

# EchoRisk-MICCAI: AI for Cardiac Function Estimation, Assessment and Early Prediction of Therapy-Induced Cardiotoxicity from Echocardiography: Structured description of the challenge design

## CHALLENGE ORGANIZATION

### Title

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

EchoRisk-MICCAI: AI for Cardiac Function Estimation, Assessment and Early Prediction of Therapy-Induced Cardiotoxicity from Echocardiography

### Challenge acronym

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

EchoRisk-MICCAI

### 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.

This Challenge Proposal for MICCAI 2026 focuses on the automated analysis of echocardiography videos for the early detection and prediction of cancer therapy-induced cardiotoxicity in breast cancer patients. The proposed challenge builds on the CARDIOCARE project, a multidisciplinary, EU-funded initiative focused on improving cardiovascular outcomes in older and multimorbid women undergoing breast cancer therapy. CARDIOCARE has established a prospective, longitudinal clinical study across 6 European hospital sites in 5 countries, systematically collecting real-world echocardiography imaging, standardized clinical metadata and biomarkers of cardiac injury. The study is specifically designed to capture the onset and predict the progression of cardiotoxic effects from anthracyclines and HER2-targeted therapies, two treatments known to induce cardiac dysfunction. The resulting dataset provides a unique and clinically rich foundation for developing and benchmarking AI algorithms that aim to identify early subclinical changes in cardiac function, support risk stratification, and guide advanced personalized monitoring strategies in cardio-oncology. Cardiotoxicity is a critical and growing concern in oncology, representing a major dose-limiting and treatment-interrupting complication of life-saving therapies, such as anthracyclines and HER2-targeted agents. Studies have shown that up to 20–30% of patients receiving anthracyclines and 7–10% receiving trastuzumab develop some form of cardiac dysfunction. Early detection is essential, as even subclinical dysfunction is associated with increased long-term cardiovascular morbidity and mortality, particularly in older and multimorbid women; a population underrepresented in clinical trials but highly vulnerable in real-world settings. Troponin I and NT-proBNP are early surrogate biomarkers of cardiotoxicity, often rising before imaging changes, and supporting early risk stratification. Echocardiography is the standard of care for cardiac monitoring due to its non-invasive nature, widespread availability, real-time imaging capabilities,

and cost-effectiveness. However, in clinical workflows, its full potential is hampered by several limitations. First, key cardiac function parameters, such as left ventricular ejection fraction (LVEF) and global longitudinal strain (GLS) are typically derived through manual tracing or semi-automated tools, which are highly dependent on operator skill and training, leading to substantial inter- and intra-operator variability, especially in challenging patient anatomies/low-quality acquisitions. Secondly, image quality is affected by body habitus, breathing, and probe angulation, introducing noise and inconsistencies across patients and timepoints. Third, lack of standardization in acquisition protocols across institutions and manufacturers complicates the comparison and interpretation of results in longitudinal monitoring. This variability reflects real-world clinical practice and is intentionally preserved in the dataset to enable evaluation of algorithm robustness across heterogeneous acquisition environments. In multicenter studies or routine clinical practice, these issues result in limited reproducibility, often requiring repeated scans/additional imaging modalities, which may not be accessible in many settings. Moreover, the manual nature of the workflow makes it difficult to scale high-frequency monitoring in vulnerable populations, such as older patients receiving cardiotoxic chemotherapy. As a result, early signs of cardiac dysfunction are often missed, underestimated, or identified too late to prevent irreversible damage. The MICCAI community has made major contributions to medical image segmentation, cardiac function estimation, and prognostic modeling. Despite this progress, robust and deployable AI models for automated cardiac function assessment and risk prediction in echocardiography remain scarce due to the lack of standardized, large-scale, and clinically grounded benchmark specifically addressing the assessment and prediction of therapy-induced cardiotoxicity from echocardiography videos. This challenge would fill that gap by providing the first curated, multicenter dataset tailored for evaluating AI models on tasks directly tied to clinical decision-making in cardio-oncology. The CARDIOCARE echocardiography dataset is the first imaging database to incorporate explicit information on cardiotoxicity, including labels directly related to therapy-induced cardiotoxicity. The challenge will encourage the development of algorithms that are not only accurate, but robust, fair, interpretable, and capable of generalizing across different data sources, consistent with MICCAI's 2026 priorities. It will also facilitate comparisons between deep learning methods, Foundational models and traditional image analysis techniques, and hybrid approaches that leverage both imaging and clinical metadata.

**Dataset Description.** The challenge will use imaging data from CARDIOCARE prospective clinical study that includes echocardiography data from 6 major European cancer and cardiology centers across 5 countries; IEO (Italy), BOCOC (Cyprus), KSBC (Sweden), NKUA (Greece), and UOI (Greece) and IOL (Slovenia), collectively contributing over 421 patients.

**Data Types.** The dataset will consist of: (i) 2D echocardiographic sequences (grayscale DICOM videos capturing at least one full representative cardiac cycle) extracted for two standard views: apical 4-chamber and 2-chamber in multiple time points. (ii) Imaging collected over time, including baseline, early-treatment, and follow-up scans. (iii) Blood biomarkers related to cardiotoxicity (e.g., elevated troponin and/or NT-proBNP). (iv) Global longitudinal strain (GLS).

**Dataset Characteristics.** (i) Multicenter, multi-vendor acquisition, with substantial variability in ultrasound equipment and operator technique, enabling thorough robustness evaluation. (ii) Real-world clinical quality, featuring artifacts, heterogeneous framing, variable image quality, and occasional missing views, representative of routine cardiology workflows. (iii) GDPR-compliant, fully de-identified imaging, curated according to established DICOM anonymization standards. (iv) Adequate sample size (data from 421 patients across multiple timepoints; 3-month follow-ups resulting in a total number of 1528 echocardiography videos) to support AI model development.

**Proposed Challenge Tasks.** To align with MICCAI's emphasis on clinical translation, fairness, and robustness, we propose three tasks. The three tasks are intentionally defined using distinct prediction paradigms: Task 1 focuses on regression-based estimation of quantitative cardiac parameters and biomarkers, whereas Tasks 2 and 3 address classification and risk prediction problems producing calibrated probabilities of clinically defined outcomes.

(i) Task 1: Cardiac parameters and biomarkers estimation from echocardiograms. Participants

will develop AI algorithms using data from 421 patients (1528 echocardiography videos, across 5 different follow-up timepoints) to estimate key cardiac parameters and biomarkers directly from echocardiography videos. Expected outputs include LV end-diastolic and end-systolic volumes, ejection fraction, and blood biomarkers related to cardiotoxicity, i.e., elevated troponin and elevated NT-proBNP (when available). This task may involve segmentation, tracking, and video-based regression and can be partially aligned with EchoNet Dynamic data (<https://echonet.github.io/dynamic/index.html>). (ii) Task 2: Assessment of LV cardiac dysfunction (as defined by clinically accepted EF and GLS thresholds). Participants will develop predictive models that identify patients with cardiac dysfunction during or after therapy using data from 421 patients (1528 echocardiography videos, across 5 different follow-up timepoints). Participants are required to output a calibrated probability of LV dysfunction per examination. Ranking will be based on AUC-ROC as the primary metric, with balanced accuracy used as a tie-breaker. Sensitivity at a fixed specificity (90%) will be reported as a secondary clinically interpretable metric. (iii) Task 3: Early Prediction of Therapy-Induced Cardiotoxicity. Participants will develop predictive models that predict cardiotoxicity from baseline echocardiography videos at a future time point using data from 254 patients. Participants must output a baseline cardiotoxicity risk probability. Ranking will be based on AUC-ROC as the primary metric, with balanced accuracy and sensitivity used as tie-breakers. For contextualization, at least one literature-based reference baseline will be defined per task (e.g., EchoNet-style video models for functional estimation and ICOS-based risk modeling for cardiotoxicity prediction). More specifically, Task 1 will include an EchoNet-style reference baseline for functional estimation, Task 2 a guideline-aligned LVEF <50% dysfunction baseline, and Task 3 an HFA-ICOS-style clinical risk baseline. Each task has its own evaluation protocol, statistical analysis plan, leaderboard, and awards, and is evaluated independently. No aggregated cross-task ranking will be computed. This structure allows focused methodological contributions while ensuring fair, transparent, and task-specific evaluation. Cancer therapy-related cardiotoxicity remains a major cause of morbidity and treatment interruption in oncology. Although echocardiography is routinely used for monitoring, clinical decisions are constrained by measurement variability and delayed recognition of subclinical dysfunction. The EchoRisk-MICCAI challenge aligns algorithm evaluation with clinically actionable thresholds rather than abstract metrics alone while retaining standard benchmarking metrics for fair comparison. For functional estimation, performance is interpreted within guideline-informed margins. Absolute LVEF errors within 5 percentage points are considered clinically acceptable, while deviations beyond 10 percentage points may alter management around key cut-offs. These thresholds are consistent with commonly reported inter-observer variability in routine echocardiography and therefore provide clinically interpretable reference points for algorithm evaluation. A relative GLS reduction greater than 15 percent is treated as clinically meaningful. For dysfunction detection and early risk prediction, evaluation extends beyond AUC to clinically relevant operating points, prioritizing sensitivity to minimize missed high-risk patients. By embedding decision-oriented thresholds and real-world variability into evaluation, the challenge benchmarks models according to their potential to influence clinical management and prevent irreversible cardiac dysfunction. The dataset is being collected under approved ethics protocols of the CARDIOCARE project consortium, following: (i) fully anonymized processes according to GDPR and institutional guidelines and, (ii) adequate processing to remove identifiable metadata. Dataset release to challenge participants will occur only after completion of the required secondary-use approvals from participating institutions. The final dataset release plan will follow MICCAI standards, including training/validation/testing partitions with secure test-set handling. All reference measurements are derived from routine clinical workflows across multiple centers and vendors and therefore reflect real-world acquisition and measurement variability rather than idealized research-grade annotations. For training purposes, we encourage participants to use other public datasets and echocardiography Foundational models. Participants using external public datasets or foundation models for pretraining or model development will be required to disclose them, and external retrieval

or internet access during inference will not be permitted. Organizing Team Qualifications. The organizing team combines expertise in: (i) Cardiovascular imaging, (ii) Oncology and cardio-oncology clinical practice, (iii) Medical AI research, including segmentation, video analysis, prognostic modeling, multimodal learning, and Foundational models, (iv) Responsible AI, robustness evaluation, and fairness frameworks and, (v) Large-scale challenge organization and data curation, to ensure the challenge is scientifically rigorous and clinically aligned.

