The PANORAMA Study Protocol: Pancreatic Cancer Diagnosis - Radiologists Meet AI

Natália Alves[1,2*], Megan Schuurmans[1,2*], Dawid Rutkowski[3], Derya Yakar[2], Ingfried Haldorsen[4], Marjolein Liedenbaum[4], Anders Molven[5], Pierpaolo Vendittelli[1], Geert Litjens[1], John Hermans[1], Henkjan Huisman[1]

1) Department of Medical Imaging, Radboud University Medical Center, Nijmegen, The Netherlands
2) Department of Radiology, University Medical Center Groningen, Groningen, The Netherlands
3) Division of Surgery & Oncology, Karolinska Institutet, Stockholm, Sweden
4) Mohn Medical Imaging and Visualization Centre, Department of Radiology, Haukeland University Hospital, Bergen, Norway
5) Department of Pathology, Haukeland University Hospital, Bergen, Norway

*These authors contributed equally

1. Introduction

Pancreatic ductal adenocarcinoma (PDAC) is estimated to become the second leading cause of cancer-related deaths in Western countries by 2030 (*Siegel et al., 2020*). Due to the lack of early, disease-specific symptoms, 80–85% of patients are diagnosed in advanced stages (*Ryan et al., 2014*). However, stage I patients present a significantly more favorable prognosis than stage IV patients (median survival: 26 vs. 4.8 months), making early detection the current most effective approach to improving outcome (*Schwartz et al., 2021*).

Abdominal contrast-enhanced computed tomography (CECT) scans are the most common type of CT examination to evaluate diseases and determine treatment response, accounting for 40%–52% of all CT examinations performed worldwide (*Prokop et al., 2022; Schöckel et al., 2020; Gonzales et al., 2015*). CECT is the first line of diagnosis for PDAC, as there are no validated diagnostic biomarkers (*Grossberg et al., 2020*). Studies have shown that in 16%–84% of diagnosed patients, cancer signs (such as pancreatic duct cutoff/dilatation and pancreatic atrophy) can be retrospectively seen on pre-diagnostic CECT scans (3–36 months before clinical diagnosis) *(Singh et al., 2020; Toshima et al., 2021)*. Secondary imaging signs and early subtle focal lesions can be overlooked in

clinical practice, where most abdominal CECT scans are acquired for non-pancreatic-specific indications, leading to delayed diagnosis and reduced survival.

Radiologists' performance at detecting PDAC on routine abdominal CECT is still underexplored. Current research is limited to small reader- and case samples, usually from single centers, and most studies do not stratify reader performance based on tumor size/stage. Artificial intelligence (AI) is starting to achieve expert performances in cancer diagnosis across many domains (*Esteva et al., 2017*; *McKinney et al., 2020*), and an increasing number of publications study AI for PDAC diagnosis on CECT (*Cao, K. et al., 2023*; *Korfiatis et al., 2023*; *Chen et al., 2023*; *Park et al., 2023*). However, most studies do not focus on early detection and consider private, single-center data sets. Furthermore, current AI research largely lacks clinical comparison to radiologists.

The PANORAMA study is a new prospectively designed multi-center study to assess the performance of both radiologists and AI at PDAC detection in routine abdominal CECT scans. PANORAMA's goals are threefold: 1) to establish the clinical baseline performance of radiologists at PDAC detection through a large-scale, international reader study, 2) to establish the state-of-the-art AI performance at PDAC detection through an international AI grand-challenge, and 3) to compare AI and radiologists, with the end goal of obtaining substantial evidence to start implementing AI to help find PDAC earlier. The primary hypothesis of the PANORAMA study is that AI is at least non-inferior to radiologists in the reader study at PDAC detection on CECT scans.

2. Materials and Methods

2.1. Data Set

The PANORAMA data set consists of a retrospective, multi-center, cancer-enriched set of routine abdominal and upper-abdominal CECT scans.

The data set contains images originating from at least two centers in The Netherlands (Radboud University Medical Center (RUMC), and University Medical Center Groningen (UMCG)), one center in Norway (Helse Bergen HF Haukeland University Hospital (HUH)), and one center in Sweden (Karolinska Institute (KSKI)).

Additionally, the training set includes three publicly available data sets from the United States of America (USA), namely the Medical Segmentation Decathlon (MSD) pancreatic cancer (training) dataset (*Antonelli, et al. 2022*) from Memorial Sloan Kettering Cancer Center (New York, USA), the National Institute of Health's Clinical Proteomic Tumor Analysis Consortium Pancreatic Ductal Adenocarcinoma (TCIA-CPTAC) data set, and pancreas CT data set (NIH-CT) (*Roth, et al. 2016*).

All images are CECT scans in the portal-venous phase, as this is the most commonly acquired phase in routine abdominal imaging. All exams are from patients undergoing CECT without a history of pancreatic cancer treatment, and without any prior positive PDAC histopathology findings.

The data set will be sampled into 3 splits (as summarized in Table 1) according to the following use cases:

- **Training and development set:** Used to train and develop AI models (made publicly available under a non-commercial CC BY-NC 4.0 license). Includes currently public data from the USA (MSD, TCIA, NIH-CT) and new cases from two centers in The Netherlands (RUMC, UMCG).

- **Validation and tuning cohort:** Used for AI model tuning and selection. This data set will remain secret to allow unbiased AI development.

- **Testing Cohort**: Used to benchmark AI, and radiologists, and test all hypotheses. Includes internal testing data (unseen cases from two seen center {RUMC, UMCG}) and external testing data (unseen cases from (at least) two unseen centers {HUH, KSKI}). This data set will remain secret to allow unbiased AI and radiologist's assessment.

Table 1. Overview of the PANORAMA dataset.

Data split	**Public training and development set**	**Hidden validation and tuning cohort**	**Hidden testing cohort**	**Total**
Data source	RUMC, UMCG, TCIA, NIH-CT, MSD (The Netherlands, USA)	RUMC, UMCG (The Netherlands)	RUMC, UMCG, HUH, KSKI (The Netherlands, Norway, Sweden)	
No. Cases	1800*	100	400*	**2300**
* Tentative numbers				

RUMC – Radboud University Medical Center; TCIA – The Cancer Imaging Archive; NIH – National Institute of Health; MSD – Medical Segmentation Decathlon; UMCG – University Medical Center Groningen; KSKI – Karolinska Institute; HUH – Helse Bergen HF Haukeland University Hospital.

2.1. Study Population

To meaningfully validate AI and radiologists at PDAC detection towards clinical translation, it is essential to analyze performance across the patient population encountered in clinical routine. For instance, while surgical resection can provide the most comprehensive tissue specimen to facilitate accurate histopathology grading, a cohort of solely resected patients deviates substantially from the distribution of patients encountered during clinical routine. A clinically representative data set must also include cases that did not undergo surgical resection as well as cases without PDAC (negative cases). It is also important to ensure that negative cases include both cases with a "normal" pancreas (absence of pancreatic alterations) and cases with non-PDAC lesions (such as pancreatitis and pancreatic cysts), as both scenarios will be routinely encountered in the clinical workflow where AI would be deployed.

The exact characteristics and patient-distribution per split (train/validation/testing) will be carefully considered and recorded. However, this information will remain blinded for this purpose of this protocol to prevent bias.

2.2. Reference Standard

The hidden testing cohort has the highest-quality reference standard for all cases, to optimally validate AI and radiologists. As histopathology analysis is the gold standard for PDAC diagnosis confirmation, all positive PDAC cases in the hidden testing cohort will have histopathology ground truth, either through surgical resection or biopsy assessment. For the negative cases, the ground-truth label will be established through histopathology assessment (for cases with non-PDAC pancreatic lesions such as cysts and intraductal mucinous neoplasms) and/or follow-up data. Only patients who do not develop PDAC within 36 months after their initial CT scan will be included as negative cases.

