

Multimodal survival and recurrence prediction in head and neck oncology: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Multimodal survival and recurrence prediction in head and neck oncology

Challenge acronym

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

HANCOTHON

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.

Clinical endpoints, such as predicting survival and recurrence risk, are critical in guiding oncological treatment decisions. In practice, clinicians combine diverse patient information to tailor treatment strategies. However, classical AI tools often fall short by relying on a single source of data, limiting their utility in complex clinical scenarios, especially in treatment planning and decision-making. The novel HANCOCK dataset addresses this gap by offering a comprehensive multimodal resource, encompassing six data types: clinical data, pathology reports, histopathology images (primary tumor and lymph nodes), tissue microarrays, tabular blood data, and free-text surgery reports. With data from 763 head and neck cancer patients collected at a single center to minimize technical biases, HANCOCK places the focus squarely on patient-specific insights. This challenge invites participants to harness the full potential of these diverse modalities, maximizing their predictive power for critical endpoints like survival and recurrence prediction for precision oncology. Challenge participants need to derive strategies for how to fuse the different data streams and combine large image datasets with structured and free text data in order to perform two binary classification tasks.

The two tasks use the same data to (i) determine the 5-year-survival to potentially help in therapy and adjuvant therapy prediction and (ii) estimate the recurrence risk to ensure timely follow-up intervals for monitoring purposes.

Challenge keywords

List the primary keywords that characterize the challenge.

Multimodal, Cancer, Histopathology, Tissue Micro Arrays, Records, Survival, Recurrence

Year

2025

Novelty of the challenge

Briefly describe the novelty of the challenge.

This challenge has an unprecedented amount of data for patients, as well as a large corpus of data (~ 4 TB in total) in a unique multimodal setting. Bringing information from these various sources together is highly challenging and key for future multimodal decision-making AI systems.

Task description and application scenarios

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

Predicting clinically relevant endpoints such as 5-year survival and 2-year recurrence risk is crucial for improving patient outcomes in oncology, particularly in head and neck cancer, where prognosis remains challenging. These endpoints address key aspects of patient care:

5-Year Survival (Binary Prediction)

=====

5-Year survival prediction is a critical marker for assessing the effectiveness of initial treatment strategies. Accurately identifying patients at higher risk of mortality within five years enables clinicians to tailor more aggressive treatment approaches or palliative care for high-risk patients. In addition, we would be able to monitor survivors more closely during this critical period, improving early detection of complications. Further, we can enhance resource allocation by identifying those who may benefit most from advanced therapies or clinical trials.

2-Year Recurrence Risk (Binary Prediction)

=====

Recurrence within five years is a major concern in head and neck cancer due to its high impact on morbidity and mortality. Predicting recurrence risk helps in customizing follow-up plans: Patients with high recurrence risk may require intensive monitoring (e.g., more frequent imaging or biomarker analysis) to detect and address relapses at an early, potentially treatable stage. Studies show that ~88% of recurrences occur in the first two years after primary treatment (surgery) [<https://pubmed.ncbi.nlm.nih.gov/37162461/>]. Knowing the high recurrence risk would immediately affect the patient's follow-up intervals.

We further can allow early adjuvant therapy decisions: Understanding recurrence likelihood allows clinicians to recommend additional treatments post-surgery, such as radiotherapy or chemotherapy, to reduce relapse probability. In addition, we can provide psychological and social support. Early knowledge of recurrence risk can prepare patients and families for potential outcomes, ensuring appropriate counseling and psychosocial support.

By focusing on these endpoints, the challenge aligns with key clinical priorities, emphasizing actionable insights that improve individualized treatment planning and long-term care strategies in head and neck cancer patients.

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

None

Duration

How long does the challenge take?

2 Hours

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

We will present and discuss the dataset, give participants the ability to present their solutions and announce the winner

Expected number of participants

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

We anticipate a number of submissions comparable to other histopathology challenges. For instance, MIDOG 2021 received 46 user submissions, and MIDOG 2022 followed a similar trend. Therefore, we estimate participation to be approximately 50 individuals. Additionally, we have received interest in the challenge from various institutions, including universities across Germany (e.g., TU Munich and Flensburg) and organizations such as Microsoft Health.

