling,
- inter-algorithm variability,

- common problems/biases of the submitted methods, or
- ranking variability.

See previous question.

TASK 14: Language - Classifying kidney abnormality

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

See Task 1.

Cohorts

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

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

For the final application, reports will be collected from patients that underwent a computed tomography (CT) scan of the kidney. The final application extracts information about the CT image corresponding to the report. For example, discrimination of healthy kidney tissue from renal cell carcinoma and the identification of additional abnormalities such as cysts and kidney stones. These extracted anomalies reduce manual labeling time of reports and provide labels to implement a classifier trained to detect these abnormalities in CT images.

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

Subjects that underwent a computed tomography (CT) scan of the kidney.

Imaging modality(ies)

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

Not applicable.

Context information

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

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

No further information given.

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

No further information given.

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.

Any form of significant kidney abnormalities that are mentioned in radiology reports.

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

The algorithm needs to classify whether any form of significant kidney abnormality is mentioned in the radiology report.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

Not applicable

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

The radiology reports are manually written by radiologists.

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

The data was acquired at Radboud University Medical Center, Nijmegen, The Netherlands.

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).

Not applicable.

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.

One case is a single radiology report.

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

Development phase: during this phase, at least 32 cases from the challenge cohort will be released. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset will consist of up to 100 cases, with the final number depending on available computational resources. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the 183 test cases of the DRAGON challenge.

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

See Task 1.

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

Cases were selected to form a balanced dataset with 300 normal and 300 abnormal cases by randomly sampling from a internal dataset of about 300.000 cases and performing the classification, or opting to skip a case when classification was unclear. After 300 normal cases had been reached, additional normal cases were also discarded until 300 abnormal cases were found. Abnormalities include renal cell carcinoma, angiomyolipoma, cysts, kidney stones, conjoined kidneys, cases with partial or full nephrectomy, and several other rare abnormalities.

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.

Cases were manually analyzed by a PhD candidate and labeled as either normal or abnormal.

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.

Not applicable.

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.

Cases were analyzed by a native Dutch speaking PhD candidate.

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

Not applicable each case was annotated by one/same rater

Data pre-processing method(s)

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

The reports are all anonymized during preprocessing by an in-house developed algorithm called (HIPS) in which personally identifiable information such as dates, person names, report identifiers, places, telephone numbers, hospitals, patient ages, and full names of medical practitioners are replaced by sensible but fake surrogates. This is to protect the patient and doctor's privacy. The reports provided to the participating teams therefore do not include any sensitive information.

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.

Cases were analyzed by a native Dutch speaker

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

None.

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)

Area Under the Receiver Operating Characteristic Curve (AUC).

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

The labels for this task are binary. The frequency of the labels is balanced. Therefore, we use the AUC to evaluate model performance.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 15: Language - Predicting Hip Kellgren-Lawrence score

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

See Task 1.

Cohorts

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

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

Patients who underwent hip X-ray with up to three X-ray images, without bilateral hip replacement and over 30 years old.

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

Patients who underwent hip X-ray at Radboudumc between 1st January 2002 and 31st December 2022. Studies with up to three X-ray images, without bilateral hip replacement and with more than one keyword of 'joint gap', 'degeneration', and 'coxartr' in the report were included. Cases from patients younger than 30 at the time of the X-ray were excluded, given that hip osteoarthritis is not common at this age and thus the reports refer mostly to other aspects, such as hip dysplasia.

Imaging modality(ies)

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

Not applicable.

Context information

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

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

No further information given.

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

No further information given.

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.

Radiology reports for hip osteoarthritis.

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

The algorithm should predict the Kellgren-Lawrence score based on the radiology report.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

Not applicable

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

The radiology reports are manually written by radiologists.

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

The data was acquired at Radboud University Medical Center, Nijmegen, The Netherlands.

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).

Not applicable.

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 is one radiology report.

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

Development phase: during this phase, at least 32 cases from the challenge cohort will be released. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset will consist of up to 100 cases, with the final number depending on available computational resources. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the 172 test cases of the DRAGON challenge.

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

See Task 1.

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

Distribution of classes reflects real-world distribution.

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.

All reports for the test set were manually annotated by a trained investigator (M.N.P.). The reports that were difficult to annotate were revised by a radiologist (M.J.C.M.R.).

The reports for model development were annotated with ChatGPT (with GPT-4 on 13 June 2023) inputting the prompt below*, which was revised by the radiologist too. From the development data, 3 cases were excluded, due to missing annotations (1), or having a duplicate report (2). This included 4803 cases (4127 patients) for development.

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.

*Prompt inputted to GPT-4 to peform the annotations:

You are a helpful Dutch medical assistant. You will help the user assess hip osteoarthritis in radiology reports using the Kellgren Lawrence scale. Below is a brief description of the Kellgren Lawrence scale:

- 0: no radiographic core features of osteoarthritis, no joint gap narrowing, no bone abnormalities. Keywords: no coxarthrosis

- 1: possible joint gap narrowing, possible osteophyte formation. Keywords: no obvious coxarthrosis

- 2: obvious osteophyte formation, possible joint gap narrowing. Keywords: minimal coxarthrosis, incipient coxarthrosis, mild coxarthrosis, minor coxarthrosis

- 3: moderate osteophyte formation, marked joint gap narrowing and some sclerosis, possible degenerative bone defects. Keywords: moderate coxarthrosis

- 4: large definite osteophytes, definite joint gap narrowing and severe sclerosis, definite degenerative bone defects. Keywords: end-stage coxarthrosis, severe coxarthrosis, substantial coxarthrosis, strong coxarthrosis, obvious degeneration, obvious osteophyte formation

- not applicable: there is not enough information in the report to give an assessment - prosthesis: the patient has a hip prosthesis. Keywords: THP, THR, total hip replacement, total hip prosthesis

Always answer in the following form:

left:

right:

Always choose from one of the following options for the left and right grade: [0, 1, 2, 3, 4, not applicable, prosthesis].

Keep your motivation short but complete and clear. For example:

left: 3

right: 4

left: prosthesis

right: 4

Below are some more important things to watch out for:

'bdz' stands for bilateral.

Cox osteoarthritis and pelvic osteoarthritis are synonyms.

If a single condition is listed for a grade, that is enough to get that grade. 'Severe sclerosis' is therefore grade 4, even though joint gap narrowing is not mentioned.

Joint gap narrowing always results in grade 3 or above, 2 or below is not possible.

Cox osteoarthritis always results in grade 2 or higher, 1 or lower is not possible.