The public training cohort is large and representative to optimally train clinically relevant AI algorithms. The PANORAMA public training and development data set will contain a combination of previous public data sets and at least 1400 new public cases derived from RUMC and UMCG (The Netherlands), making it the largest publicly available data set in the field. Positive cases are confirmed based on histopathology when available, or based on radiology reports and follow-up from clinical routine if no biopsy or resection was performed. Negative cases will be confirmed through clinical reports. For cases deriving from existing publicly available data sets, the corresponding ground-truth will be considered. For cases in the MSD data set, differentiations between PDAC and non-PDAC cases will be made as reported by *Suman et al. 2021*. All training cases will carry patient/image-level annotations, of which, about half of the cases will also include expert-derived tumor delineations, while the remainder will include AI-derived delineations (based on a re-trained version of the method proposed by *Alves, et al. 2022*). A subset of 100 scans with the same reference standard of the training cohort will be used as the hidden validation and tuning cohort.

2.3. AI study

The PANORAMA: AI study aims to evaluate the performance of modern AI algorithms at patient-level diagnosis and lesion-level detection of PDAC in abdominal CECT scans. Similar to radiologists, the objective of AI developed in this study is to read CECT exams and produce an overall patient-level score for PDAC diagnosis and the lesion location, as depicted in Figure 1. The AI outputs will be compared to the ground truth of the tuning and validation set and hidden test set to compute AI performance (see section 3.1).

AI algorithms for the PANORAMA study will be developed in the context of an international AI Grand Challenge hosted at the grand-challenge.org platform. Grand Challenges are the gold-standard for assessing and comparing state of the art AI algorithms for a given task in a controlled and unbiased environment.

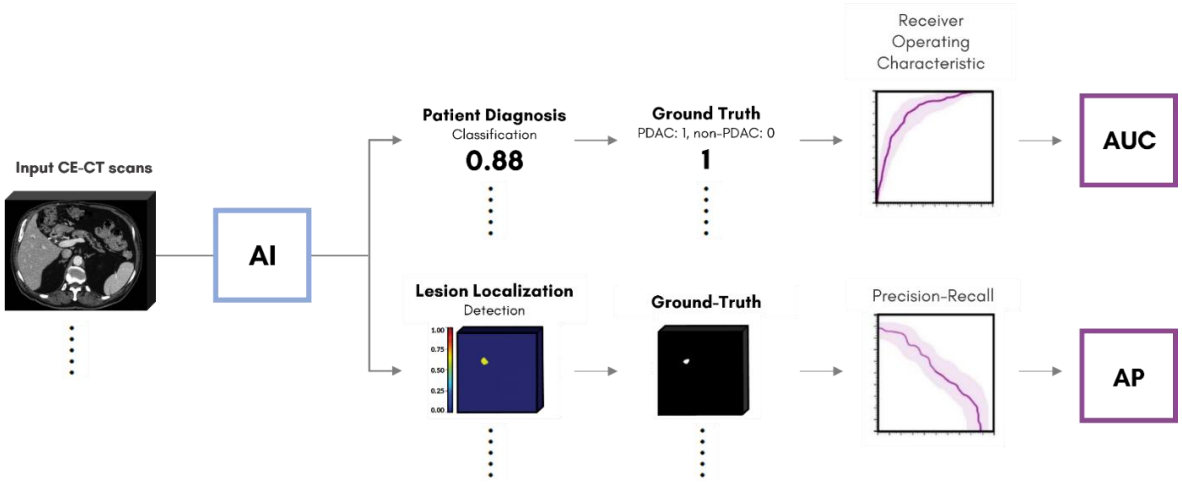

Figure 1. Overview of the AI workflow.

2.4. Grand Challenge Structure

The PANORAMA: AI study Grand Challenge takes place in two phases:

1. **Development Phase (Duration: 4–6 months):** Anyone can participate in this phase of the challenge. Interested teams can create an account on grand-challenge.org, and register for the PANORAMA24 challenge at panorama.grand-challenge.org. Afterward, they will be provided access to download the public training dataset, and in turn, they can start developing and training AI models. During the evaluation, algorithms are executed on the grand-challenge.org platform, their performance is estimated on the hidden validation and tuning cohort, and team rankings are updated accordingly on a live, public leaderboard. Facilitating validation in such a manner ensures that any image used for evaluation remains truly unseen and that AI predictions cannot be tampered with —allowing for bias-free performance estimation.

2. **Testing Phase (Duration: 1 month):** At the end of the open development phase, each registered team can choose to submit a single AI algorithm for evaluation on the hidden testing cohort. Based on their performance on this cohort, new rankings will be drawn and the top 5 AI algorithms of the PANORAMA Grand Challenge will be determined and announced. To qualify as one of these top teams, participants must also submit a short paper on their methodology (2-3 pages) and a public/private URL to their source code on GitHub —to ensure fairness, traceability and reproducibility of all proposed solutions.

2.5. Reader Study

Parallel to the AI study, the PANORAMA: Reader Study aims to evaluate the performance of abdominal radiologists with various levels of expertise at patient-level diagnosis of PDAC in abdominal CECT scans. Radiologists receive the same information as AI and provide a patient-level likelihood score for the presence of PDAC between 0 and 100 (with incremental steps of 1), together with a point coordinate at the center of the tumor.

Standardized reporting and data systems (RADS) have been widely adopted for imaging-based cancer diagnosis across many domains such as liver (LI-RADS), colon (C-RADS),

breast (BI-RADS), and prostate (PI-RADS). However, in the absence of large-scale multi-reader multi-case studies there are currently no categorical scales for PDAC diagnosis on CECT. To investigate the potential for standardized reporting on PDAC diagnosis, we ask readers to classify cases into 5 different risk levels, defined in line with clinically validated RADS for other cancer diseases (*An, et al., 2019*).

- Category 1: Very low PDAC suspicion.
- Category 2: Low PDAC suspicion
- Category 3: Equivocal
- Category 4: High PDAC suspicion
- Category 5: Very high PDAC suspicion

These discrete categories will be used to establish the clinical operating point at which sensitivity and specificity of radiologists will be compared to AI algorithms.

2.6. Reader Study Structure

The PANORAMA: Reader Study is hosted on grand-challenge.org/reader-studies, which supports abdominal CT viewers and annotation tools, using the 400 cases on the hidden testing cohort. Primary costs of the study are reader time and effort. As reading all 400 cases is too labor-intensive, we opt for a split-plot design (*Obuchowski et al., 2009*; *Obuchowski et al., 2012*; *Chen et al., 2018*). By doing so, we preserve a sample size of 400 cases, while the individual workload is reduced to a maximum of 100 cases (as illustrated in Figure 2). Prior to the start of the study, readers are provided a detailed guide on the annotation workflow, including all tools made available on the platform (e.g., navigation, zooming, windowing, measuring scale). They are also provided access to a practice session with 5 example cases (from the MSD data set), to get familiarized with both the reading interface and the expected workflow. Afterwards, each reader is provided a unique, password-protected, non-transferable URL to access their instance of the reader study (with all 100 allotted cases). Cases are made available sequentially and cannot be revisited post-assessment. Readers are expected to complete their assessments in 3-5 months.