### Challenge keywords

List the primary keywords that characterize the challenge.

Echocardiography video analysis, Therapy-induced cardiotoxicity, Cardiac function assessment, Real-world clinical data

### Year

2026

### Novelty of the challenge

Briefly describe the novelty of the challenge.

First challenge in therapy-induced cardiotoxicity from echocardiography: Introduction of the first curated, multicenter, longitudinal echocardiography video dataset explicitly labeled for cancer therapy-induced cardiotoxicity, enabling standardized benchmarking of AI methods for early detection and prediction of cardiac dysfunction in cardio-oncology.

Clinically grounded integration: Unlike existing echocardiography benchmarks focused on generic cardiac function estimation, the challenge integrates real-world, multi-vendor imaging with longitudinal follow-up and cardiotoxicity-specific clinical endpoints and biomarkers, directly linking AI outputs to clinically actionable decisions.

Forward-looking task design: The proposed tasks move beyond retrospective assessment by emphasizing early prediction from baseline imaging, robustness across centers, and multimodal learning, addressing a critical unmet need for deployable and clinically reliable AI in echocardiography.

### Task description and application scenarios

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

**Task 1 – Automated Estimation of Cardiac Functional Parameters and Biomarkers from Echocardiography:** This task supports the deployment of scalable and standardized echocardiography analysis by enabling automated estimation of key cardiac parameters (e.g., LV volumes, LVEF, GLS) and cardiotoxicity-related biomarkers directly from routine echocardiography videos. In clinical practice, such automation reduces operator dependency and variability, facilitates high-throughput cardiac monitoring during cancer therapy, and supports consistent longitudinal assessment across centers and vendors.

**Task 2 – Automated Assessment of Left Ventricular Dysfunction from Longitudinal Echocardiography:** By focusing on the assessment of left ventricular dysfunction using clinically accepted LVEF and GLS thresholds, this task directly addresses the need for early and reliable detection of cardiac impairment during or after cancer therapy. The resulting models can support clinical decision-making by flagging patients at risk, enabling timely

referral to cardio-oncology services, therapy adjustment, or initiation of cardioprotective strategies.

### Task 3 – Early Prediction of Therapy-Induced Cardiotoxicity from Baseline Echocardiography:

This task targets prospective risk stratification by predicting future cardiotoxicity from baseline echocardiography videos, reflecting a real-world clinical scenario at therapy initiation. Such predictive models have direct application in personalized treatment planning, intensified surveillance of high-risk patients, and prevention of irreversible cardiac damage, supporting proactive rather than reactive management in cardio-oncology.

## FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

### Workshop

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

N/A

### Duration

How long does the challenge take?

Half day

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

A half-day duration is requested to ensure adequate time for the effective presentation, discussion, and scientific exchange around this clinically and technically complex challenge. The proposed challenge includes three distinct but interrelated tasks; cardiac parameter and biomarker estimation, assessment of left ventricular dysfunction, and early prediction of therapy-induced cardiotoxicity, each addressing different methodological aspects (segmentation, video analysis, multimodal learning, and prognostic modeling) and levels of clinical relevance. A half-day format challenge allows sufficient time to clearly introduce the clinical background, dataset characteristics, and evaluation protocols for each task, ensuring that results are interpreted correctly and consistently by the audience. The proposed half-day slot will enable: (i) presentation of top-performing methods across all tasks, (ii) comparative analysis and cross-task insights, (iii) a moderated discussion involving clinicians and AI researchers on clinical translation and deployment challenges, (iv) interaction with participants, including Q&A; and feedback on future benchmark extensions.

### 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.

Based on participation levels observed in comparable MICCAI cardiac imaging challenges, we estimate that the proposed EchoRisk-MICCAI Challenge will attract approximately 25–40 participating teams. Previous MICCAI challenges in cardiac imaging and echocardiography provide a strong empirical basis for this estimate. The ACDC Challenge (MICCAI 2017), focusing on automated cardiac diagnosis from cine MRI, reported approximately 30 participating teams. The CAMUS Challenge (MICCAI 2020), which addressed echocardiography-based left ventricular segmentation, attracted more than 20 international teams, despite being limited to a single modality and task. The M&Ms; Challenge (MICCAI 2020), centered on multi-center and multi-vendor cardiac MRI segmentation, reported around 40 registered teams, reflecting strong interest in clinically grounded, heterogeneous datasets. Similar participation levels have also been observed in more recent cardiac challenges

focusing on robustness, generalization, and clinical translation. The proposed challenge is expected to attract comparable or greater interest for several reasons. Foremost, it addresses a timely and clinically unmet need: the early detection and prediction of cancer therapy-induced cardiotoxicity, a problem of growing importance that has not previously been the focus of a MICCAI challenge. Second, it is based on a unique, multicenter, real-world echocardiography dataset with explicit cardiotoxicity labels and longitudinal follow-up, which is currently not available in any existing public benchmark. Third, the challenge design spans multiple tasks (cardiac parameter estimation, dysfunction assessment, and early prediction), appealing to a broad spectrum of researchers working on segmentation, video analysis, multimodal learning, foundation models, and explainable AI. In terms of potential participants, we expect strong engagement from: (i) academic research groups active in cardiac imaging, echocardiography, and medical video analysis, (ii) teams that have previously participated in MICCAI cardiac challenges (iii) groups working on foundation models and multimodal learning applied to medical imaging, (iv) early-career researchers and PhD students seeking clinically meaningful benchmark problems.

### **Publication and future plans**

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

Individual challenge papers will be published in the LNCS proceedings as part of the STACOM workshop, ensuring timely dissemination of methodological advances and reproducibility of results within the MICCAI community. Beyond the workshop proceedings, we plan to coordinate a comprehensive challenge summary paper that consolidates the results across all tasks, including comparative performance analysis, robustness and generalization assessment across centers and vendors, and insights into clinically relevant failure modes. This paper will go beyond leaderboard reporting and focus on methodological trends, clinical interpretability, and translational relevance of AI models for cardiotoxicity detection and prediction. The challenge summary paper will be submitted to a high-impact, peer-reviewed journal with strong visibility in both the medical imaging and clinical communities, such as Nature Medicine, Nature Machine Intelligence, Medical Image Analysis, IEEE Transactions on Medical Imaging, or the Journal of Cardiovascular Magnetic Resonance. Similar challenge summary papers published in these venues have demonstrated substantial impact by defining state-of-the-art baselines, shaping subsequent research directions, and serving as reference benchmarks for years after publication. In addition to the primary publication, the challenge outcomes are expected to generate long-term impact by: (i) establishing the first standardized benchmark for AI-based assessment and early prediction of cancer therapy-induced cardiotoxicity from echocardiography, (ii) providing validated performance baselines that can be reused by future studies and regulatory-oriented evaluations, (iii) informing the design of clinically deployable and trustworthy AI models, (iv) fostering sustained collaboration between the MICCAI community and cardio-oncology clinicians, accelerating clinical translation.

### **MICCAI LNCS proceedings**

Indicate if you want to offer MICCAI Springer LNCS proceedings to the participants. Publishing a proceedings volume is optional and at the discretion of each challenge's organizers. At a minimum, organizers must ensure that a description of each participant's submission is publicly available. Organizers who wish to publish MICCAI Springer LNCS proceedings must adhere to the MICCAI Satellite events publication process.

Yes

### **Collaboration with European Society of Radiology (ESR)**



In collaboration with European Society of Radiology (ESR), we announce special clinical interest topics with associated clinicians who can help with the preparation of the proposals; the best 3 challenge proposals on these topics will get the opportunity to present their challenges at the European Congress of Radiology (ECR) 2027 in a special session. If you want to organize a challenge in collaboration with ESR on one of these topics, please reach out to the MICCAI Challenges Team ([miccai-challenges-2026@dkfz-heidelberg.de](mailto:miccai-challenges-2026@dkfz-heidelberg.de)) and we will put you in contact with the corresponding clinician.

Challenge in collaboration with ESR. Ticking 'Yes' implies that the challenge has been prepared in collaboration with the clinical contact point.

No

### **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.

At minimum, each submission will be provided with access to one dedicated GPU with at least 16 GB of VRAM, multi-core CPU resources, and at least 32 GB of system memory. These specifications represent the minimum guaranteed computational resources available for submission evaluation. Final hardware specifications (including exact GPU model, CPU configuration, and memory allocation) are currently being finalized in coordination with the hosting infrastructure and will be publicly announced no later than three months before the test submission deadline. All submissions must complete inference within the allocated computational resources. External API calls, internet access, or retrieval from external systems during inference will not be permitted.

## **TASK 1: Automated Estimation of Cardiac Functional Parameters and Biomarkers from Echocardiography**

### **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.

Echocardiography is the most widely used imaging modality for assessing cardiac structure and function in clinical practice, owing to its real-time capability, portability, and absence of ionizing radiation. In oncology patients, and particularly in breast cancer populations receiving potentially cardiotoxic therapies, echocardiography plays a central role in longitudinal monitoring for early signs of cardiac dysfunction. However, the extraction of clinically relevant cardiac functional parameters remains highly operator-dependent, time-consuming, and subject to substantial inter- and intra-observer variability, limiting its scalability for large-scale longitudinal assessment and early risk stratification.

This task aims to advance automated, video-based analysis of echocardiography by challenging participants to develop AI algorithms that directly estimate key cardiac functional parameters and cardiotoxicity-related biomarkers from echocardiography videos. Using a unique, longitudinal dataset from 421 patients (1,528 echocardiography videos acquired across five standardized follow-up timepoints), participants will estimate left ventricular end-diastolic and end-systolic volumes, ejection fraction, and clinically relevant blood biomarkers associated with cardiotoxicity, including troponin and NT-proBNP when available. The dataset reflects real-world clinical heterogeneity in acquisition conditions, patient characteristics, and disease progression, emphasizing robustness and generalizability.

From a technical perspective, this task encompasses multiple challenging aspects of medical video analysis, including cardiac chamber segmentation, temporal tracking of myocardial motion, and regression of continuous clinical parameters from noisy and variable ultrasound data. Participants may leverage spatiotemporal deep learning architectures, self-supervised or multi-task learning strategies, and uncertainty-aware modeling to bridge the gap between imaging-derived features and downstream functional and biochemical indicators. The task is partially aligned with existing benchmarks such as EchoNet-Dynamic, while extending beyond them by introducing longitudinal follow-up data and the prediction of blood-based biomarkers, thereby increasing both clinical relevance and methodological complexity.