Sclerosis always results in grade 3 or higher, 2 or lower is not possible.

Always choose the highest possible grade.

'Moderate to severe coxarthrosis' is therefore grade 4.

This is the report to review:

report

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.

Not applicable.

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

Not applicable.

Data pre-processing method(s)

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

The reports are all anonymized during preprocessing by an in-house developed algorithm called (HIPS) in which personally identifiable information such as dates, person names, report identifiers, places, telephone numbers, hospitals, patient ages, and full names of medical practitioners are replaced by sensible but fake surrogates. This is to protect the patient and doctor's privacy. The reports provided to the participating teams therefore do not include any sensitive information.

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.

All reports for the test set were manually annotated by a trained investigator, and the report that were difficult to annotate were revised by a radiologist.

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

None.

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)

Unweighted Kappa. Given there are two labels per report (left and right), the predictions for the left and right hip are concatenated. The unweighted kappa is computed on the concatenated tensor.

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

The labels for this dataset range from 0 up to and including 4 for the Kellgren-Lawrence score, and additionally include "hip prosthesis" and "not determinable". The frequency of the labels is imbalanced. The Kellgren-Lawrence score is ordinal, but "hip prosthesis" and "not determinable" are not. Therefore, we use the Unweighted Kappa to evaluate model performance.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 16: Language - Classifying colon histopathology diagnosis

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

See Task 1.

Cohorts

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

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

For the final applications, report will be collected from patients presenting for cancer diagnosis at the hospital, encompassing a variety of diagnostic procedures such as biopsies, resection excisions, diagnostic imaging, and lab tests. The increasing number of patients due to cancer screening programs has significantly increased pathologist workloads, and visual inspection of diagnostic images and samples is time-consuming and susceptible to inter-observer variability. AI-based computer-aided diagnosis can enhance histological analysis, reducing time, effort, and subjective bias. By pre-screening tissue samples and diagnostic images to exclude normal or benign cases, AI can assist in the accurate classification of cancers into clinically relevant categories, improving diagnostic efficiency and reducing pathologists' workloads.

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

The challenge cohort includes pathology reports from patients who received a histopathology diagnosis of colon biopsies between 1 January 2000 and 31 December 2009 at Radboud University Medical Center. For patients with multiple visits in this period, the first visit was selected.

Imaging modality(ies)

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

Reports refer to histopathology whole-slides (hematoxylin-eosin stained)

Context information

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

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

No further information given.

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

No further information given.

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.

Pathology reports derived from colon biopsy samples.

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

The algorithm target is to extract relevant diagnostic information from the pathology reports. Specifically, for each block of the report, the algorithm needs to determine whether the specimen was obtained from 1) biopsy or polypectomy, and whether the pathologist rated the specimen as 2) hyperplastic polyps, 3) low-grade dysplasia (lgd), 4) high-grade dysplasia (hgd) 5) cancer, 6) serrated polyps, or 7) non-informative (ni). Each of these seven properties is binary, and multiple can be present per block.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

Not applicable

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

Not applicable.

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

The data was acquired at Radboud University Medical Center, Nijmegen, The Netherlands.

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).

Not applicable.

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.

One case refers to a single pathology report.

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

Development phase: during this phase, at least 32 cases from the challenge cohort will be released. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset will consist of up to 100 cases, with the final number depending on available computational resources. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the 1177 test cases of the DRAGON challenge

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

See Task 1.

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

Distribution of classes reflects real-world distribution.

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.

Annotations were manually made by native Dutch analysists under the supervision of two gastrointestinal pathologists.

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.

Not applicable.

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

The two gastrointestinal pathologists respectively have 13 and 21 years of experience.

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

Not applicable.

Data pre-processing method(s)

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

The reports are all anonymized during preprocessing by an in-house developed algorithm called (HIPS) in which personally identifiable information such as dates, person names, report identifiers, places, telephone numbers, hospitals, patient ages, and full names of medical practitioners are replaced by sensible but fake surrogates. This is to protect the patient and doctor's privacy. The reports provided to the participating teams therefore do not include any sensitive information.

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.

No further possible error sources besides manual annotation mistakes which is mitigated by the experience level of the annotators.

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

None.

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 performance will be assessed by calculating the Area Under the Receiver Operating Characteristic Curve (AUC) for each class individually. The overall model performance is defined as the average of the individual AUC values.

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

AUC is particularly suitable as it is less affected by class imbalance compared to accuracy. By capturing both sensitivity and specificity, AUC reflects the algorithm's ability to correctly identify true positives and true negatives, which is essential for reliable clinical diagnosis.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

- combining algorithms via ensembling,

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

See previous question.

TASK 17: Language - Predicting lesion size measurements

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

See Task 1.

Cohorts

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

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

For the final applications, reports will originate from patients presenting for cancer diagnosis at the hospital, encompassing a variety of diagnostic procedures such as biopsies, resection excisions, diagnostic imaging, and lab tests.

If a lesion is present the reports may include size measurement of this lesion which is an important risk factor for malignancy. And should therefore be extracted from the reports. However, manually extracting the size from the radiology reports is time-consuming, limiting the size of the datasets. This offers a clear opportunity for AI to automatically extract the size of the lesion of unstructured diagnostic reports.

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

The challenge cohort is a unique combination of 3 individual datasets:

Dataset is created by Hendrix et al. [1] and includes chest CT reports from two Dutch hospitals (period 1 January 2008 to 31 December 2019) for nodule trend analysis. All pulmonary nodules included have a maximum diameter of 30 mm, regardless of morphology, type (i.e., solid, part-solid, non-solid, calcified, or perifissural), and malignancy status.

Dataset: 253 radiology reports of the Radboudumc and 144 reports of the JBZ hospitals in the Netherlands were selected from archives containing thorax-abdomen CT scans conducted between 2000-2020.

Dataset: 2343 consecutive patients undergoing pancreatic and liver CECT scans between 1 January 2011 and 31 December 2022 at Radboud University Medical Center were included. Additionally, all CT scans for 1076 consecutive patients who underwent histopathology analysis of pancreas specimens or distant metastasis of pancreatic cancer between 1 January 2006 and 31 December 2022 at Radboud University Medical Center were retrospectively collected. This resulted in 3419 selected patients. This is the dataset used also in the PANORAMA study [2].