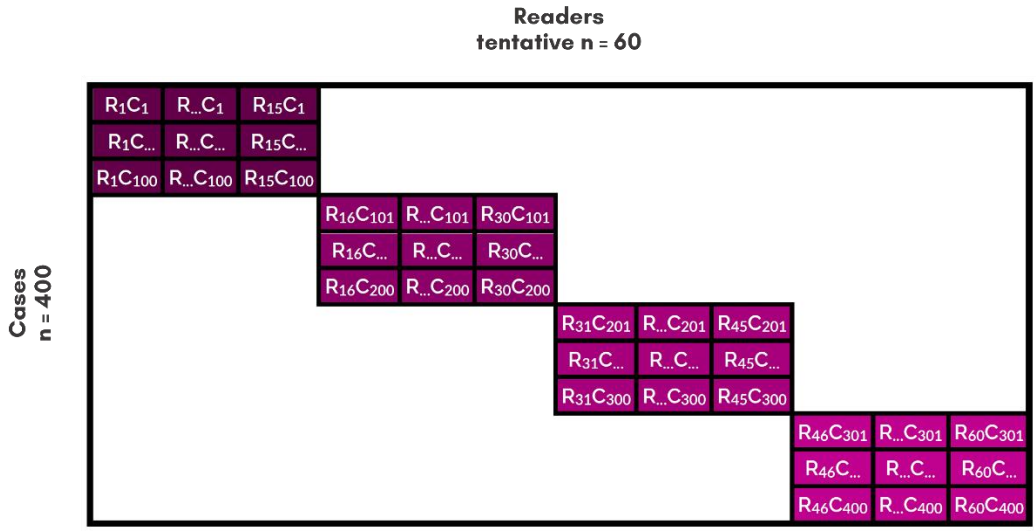

Figure 2. Tentative study design for the distribution of readers and cases in a 4x4 split-plot configuration. All 60 readers and 400 cases are divided into 4 blocks, in a stratified

manner, that takes reader and case distributions into account to minimize any potential differences between separate blocks. Each block of readers reads their own set of cases. As this study design is reader-dependent, it is susceptible to changes based on the final outcomes of reader recruitment.

3. Experimental Analysis

3.1. Performance Metrics and Comparisons

The key performance metrics used to evaluate AI and radiologists have been summarized in Figure 3. Patient-level diagnosis performance will be evaluated through receiver operating characteristic (ROC) curve analysis, using the area under the curve (AUROC) as a summary metric. AUROC will be derived for AI and clinicians from the respective continuous patient-level cancer likelihood scores. To compare AI sensitivity/specificity with readers, AI cancer continuous predictions will be binarized to match the sensitivity/specificity of radiologists at clinically relevant operating points.

Lesion-level localization performance will be evaluated through precision-recall curve analysis, with the average precision (AP) as a summary metric. AP will be derived for AI from the lesion-level predictions extracted from the tumor likelihood maps. To compare AI precision/recall with readers, an operating point will be defined based on the point-annotations provided in the reader study.

Algorithms in the PANORAMA: Grand Challenge will be independently ranked according to patient-level diagnosis performance using AUROC (diagnosis rank) and lesion-level localization performance using AP (localization rank). The final overall ranking score for each algorithm will be obtained by averaging the diagnosis and localization ranks.

The PANORAMA organizers will publicly release functions to compute all performance metrics discussed in this section, in a GitHub repo.

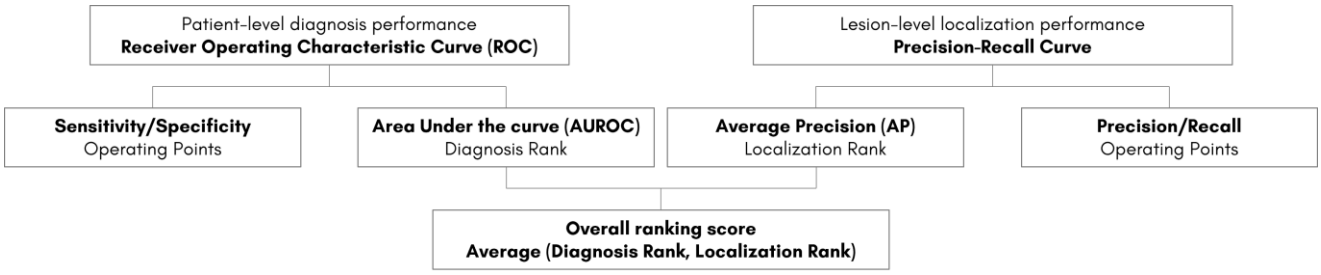

Figure 3. Performance metrics for AI and radiologists' evaluation. Overall ranking score is only used to evaluate different AI algorithms with respect to each other and facilitate the validation and testing leaderboards of the PANORAMA: Grand Challenge. When comparing AI performance to that of radiologists at specific operating points, AI predictions are binarized to match the sensitivity, specificity, precision or recall of radiologists.

3.2. Hit-Criterium for Lesion Detection

In this section, we define the criterion to establish hit/miss between lesion detections (predicted by AI or radiologists) and lesion annotations (established by the reference standard defined in Section 2.2). A "hit criterion" is a condition that must be satisfied for each predicted lesion to count as a hit or true positive. In line with recent literature for pancreatic cancer detection AI, we opt for a hit criterion based on object overlap for the 3D detections predicted by AI:

- **True Positive**: A candidate lesion is considered a true positive if the intersection over union (IoU) with the ground truth (calculated in 3D between the whole extracted candidate lesion volume and the tumor ground truth volume) was at least 0.1. This threshold is set in line with well-established previously published studies of the same nature considering other cancer diseases (*McKinney et al., 2020; Saha et al., 2021)* as it addresses the clinical need for object-level localization while also considering the non-overlapping nature of objects in 3D.

- **False Positives**: Predictions with no/insufficient overlap count towards false positives, regardless of their size or location.

For the point-coordinate annotations predicted by radiologists, we opt for a hit criterion based on distance/location:

- **True Positives:** For a predicted lesion coordinate to be counted as a true positive, it must reside ≤ 5 mm of the ground-truth tumor boundary (as done in *Saha et al. 2022*). Such a margin is considered to account for smaller lesions, where the ground-truth annotation spans few slices, and radiologists' point predictions can register a miss from marginal deviations (despite correct cognitive localization).

- **False Positives:** Predictions that are > 5 mm from the ground-truth annotation boundary count towards false positives.

3.3. Statistical Tests

Statistical tests are performed in the final arc of the PANORAMA study, after selecting the top-ranking AI algorithms from the AI study (Grand Challenge) and estimating the performance of all radiologists from the reader study. Each test is facilitated using AI and/or radiologists' predictions on the hidden testing cohort.

As the primary outcome of the PANORAMA study, we statistically evaluate the diagnostic performance of a state-of-the-art artificial intelligence system in comparison to radiologists at PDAC diagnosis in CECT. We evaluate performance by calculating sensitivity, specificity, and the AUROC of the AI system, according to its case-level predictions of significant cancer diagnosis. The AI system is defined as the ensemble of the top three algorithms developed for the PANORAMA Grand Challenge. We investigate and compare AI predictions relative to the radiology readings made during a multi-reader multi-case (MRMC) observer study with 400 testing cases. For these comparisons, histopathology and a follow-up period of ≥ 36 months are used to establish the reference standard. All study objectives are pre-specified in a hierarchical family and tested accordingly (Figure 4).

Across the 400 patient examinations used in the MRMC observer study, we compare the case-level diagnostic performance of the AI system to the mean of the participating radiologists. We define the test statistic as the AUROC of the AI system minus the mean AUROC of the radiologists. Differences in (mean) AUROC and two-sided 95% Wald confidence intervals are computed. Non-inferiority is concluded, if the lower boundary of the two-sided 95% confidence interval for the test statistic is greater than a margin of -0.05 (established in concordance with *Saha et al. 2022*; *McKinney et al., 2020*; *Rodriguez-Ruiz et al., 2019*). If non-inferiority is concluded, then superiority of the AI system over radiologists is tested and concluded if the lower boundary of the two-sided 95% confidence interval for the test statistic is > 0. Overall, these comparisons aim to investigate whether an AI system at CECT can offer non-inferior or superior diagnostic performance compared to radiologists.