Publication and future plans

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

The dataset is currently in revision at Nature Communications. Given the interest in multimodal patient outcome and risk prediction, we plan to coordinate a publication that summarizes the main approaches and the insights derived from the challenge. This should be placed at renowned, peer-reviewed journals, such as MedIA or in the Nature family. Additionally, we plan to offer the possibility to prepare challenge proceedings to challenge participants.

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.

No on-site evaluation is planned. We will use grand-challenge.org for the evaluation. Participants need to upload docker containers with their algorithm to the platform to submit their approaches to the challenge prior to the on-site event. We will use negotiated hardware with grand-challenge.org. On the event, we need only typical broadcasting hardware for giving talks and panel discussion with the challenge winners.

TASK 1: 2-Year Survival prediction

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.

Clinical endpoints, such as predicting survival and recurrence risk, are critical in guiding oncological treatment decisions. In practice, clinicians combine diverse patient information to tailor treatment strategies. However, classical AI tools often fall short by relying on a single source of data, limiting their utility in complex clinical scenarios, especially in treatment planning and decision-making. The novel HANCOCK dataset addresses this gap by offering a comprehensive multimodal resource, encompassing six data types: clinical data, pathology reports, histopathology images (primary tumor and lymph nodes), tissue microarrays, tabular blood data, and free-text surgery reports. With data from 763 head and neck cancer patients collected at a single center to minimize technical biases, HANCOCK places the focus squarely on patient-specific insights. This challenge invites participants to harness the full potential of these diverse modalities, maximizing their predictive power for critical endpoints like survival and recurrence prediction for precision oncology. Challenge participants need to derive strategies for how to fuse the different data streams and combine large image datasets with structured and free text data in order to perform two binary classification tasks.

The two tasks use the same data to (i) determine the 5-year-survival to potentially help in therapy and adjuvant therapy prediction and (ii) estimate the recurrence risk to ensure timely follow-up intervals for monitoring purposes.

Keywords

List the primary keywords that characterize the task.

Survival, Endpoint, Outcome, Multimodal, Accuracy, Classification

ORGANIZATION

Organizers

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

Prof. Dr. Andreas Kist (FAU Erlangen-Nürnberg)

PD Dr. Markus Eckstein (University Hospital Erlangen)

Prof. Dr. Antoniu-Oreste Gostian (Barmherzige Brüder Hospital Straubing)

Prof. Dr. Katharina Breininger (University Würzburg)

b) Provide information on the primary contact person.

Andreas Kist, andreas.kist@fau.de

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place. Not every challenge closes after the submission deadline (one-time

event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

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

One-time event with fixed conference submission deadline

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).

We will extend the HANCOCK dataset's website: <https://hancock.research.fau.eu/index>

Participation policies

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

Fully automatic

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

No additional data allowed

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

Members from the same Department as the organizer may participate but are not eligible for awards

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

We aim to have three monetary prizes for the best three participants for Task 1 and Task 2, respectively. Each task will be awarded separately. The same team can be awarded twice.

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 3 performing methods will be announced publicly during the challenge event at MICCAI.

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

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

We aim to write a peer-reviewed publication with the challenge results (see above). With each submission, we highly encourage the contributors to write a 2-4 page article and post it on arxiv. We plan to offer co-authorship of the overall challenge report to a maximum of two authors per submission given competitive performance on the challenge task(s). By default, we would include the first and last author of a given submission. We do not impose any embargo time, such that individuals can write up their own papers at any time.

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.

Algorithm container submission (type 2) on Grand Challenge.

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 allow the participants to use the ranking code on their own validation data. Validation data splits will be recommended on the website and available by download. With the original paper, we provide suggested validation splits.

We will also allow multiple submissions and running on the test data without disclosing the performance to test if the algorithm runs through without any issues. We will limit this to 3 evaluations per day, the last submission counts towards the leaderboard and the final evaluations.

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

Training data is already released. Training/validation data splits are also released.

End of March: Start of challenge, description of the task and endpoints, registration for participants and live of the challenge website. We will provide a Slack workspace with announced Q&A; sessions.

Mid of August: Deadline for registration for participants and availability of testing data.