[1]. Hendrix W, Rutten M, Hendrix N, van Ginneken B, Schaefer-Prokop C, Scholten ET, et al. Trends in the incidence of pulmonary nodules in chest computed tomography: 10-year results from two Dutch hospitals. Eur Radiol [Internet]. 2023 Jun 20 [cited 2023 Aug 17]; Available from: https://doi.org/10.1007/s00330-023-09826-3

[2] Alves N, Schuurmans M, Rutkowski D, Yakar D, Haldorsen I, Liedenbaum M, et al. The PANORAMA Study Protocol: Pancreatic Cancer Diagnosis - Radiologists Meet AI [Internet]. Zenodo; 2024 Jan [cited 2024 Mar 18]. Available from: https://zenodo.org/doi/10.5281/zenodo.10599559

Imaging modality(ies)

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

Computed Tomography (CT) and Contrast-Enhanced Computed Tomography (CECT)

Context information

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

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

No further information given.

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

No further information given.

Target entity(ies)

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

The reports describe findings based on lungs and thorax shown in CT data as well as pancreas and liver shown in CECT scans.

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.

A radiology report will be sourced from one of three datasets. The algorithm target, the specific lesion type and size that needs to be predicted, will thus depend on the dataset from which the report originates. For each report the source of the report and the prediction objective will be provided.

For the 3 datasets the task is defined as follows:

- dataset 1: Predict the largest reported diameter of pulmonary nodules described in the radiology report. When multiple sizes are described for a single lesion (e.g., the short and long axis), the size for that lesion should be averaged (e.g., 9 mm for a lesion of size 1.0 x 0.8 cm).
- dataset 2: Predict the size in mm for all RECIST target lesions. For lymph nodes the short axis should be reported.
- dataset 3: Predict the reported diameter of the PDAC in mm, as described in the radiology report. When multiple axes are measured (e.g., "12 x 34 mm"), report the longest axis size (i.e., 34). When a range is given for one axis (e.g., "1 to 2 mm"), provide the average (i.e., 1.5).

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

Not applicable

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

The radiology reports are manually written by radiologists.

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

The reports are from medical centers in the Netherlands, including: Radboud University Medical Center (Radboudumc) and Jeroen Bosch Ziekenhuis.

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).

Not applicable.

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.

One case refers to a single radiology report.

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

Development phase: during this phase, at least 32 cases from the challenge cohort will be released. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases. As this task includes three categories of reports, we will release at least 32 shots per category.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset will consist of up to 100 cases, with the final number depending on available computational resources. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the 298 test cases of the DRAGON challenge.

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

See Task 1.

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

Distribution of test cases from the 3 individual datasets: 32 (pulmonary node cohort) + 119 (RECIST lesions cohort) + 147 (pancreatic ductal adenocarcinoma cohort). The individual datasets all follow real-world distribution.

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.

For all individual cohorts, the annotations were performed in a previous study:

Pancreatic and cohort: cases were manually annotated by one of three trained investigators (N.A., M.S., I.M.E.S) as part of the PANORAMA study (for more information see: https://zenodo.org/doi/10.5281/zenodo.10599559f)

Pulmonary nodule cohort: Each report was given a label indicating (1) whether a pulmonary nodule was reported and (2) the diameter of the largest reported nodule if available (otherwise, a missing value was registered). All pulmonary nodules were included with a maximum (For more information see: https://doi.org/10.1007/s00330-023-09826-3)

The reports were manually annotated by a trained investigator to determine lesion sizes.

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

See previous question.

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.

- Pulmonary nodules cohort: The cases were annotated by an experienced radiologist (32 years of experience).

- The RECIST lesion cohort: The reports were manually annotated by a trained investigator (M.J.J.d.G) under the supervision of a radiologist (E.Th.S.) to determine the lesion sizes.

- Pancreatic and liver cohort: The cases were manually annotated by one of three trained investigators (N.A., M.S., I.M.E.S.).

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

Not applicable each case was annotated by one/same rater

Data pre-processing method(s)

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

The reports are all anonymized during preprocessing by an in-house developed algorithm called (HIPS) in which personally identifiable information such as dates, person names, report identifiers, places, telephone numbers, hospitals, patient ages, and full names of medical practitioners are replaced by sensible but fake surrogates. This is to protect the patient and doctor's privacy. The reports provided to the participating teams therefore do not include any sensitive information.

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.

No further possible error sources besides manual annotation mistakes which is mitigated by the experience level of the annotators.

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

None.

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 labels of this tasks are floating values ranging from:

1-30mm for lesions in the pulmonary nodules (95% of values between 2 and 28 mm)

4- 166mm for RECIST lesions (95% of values between 6 and 94 mm)

6- 130mm (95% of values between 12 and 83 mm) for the diameter in PDAC.

The ranges are based on development data. The Robust Symmetric Mean Absolute Percentage Error (RSMAPE) with an epsilon of 4mm is used to evaluate model performance. This evaluates the lesion size prediction in a diagnostically relevant manner.

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

The metric RSMAPE provides a balanced assessment by combining relative errors with a margin of tolerance, making it diagnostically relevant. This metric ensures that small errors within the acceptable clinical range do not overly penalize the model, reflecting practical significance in medical diagnostics.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 18: Language - Predicting prostate volume, PSA, and PSA density

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

See Task 1.

Cohorts

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

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

The target cohort consists of radiology reports obtained from patients who have a suspicion for prostate cancer (elevated PSA levels, and/or abnormal digital rectal examination findings, and/or lower urinary tract symptoms)

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

Patients who had a suspicion for prostate cancer (elevated PSA levels, and/or abnormal digital rectal examination findings, and/or lower urinary tract symptoms) between 1 January 2021 and 31 December 2022 at University Medical Center Groningen, between 1 January 2012 and 17 February 2023 at Antoni van Leeuwenhoek Ziekenhuis and all cases with a radiology report from Radboudumc.

Imaging modality(ies)

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

Radiology MRI reports

Context information

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

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

No further information given.

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

No further information given.

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.

Radiology MRI reports for suspected prostate cancer

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.

1. Predict the PSA level based on the radiology report. When a range is given (e.g., "PSA: 4-5"), provide the average (i.e., 4.5).

2. Predict the prostate volume (in cm3) based on the radiology report. When dimensions are given (e.g., "3 x 4 x 5 mm"), we used the ellipsoid formula to calculate the volume, with l, w, and h the spatial dimensions in mm.

3. Predict the PSA density, which is either directly described in the radiology report, or needs to be calculated based on the PSA level and prostate volume. All required information is provided in the report, and the PSA density is related to the PSA and prostate volume as: PSA density = PSA / prostate volume.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

Not applicable

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

Multiparametric prostate MRI protocol

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.

University Medical Center Groningen (UMCG), Antoni van Leeuwenhoek Ziekenhuis and Radboudumc (RUMC)

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).