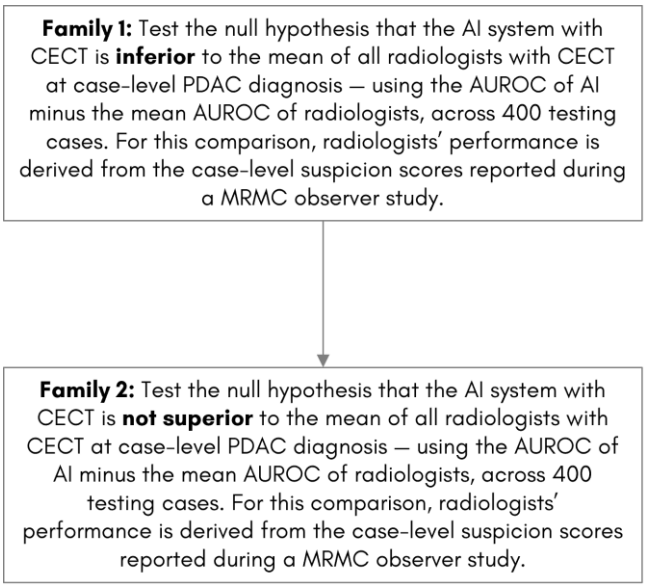

Figure 4. Flowchart illustrates the strategic plan to test study objectives. If the null hypothesis of family 1 is rejected with an alpha value of 0.05, AI non-inferiority is confirmed, and we test family 2. If the null hypothesis of family 2 is then rejected with an alfa value of 0.05, AI superiority is confirmed. If the null hypothesis of family 1 is not rejected with an alpha value of 0.05 we do not test family 2.

4. Discussion and future directions

The PANORAMA study is a prospectively designed study to compare state-of-the-art AI algorithms and radiologists with varying levels of expertise in PDAC detection on routine CECT scans. The PANORAMA study protocol was established in collaboration with a multi-disciplinary and international expert scientific advisory board. By prospectively publishing the study protocol before starting the AI Grand Challenge and the Reader Study, we aim to increase the transparency and accountability of our study and, consequently, its impact on the scientific community.

Within the PANORAMA study, we will host the first international Grand Challenge for PDAC detection/diagnosis. Grand Challenges can address the lack of trust, scientific evidence, and adequate validation among AI solutions (*Leeuwen et al., 2021*) by providing

the means to compare algorithms against each other using common training and testing data. By opening AI development to the scientific community, we aim to encourage diverse and creative solutions, and by providing a common and publicly available validation environment, we aim to foster accountable model development and unbiased AI assessment.

The PANORAMA study encompasses a large-scale international reader study with radiologists from varying levels of expertise at reading CECT scans for PDAC detection. This will be the first feasibility study for introducing a standardized reporting system in PDAC diagnosis and will provide unprecedented evidence for the baseline radiologists' performance at reading CECT for PDAC diagnosis. CECT is the most widely used modality for diagnosing abdominal disease in general and pancreatic disease specifically, accounting for over 120 million scans yearly worldwide (*Prokop et al., 2022; Gonzales et al., 2015*). Since it has been recently shown that incidental- or screening-detected PDAC patients have a median overall survival of 9.8 years compared to only 1.5 years for those diagnosed with standard clinical workflows (*Dbouk et al., 2022*), a reliable AI system for PDAC detection that can run in the background for CECT examinations holds great potential for early detection and improving patient outcomes.

PANORAMA provides a large public data set together with reliable experiments and analysis with the goal of showing that AI can reliably detect PDAC on CECT scans, thus providing substantial evidence to detect incidental PDAC earlier, improving patient survival.

Annex 1) Pancreatic Cancer Diagnosis: Radiologists Meet AI The PANORAMA Study BIAS Assessment

This annex is a checklist to demonstrate compliance to the MICCAI community guidelines for organizing biomedical image analysis challenges (*Maier-Hein, et al 2020*). The purpose of the checklist is to standardize and facilitate the PANORAMA study review process and raise interpretability and reproducibility of challenge results by making relevant information explicit.

Item 1: Title

a) Use the title to convey the essential information on the challenge mission

Pancreatic cancer diagnosis: radiologists meet AI

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

PANORAMA

Item 2: 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.

Pancreatic ductal adenocarcinoma (PDAC) is estimated to become the second leading cause of cancer-related deaths in Western countries by 2030. Due to the lack of early, disease-specific symptoms, 80–85% of patients are diagnosed in advanced disease stages. However, early-stage patients present a significantly more favorable prognosis, making early detection the current most effective approach to improving outcome (*Schwartz et al., 2021*).

Abdominal contrast-enhanced computed tomography (CECT) scans are usually the first line of diagnosis for PDAC, as there are no validated early diagnostic biomarkers. Studies have shown that in 16%–84% of cases, cancer signs (such as pancreatic duct cutoff/dilatation and pancreatic atrophy) can be retrospectively seen on pre-diagnostic CECT scans (3–36 months before clinical diagnosis) *(Singh et al., 2020; Toshima et al., 2021)*. Such secondary signs and early subtle focal lesions can be overlooked in clinical practice, where most abdominal CECT scans are acquired for non-pancreatic-specific indications, leading to delayed diagnosis and reduced survival.

Radiologists' performance at detecting PDAC on routine abdominal CECT is still underexplored. Current research is limited to small reader and case samples, usually from single centers, and most studies do not stratify reader performance based on tumor

size/stage. Artificial intelligence (AI) is starting to achieve expert performances in cancer diagnosis across many domains (*Esteva et al., 2017; McKinney et al., 2020*), and an increasing number of publications study AI for PDAC diagnosis on CECT (*Korfiatis et al., 2023; Chen et al., 2023; Park et al., 2023*). However, most studies do not focus on early detection and consider private, single-center data sets. Furthermore, current AI research largely lacks clinical comparison to radiologists.

The PANORAMA study is a new prospectively designed, confirmatory multi-center study to assess the performance of both radiologists and AI at PDAC detection in routine abdominal CECT scans. PANORAMA's goals are threefold: 1) to establish the clinical baseline performance of radiologists at PDAC detection through a large-scale, multi-institutional reader study, 2) to establish the state-of-the-art AI performance at PDAC detection through an international AI grand-challenge, and 3) to compare AI and radiologists.

Key aspects of the PANORAMA study design have been established in conjunction with an international scientific advisory board of 14 experts in pancreas AI, pathology, radiology, surgery and patient experience – to unify and standardize present-day guidelines, and to ensure meaningful validation of AI towards clinical translation (*Reinke et al., 2021*).

Item 3: Keywords List the primary keywords that characterize the challenge

Pancreatic cancer; artificial intelligence; computed tomography; radiologists; computer-aided detection and diagnosis.

CHALLENGE ORGANIZATION

Item 4: Organizers

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

Natália Alves[1,2*], Megan Schuurmans[1,2*], Dawid Rutkowski[3], Matthias Löhr[3], Derya Yakar[2], Ingfried Haldorsen[4], Marjolein Liedenbaum[4], Anders Molven[5], Pierpaolo Vendittelli[1], Geert Litjens[1], John Hermans[1], Henkjan Huisman[1]

1) Department of Medical Imaging, Radboud University Medical Center, Nijmegen, The Netherlands
2) Department of Radiology, University Medical Center Groningen, Groningen, The Netherlands
3) Division of Surgery & Oncology, Karolinska Institutet, Stockholm, Sweden
4) Mohn Medical Imaging and Visualization Centre, Department of Radiology, Haukeland University Hospital, Bergen, Norway
5) Department of Pathology, Haukeland University Hospital, Bergen, Norway

*These authors contributed equally

b) Provide information on the primary contact person.

Natália Alves: natalia.alves@radboudumc.nl.
Megan Schuurmans: megan.schuurmans@radboudumc.nl.

Item 5: Lifecycle 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. Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

After completion of the 2024 edition of the PANORAMA: AI Grand Challenge the Open Development Phase – Validation and Tuning leaderboard will remain live for continuous submissions. Interested teams can make submissions to the Open Development Phase – Testing leaderboard upon request (with supporting documents, e.g., institutional e-mail address, associated publication, etc.). This step is necessary to preserve the integrity of the hidden testing cohort (by avoiding overfitting). It also ensures traceability and verification of all post-challenge solutions that claim to match/outperform prior submissions on the testing dataset. The PANORAMA: AI Grand Challenge will not be a one-time event. Future iterations may explore the effects of additional information (e.g., using other contrast phases or adding magnetic resonance imaging) and more rigorous testing (e.g., larger number of external testing data centers) on AI performance and generalization.