The envisioned impact of this task is twofold. Biomedically, it aims to enable scalable, objective, and reproducible assessment of cardiac function and pathology in patients undergoing cancer therapy, facilitating earlier detection of cardiotoxic effects, and supporting personalized treatment decisions. Technically, the task promotes the development of advanced AI methods for end-to-end echocardiography video analysis that integrate functional imaging and surrogate biomarker estimation. By bridging imaging, functional assessment, and biochemical risk indicators within a unified framework, this challenge seeks to accelerate the translation of AI-driven echocardiography analysis into routine clinical workflows and longitudinal cardio-oncology care.



**Keywords**

List the primary keywords that characterize the task.

Echocardiography, cardiac function estimation, longitudinal analysis, cardio-oncology

**ORGANIZATION****Organizers**

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

Kostas Marias

Foundation for Research and Technology - Hellas

Manolis Tsiknakis

Hellenic Mediterranean University

Grigorios Kalliataakis

Foundation for Research and Technology

Georgios Manikis

Foundation for Research and Technology - Hellas

Georgia Karanasiou

University of Ioannina

Eleni Georga

University of Ioannina

Katerina Naka

University Hospital of Ioannina

Dorothea Tsekoura

National and Kapodistrian University of Athens

Gerasimos Filippatos

National and Kapodistrian University of Athens

Anastasia Constantinidou

Bank of Cyprus Oncology Centre

Andri Papakonstantinou

Karolinska University Hospital

Ketti Mazzocco

Istituto Europeo di Oncologia IRCCS Milano

Domen Ribnikar  
Institute of Oncology Ljubljana

Dimitrios Fotiadis  
University of Ioannina

b) Provide information on the primary contact person.

Kostas Marias, kmarias@ics.forth.gr  
Foundation for Research and Technology - Hellas

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

Yes, clinicians are integral members of the organizing team and play a central role in the clinical design, validation, and interpretation of the challenge. Their involvement ensures that the challenge addresses clinically relevant questions, reflects real-world clinical practice, and that the chosen evaluation metrics and outcomes are meaningful, interpretable, and impactful for patient care.

Katerina Naka (University Hospital of Ioannina) – Clinical lead for cardiac imaging data curation and annotation. She oversees the clinical validation of echocardiographic datasets, contributes to defining clinically relevant endpoints, and supports interpretation of challenge outcomes from a cardiology perspective.

Dorothea Tsekoura (National and Kapodistrian University of Athens) – Clinical advisor for cardio-oncology pathways and patient stratification. She contributes to the definition of inclusion/exclusion criteria, clinical metadata harmonization, and ensures alignment with current cardio-oncology guidelines.

Gerasimos Filippatos (National and Kapodistrian University of Athens) – Senior clinical advisor and scientific oversight. He provides high-level clinical guidance on heart failure and therapy-related cardiac dysfunction, ensuring the clinical credibility and translational relevance of the challenge.

Anastasia Constantinidou (Bank of Cyprus Oncology Centre) – Oncology clinical advisor. She supports the integration of oncological treatment context, therapy exposure variables, and clinically meaningful outcomes related to cancer therapy-related cardiotoxicity.

Andri Papakonstantinou (Karolinska University Hospital) – Clinical expert for international validation and benchmarking. She supports interpretation of results in a broader clinical context.

Ketti Mazzocco (Istituto Europeo di Oncologia IRCCS, Milan) – Clinical advisor for patient-centered outcomes. She provides input on clinically relevant outcome measures and supports alignment with patient-reported and longitudinal assessment perspectives.

Domen Ribnikar (Institute of Oncology Ljubljana) – Clinical advisor for real-world oncology practice. He advises on translational applicability of the developed methods.

## 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).

29th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2026)

b) Report the platform used to run the challenge.

The challenge will be run on the Synapse platform (<https://www.synapse.org>). Synapse provides a robust and widely used infrastructure for running biomedical data challenges, supporting secure data hosting, participant registration, submission management, leaderboard evaluation, and transparent benchmarking. The platform is particularly well suited for challenges involving sensitive clinical data, offering fine-grained access control, governance mechanisms, and compliance with ethical and data-use requirements. The use of Synapse ensures a secure, transparent, and reproducible challenge execution, aligned with MICCAI policies and best practices for clinical AI benchmarking.

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

Yes

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

The challenge will be hosted on Synapse (Sage Bionetworks), which will serve as the central platform for data access, documentation, submissions, and leaderboards. A public placeholder Synapse project has been created and will be populated with all challenge materials upon acceptance: [Synapse Project URL <https://www.synapse.org/Synapse:syn72001386>]

### Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

Fully Automatic.

All submitted methods must operate in a fully automatic manner at inference time, with no user interaction or

manual intervention on validation or test data. This includes no manual selection of frames, regions of interest, contours, or case-specific parameter tuning. During training, participants are allowed to perform standard data curation, preprocessing, and augmentation on the provided training set, as well as pre-train models using external public datasets or foundation models. Participants using external public datasets or foundation models for pretraining or model development will be required to disclose them in their submission description. External retrieval, API calls, or internet access during inference will not be permitted.

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 may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

#### **Publicly available data is 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.

#### **May participate but not eligible for awards and not listed in leaderboard**

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

The challenge will recognize outstanding contributions based on the official final leaderboard ranking. The top five teams will be invited to present their methods and results through oral presentations (in person or online) during the associated MICCAI 2026 challenge session/workshop. In addition, selected teams will be invited to contribute to a joint challenge summary paper, which will be submitted to a high-impact peer-reviewed journal following completion of the challenge. Official certificates of achievement will be awarded to all top-ranked teams. To further acknowledge excellence, monetary awards totaling €1,200 will be granted and distributed as follows: €600 for 1st place, €400 for 2nd place, and €200 for 3rd place. Overall, this award policy aims to promote scientific excellence, reproducibility, and active community engagement, while ensuring a transparent and fair evaluation process in line with MICCAI challenge best practices. Rankings are computed and presented separately for each task. No aggregated cross-task ranking will be generated. Awards are granted per task based exclusively on the respective task leaderboard derived from hidden test-set performance.

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.

**One official leaderboard will be generated per task based on performance on the hidden test set. Test-set results will be released at the official challenge result announcement.**

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).

Upon completion of the challenge, the organizers will coordinate a comprehensive peer-reviewed publication summarizing the challenge design, key analyses, and results. The paper will be a collaborative effort involving the

challenge organizing team, clinical and technical collaborators, and the top-performing participating teams. Each eligible team will be invited to nominate up to three co-authors, provided that the team has made a valid submission and complied with all challenge rules. Participating teams remain free to publish their own methods and results independently; however, to preserve the integrity of the challenge outcomes, an embargo period of three months following the release of the official challenge results will apply. The organizers reserve the right to exclude teams from co-authorship in the challenge paper in cases of rule violations or non-compliant submissions. All teams submitting a valid final test-set submission are eligible to submit a workshop paper to the associated LNCS proceedings, subject to standard peer review. Top-ranked teams per task will additionally be invited to contribute to the joint challenge summary paper, provided they comply with challenge rules and submission requirements.

### 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.

Submissions will be accepted in the form of Docker containers via the Synapse platform. Detailed submission instructions and technical requirements will be provided at the time of the official challenge announcement. All submissions are evaluated automatically on the challenge server. During the development phase, submissions are evaluated on the validation set with visible ground truth labels and publicly displayed metrics. During the final evaluation phase, submitted containers are executed on the hidden test set, and no test performance metrics are visible until official results are announced.

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.

Participants may submit up to three development runs on the validation dataset to verify correctness and obtain performance feedback. Validation metrics are returned privately to the submitting team and are intended exclusively for model development and debugging purposes. Validation results are not used for official ranking and will not be used to determine awards. Final rankings are based exclusively on performance on the hidden test set. Only the final submission per team before the deadline will be evaluated on the test set and used for official ranking.

### 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

April 1, 2026: Challenge website opens for registration; release of training and validation data

April 10, 2026: Submission system opens for validation submissions

July 15, 2026: Submission system opens for test submissions

August 20, 2026: Registration and Docker submission deadline

October 7, 2026: Release of final results during the MICCAI Annual Meeting

January 31, 2027: Publication of the challenge summary

## Ethics approval

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

The dataset used in the EchoRisk-MICCAI Challenge originates from the CARDIOCARE project (Horizon 2020, Grant Agreement No. 945175), a prospective, multicenter clinical study conducted across six European hospital sites in five countries. The CARDIOCARE study received ethical approval from the relevant Institutional Review Boards (IRBs) / Ethics Committees at each participating clinical center, in accordance with national regulations, the Declaration of Helsinki, and the EU General Data Protection Regulation (GDPR). Ethics approvals were obtained locally at each recruiting institution prior to patient enrollment and data collection, covering: (i) prospective acquisition of echocardiography imaging, (ii) collection of clinical and biomarker data, and (iii) longitudinal follow-up of breast cancer patients undergoing potentially cardiotoxic therapies. Imaging and associated metadata were fully de-identified following established DICOM anonymization standards prior to any secondary processing. For the purposes of the MICCAI 2026 Challenge, formal secondary-use authorization for release of the fully anonymized dataset is being processed across participating institutions. Approvals have already been received by multiple centers, while the remaining institutions are in the final stages of institutional confirmation following completion of the required review procedures. The dataset will be released exclusively in anonymized form and strictly for non-commercial scientific research. Secondary-use authorization is treated as a key prerequisite in the release process, and the current center-level approval status is reported transparently in the "Further Comments" section of this proposal.

## 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)



Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

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

### 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 related to the EchoRisk-MICCAI training dataset, evaluation procedures, and challenge scoring metrics will be made publicly available through the EchoRisk-MICCAI GitHub repository. The evaluation code will be released prior to the official start of the challenge, enabling participants to test their algorithms locally and to clearly understand the evaluation and ranking process.

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

All participants are required to submit their algorithms in the form of a self-contained Docker container for evaluation. The submitted code must be able to run without requiring additional manual setup or external dependencies beyond those specified by the organizers. For the purpose of reproducibility and transparency, participants are strongly encouraged to provide links to their source code on a public repository (e.g., GitHub, GitLab, or a similar platform); however, public code release is not a condition for participation during the challenge.