All imaging acquisitions and reporting were performed by trained MRI radiographers.

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.

One case refers to one report.

With labels for the PSA level, provided as floating point values ranging from 0 to 870 ng/mL

2. With labels for the prostate volume, provided as floating point values ranging from 4.0 to 470cm3

3. With labels for the PSA density, provided as floating point values ranging from 0 to 171 ng/mL2

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

Development phase: during this phase, at least 32 cases from the challenge cohort will be released. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases. As this task includes three categories of reports, we will release at least 32 shots per category.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset will consist of up to 100 cases, with the final number depending on available computational resources. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the 6236 test cases of the DRAGON challenge (2046 cases for PSA level, 2170 cases for prostate volume, 2020 cases for PSA density).

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

See Task 1.

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

Distribution of classes reflects real-world distribution.

1. The labels for extracting the PSA level are floating point values ranging from 0 to 870 ng/mL, with 95% of values between 2 and 35 ng/mL (ranges based on the development data).

2. The labels for prostate volume extraction are floating point values ranging from 4.0 to 470 cm3, with 95% of values between 24 and 181 cm3 (ranges based on the development data).

3. The labels for the PSA density extraction are floating point values ranging from 0 to 171 ng/mL2, with 95% of values between 0.03 and 0.6 ng/mL2 (ranges are based on the development data).

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.

Findings were manually annotated by trained investigators.

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.

Not applicable.

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.

Not applicable.

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

Not applicable.

Data pre-processing method(s)

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

The reports are all anonymized during preprocessing by an in-house developed algorithm called (HIPS) in which personally identifiable information such as dates, person names, report identifiers, places, telephone numbers, hospitals, patient ages, and full names of medical practitioners are replaced by sensible but fake surrogates. This is to protect the patient and doctor's privacy. The reports provided to the participating teams therefore do not include any sensitive information.

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.

No further possible error sources besides manual annotation mistakes which is mitigated by the experience level of the annotators.

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

None.

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)

To evaluate model performance, we used the Robust Symmetric Mean Absolute Percentage Error (RSMAPE):
- with an epsilon of 0.04 ng/mL for PSA level
- with an epsilon of 4 cm3 for prostate volume
- with an epsilon of 0.04 ng/mL2 for PSA density

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

The RSMAPE with specific epsilon values was chosen due to its robustness against outliers and balanced treatment of over- and under-estimations. In biomedical applications, precise measurements are critical. The chosen epsilon values (4 cm^3 for prostate volume, 0.04 ng/mL for PSA level, and 0.04 ng/mL^2 for PSA density) reflect typical clinical measurement variability and ensure the metrics are sensitive to clinically meaningful discrepancies.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 19: Language - Anonymizing report

SUMMARY

Abstract

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

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

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

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

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

Examples:

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

See Task 1.

Challenge venue and platform

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

See Task 1.

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

See Task 1.

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

See Task 1.

Participation policies

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

See Task 1.

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

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

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

See Task 1.

e) Define the policy for result announcement.

Examples:

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

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

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

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

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

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

Examples:

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

See Task 1.

Code availability

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

See Task 1.

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

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

See Task 1.

Cohorts

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

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

The target cohort for the anonymization task consists of medical records such as pathology and radiology reports or electronic health records (EHRs) from real patients treated as a hospital.

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

The challenge cohort is part of the DRAGON challenge and sourced from two hospitals (Radboudumc, Antoni van Leeuwenhoek Ziekenhuis) and includes reports from multiple studies:

Radboudumc:

a. 1600 reports (1574 patients) were sampled randomly from patients with a thorax-abdomen or thorax CT scan at Radboudumc between 1 January 2008 and 31 December 2019. 65 duplicate reports were excluded to prevent overlap of reports between the dataset splits, resulting in 1535 included cases (1514 patients).

b. Basal cell carcinoma pathology reports. 13183 cases (6555 patients) were selected from patients who underwent a skin cancer biopsy between 1 January 2003 and 31 December 2022 at Radboudumc with the diagnostic code for BCC. 756 cases (685 patients) were sampled.

c. Prostate pathology reports. 12,437 consecutive cases (10,402 patients) were selected from patients who underwent histopathological evaluation of prostate tissue between 1 January 2012 and 12 February 2024 at Radboudumc. 413 cases (412 patients) were randomly sampled.

d. Prostate pathology procedure reports. 1293 consecutive cases (1180 patients) were selected from patients who underwent an MR-guided prostate biopsy between 1 January 2014 and 31 December 2020 at Radboudumc. 515 cases (498 patients) were randomly sampled. 6 duplicate reports were excluded to prevent overlap of reports between the dataset splits, resulting in 509 included cases (492 patients).

e. Lung pathology reports. 1019 consecutive cases (902 patients) were selected from patients suspected of having non-small cell lung cancer between 1 January 2016 and 31 December 2022 at Radboudumc. 284 cases (273 patients) were randomly sampled.

2. Antoni van Leeuwenhoek Ziekenhuis.

a. 7581 consecutive cases (5017 patients) were selected from patients who received a prostate histopathological evaluation between 1 January 1995 and 17 February 2023 at Antoni van Leeuwenhoek Ziekenhuis. 462 cases (441 patients) were randomly sampled.

b. 9766 consecutive cases (7841 patients) were selected from patients who had a suspicion of prostate cancer (elevated PSA levels, and/or abnormal digital rectal examination findings, and/or lower urinary tract symptoms) between 1 January 2012 and 17 February 2023 at Antoni van Leeuwenhoek Ziekenhuis. 273 cases (269 patients) were randomly sampled. Additionally, 250 consecutive cases (250 patients) between 14 October 2015 and 3 March 2016 were selected. 9 cases were selected in both sets and only included once. This included 514 cases (498 patients)

Imaging modality(ies)

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

Reports refer to histopathology whole-slides (hematoxylin-eosin stained) slides,

computed tomography (CT) images and immunohistochemistry (IHC) slides

Context information

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

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

No further information given.

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

No further information given.

Target entity(ies)

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

The radiology reports are derived from lungs and thorax shown in CT scans, prostate biopsies, skin biopsies and lung biopsies.

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.

Anonymization is an important first step in processing medical data to be used in further research. To protect patient and doctor's privacy, personally identifiable information (PII) should be removed and/or replaced with non-identifiable information.