Item 6: Challenge venue and platform

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

This challenge is not associated with any conference.

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

https://grand-challenge.org/

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

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

Item 7: Participation policies

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

This challenge only supports the submission of fully automated methods in Docker containers. It is not possible to submit semi-automated or interactive methods. All Docker containers submitted to the challenge will be executed in an offline setting (i.e., they will not have access to the internet, and cannot download/upload any resources). All necessary resources (e.g., pre-trained AI model weights) must be encapsulated in the submitted containers a priori.

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.

Use of pre-trained AI models on computer vision and/or medical imaging datasets (e.g., ImageNet, Medical Segmentation Decathlon), and use of any other dataset besides the PANORAMA training datasets, is allowed, only as long as such data and respective annotations and/or models are published under a permissive license (within 3 months of the Open Development Phase deadline), and participants clearly state their source and use-case, in each submission.

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

Members of all sponsoring or organizing entities (i.e., Radboud University Medical Center, University Medical Center Groningen, Karolinska Institute) can freely participate in the challenge, but are not eligible for the final ranking in the Closed Testing Phase or any of the prizes.

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

Money prizes will be awarded to the top five performing teams in the testing phase leaderboard:

- 1st place: €1000
- 2nd place: €500
- 3rd place: €250
- 4th place: €150
- 5th place: €100

e) Define the policy for result announcement.
Examples:

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

At the end of the Open Development Phase, all AI algorithm submissions and their respective performance will be announced publicly via its leader board. At the end of Closed Testing Phase, all AI algorithm submissions and their respective performance will be announced publicly via its leader board.

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).

Up to 3 members from each team responsible for one of the top-performing 5 AI algorithms will be invited to join the PANORAMA challenge paper as a consortium author.

Research using the PANORAMA public training dataset, is requested to cite this document, which will be published on Zenodo with a corresponding DOI. Once a study protocol and/or a challenge paper has been published, they are requested to refer to those publication(s) instead.

Item 8: 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.

The PANORAMA: AI Grand Challenge takes place in two phases:

5. **Development Phase** (Duration: 4–6 months): Anyone can participate in this phase of the challenge. Interested teams can create an account on grand-challenge.org, and register for the PANORAMA23 challenge at panorama.grand-challenge.org. Afterwards, they will be provided access to download the public training dataset, and in turn, they can start developing and training AI models using their private or public compute resources. Participants can also use additional data to train their models, but such data must be publicly available under a permissive license, and its source must be clearly stated. Participants can upload and submit their trained algorithms (in Docker containers) for evaluation (similar to the PI-CAI, AIROGS and CoNIC2022 challenges) a maximum of five times throughout the challenge. During evaluation, algorithms are executed on the grand-challenge.org platform, their performance is estimated on the hidden validation and tuning cohort, and team rankings are updated accordingly on a live, public leaderboard. Facilitating validation in such a manner, ensures that any image used for evaluation remains truly unseen, and that AI predictions cannot be tampered with —allowing for bias-free performance estimation.

6. **Testing Phase** (Duration: 1 month): At the end of the open development phase, each registered team can choose to submit a single AI algorithm (presumably their top-performing model) for evaluation on the hidden testing cohort. Based on their performance on this cohort, all new rankings will be drawn and the top 5 AI algorithms of PANORAMA23 Grand Challenge will be determined and announced. To qualify as one of these top teams, participants must also submit a short paper on their methodology (2–3 pages) and a public/private URL to their source code on GitHub — to ensure fairness, traceability and reproducibility of all proposed solutions.

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 answer to Item 8(a).

Item 9: 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

Tentative Timeline:

- February 26: Release of public training cases.
- March 18: Accepting submissions for the Development Phase – Tuning Leaderboard.
- July 1: Accepting submissions for the Testing Phase Leaderboard.
- July 8: Closing submissions for the Testing Phase Leaderboard.
- August 12: Public announcement of the Testing Phase Leaderboard and the PANORAMA: Grand Challenge winners.

Item 10: 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.

The institutional review boards of Radboud University Medical Center (RUMC), University Medical Center Groningen (UMCG), and Karolinska Intitutet (KI) have waived the need for informed patient consent, for the retrospective scientific use of anonymized clinical data in this study.

MISSION OF THE CHALLENGE

Item 14: 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

Main fields of application: diagnosis, research, risk stratification

State the task category(ies).
Examples:

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

Task categories: classification, detection, localization, prediction.

Item 16: 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 gender 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.

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

The goal of the PANORAMA challenge is to detect PDAC in routine contrast-enhanced abdominal computed tomography (CECT) scans. Target and challenge cohorts of the PANORAMA: AI Grand Challenge are similar as they both represent the same patient population: patients undergoing CECT. This target population includes patients with a healthy pancreas, patients with non-PDAC pancreatic lesions/abnormalities, and patients with PDAC, which are all represented in the challenges training, hidden validation, and

secrete testing cohorts. Nonetheless, deviations in the challenge cohorts with respect to the target cohort do exist, due to the following factors:

- The challenge cohorts are enriched with additional positives given the low incidence of PDAC in the general population to ensure the development of meaningful and discriminative AI algorithms.
- By sampling one study per patient, we increase the diversity of benign and malignant findings in the hidden testing and validation cohorts. However, in clinical practice or the target cohort, multiple studies from the same patient can be encountered.
- Excluding cases from the training datasets and the hidden testing and validation cohorts, that cannot be annotated due to incomplete imaging, poor scan quality or artifacts and ambiguous diagnostic reports.
- Including only positive cases with a histopathology confirmation of PDAC in the secret testing cohort.

Item 17: Imaging modality(ies)

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

Contrast-enhanced CT: Axial, portal-venous CT scan. All cases used for the reader study will also include clinical information as sex and age at time of imaging. Note: sagittal and coronal view CT scans are not available for all cases during any stage of the challenge. Furthermore, all other perfusion phases (I.e., arterial-, parenchyma-, delayed-phase) are not available for all cases during any stage of the challenge.

Item 18: Context information

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

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

- CT scanner vendor and model for all patients in the public training, hidden validation, and secreting testing cohort.
- Automatically generate segmentations obtained with a previously validated AI algorithm (*Alves et al., 2022*) for the following structures of interest for all patients in the public training cohort:
    - Pancreatic parenchyma.
    - Pancreatic duct.
    - Common bile duct.
    - Pancreas-surrounding veins (portal vein, superior mesenteric vein, and splenic vein).
    - Pancreas-surrounding arteries (aorta, superior mesenteric artery, celiac trunk, hepatic artery, and splenic artery.

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

- Patient age (unit: years) for all patients in the public training, hidden validation, and secreting testing cohort.
- Patient sex assigned at birth (male/female).

Item 19: 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.

Pancreas shown in contrast-enhanced CT (portal-venous phase) of the abdomen or upper abdomen.

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.

Pancreatic ductal adenocarcinoma.

Item 20: 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 (parameter 26), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find liver segmentation algorithm for CT images that processes CT images of a certain size in less than a minute on a certain hardware with an error that reflects inter- rater variability of experts.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images. Corresponding metrics are listed below (parameter 26).

The properties to be optimized for AI algorithms in the PANORAMA: AI Grand Challenge are: patient-level diagnosis (or classification) of PDAC in routine CECT; and lesion-level localization of PDAC lesions in CECT (Figure 1).