### 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.

CARDIO CARE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 945175. No other conflicts of interest. Test images will only be accessible to the challenge organizers.

## 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

Decision support, Research, CAD, Assistance

### Task category(ies)

State the task category(ies)

Examples:

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

Classification, Detection, Segmentation, 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.

The target cohort for this challenge comprises elderly and multimorbid women diagnosed with breast cancer who are undergoing potentially cardiotoxic cancer therapies, including anthracyclines and HER2-targeted agents. These patients represent a clinically vulnerable population with an elevated risk of therapy-induced cardiotoxicity due to age-related cardiovascular changes and the high prevalence of pre-existing comorbidities. The target cohort is characterized by a need for frequent, longitudinal cardiac assessment to enable early detection of subclinical cardiac dysfunction, risk stratification, and timely intervention. Given that older breast cancer patients are underrepresented in clinical trials but highly prevalent in real-world clinical practice, this cohort represents a critical target population for deployable, scalable, and equitable AI-based decision support tools in cardio-oncology. The target cohort reflects the intended clinical deployment population of AI systems developed

through this challenge, namely elderly and multimorbid women undergoing potentially cardiotoxic breast cancer therapy. While the challenge dataset is derived from a prospective clinical study population with similar characteristics, it may not perfectly mirror all demographic or comorbidity distributions of the broader real-world population. The target cohort therefore represents the intended application setting, whereas the challenge cohort represents the available prospective clinical dataset used for benchmarking.

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

The challenge cohort corresponds to patients enrolled in the CARDIOCARE prospective multicenter study and represents the empirical dataset available for benchmarking. This cohort includes breast cancer patients undergoing potentially cardiotoxic therapy, with longitudinal cardiac monitoring. While it reflects the clinical scenario described in the target cohort, it should be interpreted as a study population rather than a complete epidemiological representation of all elderly and multimorbid breast cancer patients. The proposed challenge will release 2D echocardiography video data acquired from 421 breast cancer patients enrolled in the CARDIOCARE prospective clinical study. The dataset originates from six European cancer and cardiology centers across five countries (Italy, Cyprus, Sweden, Greece, and Slovenia), ensuring diversity in patient populations, acquisition protocols, and ultrasound equipment. All data represent real-world clinical scenarios, including patients undergoing potentially cardiotoxic cancer therapies and exhibiting a spectrum of normal cardiac function, subclinical changes, and therapy-induced cardiac dysfunction. The imaging dataset consists of grayscale DICOM echocardiography videos capturing at least one representative cardiac cycle, acquired in standard apical 4-chamber and 2-chamber views, and collected at multiple longitudinal timepoints, including baseline and follow-up examinations. To reflect routine clinical practice, the data include variability in image quality, framing, patient anatomy, and operator technique, as well as occasional missing or suboptimal views. Echocardiography data were acquired using multi-vendor ultrasound systems from major manufacturers commonly used in clinical practice. In addition to imaging, the challenge cohort includes cardiotoxicity-related clinical labels, such as left ventricular ejection fraction, global longitudinal strain, and blood biomarkers (e.g., troponin and NT-proBNP), enabling clinically grounded evaluation of automated analysis and prediction methods. Ethical approvals for data acquisition were obtained at all participating centers, and all data are fully de-identified in compliance with GDPR and institutional regulations. Performance results should therefore be interpreted within the context of the CARDIOCARE study population, and external validation in additional cohorts would be required prior to broader clinical deployment.

### **Imaging modality(ies)**

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

The challenge is based on Cardiac ultrasound (echocardiography)

### **Context information**

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

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

Additional information provided alongside the imaging data includes clinically relevant quantitative and biomarker measurements, specifically: (i) Left ventricular ejection fraction (LVEF), (ii) Cardiac troponin levels, (iii) N-terminal pro-B-type natriuretic peptide (NT-proBNP); these parameters are directly related to cardiac function and myocardial injury and are provided to support clinically meaningful model development and evaluation.

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

No additional patient-level context information (e.g., demographics, medical history, or clinical background) will be provided beyond the imaging data and the specified cardiac measurements.

### Target entity(ies)

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

The data for this challenge originate from cardiac ultrasound (echocardiography) imaging of the heart, with a primary focus on the left ventricle and associated cardiac chambers as visualized in standard apical 4-chamber and apical 2-chamber views. These views provide essential anatomical and functional information for assessing left ventricular size, systolic function, myocardial deformation, and early signs of therapy-induced cardiotoxicity. In the target biomedical application, echocardiography would be acquired from breast cancer patients undergoing potentially cardiotoxic therapies as part of routine cardiac monitoring before, during, and after treatment. The imaging focuses on longitudinal assessment of cardiac structure and function to enable early detection of subclinical dysfunction and risk stratification in real-world cardio-oncology practice.

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

For Task 1, participating algorithms are designed to analyze 2D echocardiography videos of the left ventricle and to predict key cardiac functional parameters, including left ventricular ejection fraction, end-diastolic volume, and end-systolic volume, directly from imaging data. The algorithms focus on the left ventricular cavity and its temporal dynamics across the cardiac cycle, and may implicitly or explicitly rely on segmentation, tracking, or video-based regression techniques. In addition, the algorithms aim to predict cardiotoxicity-related biomarker status, such as elevated troponin and NT-proBNP levels, when available.

### 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,Consistency

## 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).

#### Ultrasound System

- Manufacturer: GE HealthCare
- Model: Vivid iq portable cardiovascular ultrasound system

#### System Configurations

- Standard Vivid iq
- Vivid iq Premium configuration
- Vivid iq Ultra Edition (latest release with enhanced AI and workflow tools)

#### Typical Hardware / Software Revisions in Clinical Use

- Vivid iq v204
- Vivid iq v206 (These correspond to commonly deployed hardware and software revision levels in routine clinical practice.)

#### Key Technical Specifications

- 15.6-inch touchscreen display
- Approximate weight: 5.2 kg (with battery)
- Portable laptop-style form factor
- Imaging modalities: B-mode, Color Doppler, Pulsed-Wave Doppler, Continuous-Wave Doppler, 2D and limited 4D imaging
- Integrated AI-based quantification tools (Ultra Edition), including automated ejection fraction and strain analysis

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

#### Patient Preparation and Setup

Patient positioned supine or in the left lateral decubitus position

Continuous ECG tracing displayed during acquisition

Image optimization performed for gain, depth, and sector width

#### Standard Imaging Views Acquired

Parasternal long-axis view

Parasternal short-axis views at basal, mid, and apical levels

Apical four-chamber view (cine loops and end-diastolic/end-systolic frames)

Apical two-chamber view (cine loops and end-diastolic/end-systolic frames)

Apical three-chamber view (cine loops and end-diastolic/end-systolic frames)

## Left Ventricular Systolic Function Assessment

Visual assessment of global and regional function across all views

Left ventricular ejection fraction (LVEF) measured using the biplane Simpson method from apical four- and two-chamber views

Endocardial borders traced at end-diastole and end-systole

Foreshortening avoided during acquisition and analysis

LVEF reported as a percentage

## Left Ventricular Volumes and Dimensions

LV end-diastolic volume (LVEDV) and end-systolic volume (LVESV) derived using biplane Simpson method

LV end-diastolic and end-systolic diameters measured in parasternal long-axis view

Septal and posterior wall thickness recorded

## Global Longitudinal Strain (GLS)

Acquired from apical four-, two-, and three-chamber views

Frame rates between 50–90 frames per second

Adequate tracking quality required across all myocardial segments

GLS reported as a percentage

## Regional Wall Motion Analysis

Segmental assessment using the standardized 17-segment LV model

Wall motion graded for each segment

## Left Ventricular Diastolic Function Assessment

Mitral inflow E and A wave velocities

E-wave deceleration time



Tissue Doppler imaging of septal and lateral  $e'$  velocities

$E/e'$  ratio

Left atrial volume index

Tricuspid regurgitation velocity

Supporting Findings and Additional Assessments

LV mass and geometric pattern

Presence of hypertrophy or dilation

Pericardial effusion

Valvular disease affecting LV loading conditions

Reporting Standards

Quantitative values reported with measurement method specified

Image quality statement included

Comparison with prior studies when available

Integrated clinical interpretation provided for patient management

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

The challenge dataset was collected across multiple clinical sites to ensure diversity, robustness, and generalizability. Participating institutions include: (i) Istituto Europeo di Oncologia (IEO), Italy; (ii) Bank of Cyprus Oncology Centre (BOCOC), Cyprus; (iii) Karolinska University Hospital / Karolinska Institutet (KSBC), Sweden; (iv) National and Kapodistrian University of Athens (NKUA), Greece; (v) University Hospital of Ioannina (UOI), Greece; and (vi) Institute of Oncology Ljubljana (IOL), Slovenia.

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 data were acquired by trained and certified cardiac sonographers with clinical experience in transthoracic echocardiography. Data acquisition was conducted under the oversight of cardiologists and senior clinical advisors, who ensured adherence to standardized imaging protocols, clinical quality control, and consistency across participating sites.

## Training and test case characteristics

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

Examples:

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

A case in this challenge refers to the complete set of data required to produce one algorithm output that is compared against a corresponding reference result. For Task 1, a case corresponds to a single echocardiography examination from one patient at a specific timepoint.

Each case includes one or more 2D echocardiography video sequences (grayscale DICOM format) capturing at least one representative cardiac cycle in standard apical views (apical 4-chamber and/or apical 2-chamber), together with associated metadata as provided by the organizers. The expected algorithm outputs for each case include left ventricular end-diastolic volume, end-systolic volume, and ejection fraction, and, when available, cardiotoxicity-related biomarker status (e.g., elevated troponin and/or NT-proBNP).

For training cases, reference annotations and labels (e.g., cardiac functional parameters and biomarkers) are provided to participants. For validation and test cases, the input echocardiography videos are provided, while the corresponding reference values are withheld by the organizers and used exclusively for evaluation and ranking. Each case is evaluated independently, enabling standardized and fair comparison of algorithm performance across patients and timepoints.

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

The dataset is split at the patient level into training (236 cases), validation (59 cases), and test (126 cases) sets to prevent information leakage across longitudinal follow-up timepoints. For the training and validation sets, full reference annotations are provided to participants. The validation set is intended for model tuning and internal evaluation. The test set consists of held-out cases with reference labels strictly withheld from participants. Test labels are accessible only to the organizers and are used exclusively for final evaluation and official ranking. No test-set ground truth will be released prior to the official conclusion of the challenge.

c) How much of the data are already annotated (stratified by train test in percentage)?