The goal of the large language model (LLM) is to accurately predict and annotate words that constitute personally identifiable information (PII) in reports using BIO tags. The types of PII to be identified include dates (e.g., "12-3-2004"), person names (e.g., "Joeran Bosma" or "JSB"), report identifiers (e.g., "T12-345678"), places (e.g., "Radboudumc, Nijmegen" or "France"), personally identifying numbers (e.g., "06-12345678" or "012345"), clinical trial names (e.g., "4M trial"), hospital accreditation numbers (e.g., "M123"), times (e.g., "12:34"), and patient ages (e.g., "86"). Each word is annotated with BIO tags (B for the beginning of an entity, I for inside an entity, and O for outside any entity) to classify whether it belongs to a piece of identifiable information.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

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

Not applicable

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

Not applicable

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

The challenge cohort is sourced from two hospitals: Radboud University Medical Center (Radboudumc) in Nijmegen and Antoni van Leeuwenhoek Ziekenhuis in Amsterdam.

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).

Not applicable.

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.

One case is a radiology/pathology report.

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

Development phase: during this phase, at least 32 cases from the challenge cohort will be released. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset will consist of up to 100 cases, with the final number depending on available computational resources. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the 1037 test cases of the DRAGON challenge.

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

See Task 1.

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

Distribution of classes reflects real-world distribution.

Annotation characteristics

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

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

The reports were annotated semi-automatically, by processing the reports with a tool developed in-house (i.e., a rule-based system). Specific for this task, the automatically annotated reports are manually verified and corrected by 3 trained investigators. We used the annotation software Doccano for manually verifying and correcting annotations. Doccano is an open-source text annotation tool for humans. It provides annotation features for text classification and sequence labeling. The manually verified annotations are then converted into BIO tags.

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.

There was a guideline using examples for the annotation process provided. The classes to be annotated were dates (e.g., "12-3-2004") annotated as <DATUUM>, person names (e.g., "Joeran Bosma" or "JSB") annotated as <PERSOON>, report identifiers (such as T-numbers used in pathology, e.g., "T12-345678") annotated as <RAPPORT_ID>, places (hospitals and patient-specific places, e.g., "Radboudumc, Nijmegen" or "France") annotated as <PLAATS>, personally identifying numbers (such as telephone numbers, patient identifiers, e.g. "06-12345678" or "012345") annotated as <TELEFONNUMMER>, clinical trial names (e.g., "4M trial") annotated as <STUDIE_NAAM>, hospital accreditation numbers (e.g., "M123") annotated as <ACCREDATIENUMMER>, times (e.g., "12:34") annotated as <TIJD> and patient ages (e.g., "86") annotated as <LEEFTIJD>

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

The annotators were all Dutch speaking trained investigators.

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

Not applicable.

Data pre-processing method(s)

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

The reports are all anonymized during preprocessing by an in-house developed algorithm called (HIPS) in which personally identifiable information such as dates, person names, report identifiers, places, telephone numbers, hospitals, patient ages, and full names of medical practitioners are replaced by sensible but fake surrogates. This is to protect the patient and doctor's privacy. The reports provided to the participating teams therefore do not include any sensitive information.

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.

It is possible that during the manual annotation process some PII is missed by the annotators. This is mitigated by applying a baseline anonymization model from the DRAGON challenge on the annotated test cases and manually checking those cases where the predictions of the model and the original annotations differ.

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

None.

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)

For anonymization, a few-shot prompting approach is used where the LLM is prompted to tag PII based on specific categories. The model's performance is assessed by comparing the tagged entities to the ground truth, ensuring that each predicted entity matches the annotated PII data exactly in terms of type and position. The macro F1 metric is used to evaluate the predictions. This metric only counts predictions as correct when they perfectly overlap with ground truth annotations.

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

F1 score is a good metric for this task as it considers both precision and recall. It provides a balanced evaluation of the model's performance on identifying PII, ensuring that both the correctness of predicted entities and the coverage of actual entities are taken into account. This gives a more accurate estimate of the model's ability to anonymize sensitive information.

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.

See Task 1.

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

See Task 1.

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

See Task 1.

Statistical analyses

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.

See Task 1.

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

See previous question.

Further analyses

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

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

See previous question.

TASK 20: Vision-Language - Generating caption from WSI

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

See Task 1.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

See Task 1.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

See Task 1.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

See Task 1.

c) Provide the URL for the challenge website (if any).

See Task 1.

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

See Task 1.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

See Task 1.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

See Task 1.

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

See Task 1.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation

- Tracking

See Task 1.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

The target cohort consists of patients presenting for cancer diagnosis at the hospital and undergo procedures such as biopsies or resections. These patients' tissue samples are analyzed through whole-slide imaging (WSI), and detailed pathology reports are generated based on the visual inspection of these images. The final biomedical application aims to generate accurate diagnostic captions from WSI, assisting pathologists in making reliable diagnoses. The motivation for this application is to alleviate the increasing workload on pathologists due to rising cancer screening programs, enhance diagnostic accuracy, and reduce inter-observer variability through AI-assisted analysis of digital pathology images.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

The challenge cohort includes pathology reports and whole-slide images from two distinct datasets: the colon dataset and the skin dataset. The colon dataset includes data from patients who received a histopathology diagnosis of colon biopsies between 1 January 2000 and 31 December 2009 at Radboudumc. For patients with multiple visits in this period, the first visit was selected. The skin dataset comprises pathology reports and images from patients who underwent a skin biopsy between 1 January 2003 and 31 December 2022 at Radboudumc with the diagnostic code for BCC.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Histopathology whole-slides (hematoxylin-eosin stained)

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

No further information given.

b) ... to the patient in general (e.g. sex, medical history).

No further information given.

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

The pathology reports derived from skin or colon biopsy samples.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

The algorithm target is to generate a caption that encapsulates the relevant clinical diagnosis from whole-slide images (WSI) of skin and colon tissue samples. This involves analyzing the WSI to identify key histopathological features indicative of various diagnoses, such as basal cell carcinoma, hyperplastic polyps, dysplasia, and other relevant conditions. The algorithm must then succinctly summarize these findings in a coherent and clinically meaningful caption.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

The data was acquired in a single center in the Netherlands (Radboud University Medical Center, Nijmegen) and the images to which the pathology reports refer were acquired using a 3DHISTECH P1000 scanner at 40X magnification (0.25 micron/pixel resolution).

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Not applicable

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

The images were acquired at Radboud University Medical Center, Nijmegen, The Netherlands.

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).

Not applicable.

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.

One case refers to a pair of whole-slide images and captions.

b) State the total number of training, validation and test cases.