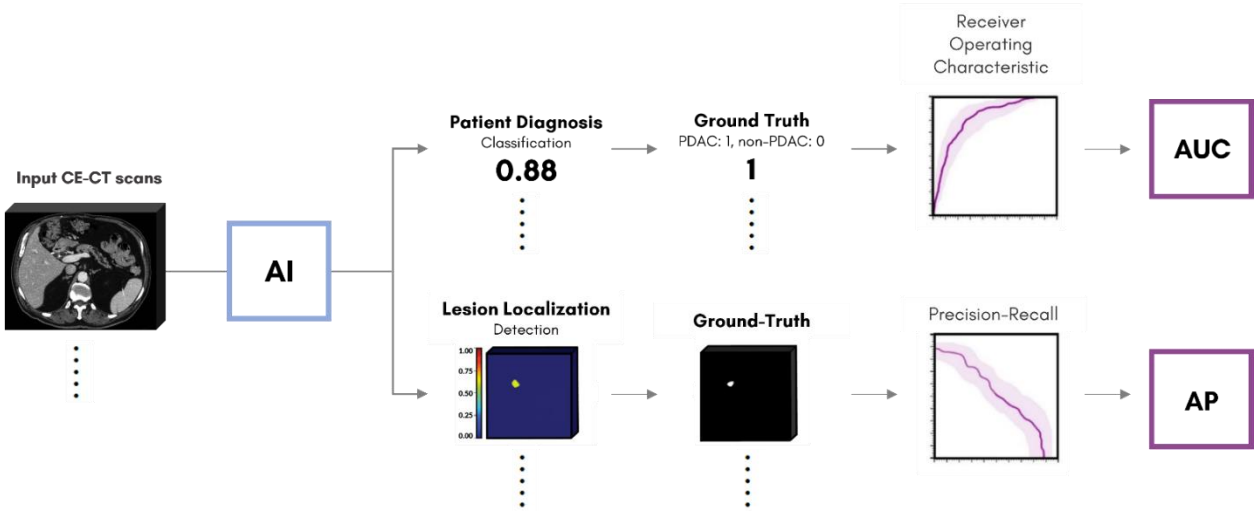

Figure 1. Overview of the AI workflow. AUC: area under the receiver operating characteristic curve; AP: average precision. AI algorithms developed for the PANORAMA: AI Grand Challenge take an abdominal CECT scan in the portal-venous phase as input and produce a patient diagnosis score and a lesion localization map. The patient diagnosis score is a floating-point number between 0 and 1. The lesion localization map is a detection map with the same dimensions of the input scan where all voxels belonging to the detected lesion are assigned a lesion likelihood value between 0 and 1.

CHALLENGE DATA SETS

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 sets will contain CECT scans acquired with several CT scanners in order to approximate real world clinical practice. These scanners include but are not limited to CT scanners from Siemens, Philips, Toshiba, GE Healthcare, and Canon.

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

All included scans are acquired with a contrast-enhancement protocol in the portal-venous phase. Series on the portal venous phase will be considered.

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.

This retrospective study includes abdominal CECT exams, acquired between 2011-2021, at two Dutch centers (Radboud University Medical Center (RUMC), University Medical Center Groningen (UMCG)), and one Swedish center (Karolinska Institute (KSKI)). Additionally, the training set includes three publicly available data sets, namely the Medical Segmentation Decathlon (MSD) pancreatic cancer (training) dataset (*Antonelli, et al. 2022*), the Cancer Imaging Archive Clinical Proteomic Tumor Analysis Consortium Pancreatic Ductal

Adenocarcinoma (TCIA-CPTAC) data set, and the National Institute of Health (NIH) pancreas CT data set (*Roth, et al. 2016*).

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 CT radiographers.

Item 22: 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) (parameter 21) and may include context information (parameter 18). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Training and test cases both represent CECT scans of the abdomen containing the pancreas. All training, validation and testing cases carry expert-derived image-level binary annotations for the presence/absence of PDAC. All validation/testing cases and between 80%–100% of public training cases also carry expert-derived voxel-level lesion delineations of PDAC lesions. Remaining public training cases carry AI-derived voxel-level lesion delineations of PDAC lesions obtained using a publicly available and previously validate AI algorithm (*Alves et al., 2022*).

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 answers to Item 21) and may include context information (see answers to Item 18). Both training and test cases are annotated in accordance with the reference standard detailed in Item 23.

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

The PANORAMA: AI Grand Challenge data set will be sampled into 4 splits according to the following use cases:

- **Public training and development set:** Contains around 1800 cases originating from RUMC, UMCG, MSD, TCIA, NIH. Used by all participants and researchers, to train and develop AI models (made available under a non-commercial CC BY-NC 4.0 license).

- **Hidden validation and tuning cohort:** Contains around 100 cases from RUMC, UMCG. Used for a live, public leaderboard that enables model tuning and selection over the course of the grand challenge.

- **Hidden Testing Cohort**: Contains around 400 cases from RUMC, UMCG, HUH, and KSI. Used to benchmark AI, and radiologists, and test all hypotheses. Includes internal testing data (unseen cases from seen centers {RUMC, UMCG}) and external testing data (unseen cases from (at least) two unseen centers {HUH, KSKI}).

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

A total of 500 cases (400 testing, 100 validation) is used for evaluation, considering both practicality (numbers of cases in our multi-center cohort, for which it would be feasible to acquire at least 36 months of follow-up data) and viability (minimum number of cases need for a reliable AI performance benchmark at present-time and through the coming years). All remaining cases in the cohort are used to create the training datasets.

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.

To meaningfully validate AI and radiologists at PDAC detection towards clinical translation, it is essential to analyze performance across the patient population encountered in clinical routine. For instance, while surgical resection can provide the most comprehensive tissue specimen to facilitate accurate histopathology grading, a cohort of solely resected patients is heavily biased, deviating substantially from the distribution of patients encountered during clinical routine. A clinically representative data set must also include cases that did not undergo surgical resection (biopsy/radiology-confirmed PDAC) as well as cases without PDAC (negative cases). It is also important to ensure that negative cases include both cases without pancreatic alterations and cases with non-PDAC lesions, as both scenarios will be routinely encountered in the clinical workflow where AI would be deployed.

Item 23: 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.

The hidden testing cohort has the highest-quality reference standard for all cases, to optimally validate AI and radiologists. As histopathology analysis is the gold standard for PDAC diagnosis confirmation, all positive PDAC cases in the hidden testing cohort will have histopathology ground truth, either through surgical resection or biopsy assessment. For the negative cases, the ground-truth label will be established through histopathology assessment (for cases with non-PDAC pancreatic lesions such as cysts and intraductal

mucinous neoplasms) and/or follow-up data. Only patients who do not develop PDAC within 36 months after their initial CT scan will be included as negative cases. For positive cases, expert-derived tumor delineations will be available.

The public training cohort is large and representative to optimally train clinically relevant AI algorithms. The PANORAMA public training and development data set will contain a combination of previous public data sets and at least 1400 new public cases derived from RUMC and UMCG (The Netherlands). Positive cases are confirmed based on histopathology when available, or based on radiology reports and follow-up from clinical routine if no biopsy or resection was performed. Negative cases will be confirmed through clinical reports. For cases deriving from existing publicly available data sets, the corresponding ground-truth will be considered. For cases in the MSD data set, differentiations between PDAC and non-PDAC cases will be made as reported by *Suman et al. 2021*. All training cases will carry patient/image-level annotations, of which, about half of the cases will also include expert-derived tumor delineations, while the remainder will include AI-derived delineations (based on a re-trained version of the method proposed by *Alves, et al. 2022*). We leave it up to the participants to formulate the most effective training strategy to utilize some/all of this data (as they best see fit), to develop their models. A subset of 100 scans with the same reference standard of the training cohort will be used as the hidden validation and tuning cohort.

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.

Each annotation is derived using CECT scans in the portal-venous phase (axial view) and diagnostic reports (radiology, pathology). Lesion delineations are created. Annotators were asked to provide the delineation of tumor lesions (identified through the diagnostic reports) using ITK-SNAP v3.80.

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.

Lesion delineations were performed by one of 4 investigators. All lesion delineations in the validation and testing sets were manually reviewed by an expert radiologist with over 20 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.

Multiple annotations were not considered for the same case.

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

Contours of all PDAC lesion delineations are susceptible to annotation errors, due to the following factors:

- Exact spatial extent of PDAC lesions cannot be clearly estimated on CECT.
- Inter-reader variability among annotators.