31st of August: Deadline for docker container submission and 2+ page preprint abstract submission

23rd-27th of September: Announcement of winners and release of the final results

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).

We received ethics approval by the local ethics board of the University of Erlangen, Medical Faculty: #23-22-Br

Data usage agreement

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

Examples:

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

CC BY (Attribution)

Code availability

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

Code will be provided together with the challenge openly on Github.

Any code related to the HANCOCK dataset is already available:

https://github.com/ankilab/HANCOCK_MultimodalDataset

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

We will provide links to the docker containers the participants submitted, given their consent. Participants will be encouraged to give permission for this and make their code publicly available on Github. We plan to include the participants' Github links on the challenge Github repo.

Conflicts of interest

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

The authors do not have any conflict of interest w.r.t. the challenge.

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

CAD, Decision support, Intervention planning, Treatment planning

Task category(ies)

State the task category(ies)

Examples:

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

Classification, Prediction

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 suffering from squamous head and neck cancer (HNCSS), ideally generalized to multiple hospitals

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

The data was acquired from the Department of Otorhinolaryngology and Head and Neck Surgery and from the Pathological Institute of the University Hospital in Erlangen. All data was collected and published following the local ethics committee vote (#23-22-Br). Retrospective, multimodal data was gathered from patients who were diagnosed with head and neck cancer between 2005 and 2019. Only patients who had a curative first treatment were included. The cohort does not provide full data for each patient. The challenge participants need to figure out strategies to overcome these limitations, in the worst case, just dropping patient records.

Imaging modality(ies)

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

Digital Pathology: Histopathology and Tissue Micro Arrays
(see below)

Context information

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

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

Image data contains information about the tumor progression (primary tumor) and seeding in the adjacent lymph nodes. Further, TMAs contain cores from the primary tumor and in some cases also from the lymph node and are stained against a variety of antigens associated with head and neck cancer pathology, namely CD3, CD8, CD56, CD68, CD163, PD-L1, and MHC-1. For example, PD-L1 pathways have been shown to be key for targeted therapeutics [<https://pmc.ncbi.nlm.nih.gov/articles/PMC7531035/>].

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

The modalities in our dataset can be categorized into image data (histopathological images), structured data (clinical, pathological, and blood data), and free text (surgery reports). The full HANCOCK dataset has a very rich repertoire of patients in general. We have a typical age distribution for head and neck cancer. 80% are male and 20% are female (also as described in the literature), 47% are smokers, 28% are non smokers and the remaining 25% have smoked at some point. The cohort does not provide full data for each patient. The challenge participants need to figure out strategies to overcome these limitations, in the worst case, just dropping patient records. More information can be received at the dataset website: <https://hancock.research.fau.eu/index>

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.

Histopathology from primary patient tumor resection (Head&Neck, primary surgery) and adjacent lymph node depending on the patient and disease progression.

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.

5-year survival given the patient data

We chose 5-year survival as clinically highly significant and known across cancer types.

Assessment aim(s)

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

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

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

Accuracy, F1

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).

* 3DHistech P1000

* Aperio Leica Biosystems GT450

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

Tissue samples of the respective patients were collected from the pathological archive of the University Hospital in Erlangen. The samples originate from the primary tumor and, if present, positive lymph nodes that had been resected. The tissue samples had been fixed in formalin, embedded in paraffin, and routinely stained with HE. The 709 primary tumor sections were scanned using a 3DHistech P1000 at 82.44x magnification and with a resolution of 0.1213 $\mu\text{m}/\text{px}$. A single slide was available for 701 cases whereas two slides were available for eight cases. The 396 lymph node sections were scanned using an Aperio Leica Biosystems GT450 at 40x magnification with 0.2634 $\mu\text{m}/\text{px}$ and using 3DHistech P1000 at 51.42x magnification with 0.1945 $\mu\text{m}/\text{px}$. All digitized WSIs were stored in the pyramidal Aperio file format (.svs).