Development phase: given that this task includes cases from two separate cohorts, during this phase, at least 32 cases per cohort will be released. The goal of this data is for participants to develop and debug their fine-tuning pipeline locally on their own machines. While the primary focus is on refining and testing these pipelines, participants may also use this data for model training if they wish. However, the emphasis remains on preparing their pipeline for the later stages of the challenge.

On platform (data is sequestered): after the development phase, participants will transition to the platform, where they will be provided with a limited number of additional labeled examples, called "hidden shots." These hidden shots, consisting of at least 32 cases per cohort, are sequestered. Participants can use these cases to fine-tune their models on the platform before running inference on the test sets. They won't have direct access to these cases.

Test phases: a sequestered dataset will be used for testing during the experimental and final test phases. The total challenge dataset comprises more than 1000 cases per cohort. A subset will be selected for testing based on computational constraints. This selected subset will be divided into two non-overlapping sets, both derived from the same distribution.

- experimental-test phase: after fine-tuning the models with the hidden shots on the platform, participants will evaluate their models on the experimental-test dataset. This dataset will consist of roughly 20% of the total test set, ensuring a balanced distribution of 50% cases from the skin cohort and 50% cases from the colon cohort. Participants can evaluate their models multiple times during this phase.

- final-test phase: at the end of the challenge, participants will be allowed to select only one model, which will be run once on the remaining 80% cases of the test dataset. The exact number of cases will depend on computational constraints.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

See Task 1.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Not applicable.

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.

Images are always connected to a single report.

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.

Not applicable.

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.

Not applicable.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Not applicable.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

The reports are all anonymized during preprocessing by an in-house developed algorithm called (HIPS) in which personally identifiable information such as dates, person names, report identifiers, places, telephone numbers, hospitals, patient ages, and full names of medical practitioners are replaced by sensible but fake surrogates. This is to protect the patient and doctor's privacy. The reports provided to the participating teams therefore do not include any sensitive information.

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.

Not applicable.

b) In an analogous manner, describe and quantify other relevant sources of error.

None.

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 chosen metric to assess the property of algorithm accuracy is ROUGE (Recall-Oriented Understudy for Gisting Evaluation). ROUGE evaluates the similarity between generated captions and reference captions, reflecting the algorithm's ability to produce clinically relevant diagnostic summaries.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

ROUGE was selected as the metric for assessing algorithm performance due to its relevance in evaluating the quality of generated text, including diagnostic captions. ROUGE provides a quantitative measure of the similarity between the generated captions and reference captions, allowing for an objective assessment of the algorithm's performance.

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.

See Task 1.

b) Describe the method(s) used to manage submissions with missing results on test cases.

See Task 1.

c) Justify why the described ranking scheme(s) was/were used.

See Task 1.

Statistical analyses

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.

See Task 1.

b) Justify why the described statistical method(s) was/were used.

See previous question.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

See previous question.

TASK 21: Vision - optional tasks

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

These optional tasks are not yet fully defined due to various constraints and ongoing discussions. We are currently negotiating Data Transfer Agreements (DTAs) and exploring partnerships with additional medical centers. While these collaborators have not yet fully confirmed their commitment due to time constraints, we are optimistic about their participation, which would significantly enrich the challenge.

The optional tasks are designed to enhance the diversity and scope of our challenge. They will introduce a wealth of previously unreleased private data, offering unique opportunities for innovation and discovery. However, due to the pending DTAs, we are unable to provide specific details about the data at this moment. As a result, these tasks remain optional for now.

Should our proposal be accepted, we anticipate that these new collaborators will join and contribute not only to the existing tasks but also to the development of these optional tasks. Their involvement will bring additional depth and breadth to the challenge, facilitating groundbreaking advancements in the field.

In summary, while the current proposal focuses on well-defined tasks, the inclusion of these optional tasks, contingent on finalizing DTAs and securing partner commitments, promises to significantly enhance the challenge by leveraging diverse and previously unavailable data.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

See Task 1.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

See Task 1.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

See Task 1.

c) Provide the URL for the challenge website (if any).

See Task 1.

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

See Task 1.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

See Task 1.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)

- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

See Task 1.

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

See Task 1.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

See Task 1.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

1. Predict slide level tumor pleomorphism score
The target cohort are patients undergoing breast cancer surgical resection procedures,resulting in the collection of digital pathology whole-slide images.

2. Detect and classify cells in PD-L1 staining
The target cohort for the task comprises patients diagnosed with non-small cell lung cancer (NSCLC) whose pathology slides have been stained with the PD-L1 immunohistochemical marker. These patients may have undergone either biopsy or surgical resection procedures, resulting in the collection of digital pathology whole-slide images.

3. Bosniak classification in kidney CT (5 classes)
The target cohort consists of patients who have been identified with cystic renal masses on CT scans

4. Pancreatic cancer yes / no
The target cohort are patients undergoing a CT scan of the pancreas.

5. Detect cysts in kidney CT
The target cohort consists of patients undergoing a CT scan of the kidneys.

6. Classify and Detect incidental findings in lung CT

The target cohort is participants included in lung screening programs, typically high risk participants with at least 20 years smoking history, aged between 50 – 80 years old.

7. Detect lymph nodes in CT
The target cohort comprises patients who are suspected of having or have been diagnosed with conditions that affect the lymph nodes.

8. Pulmonary lobe segmentation in CT scans
The target cohort comprises patients undergoing a chest CT, including patients with a suspicion of pulmonary diseases such as COVID-19.

9. Segmentation of primary Gross Tumour Volume for esophageal cancer radiotherapy
The target cohort comprises patients with confirmed diagnosis of esophageal cancer (either squamous cell or adenocarcinoma) undergoing chemo-radiotherapy. No prior surgical resection. Has radiotherapy treatment planning CT in the supine position.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

1. Predict slide level tumor pleomorphism score
The challenge cohort comprises 128 regions of interest (ROI) of breast cancer surgical resections with hematoxylin and eosin (H&E;) and the corresponding ROI-level tumor pleomorphism score. The ROIs and labels are private data and sourced from the Radboud university medical center. The ROI-level tumor pleomorphism scores are floating values ranging from 1 to 3.

2. Detect and classify cells in PD-L1 staining
The challenge cohort comprises 152 digital pathology whole-slide images of non-small cell lung cancer patients (one slide per patient) stained with PD-L1 immunohistochemical marker. Patients originated from one of three clinical centers: 73 patients from Radboud University Medical Center (Nijmegen, Netherlands), 54 patients from the Sacro Cuore Don Calabria Hospital of Negrar (Verona, Italy) and 25 patients from the Netherlands Cancer Institute-Antoni van Leeuwenhoek Hospital (Amsterdam, Netherlands). Most samples will contain biopsies, but several will contain surgical resections.