Note that we primarily investigate image-level classification and lesion-level detection (using a lenient hit criterion, as detailed in Item 26(a)) performance in the PANORAMA: AI Grand Challenge. We do not evaluate segmentation or the exact spatial extent of PDAC, predicted by radiologists or AI. Thus, annotation uncertainty along lesion boundaries, due to the factors listed above, have negligible impact (if any) on the outcomes of this study.

ASSESSMENT METHODS

Item 26: 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) (parameter 20). State which metric(s) were used to compute the ranking(s) (if any).

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

The key performance metrics used to evaluate AI and radiologists have been summarized in Figure 2. Patient-level diagnosis performance will be evaluated through receiver operating characteristic (ROC) curve analysis, using the area under the curve (AUROC) as a summary metric. AUROC will be derived for AI and clinicians from the respective continuous patient-level cancer likelihood scores. To compare AI sensitivity/specificity with readers, AI cancer continuous predictions will be binarized to match the sensitivity/specificity of radiologists at clinically relevant operating points. These operating points will be derived from the categorical labels in the reader study.

Lesion-level localization performance will be evaluated through precision-recall curve analysis, with the average precision (AP) as a summary metric. AP will be derived for AI from the lesion-level predictions extracted from the tumor likelihood maps. To compare AI precision/recall with readers, an operating point will be defined based on the point-annotations provided in the reader study.

Algorithms in the PANORAMA: Grand Challenge will be independently ranked according to patient-level diagnosis performance using AUROC (diagnosis rank) and lesion-level localization performance using AP (localization rank). The final overall ranking score for each algorithm will be obtained by averaging the diagnosis and localization ranks.

The PANORAMA organizers will publicly release functions to compute all performance metrics discussed in this section, in a GitHub repo.

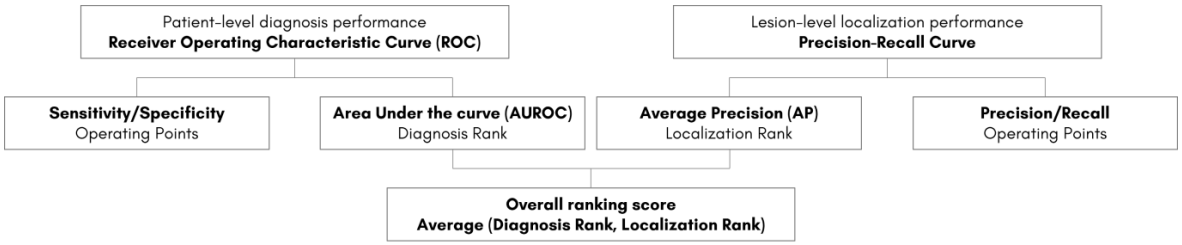

Figure 2. Performance metrics for AI and radiologists' evaluation. Overall ranking score is only used to evaluate different AI algorithms with respect to each other and facilitate the validation and testing leaderboards of the PANORAMA: Grand Challenge. When comparing AI performance to that of radiologists at specific operating points, AI predictions are binarized to match the sensitivity, specificity, precision or recall of radiologists.

Secondary Performance Metrics

Intersection over Union (IoU) is used for spatial congruence analysis of AI detections (but not for validation or testing, given that IoU cannot accurately evaluate detection or diagnosis performance (*Reinke et al., 2020*)).

Hit Criterion

A "hit criterion" is a condition that must be satisfied for each predicted lesion to count as a hit or true positive. In line with recent literature for pancreatic cancer detection AI, we opt for a hit criterion based on object overlap for the 3D detections predicted by AI:

- **True Positive**: A candidate lesion is considered a true positive if the intersection over union (IoU) with the ground truth (calculated in 3D between the whole extracted candidate lesion volume and the tumor ground truth volume) was at least 0.1. This threshold is set in line with well-established previously published studies of the same nature considering other cancer diseases (*McKinney et al., 2020; Saha et al., 2021)* as it addresses the clinical need for object-level localization, while also taking into account the non-overlapping nature of objects in 3D.

- **False Positives**: Predictions with no/insufficient overlap count towards false positives, regardless of their size or location.

For the point-coordinate annotations predicted by radiologists, we opt for a hit criterion based on distance/location:

- **True Positives:** For a predicted lesion coordinate to be counted as a true positive, it must reside ≤ 5 mm of the ground-truth tumor boundary (as done in *Saha et al. 2022*). Such a margin is considered to account for smaller lesions, where the ground-truth annotation spans few slices, and radiologists' point predictions can register a miss from marginal deviations (despite correct cognitive localization).

- **False Positives:** Predictions that are > 5 mm from the ground-truth annotation boundary count towards false positives.

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

See answers to Item 20 and Item 26(a).

Item 27: 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 answers to Item 20 and Item 26(a).

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

Submissions with missing results will be disqualified, and not presented on any leaderboard.

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

For patient pathway planning, both patient-level risk stratification and lesion-level detection are instrumental. The two are linked and important for the diagnostic process. Therefore, we opted for equal weighting between AUROC rank and AP rank.

Item 28: 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
- indication of any software product that was used for all data analysis methods.

Statistical tests are performed in the final arc of the PANORAMA study, after selecting the three top-ranking AI algorithms from the AI study (Grand Challenge) and estimating the performance of all radiologists from the reader study. Each test is facilitated using AI and/or radiologists' predictions on the hidden testing cohort.

As the primary outcome of the PANORAMA study, we statistically evaluate the diagnostic performance of a state-of-the-art artificial intelligence system in comparison to radiologists at PDAC diagnosis in CECT. We evaluate performance by calculating sensitivity, specificity and the AUROC of the AI system, according to its case-level predictions of significant cancer diagnosis. The AI system is defined as the ensemble of the top three algorithms developed for the PANORAMA23 Grand Challenge. We investigate and compare AI predictions relative to the radiology readings made during a multi-reader multi-case (MRMC) observer study with 400 testing cases. For these comparisons, histopathology and a follow-up period of ≥ 1 year are used to establish the reference standard. All study objectives are pre-specified in a hierarchical family and tested accordingly (as shown in Figure 3).

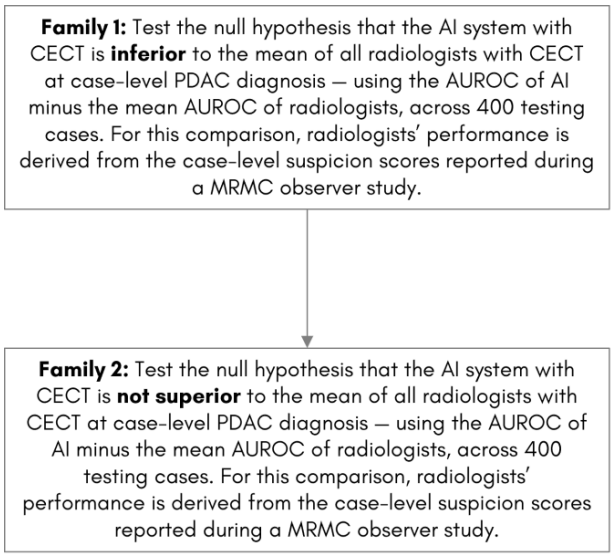

Figure 3. Flowchart illustrates the strategic plan to test study objectives. If the null hypothesis of family 1 is rejected with an alpha value of 0.05, AI non-inferiority is confirmed and we test family 2. If the null hypothesis of family 2 is then rejected with an alfa value of 0.05, AI superiority is confirmed. If the null hypothesis of family 1 is not rejected with an alpha value of 0.05 we do not test family 2.