All data included in the challenge are already annotated at the case level with clinically derived reference labels. Training set ( $\approx 56\%$ , 236 cases): 100% annotated. Reference labels include left ventricular ejection fraction, and, when available, cardiotoxicity-related biomarkers (e.g., troponin, NT-proBNP).

Validation set ( $\approx 14\%$ , 59 cases): 100% annotated. Reference labels are provided to participants for model development and validation.

Test set ( $\approx 30\%$ , 126 cases): 100% annotated. Reference labels are available to the organizers only and remain hidden from participants until the challenge concludes. These annotations are used exclusively for final evaluation and ranking.

Overall, 100% of the dataset is annotated, with annotations distributed across training, validation, and test sets according to the defined split.

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

The dataset includes 421 cases, reflecting the size of the CARDIOCARE multicenter study and providing sufficient diversity for robust algorithm development. Data are split at the patient level into 236 training cases ( $\approx 56\%$ ), 59 validation cases ( $\approx 14\%$ ), and 126 test cases ( $\approx 30\%$ ) to avoid information leakage across longitudinal follow-up timepoints. The combined training and validation sets ( $\approx 70\%$ ) support effective model training and tuning across heterogeneous clinical conditions, while the relatively large held-out test set ( $\approx 30\%$ ) enables rigorous and unbiased evaluation of generalization in a real-world, multicenter, and multi-vendor setting.

e) 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.

A subset of cases is annotated with therapy-induced cardiotoxicity status based on longitudinal clinical follow-up, while other cases do not carry this specific outcome annotation due to differences in follow-up timing and clinical data availability. Importantly, this does not affect the core challenge tasks, which are defined to be fully supported by the available annotations in each split. The presence of partially annotated outcomes reflects real-world clinical data collection scenarios and introduces additional flexibility and novelty, enabling the exploration of methods that can leverage incomplete labels, weak supervision, or semi-supervised learning strategies.

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

The challenge will include a substantial proportion of unseen and previously unpublished data derived from the ongoing CARDIOCARE prospective multicenter clinical study. All cases used for the challenge have been collected specifically within this study and have not been released in any prior public dataset or benchmark. In particular, all test cases (126 cases,  $\approx 30\%$  of the dataset) will consist of unseen and unpublished data, withheld entirely from participants during the challenge and used exclusively for final evaluation and ranking. In addition, portions of the training and validation data correspond to newly curated, real-world echocardiography acquisitions that have not been previously disseminated.

### 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 annotations for this challenge correspond to clinically derived measurements routinely used in cardio-oncology practice, rather than synthetic or in silico ground truth. For training, validation, and test cases, left ventricular ejection fraction (LVEF) and global longitudinal strain (GLS) were measured as part of standard clinical echocardiography workflows at the participating centers, using vendor-provided or clinically validated software tools in accordance with local and international guidelines. Blood biomarkers related to cardiotoxicity, including troponin and NT-proBNP, were obtained through routine clinical laboratory testing performed at the corresponding follow-up timepoints, following standard hospital protocols. These biomarker values were used to derive cardiotoxicity-related labels where applicable. No centralized re-annotation or adjudication was performed

for the purposes of the challenge; therefore, reference measurements reflect real-world inter-reader variability across participating centers.

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 challenge-specific annotation instructions were provided to the annotators. All reference measurements were obtained as part of routine clinical practice, following standard echocardiography acquisition and reporting protocols at the participating centers. Cardiac function parameters (e.g., LVEF and GLS) and blood biomarkers (e.g., troponin and NT-proBNP) were measured according to local clinical workflows and established clinical guidelines, using vendor-provided or clinically validated tools. The same procedures apply to the training, validation, and test cases. As reference annotations were generated during routine patient care rather than through a dedicated annotation campaign, no additional annotator training phase or challenge-specific annotation software was required.

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

All reference annotations were generated during routine clinical care by medically trained professionals, including cardiac sonographers and cardiologists, using vendor-provided or clinically validated echocardiography analysis software. Cardiotoxicity-related blood biomarkers were measured by certified clinical laboratories following standard hospital protocols. The same annotation process applies to the training, validation, and test cases.

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.

No challenge-specific pre-processing is applied to the raw data prior to release. Training and validation data are provided in their original DICOM format, preserving the native image quality, acquisition characteristics, and metadata as acquired in clinical practice. The test data are similarly provided without pre-processing, with reference labels retained by the organizers for evaluation. Participants are free to apply their own pre-processing, normalization, or data augmentation strategies as part of their algorithm development, provided that all processing at test time is fully automatic and compliant with the challenge rules.

### **Sources of error**

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

Although no manual annotation campaign was conducted specifically for the challenge, reference values for LVEF, LV volumes, and GLS were obtained during routine clinical workflows. As such, they are subject to known inter- and intra-observer variability inherent to echocardiographic measurements. In routine practice, LVEF variability is

commonly reported in the range of  $\pm 5$  percentage points, particularly when derived using biplane Simpson's method, as summarized in the chamber quantification recommendations by Lang et al. (2015) and confirmed in reproducibility analyses by Thavendiranathan et al. (2013). Variability in LV volume estimation may range between 10–15 ml depending on image quality and tracing consistency, consistent with findings reported by Thavendiranathan et al. (2013). GLS measurements may vary depending on vendor-specific implementations and tracking quality, with relative differences of approximately 1–2 percentage points across repeated measurements. These sources of variability reflect real-world clinical conditions rather than annotation error. The dataset intentionally preserves this heterogeneity across centers, operators, and vendors to provide a realistic upper bound for achievable algorithmic performance. Accordingly, clinically acceptable error margins are defined relative to this inherent measurement variability.

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

Multi-vendor and multi-site acquisition variability constitutes the primary additional source of error. Differences in ultrasound equipment settings, image resolution, frame rate, and operator technique across the six participating centres may introduce systematic biases. Furthermore, temporal variability in cardiac function due to treatment timing, patient hydration status, and heart rate at the time of acquisition may affect LVEF measurements independently of true dysfunction status. Site-specific differences in image quality and patient positioning may also contribute to performance variation across centres.

## ASSESSMENT METHODS

### Metric(s)

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

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

For the estimation of left ventricular ejection fraction (LVEF) and left ventricular volumes, algorithm performance will be evaluated using Mean Absolute Error (MAE), which serves as the primary ranking metric due to its robustness and direct clinical interpretability. Root Mean Squared Error (RMSE) and the coefficient of determination ( $R^2$ ) will be reported as secondary metrics to provide additional insight into error distribution and model fit. Lower MAE indicates better performance. MAE is reported in clinically meaningful units (percentage points for LVEF and milliliters for volumes), facilitating direct comparison with routine clinical variability. In addition to continuous regression metrics, performance will be evaluated relative to clinically actionable error thresholds. For LVEF, an absolute error  $\leq 5$  percentage points is considered clinically acceptable, reflecting typical inter- and intra-observer variability in routine echocardiography, as described in the chamber quantification recommendations by Lang et al. (2015) and reproducibility analyses by Thavendiranathan et al. (2013). Errors exceeding 10 percentage points are likely to influence treatment decisions, particularly around clinically relevant cut-offs such as 50% and 40%. We therefore additionally report the proportion of cases with absolute LVEF error  $\leq 5\%$  and  $\leq 10\%$ , as well as performance stratified within the clinically critical range of LVEF 45–55%. For left ventricular end-diastolic and end-systolic volumes, absolute errors within approximately 10–15 ml are considered acceptable in longitudinal follow-up, consistent with reproducibility findings reported by Thavendiranathan et al. (2013). Accordingly, we report the proportion of cases within 10 ml (typical reproducibility margin) and 20 ml

(clinically meaningful deviation) absolute error margins. For global longitudinal strain (GLS), a relative deterioration >15% from baseline is widely regarded as clinically meaningful in cardio-oncology, and we additionally report agreement in detecting clinically relevant strain worsening. Agreement will be quantified using sensitivity, specificity, and F1-score for detecting clinically relevant GLS deterioration defined as a relative reduction greater than 15% from baseline. Binary labels will be derived from reference and predicted GLS values using the same threshold. These threshold-based metrics are reported for clinical interpretability and do not affect the official leaderboard ranking, which is based on MAE.

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

Mean Absolute Error (MAE) is used as the primary metric because it provides a clinically interpretable measure of error in absolute units (percentage points for LVEF and milliliters for volumes), which directly reflects routine clinical variability. RMSE and  $R^2$  are reported as complementary metrics to capture larger errors and overall model fit, supporting a comprehensive yet clinically meaningful evaluation.

### 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. Ideally, provide the ranking scheme as a concrete pseudo code.

The final ranking is based on performance on the hidden test set using Mean Absolute Error (MAE). For each submitted algorithm, MAE is computed separately for left ventricular ejection fraction, end-diastolic volume, and end-systolic volume by averaging the absolute error across all test cases. Algorithms are ranked independently for each target. The final challenge ranking is obtained by summing the individual ranks across the three targets, ensuring that differences in measurement scale do not bias the evaluation. Ties are resolved using LVEF MAE as the primary criterion, followed by end-diastolic and end-systolic volume MAE. The three tasks of the challenge are evaluated independently. Each task has its own evaluation metrics, ranking procedure, and leaderboard. Participation in Task 1 does not require participation in Tasks 2 or 3. Teams may choose to participate in one, two, or all three tasks. Rankings are computed separately for each task, and performance in one task does not influence ranking in another. Awards and recognitions are granted on a per-task basis according to the official leaderboard of each task. No aggregated cross-task ranking will be computed.

In addition to participant submissions, a predefined reference baseline will be included for benchmarking purposes. To strengthen interpretability and long-term benchmark value, Task 1 will include a concrete reference baseline based on the EchoNet-Dynamic framework (Ouyang et al., 2020). The official publicly available implementation will be trained on the Task 1 training set using the official data partitions for EF estimation. In addition, an EchoNet-style video-based segmentation baseline will be trained to derive EDV and ESV via the standard biplane Simpson method. Baseline performance will be evaluated under identical data partitions and computational constraints as participant submissions. Baseline results will be reported alongside submitted methods on the final leaderboard to provide a transparent performance anchor and facilitate comparison with prior literature.

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

Participating teams are required to submit Docker containers that automatically process all test cases on the challenge evaluation server. If a submission fails to produce a valid output for one or more test cases, those cases will be assigned the worst possible metric value for the corresponding task and metric. This policy ensures fair comparison across submissions and discourages partial or non-robust solutions.