Additionally, TMAs were created from the paraffin-embedded primary tumor blocks. The TMA cores with a diameter of 1.5 mm were extracted from the tumor center and the tumor invasion front. They were stained using

HE and they were stained for specific immune cell populations using the IHC markers CD3, CD8, CD56, CD68, CD163, PD-L1, and MHC-1. From each patient, at least two cores were collected per origin and marker. This resulted in 368 TMAs, each with cores arranged in 12 rows by 6 columns. The TMAs were scanned using a 3DHistech P1000 at 82.44x magnification with a resolution of 0.1213 $\mu\text{m}/\text{px}$.

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 Hospital Erlangen (Institute of Pathology, Director Prof. Hartmann, and Institute for Otolaryngology, Head & Neck Surgery, Director Prof. Iro)

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).

Clinical experts (board-certified pathologists, consultant/"Fachärzte" for ENT, molecular biology labs, trained technical assistance)

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 all information that is available for one particular patient, i.e. all multimodal information for a given patient. The information distribution will be available on the challenge website.

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

763 cases in total in HANCOCK and ~50 cases for the test set, aiming to match age- and gender-distribution of the HANCOCK dataset as good as possible. The training and validation split can be done in various ways (also shown in the HANCOCK dataset pre-print publication, we provide the splits online for the challenge participants' convenience). Challenge participants are responsible to define their own training/validation split, as the data is already publicly available.

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

The train/validation is due to the nature of the HANCOCK dataset (all public available data, suggested train/validation splits online available) and test cases are the ones we can acquire with the same standards as in the original HANCOCK dataset

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

We aim for a class balance according to the Task 1 (25 survivors and 25 non-survivors) in the test dataset, despite the class imbalance in the training dataset. The gender and age-distribution should be similar to the HANCOCK dataset.

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

As test dataset, we provide unseen, unpublished data with the same data quality criteria as the original HANCOCK dataset

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 reference annotation is per report to the clinics. We do not have the exact death date, but rather individual information events that include if the patient passed away.

Of note, we provide partial annotations of cancer regions in the histopathology data (1 annotator, supervised by board-certified pathologist). These annotations do not claim to be in any form complete or fully correct, and are not included in any way in the evaluation of the challenge.

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.

No annotations are relevant to this challenge - N/A.

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

N/A

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

N/A

Data pre-processing method(s)

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

Data was converted to public standards, cleaned and streamlined for usability. Detailed information is provided in the Github repository and the HANCOCK dataset pre-print publication. We will further provide pre-computed embeddings for whole-slide images and TMAs from recent foundation models.

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.

N/A

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

The data stem from various sources and can contain copy&paste; errors, mislabeling (Histopathology and TMA data belong to different patients) and limits in timeliness (knowledge of patient's death and death date can be significantly different).

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 assess the performance of algorithms in predicting 5-year survival outcomes for head and neck cancer patients, we propose using accuracy as the primary evaluation metric. Accuracy measures the proportion of correctly classified patients into survivor and non-survivor categories, providing an intuitive and interpretable assessment of overall predictive performance. We will further compute the F1 score based on the harmonic mean of precision and recall. We will use the average of F1 score and Accuracy as the metric for ranking. Using the metric assessment tool from [<https://metrics-reloaded.dkfz.de/>], we also evaluate Balanced Accuracy and announce the final choice by the time the challenge is announced. We currently plan to use a balanced test dataset, making balanced accuracy obsolete. In case of any deviations due to unforeseen circumstances, we will openly communicate the decision of the metric.

Additionally, to capture the balance between sensitivity and specificity in the classification, we will also compute the Area Under the Receiver Operating Characteristic Curve (AUC). The AUC reflects the algorithm's ability to distinguish between survivors and non-survivors across different classification thresholds. These metrics collectively ensure that the evaluated algorithms not only perform well on average but also achieve a robust balance in correctly identifying both positive and negative survival outcomes, aligning with the clinical importance of minimizing misclassification risks. Rankings of submitted algorithms will be based on their accuracy, with AUC serving as a complementary metric to evaluate discriminatory power.

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

The task is binary classification and accuracy is a well-founded metric for assessing classification performance. The F1 score has also been very established as a combinatorial metric from precision and recall. AUC is also very accepted in the community and will be reported, but not used for ranking.

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.

For each case, we get a label (survivor, non-survivor) to compute an overall confusion matrix. The gained accuracy and F1 score is used for ranking.

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

Missing results are punished by taking the opposite as guess (if survivor, then the missing case is set to non-survivor).