Across the 400 patient examinations used in the MRMC observer study, we compare the case-level diagnostic performance of the AI system to the mean of the participating radiologists. We define the test statistic as the AUROC of the AI system minus the mean AUROC of the radiologists. Differences in (mean) AUROC and two-sided 95% Wald confidence intervals are computed. Non-inferiority is concluded, if the lower boundary of the two-sided 95% confidence interval for the test statistic is greater than a margin of -0.05 (established in concordance with *Saha et al. 2022*; *McKinney et al., 2020*; *Rodriguez-Ruiz et al., 2019*). If non-inferiority is concluded, then superiority of the AI system over radiologists is tested and concluded if the lower boundary of the two-sided 95% confidence interval for the test statistic is > 0. Overall, these comparisons aim to investigate whether an AI system at CECT can offer non-inferior or superior diagnostic performance compared to radiologists.

Multi Reader Multi Case Analysis

The P value for a superiority or inferiority test in the context of MRMC analysis can be calculated by estimating the variance of the difference between the AUROC of the standalone AI system and the mean AUROC of the readers. To get a P value for the null hypothesis that the AUROC of the standalone AI is inferior to the mean AUROC of the readers versus the alternative hypothesis that the AUROC of the standalone AI is non-inferior to the mean AUROC of the readers, a Wald test with prespecified non-inferiority margin δ of -0.05 will be performed:

$$Z = \frac{\hat{\theta} - \bar{X} - \delta}{\sqrt{\widehat{Var}\left(\hat{\theta} - \bar{X}\right)}}$$

Where $\hat{\theta}$ is the estimated AUROC of the standalone AI system and $\bar{X}$ is the estimated mean AUROC of the readers. The variance of the difference between the AUROC of the standalone AI system and the mean AUROC of the readers is calculated as follows:

$$\widehat{Var}\left(\hat{\theta} - \bar{X}\right) = \mathrm{SE}^2(\hat{\theta}) + \mathrm{SE}^2(\bar{X}) - 2r\ \mathrm{SE}(\hat{\theta})\ \mathrm{SE}(\bar{X})$$

The standard error for the standalone AI system, $SE(\hat{\theta})$, and the standard error for the readers, $SE(\bar{X})$, are calculated using the analysis of variance Obuchowski-Rockette (ANOVA OR) method (Obuchowski, et al., 1995). r is the Pearson correlation coefficient between the AUROC of the standalone AI system and the mean AUROC of the readers. To estimate r, 1000 bootstrap samples are performed, with replacement of cases and readers. For each bootstrap sample, $\hat{\theta}$ and $\bar{X}$ are stored. Then, the correlation between $\hat{\theta}$ and $\bar{X}$ from the bootstrap samples is calculated to give r. For superiority testing, the equations remain unchanged, except that δ is set to zero. The P value is calculated from the Z score using a one-tailed P value.

Item 29: 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
- ranking variability

Reader Study

Parallel to the AI study, the PANORAMA: Reader Study aims to evaluate the performance of abdominal radiologists with various levels of expertise at patient-level diagnosis of PDAC in abdominal CECT scans. Radiologists receive the same information as AI and provide a patient-level likelihood score for the presence of PDAC between 0 and 100 (with incremental steps of 1), together with a point coordinate at the center of the tumor. Readers will answer the binary question: would you recommend this case for PDAC

assessment? (0-No, 1-Yes). Recommendation for PDAC assessment means requesting urgent additional imaging (for instance endoscopic ultrasound or MRI) and/or lesion biopsy, and directing the case to the multidisciplinary pancreatic tumor board. Recommendation for PDAC assessment does not include non-urgent follow-up such as recommending follow-up imaging over one month from the present examination, as this would not be a suitable clinical course for a patient with PDAC suspicion due to the aggressiveness and fast pace of the disease.

Standardized reporting and data systems (RADS) have been widely adopted for imaging-based cancer diagnosis across many domains such as liver (LI-RADS), colon (C-RADS), breast (BI-RADS), and prostate (PI-RADS). However, in the absence of large-scale multi-reader multi-case studies there are currently no categorical scales for PDAC diagnosis on CECT. To investigate the potential for standardized reporting on PDAC diagnosis, we ask readers to classify cases into 5 different risk levels, defined in line with clinically validated RADS for other cancer diseases (*An, et al., 2019*).

- **Category 1**: Very low PDAC suspicion.
- **Category 2:** Low PDAC suspicion
- **Category 3:** Equivocal
- **Category 4:** High PDAC suspicion
- **Category 5:** Very high PDAC suspicion

These discrete categories will be used to establish the clinical operating point at which sensitivity and specificity of radiologists will be compared to AI algorithms.

The PANORAMA: Reader Study is hosted on grand-challenge.org/reader-studies, which supports abdominal CT viewers and annotation tools, using the 400 cases on the hidden testing cohort. Primary costs of the study are reader time and effort. As reading all 400 cases is too labor-intensive, we opt for a split-plot design (*Obuchowski et al., 2009*; *Obuchowski et al., 2012*; *Chen et al., 2018*). By doing so, we preserve a sample size of 400 cases, while the individual workload is reduced to a maximum of 100 cases (as illustrated in Figure 4). Prior to the start of the study, readers are provided a detailed guide on the annotation workflow, including all tools made available on the platform (e.g., navigation, zooming, windowing, measuring scale). They are also provided access to a practice session with 5 example cases (from the MSD data set), to get familiarized with both the reading interface and the expected workflow. Afterwards, each reader is provided a unique, password-protected, non-transferable URL to access their instance of the reader study (with all 100 allotted cases). Cases are made available sequentially and cannot be revisited post-assessment. Readers are expected to complete their assessments in 3-5 months.

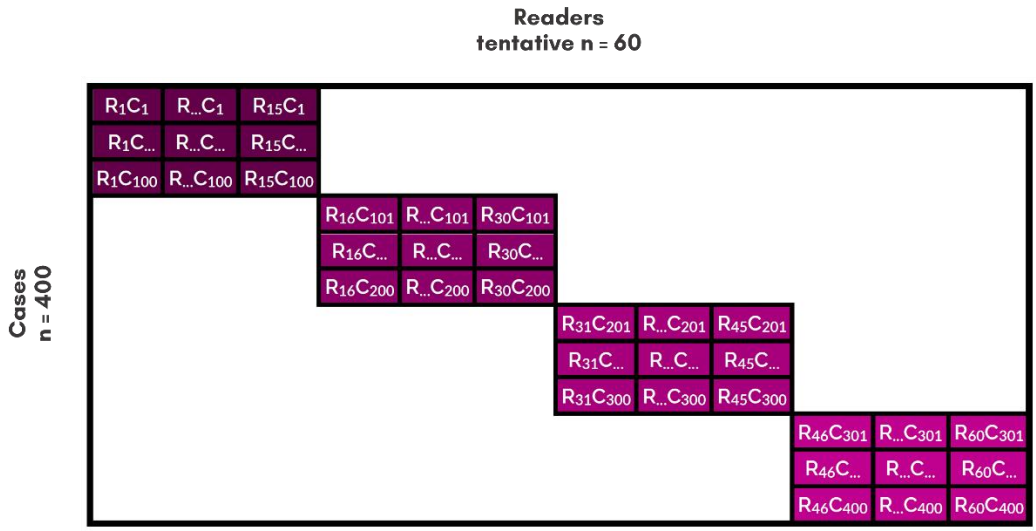

Figure 4. Tentative study design for the distribution of readers and cases in a 4x4 split-plot configuration. All 60 readers and 400 cases are divided into 4 blocks, in a stratified manner, that takes reader and case distributions into account to minimize any potential differences between separate blocks. Each block of readers reads their own set of cases. As this study design is reader-dependent, it is susceptible to changes based on the final outcomes of reader recruitment.

Scientific Advisory Board

Nancy Obuchowski, Ph.D.; Elliot K. Fishman, M.D.; Caroline Verbeke, M.D., Ph.D; Namkug Kim, Ph.D.; Steven Gallinger, M.D; Celso Matos, M.D., Ph.D.; Garima Suman, M.D.; Klaus H. Maier-Hein, Ph.D.; Horst Hahn, Ph.D.; Matthias Löhr, M.D., Ph.D.; Weichung Wang, Ph.D.; Ali Stunt;