3. Bosniak classification in kidney CT (5 classes)
The challenge cohort comprises patients from Radboud University Medical Center. The exact number of cases and details remain to be finalized.

4. Pancreatic cancer yes / no
The challenge cohort comprises CT scans of the pancreas collected from five centers. It includes 2,238 public training cases and 1,000 private testing cases.

5. Detect cysts in kidney CT
The challenge cohort comprises patients from Radboud University Medical Center. The exact number of cases and details remain to be finalized.

6. Classify and Detect incidental findings in lung CT
The challenge cohort comprises a private dataset of about 100 NLST participants, both those diagnosed with and without lung cancer.

7. Detect lymph nodes in CT
The challenge cohort comprises 454 axillary lymph nodes annotated on 90 CT scans from 72 patients from Radboud University Medical Center, Nijmegen, the Netherlands.

8. Pulmonary lobe segmentation in CT scans
The challenge cohort comprises 470 chest CT scans of patients with a suspicion of COVID-19 from Radboud University Medical Center, Nijmegen, the Netherlands.

9. Segmentation of primary Gross Tumour Volume for esophageal cancer radiotherapy
The challenge cohort consists of the single phase 3D treatment planning CT scan images of 420 esophageal cancer patients treated at MAASTRO radiotherapy clinic will be provided, for validation of the model, there will be 370 unseen subjects from this cohort.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Not applicable.

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

No further information given.

b) ... to the patient in general (e.g. sex, medical history).

No further information given.

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.

Not applicable

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.

1. Predict slide level tumor pleomorphism score: The target of the algorithm is to predict the ROI-pleomorphism score in the whole slide image.

2. Pancreas yes/no: The target of the algorithm is the accurate
classification of the presence/absence of pancreatic cancer and detecting its location in the CT scan if present.

3. Detect and classify cells in PD-L1 staining: The algorithm target is to detect three cell types within NSCLC WSIs (PD-L1 positive tumor cells, PD-L1 negative tumor cells and an 'other cells' type) and predict the coordinates of all cells within a patch.

4. Bosniak classification in kidney CT (5 classes): The target of the algorithm is to provide a Bosniak classification score for cystic renal mass.

5. Classify and Detect Incidental Findings: The target of the algorithm is to provide bounding box coordinates and class labels for 9 specific incidental findings for lung cancer screening recommended by the ERS/ESTS/ESTRO/ESR/ESTI/EFOMP statement on management of incidental findings from low dose CT screening for lung cancer. These incidental findings refer to: aortic valve disease, bronchiectasis, consolidation, coronary artery calcification, emphysema, interstitial lung abnormalities, mediastinal lymph nodes, mediastinal mass, thyroid abnormalities

6. Detect cysts in kidney CT: The algorithm target is to detect cysts in abdominal CT scans. Task remains to be finalized.

7. Pulmonary lobe segmentation in CT scans: The algorithm target is to segment lobes in CT scans from patients with diseased lungs.

8. Detect lymph nodes in CT: The algorithm target is to detect lymph nodes in CT scans and predict the coordinates and diameter.

9. Segmentation of primary Gross Tumour Volume for esophageal cancer radiotherapy: Geometrically accurate segmentation of the GTV relative to the physician reference.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Not applicable

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Not applicable

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.

Not applicable

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).

Not applicable

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.

Not applicable

b) State the total number of training, validation and test cases.

Not applicable

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

See Task 1.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Not applicable.

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.

Not applicable.

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.

Not applicable.

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.

Not applicable.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Not applicable.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Not applicable.

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.

Not applicable.

b) In an analogous manner, describe and quantify other relevant sources of error.

Not applicable.

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)

Not applicable.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

Not applicable.

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.

See Task 1.

b) Describe the method(s) used to manage submissions with missing results on test cases.

See Task 1.

c) Justify why the described ranking scheme(s) was/were used.

See Task 1.

Statistical analyses

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.

See Task 1.

b) Justify why the described statistical method(s) was/were used.

See previous question.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

See previous question.

TASK 22: Language - optional tasks

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

These optional tasks are not yet fully defined due to various constraints and ongoing discussions. We are currently negotiating Data Transfer Agreements (DTAs) and exploring partnerships with additional medical centers. While these collaborators have not yet fully confirmed their commitment due to time constraints, we are optimistic about their participation, which would significantly enrich the challenge.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

See Task 1.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

See Task 1.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

See Task 1.

c) Provide the URL for the challenge website (if any).

See Task 1.

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

See Task 1.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

See Task 1.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

See Task 1.

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

See Task 1.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration

- Retrieval
- Segmentation
- Tracking

See Task 1.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

1. Presence/absence of incidental findings: The target cohort consists of patients that underwent a CT abdomen-thorax for a clinical problem.

2. Conclusion generation from report: For the final applications, reports will be collected from patients presenting for cancer diagnosis at the hospital, encompassing a variety of diagnostic procedures such as biopsies, resection excisions, diagnostic imaging, and lab tests.

3. Report translation: same as above

4. Report restructuring: same as above

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

1. Presence/absence of incidental findings
The challenge cohort consists of about 100 CT Thorax abdomen scans from RadboudUMC, labeled by 3 expert annotators.

2. Conclusion generation from report: The challenge cohort includes pathology reports from two distinct datasets: the colon dataset and the skin dataset. The colon dataset includes data from patients who received a histopathology diagnosis of colon biopsies between 1 January 2000 and 31 December 2009 at Radboudumc. For patients with multiple visits in this period, the first visit was selected. The skin dataset comprises pathology reports from patients who underwent a skin biopsy between 1 January 2003 and 31 December 2022 at Radboudumc with the diagnostic code for BCC.

3. Report translation

4. Report restructuring
The challenge cohort consists of a collection of older reports from Radboud University Medical Center and other medical institutions, which currently lack a standardized structure. These reports encompass a wide range of samples from multiple organs. The standardized structure includes dividing the raw text into sections on Clinical

Question, Microscopy, Macroscopy, and Conclusions. A common structure across different medical institutions can enhance the consistency and comparability of the reports, thereby fostering better collaboration and knowledge sharing among medical professionals

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Not applicable.

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

No further information given.

b) ... to the patient in general (e.g. sex, medical history).

No further information given.

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.

Not applicable

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.

1. Conclusion generation from report: The algorithm target is to generate the conclusion section of the report based on the microscopy section's content. This involves synthesizing the findings from the microscopic examination into a coherent conclusion that summarizes the diagnosis and relevant observations.