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

The proposed ranking scheme is designed to reflect overall clinical performance in estimating key cardiac functional parameters. By evaluating left ventricular ejection fraction, end-diastolic volume, and end-systolic volume separately and combining their individual ranks, the ranking avoids bias introduced by differing measurement scales and ensures balanced assessment across all clinically relevant outputs. This approach aligns with routine echocardiographic practice, where accurate and consistent estimation of multiple interrelated parameters is required for reliable cardiac function assessment and longitudinal monitoring. Summing ranks provides a fair, transparent, and robust comparison of algorithm performance across heterogeneous real-world data.

## Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

Statistical analysis is designed to ensure robust and unbiased evaluation of submitted algorithms. Performance is assessed using case-level regression metrics that are aggregated across the test set, and final rankings are computed using a rank-based aggregation scheme to avoid bias from differing measurement scales. Missing or invalid outputs are handled consistently by assigning worst-case metric values. No distributional assumptions are required, as the evaluation relies on non-parametric, rank-based methods. The presence of real-world measurement variability defines a practical ceiling for algorithmic performance and is considered when interpreting statistical differences between methods.

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

The statistical methods for Task 1 are designed to provide precise and reliable estimates of algorithm performance in automated cardiac parameter and biomarker estimation from ultrasound videos. Performance metrics are computed at the case level and aggregated across the full test set. To quantify the precision of the estimated performance, summary statistics including mean, median, standard deviation, and confidence intervals are reported for each metric. Where applicable, confidence intervals are estimated using non-parametric bootstrapping across test cases to avoid distributional assumptions.

Provide a description of how variability of the performance of individual algorithms across tests cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

Performance variability across cases is assessed by analyzing metric distributions over all test samples. Case-level errors are examined to capture heterogeneity arising from differences in image quality, acquisition protocols, patient anatomy, and disease severity. Variability is summarized using standard deviation, interquartile range, and percentile-based statistics, providing insight into algorithm robustness across diverse real-world ultrasound data.

Provide a description of how variability of rankings is assessed.

To ensure robustness of the final rankings, variability in algorithm rankings is evaluated across individual target parameters and metrics. Rankings are computed separately for each estimated cardiac parameter, and an

aggregated rank-based scheme is used to obtain overall challenge rankings. Rank stability is assessed using rank correlation measures allowing comparison of rankings across metrics and ensuring that overall rankings are not driven by a single parameter or metric.

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

To formally assess whether observed differences in performance between submitted algorithms are statistically significant, we will perform case-level paired statistical analyses on the primary ranking metric (Mean Absolute Error, MAE). For each pair of algorithms, absolute errors will be computed per test case and compared using a two-sided Wilcoxon signed-rank test, which accounts for the paired nature of the evaluation (all algorithms are tested on the same cases) and does not assume normality of the error distribution. This non-parametric approach is appropriate given the heterogeneous and potentially skewed distribution of echocardiographic estimation errors. When more than two algorithms are compared, pairwise comparisons will be conducted with adjustment for multiple testing using the Holm-Bonferroni procedure to control the family-wise error rate. In addition to reporting adjusted p-values, we will provide: 95% confidence intervals for the paired difference in MAE estimated via non-parametric bootstrap resampling; effect size estimates (median paired MAE difference); rank stability analyses derived from bootstrap iterations. In addition to continuous MAE comparisons, statistical analyses will also be performed on threshold-based performance indicators (e.g., proportion of cases within 5% absolute LVEF error or detection of >15% GLS deterioration). Differences in these clinically anchored proportions will be assessed using paired tests for binary outcomes (e.g., McNemar's test), ensuring that improvements in clinically meaningful performance are formally evaluated rather than inferred solely from continuous error differences.

Provide a description of the missing data handling.

No missing ground-truth data is expected for Task 1, as all test cases undergo quality control and verification prior to release. In the event that a submission fails to produce a valid output for one or more test cases, those cases are handled consistently by assigning predefined worst-case metric values, ensuring fair and unbiased comparison across all submissions.

Indicate any software product that is used for all data analysis methods.

All statistical analyses, metric computations, and visualizations are performed using standard scientific computing software, primarily Python-based libraries such as NumPy, SciPy, Pandas, and Matplotlib. These tools provide comprehensive support for statistical analysis, bootstrapping, non-parametric testing, and rank aggregation.

### 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.

N/A

## **TASK 2: Automated Assessment of Left Ventricular Dysfunction from Longitudinal Echocardiography**

### **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.

Echocardiography remains the gold standard in clinical practice for evaluating cardiac anatomy and physiology, favored for its non-invasive nature, portability, and immediate results. It is particularly vital for the serial monitoring of oncology patients—specifically those with breast cancer—undergoing treatments with cardiotoxic risks. Despite its utility, the manual interpretation of echocardiograms is labor-intensive and prone to significant observer bias. These limitations hinder the modality's scalability for widespread longitudinal tracking and early risk stratification.

The primary objective of this task is to drive innovation in the automated analysis of echocardiographic imaging through video-based artificial intelligence approaches. Specifically, participants are invited to design and develop AI algorithms capable of automatically detecting and evaluating left ventricular dysfunction directly from echocardiography videos. To support this endeavor, a distinctive dataset has been curated, comprising echocardiographic recordings from 421 patients. In total, the dataset encompasses 1,528 echocardiography videos, which were systematically acquired at predefined, standardized follow-up intervals.

The significance of this challenge lies in advancing the automated detection of Left Ventricular Dysfunction directly from raw echocardiography videos, addressing the critical clinical bottleneck of operator dependency. By shifting the focus from static image measurements to full-motion video analysis, this task promotes the development of spatiotemporal AI models capable of capturing complex cardiac dynamics and subtle wall motion abnormalities that single frames may miss. This approach offers a standardized, objective alternative to manual interpretation, significantly reducing the high inter-observer variability inherent in current practice. Ultimately, this promotes the creation of scalable, real-time screening tools that can assist clinicians in delivering consistent and accurate cardiac risk assessments.

#### **Keywords**

List the primary keywords that characterize the task.

Echocardiography, diagnosis, AI, automated cardiac assessment, left ventricular

### **ORGANIZATION**

#### **Organizers**

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

Kostas Marias  
Foundation for Research and Technology - Hellas

Manolis Tsiknakis  
Hellenic Mediterranean University

Grigorios Kalliataakis  
Foundation for Research and Technology

Georgios Manikis  
Foundation for Research and Technology - Hellas

Georgia Karanasiou  
University of Ioannina

Eleni Georga  
University of Ioannina

Katerina Naka  
University Hospital of Ioannina

Dorothea Tsekoura  
National and Kapodistrian University of Athens

Gerasimos Filippatos  
National and Kapodistrian University of Athens

Anastasia Constantinidou  
Bank of Cyprus Oncology Centre

Andri Papakonstantinou  
Karolinska University Hospital

Ketti Mazzocco  
Istituto Europeo di Oncologia IRCCS Milano

Domen Ribnikar  
Institute of Oncology Ljubljana

Dimitrios Fotiadis  
University of Ioannina

b) Provide information on the primary contact person.

Kostas Marias, kmarias@ics.forth.gr  
Foundation for Research and Technology - Hellas

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

Yes, clinicians are integral members of the organizing team and play a central role in the clinical design, validation, and interpretation of the challenge. Their involvement ensures that the challenge addresses clinically relevant questions, reflects real-world clinical practice, and that the chosen evaluation metrics and outcomes are meaningful, interpretable, and impactful for patient care.

Katerina Naka (University Hospital of Ioannina) – Clinical lead for cardiac imaging data curation and annotation. She oversees the clinical validation of echocardiographic datasets, contributes to defining clinically relevant endpoints, and supports interpretation of challenge outcomes from a cardiology perspective.

Dorothea Tsekoura (National and Kapodistrian University of Athens) – Clinical advisor for cardio-oncology pathways and patient stratification. She contributes to the definition of inclusion/exclusion criteria, clinical metadata harmonization, and ensures alignment with current cardio-oncology guidelines.

Gerasimos Filippatos (National and Kapodistrian University of Athens) – Senior clinical advisor and scientific oversight. He provides high-level clinical guidance on heart failure and therapy-related cardiac dysfunction, ensuring the clinical credibility and translational relevance of the challenge.

Anastasia Constantinidou (Bank of Cyprus Oncology Centre) – Oncology clinical advisor. She supports the integration of oncological treatment context, therapy exposure variables, and clinically meaningful outcomes related to cancer therapy-related cardiotoxicity.

Andri Papakonstantinou (Karolinska University Hospital) – Clinical expert for international validation and benchmarking. She supports interpretation of results in a broader clinical context.

Ketti Mazzocco (Istituto Europeo di Oncologia IRCCS, Milan) – Clinical advisor for patient-centered outcomes. She provides input on clinically relevant outcome measures and supports alignment with patient-reported and longitudinal assessment perspectives.

Domen Ribnikar (Institute of Oncology Ljubljana) – Clinical advisor for real-world oncology practice. He advises on translational applicability of the developed methods.

### 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).

29th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2026)

b) Report the platform used to run the challenge.

The challenge will be run on the Synapse platform (<https://www.synapse.org>). Synapse provides a robust and widely used infrastructure for running biomedical data challenges, supporting secure data hosting, participant registration, submission management, leaderboard evaluation, and transparent benchmarking. The platform is particularly well suited for challenges involving sensitive clinical data, offering fine-grained access control, governance mechanisms, and compliance with ethical and data-use requirements. The use of Synapse ensures a secure, transparent, and reproducible challenge execution, aligned with MICCAI policies and best practices for clinical AI benchmarking.

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

Yes

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

The challenge will be hosted on Synapse (Sage Bionetworks), which will serve as the central platform for data access, documentation, submissions, and leaderboards. A public placeholder Synapse project has been created and will be populated with all challenge materials upon acceptance: [Synapse Project URL <https://www.synapse.org/Synapse:syn72001386>]

### Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

Fully Automatic.

All submitted methods must operate in a fully automatic manner at inference time, with no user interaction or manual intervention on validation or test data. This includes no manual selection of frames, regions of interest, contours, or case-specific parameter tuning. During training, participants are allowed to perform standard data curation, preprocessing, and augmentation on the provided training set, as well as pre-train models using external public datasets or foundation models. Participants using external public datasets or foundation models for pretraining or model development will be required to disclose them in their submission description. External retrieval, API calls, or internet access during inference will not be permitted.