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

The better the overall accuracy and F1 score of unseen data, the better the performance of the method and likely more usable in the clinic. We also would like to ensure high precision and recall values, therefore, incorporating F1 score.

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.

To evaluate the variability of algorithm rankings, we will conduct bootstrapping with 200 resamples of the dataset. This method provides confidence intervals (CI) for the evaluation metrics (e.g., accuracy, F1 and AUC) and ensures robust ranking comparisons. Additionally, Friedman's test will be employed to assess the statistical significance of differences in performance across multiple algorithms, followed by pairwise post-hoc tests with Bonferroni correction where applicable.

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

Bootstrapping is used to assess the variability of results given a different set of cases. This aims to estimate the error for a new test set drawn from the same distribution. This allows us to report a 95% confidence interval for the accuracy the participants submitted.

Further analyses

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

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

We will especially intent to analyze inter-algorithm variability. Are always the same patients mis-diagnosed? Are there maybe clinical reasons for this?

TASK 2: 2-Year Recurrence prediction

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.

Clinical endpoints, such as predicting survival and recurrence risk, are critical in guiding oncological treatment decisions. In practice, clinicians combine diverse patient information to tailor treatment strategies. However, classical AI tools often fall short by relying on a single source of data, limiting their utility in complex clinical scenarios, especially in treatment planning and decision-making. The novel HANCOCK dataset addresses this gap by offering a comprehensive multimodal resource, encompassing six data types: clinical data, pathology reports, histopathology images (primary tumor and lymph nodes), tissue microarrays, tabular blood data, and free-text surgery reports. With data from 763 head and neck cancer patients collected at a single center to minimize technical biases, HANCOCK places the focus squarely on patient-specific insights. This challenge invites participants to harness the full potential of these diverse modalities, maximizing their predictive power for critical endpoints like survival and recurrence prediction for precision oncology. Challenge participants need to derive strategies for how to fuse the different data streams and combine large image datasets with structured and free text data in order to perform two binary classification tasks.

The two tasks use the same data to (i) determine the 5-year-survival to potentially help in therapy and adjuvant therapy prediction and (ii) estimate the recurrence risk to ensure timely follow-up intervals for monitoring purposes.

Keywords

List the primary keywords that characterize the task.

Recurrence, Endpoint, Multimodal, Accuracy, Classification

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

One-time event with fixed conference submission deadline

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).

Fully automatic

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

No additional data allowed

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

Members from the same Department as the organizer may participate but are not eligible for awards

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)

CC BY (Attribution)

Code availability

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

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

CAD, Decision support, Intervention planning, Treatment planning

Task category(ies)

State the task category(ies)

Examples:

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

Classification,Prediction

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.

see Task 1

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

see Task 1

Imaging modality(ies)

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

see Task 1

Context information

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

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

see Task 1

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

see Task 1

Target entity(ies)

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

see Task 1

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

Recurrence of cancer 2 years after initial diagnosis.

Assessment aim(s)

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

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

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

Accuracy,F1

DATA SETS**Data source(s)**

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

see Task 1

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

see Task 1

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

see Task 1

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

see Task 1

Training and test case characteristics

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

Examples:

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

see Task 1

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

see Task 1

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.

We aim for a class balance according to the Task 1, but with their respective recurrence information.

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

see Task 1

Annotation characteristics

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

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

see Task 1

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

see Task 1

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

see Task 1

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

N/A

Data pre-processing method(s)

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

see Task 1

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

see Task 1

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

see Task 1

ASSESSMENT METHODS

Metric(s)

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

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

see Task 1

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

see Task 1

Ranking method(s)

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

For each case, we get a label (gets recurrence, gets no recurrence) to compute an overall confusion matrix. The gained accuracy and F1 score is used for ranking.

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

similar to 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 Task 1

Further analyses

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

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

see Task 1

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

A publication describing the dataset and its potential use cases is currently in revision at Nature Communications and available online here as non-peer-reviewed pre-print:

<https://www.medrxiv.org/content/10.1101/2024.05.29.24308141v1>

The dataset is available here:

<https://hancock.research.fau.eu/index>

Further comments

Further comments from the organizers.

N/A