2. Report translation: The algorithm target is to automatically translate Dutch pathology reports into another language while preserving all the information and nuances present in the original text. The goal is to ensure that the translated reports accurately convey the same medical information and diagnoses as the original Dutch reports, without altering the meaning or context of the text.

3. Report restructuring: The algorithm target for report restructuring is to convert unstructured pathology reports into structured reports that follow a predefined format, typically including sections for macroscopy findings, microscopy findings, and conclusion. The algorithm aims to extract relevant information from the unstructured text and organize it into structured sections, facilitating easier interpretation and analysis of the pathology reports by clinicians and researchers.

4. Labeling incidental findings: The algorithm target for report labeling is the label 0 for the absence, and 1 for the presence of an incidental finding in the report, and will be evaluated using AUC and Cohen Kappa agreement with

human annotators

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Not applicable

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Not applicable

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.

Not applicable

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).

Not applicable

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.

Not applicable

b) State the total number of training, validation and test cases.

Not applicable

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

See Task 1.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Not applicable.

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.

Not applicable.

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.

Not applicable.

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.

Not applicable.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Not applicable.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Not applicable.

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.

Not applicable.

b) In an analogous manner, describe and quantify other relevant sources of error.

Not applicable.

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)

Not applicable.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

Not applicable.

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.

See Task 1.

b) Describe the method(s) used to manage submissions with missing results on test cases.

See Task 1.

c) Justify why the described ranking scheme(s) was/were used.

See Task 1.

Statistical analyses

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.

See Task 1.

b) Justify why the described statistical method(s) was/were used.

See previous question.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

See previous question.

TASK 23: Vision-Language - optional tasks

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

These optional tasks are not yet fully defined due to various constraints and ongoing discussions. We are currently negotiating Data Transfer Agreements (DTAs) and exploring partnerships with additional medical centers. While these collaborators have not yet fully confirmed their commitment due to time constraints, we are optimistic about their participation, which would significantly enrich the challenge.

Keywords

List the primary keywords that characterize the task.

See Task 1.

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

See Task 1.

b) Provide information on the primary contact person.

See Task 1.

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

See Task 1.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

See Task 1.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

See Task 1.

c) Provide the URL for the challenge website (if any).

See Task 1.

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

See Task 1.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

See Task 1.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

See Task 1.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

See Task 1.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

See Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

See Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

See Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See Task 1.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

See Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

See Task 1.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

See Task 1.

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

See Task 1.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

See Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

See Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

See Task 1.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration

- Retrieval
- Segmentation
- Tracking

See Task 1.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

1. Localized diagnostic heatmaps from prompts: For the final biomedical application any (Hematoxylin and eosin stained) image of (human) tissue is considered target cohort. Via textual prompting we will visualize tissue regions with high similarity to the textual prompt and therefore allowing for zero-shot localization.

2. Zero-shot cross-modal retrieval: The target cohort comprises all patients who present for cancer screening at the hospital and/or are called for diagnostic tests. The resulting medical reports referring to the scans will encompass a range of diagnostic procedures, including lab tests, resection excisions, biopsies, and diagnostic imaging, covering a variety of organ types.

3. Automatic captioning of already-segmented CT images: The target cohort is patients undergoing radiation therapy.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

1. Localized diagnostic heatmaps from prompts:The challenge cohort will consist of a collection of multi-organ whole slide images with dense annotations of various tissue types, such as, tumor, stroma, epithelial, and necrotic tissue.

2. Zero-shot cross-model retrieval: The challenge cohort will consist of a comprehensive collection of image and text pairs sampled from existing datasets. An image text pair consists of a single histopathology whole-slides (hematoxylin-eosin stained) and a corresponding captions selected from the diagnostic report. These pairs will encompass a diverse range of organs, for example, including skin, colon, prostate, lung, and pancreas.

3. Automatic captioning of already-segmented CT images: The dataset is sourced from the MAASTRO clinic and consists of 100 cases for the development phase and a total of 200 private test cases with an equal spread of brain, head and neck, chest and abdominal CT scans. Along with the scans, a matching multi-organ segmentation mask and correct captions are provided.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Not applicable.

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

No further information given.

b) ... to the patient in general (e.g. sex, medical history).

No further information given.

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.

Not applicable

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.

1. Localized diagnostic heatmaps from prompts: The algorithm target is to provide the user with a zero-shot heatmap for a textual prompt, guiding users to the relevant part of a whole slide image based on their prompt.

2. Zero-shot cross model retrieval: The target of the algorithm is to retrieve the correct corresponding text entries based on a list of image queries (image-to-text, or 'i2t') or vice versa (text-to-image, or 't2i') using a similarity measure. The algorithm can do so based on its learned aligned space of visual and language embeddings in a zero-shot setting.

3. Automatic captioning of already-segmented CT images: Multi-organ and tumor segmentations are in wide use in radiation therapy, to guide the delivery of radiation beams and avoid unwanted damage to nearby organs. However, a universally standardized nomenclature is frequently not possible; each system or each clinician adopts an idiosyncratic label for what is drawn. This means that the labels on existing segmentations are often lacking syntactic and semantic interoperability, which can hamper automatic data mining and machine interpretation of annotated imaging data.
Given a matched pair of a CT volume and its corresponding segmentation mask (with regions labeled generically as "A", "B", "C", etc.), the algorithm target is to generate a unique language-agnostic and standardized caption of each structure according to an open ontology such as the Foundational Model of Anatomy (FMA). For example, the generic label "A" should be mapped to its corresponding FMA identifier, such as "http://purl.org/sig/ont/fma/fma7088", which uniquely represents a specific anatomical structure.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

See Task 1.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Not applicable

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Not applicable

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.

Not applicable

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).

Not applicable

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.

Not applicable

b) State the total number of training, validation and test cases.

Not applicable

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

See Task 1.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Not applicable.

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.

Not applicable.

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.

Not applicable.

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.

Not applicable.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Not applicable.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Not applicable.

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.

Not applicable.

b) In an analogous manner, describe and quantify other relevant sources of error.

Not applicable.

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)

Not applicable.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

Not applicable.

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.

See Task 1.

b) Describe the method(s) used to manage submissions with missing results on test cases.

See Task 1.

c) Justify why the described ranking scheme(s) was/were used.

See Task 1.

Statistical analyses

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.

See Task 1.

b) Justify why the described statistical method(s) was/were used.

See previous question.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,

- common problems/biases of the submitted methods, or
- ranking variability.

See previous question.

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

N/A

Further comments

Further comments from the organizers.

N/A