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 may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

#### **Publicly available data is 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.

#### **May participate but not eligible for awards and not listed in leaderboard**

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

The challenge will recognize outstanding contributions based on the official final leaderboard ranking.

Top-performing teams will be invited to present their methods and results during the associated MICCAI 2026 challenge session or workshop, and selected teams will be invited to contribute to a joint challenge summary paper to be submitted to a peer-reviewed journal. Official certificates of achievement will be awarded to top-ranked teams. To further acknowledge excellence, monetary awards totaling €900 will be granted and distributed as follows: €450 for 1st place, €300 for 2nd place, and €150 for 3rd place. Rankings are computed exclusively based on performance on the hidden test set for Task 3. This award policy emphasizes excellence while ensuring a transparent and fair evaluation process in line with MICCAI challenge best practices.

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.

**One official leaderboard will be generated per task based on performance on the hidden test set. Test-set results will be released at the official challenge result announcement.**

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).

Upon completion of the challenge, the organizers will coordinate a comprehensive peer-reviewed publication summarizing the challenge design, key analyses, and results. The paper will be a collaborative effort involving the challenge organizing team, clinical and technical collaborators, and the top-performing participating teams. Each eligible team will be invited to nominate up to three co-authors, provided that the team has made a valid submission and complied with all challenge rules. Participating teams remain free to publish their own methods and results independently; however, to preserve the integrity of the challenge outcomes, an embargo period of three months following the release of the official challenge results will apply. The organizers reserve the right to exclude teams from co-authorship in the challenge paper in cases of rule violations or non-compliant submissions. All teams submitting a valid final test-set submission are eligible to submit a workshop paper to the associated LNCS proceedings, subject to standard peer review. Top-ranked teams per task will additionally be invited to contribute to the joint challenge summary paper, provided they comply with challenge rules and

submission requirements. Participating teams remain free to publish their own methods and results independently; however, to preserve the integrity of the challenge outcomes, an embargo period of three months following the release of the official challenge results will apply. The organizers reserve the right to exclude teams from co-authorship in the challenge paper in cases of rule violations or non-compliant submissions.

### 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.

Submissions will be accepted in the form of Docker containers via the Synapse platform. Detailed submission instructions and technical requirements will be provided at the time of the official challenge announcement. All submitted algorithms must operate fully automatically within the provided containerized evaluation environment and must not access external resources during test-time execution. Participants are permitted to use publicly available external datasets and pretrained models during model development; however, all external data sources and pretrained models must be explicitly disclosed in the method description submitted with the final results. Any use of private or non-public datasets must be declared and must comply with the challenge data usage policy. Participants must ensure that no direct or indirect information from the hidden test set is used during training or model selection. Any form of metadata-based case identification, retrieval, or linkage to external databases is prohibited.

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.

Participating teams may submit up to three formal submissions per task. Only the final submission will be considered for the official challenge ranking. Prior to the final test-set submission, participants may evaluate their Docker containers on the validation dataset to verify correctness and avoid submission errors. A validation leaderboard will be maintained on Synapse, providing real-time AUC-ROC scores on the validation set. Each team may submit up to three times per week to the validation leaderboard during the development phase. The test set leaderboard will be revealed only after the final submission deadline. Ground truth labels for the validation set will not be released during the challenge but may be released post-challenge for reproducibility purposes. Participants will not have access to individual validation-case labels at any point during the challenge; the only feedback provided is the aggregate AUC-ROC score returned by the Synapse leaderboard after each submission. This design prevents overfitting to the validation set while still allowing participants to verify that their Docker containers execute correctly.

### 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

April 1, 2026: Challenge website opens for registration; release of training and validation data

April 10, 2026: Submission system opens for validation submissions

July 15, 2026: Submission system opens for test submissions

August 20, 2026: Registration and Docker submission deadline

October 7, 2026: Release of final results during the MICCAI Annual Meeting

January 31, 2027: Publication of the challenge summary

## Ethics approval

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

The dataset used in the EchoRisk-MICCAI Challenge originates from the CARDIOCARE project (Horizon 2020, Grant Agreement No. 945175), a prospective, multicenter clinical study conducted across six European hospital sites in five countries. The CARDIOCARE study received ethical approval from the relevant Institutional Review Boards (IRBs) / Ethics Committees at each participating clinical center, in accordance with national regulations, the Declaration of Helsinki, and the EU General Data Protection Regulation (GDPR). Ethics approvals were obtained locally at each recruiting institution prior to patient enrollment and data collection, covering: (i) prospective acquisition of echocardiography imaging, (ii) collection of clinical and biomarker data, and (iii) longitudinal follow-up of breast cancer patients undergoing potentially cardiotoxic therapies. Imaging and associated metadata were fully de-identified following established DICOM anonymization standards prior to any secondary processing. For the purposes of the MICCAI 2026 Challenge, formal secondary-use authorization for release of the fully anonymized dataset is being processed across participating institutions. Approvals have already been received by multiple centers, while the remaining institutions are in the final stages of institutional confirmation following completion of the required review procedures. The dataset will be released exclusively in anonymized form and strictly for non-commercial scientific research. Secondary-use authorization is treated as a key prerequisite in the release process, and the current center-level approval status is reported transparently in the “Further Comments” section of this proposal.

## 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)

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

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

### 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 related to the EchoRisk-MICCAI training dataset, evaluation procedures, and challenge scoring metrics will be made publicly available through the EchoRisk-MICCAI GitHub repository. The evaluation code will be released prior to the official start of the challenge, enabling participants to test their algorithms locally and to clearly understand the evaluation and ranking process.

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

All participants are required to submit their algorithms in the form of a self-contained Docker container for evaluation. The submitted code must be able to run without requiring additional manual setup or external dependencies beyond those specified by the organizers. For the purpose of reproducibility and transparency, participants are strongly encouraged to provide links to their source code on a public repository (e.g., GitHub, GitLab, or a similar platform); however, public code release is not a condition for participation during the challenge.

### 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.

CARDIO CARE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 945175. No other conflicts of interest. Test images will only be accessible to the challenge organizers.

## MISSION OF THE CHALLENGE

### Field(s) of application

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

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis

- Research
- Screening
- Training
- Cross-phase

## Diagnosis

### Task category(ies)

State the task category(ies)

Examples:

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

Detection,Modeling,Classification

### Cohorts

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

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

The target cohort for this challenge comprises elderly and multimorbid women diagnosed with breast cancer who are undergoing potentially cardiotoxic cancer therapies, including anthracyclines and HER2-targeted agents. These patients represent a clinically vulnerable population with an elevated risk of therapy-induced cardiotoxicity due to age-related cardiovascular changes and the high prevalence of pre-existing comorbidities. The target cohort is characterized by a need for frequent, longitudinal cardiac assessment to enable early detection of subclinical cardiac dysfunction, risk stratification, and timely intervention. Given that older breast cancer patients are underrepresented in clinical trials but highly prevalent in real-world clinical practice, this cohort represents a

critical target population for deployable, scalable, and equitable AI-based decision support tools in cardio-oncology. The CARDIOCARE challenge cohort is well-aligned with this target population, as it was specifically designed to recruit elderly women with breast cancer undergoing cardiotoxic therapy across six European centres. While the cohort does not include male patients or non-breast cancer cardiotoxicity, it captures the most prevalent clinical use case for longitudinal cardiotoxicity monitoring. Generalisation to broader populations (e.g., male patients, other cancer types, or different age groups) remains an important direction for future research and is acknowledged as a limitation of the current dataset.

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

The proposed challenge will release 2D echocardiography video data acquired from 421 breast cancer patients enrolled in the CARDIOCARE prospective clinical study. The dataset originates from six European cancer and cardiology centers across five countries (Italy, Cyprus, Sweden, Greece, and Slovenia), ensuring diversity in patient populations, acquisition protocols, and ultrasound equipment. All data represent real-world clinical scenarios, including patients undergoing potentially cardiotoxic cancer therapies and exhibiting a spectrum of normal cardiac function, subclinical changes, and therapy-induced cardiac dysfunction. The imaging dataset consists of grayscale DICOM echocardiography videos capturing at least one representative cardiac cycle, acquired in standard apical 4-chamber and 2-chamber views, and collected at multiple longitudinal timepoints, including baseline and follow-up examinations. To reflect routine clinical practice, the data include variability in image quality, framing, patient anatomy, and operator technique, as well as occasional missing or suboptimal views. Echocardiography data were acquired using multi-vendor ultrasound systems from major manufacturers commonly used in clinical practice. In addition to imaging, the challenge cohort includes cardiotoxicity-related clinical labels, such as left ventricular ejection fraction, global longitudinal strain, and blood biomarkers (e.g., troponin and NT-proBNP), enabling clinically grounded evaluation of automated analysis and prediction methods. Ethical approvals for data acquisition were obtained at all participating centers, and all data are fully de-identified in compliance with GDPR and institutional regulations.

### Imaging modality(ies)

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

The challenge is based on Cardiac ultrasound (echocardiography)

### Context information

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

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

No additional information will be provided alongside the imaging data. This task deliberately restricts input to imaging data alone, without clinical metadata (e.g., age, treatment history, comorbidities). This design choice reflects the goal of evaluating the intrinsic diagnostic capacity of video-based AI models from echocardiography, independent of clinical confounders. In real-world deployment, such models would be integrated with clinical information for holistic decision-making. Challenge results should therefore be interpreted as measuring the algorithm's ability to detect LV dysfunction from imaging features alone, representing a lower bound on expected clinical utility when combined with patient context.

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

No additional patient-level context information (e.g., demographics, medical history, or clinical background) will be provided beyond the imaging data. This is intentional: the challenge aims to benchmark the ability of AI algorithms to detect LV dysfunction purely from echocardiographic video, simulating a scenario in which automated screening tools operate independently of electronic health records. This approach ensures that algorithm performance reflects imaging-derived features rather than correlations with demographic or clinical variables.

### Target entity(ies)

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

The data for this challenge originate from cardiac ultrasound (echocardiography) imaging of the left ventricle and associated cardiac chambers as visualized in standard apical 4-chamber and apical 2-chamber views of the heart. These views provide essential anatomical and functional information for assessing left ventricular size, systolic function, myocardial deformation, and early signs of therapy-induced cardiotoxicity. In the target biomedical application, echocardiography would be acquired from breast cancer patients undergoing potentially cardiotoxic therapies as part of routine cardiac monitoring before, during, and after treatment. The imaging focuses on longitudinal assessment of cardiac structure and function to enable early detection of subclinical dysfunction and risk stratification in real-world cardio-oncology practice.

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

For this task, algorithms will focus on the analysis of apical 4-chamber and apical 2-chamber for the identification of potential Left Ventricular Dysfunction from the corresponding videos.

### 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)).

AUC-ROC (Area Under the Receiver Operating Characteristic Curve), Accuracy, Precision, Recall

## 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).

#### Ultrasound System

- Manufacturer: GE HealthCare
- Model: Vivid iq portable cardiovascular ultrasound system

#### System Configurations

- Standard Vivid iq
- Vivid iq Premium configuration
- Vivid iq Ultra Edition (latest release with enhanced AI and workflow tools)

#### Typical Hardware / Software Revisions in Clinical Use

- Vivid iq v204
- Vivid iq v206 (These correspond to commonly deployed hardware and software revision levels in routine clinical practice.)

#### Key Technical Specifications

- 15.6-inch touchscreen display
- Approximate weight: 5.2 kg (with battery)
- Portable laptop-style form factor
- Imaging modalities: B-mode, Color Doppler, Pulsed-Wave Doppler, Continuous-Wave Doppler, 2D and limited 4D imaging
- Integrated AI-based quantification tools (Ultra Edition), including automated ejection fraction and strain analysis

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

#### Patient Preparation and Setup

Patient positioned supine or in the left lateral decubitus position

Continuous ECG tracing displayed during acquisition

Image optimization performed for gain, depth, and sector width

#### Standard Imaging Views Acquired

Parasternal long-axis view

Parasternal short-axis views at basal, mid, and apical levels

Apical four-chamber view (cine loops and end-diastolic/end-systolic frames)

Apical two-chamber view (cine loops and end-diastolic/end-systolic frames)

Apical three-chamber view (cine loops and end-diastolic/end-systolic frames)

## Left Ventricular Systolic Function Assessment

Visual assessment of global and regional function across all views

Left ventricular ejection fraction (LVEF) measured using the biplane Simpson method from apical four- and two-chamber views

Endocardial borders traced at end-diastole and end-systole

Foreshortening avoided during acquisition and analysis

LVEF reported as a percentage

## Left Ventricular Volumes and Dimensions

LV end-diastolic volume (LVEDV) and end-systolic volume (LVESV) derived using biplane Simpson method

LV end-diastolic and end-systolic diameters measured in parasternal long-axis view

Septal and posterior wall thickness recorded

## Global Longitudinal Strain (GLS)

Acquired from apical four-, two-, and three-chamber views

Frame rates between 50–90 frames per second

Adequate tracking quality required across all myocardial segments

GLS reported as a percentage

## Regional Wall Motion Analysis

Segmental assessment using the standardized 17-segment LV model

Wall motion graded for each segment

## Left Ventricular Diastolic Function Assessment

Mitral inflow E and A wave velocities

E-wave deceleration time

Tissue Doppler imaging of septal and lateral  $e'$  velocities

$E/e'$  ratio

Left atrial volume index

Tricuspid regurgitation velocity

Supporting Findings and Additional Assessments

LV mass and geometric pattern

Presence of hypertrophy or dilation

Pericardial effusion

Valvular disease affecting LV loading conditions

Reporting Standards

Quantitative values reported with measurement method specified

Image quality statement included

Comparison with prior studies when available

Integrated clinical interpretation provided for patient management

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

The challenge dataset was collected across multiple clinical sites to ensure diversity, robustness, and generalizability. Participating institutions include: (i) Istituto Europeo di Oncologia (IEO), Italy; (ii) Bank of Cyprus Oncology Centre (BOCOC), Cyprus; (iii) Karolinska University Hospital / Karolinska Institutet (KSBC), Sweden; (iv) National and Kapodistrian University of Athens (NKUA), Greece; (v) University Hospital of Ioannina (UOI), Greece; and (vi) Institute of Oncology Ljubljana (IOL), Slovenia.

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 data were acquired by trained and certified cardiac sonographers with clinical experience in transthoracic echocardiography. Data acquisition was conducted under the oversight of cardiologists and senior clinical advisors, who ensured adherence to standardized imaging protocols, clinical quality control, and consistency across participating sites.

## Training and test case characteristics

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

Examples:

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

A case in this challenge refers to the complete set of data required to produce one algorithm output that is compared against a corresponding reference result. For Task 2, a case corresponds to a single echocardiography examination from one patient at a specific timepoint. Each case includes one or more 2D echocardiography video sequences (grayscale DICOM format) capturing at least one representative cardiac cycle in standard apical views (apical 4-chamber and/or apical 2-chamber), together with associated metadata as provided by the organizers. The expected algorithm output for each case is a classification of left ventricular cardiac dysfunction, defined according to clinically accepted thresholds based on cardiac functional parameters (e.g., ejection fraction and/or global longitudinal strain). For training cases, reference labels indicating the presence or absence of left ventricular dysfunction are provided to participants. For validation and test cases, the input echocardiography videos are provided, while the corresponding reference labels are withheld by the organizers and used exclusively for evaluation and ranking. Each case is evaluated independently, enabling standardized and fair comparison of algorithm performance across patients and timepoints. Although the CARDIOCARE dataset is longitudinal — with each patient contributing multiple examinations at successive follow-up visits — Task 2 treats each examination independently. No cross-timepoint temporal modelling is required or expected; algorithms receive a single examination as input and produce one classification output per examination. The term “longitudinal” therefore refers to the dataset structure and the clinical monitoring context, not to the expected model architecture or input format. Specifically, left ventricular dysfunction is defined as  $LVEF < 50\%$ , consistent with the 2022 ESC Guidelines on cardio-oncology. The expected algorithm output for each case comprises: (1) a continuous probability score (0–1) indicating the likelihood of LV dysfunction, and (2) a binary classification label (0 = normal, 1 = LV dysfunction).

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

The dataset is split into training and validation (70%), and the held-out test set (30%). To avoid data leakage across longitudinal follow-up timepoints, all videos from individuals will be allocated to the training, validation or the held-out test set exclusively.

c) How much of the data are already annotated (stratified by train test in percentage)?

All videos for the training, validation and outer set will be annotated.

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

The dataset includes 421 cases, reflecting the size of the CARDIO CARE multicenter study and providing sufficient diversity for robust algorithm development. Data are split at the patient level into 236 training cases ( $\approx 56\%$ ), 59 validation cases ( $\approx 14\%$ ), and 126 test cases ( $\approx 30\%$ ) to avoid information leakage across longitudinal follow-up timepoints. The combined training and validation sets ( $\approx 70\%$ ) support effective model training and tuning across heterogeneous clinical conditions, while the relatively large held-out test set ( $\approx 30\%$ ) enables rigorous and unbiased evaluation of generalization in a real-world, multicenter, and multi-vendor setting.

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

The challenge will include a substantial proportion of unseen and previously unpublished data derived from the ongoing CARDIO CARE prospective multicenter clinical study. All cases used for the challenge have been collected specifically within this study and have not been released in any prior public dataset or benchmark. All test cases (126 cases,  $\approx 30\%$  of the dataset) will consist of unseen and unpublished data, withheld entirely from participants during the challenge and used exclusively for final evaluation and ranking.

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

The challenge will include a substantial proportion of unseen and previously unpublished data derived from the ongoing CARDIO CARE prospective multicenter clinical study. All cases correspond to baseline and follow-up echocardiography examinations collected specifically within this study and have not been released in any prior public dataset or benchmark. All test cases (76 cases,  $\approx 30\%$  of the dataset) consist of unseen and unpublished data, which are withheld entirely from participants during the challenge and used exclusively for final evaluation and ranking.

### 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 annotations for this task correspond to clinically derived measurements routinely used in cardio-oncology practice. For training, validation, and test cases, left ventricular dysfunction is defined as LVEF  $< 50\%$ , consistent with the 2022 ESC Guidelines on cardio-oncology. LVEF was measured using the biplane Simpson's method from apical four-chamber and two-chamber views as part of standard clinical echocardiography workflows at the participating centres, using vendor-provided or clinically validated software tools in accordance with local and international guidelines. The task is formulated as a binary classification problem: each case is labelled as "normal" (LVEF  $\geq 50\%$ ) or "LV dysfunction" (LVEF  $< 50\%$ ). Participants must output both a continuous predicted probability of dysfunction (range 0–1) and a binary class label for each case.

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 challenge-specific annotation instructions were provided to the annotators. All reference measurements were obtained as part of routine clinical practice, following standard echocardiography acquisition and reporting protocols at the participating centers. LVEF was measured according to local clinical workflows and established

clinical guidelines, using vendor-provided or clinically validated tools. The same procedures apply to the training, validation, and test cases. As reference annotations were generated during routine patient care rather than through a dedicated annotation campaign, no additional annotator training phase or challenge-specific annotation software was required.

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

All reference annotations were generated during routine clinical care by medically trained professionals, including cardiac sonographers and cardiologists, using vendor-provided or clinically validated echocardiography analysis software

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.

No challenge-specific pre-processing is applied to the raw data prior to release. Training and validation data are provided in their original DICOM format, preserving the native image quality, acquisition characteristics, and metadata as acquired in clinical practice. The test data are similarly provided without pre-processing, with reference labels retained by the organizers for evaluation. Participants are free to apply their own pre-processing, normalization, or data augmentation strategies as part of their algorithm development, provided that all processing at test time is fully automatic and compliant with the challenge rules.

### **Sources of error**

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

Although this challenge does not involve a dedicated manual annotation campaign, the reference labels are derived from clinical LVEF measurements, which are subject to well-documented inter-observer variability. For the biplane Simpson's method, inter-observer variability is typically reported as  $\pm 5$ –10 percentage points in the literature. This measurement variability represents an inherent ceiling on achievable algorithm accuracy and may introduce label noise, particularly for borderline cases near the LVEF = 50% binarisation threshold. To quantify this effect, we will report the proportion of cases with LVEF values within  $\pm 5\%$  of the classification threshold, as these cases are most susceptible to label uncertainty. Additionally, variability in image quality, acquisition technique, and patient positioning across the six clinical sites may introduce site-dependent biases that affect both reference measurements and algorithm performance.

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

Multi-vendor and multi-site acquisition variability constitutes the primary additional source of error. Differences in ultrasound equipment settings, image resolution, frame rate, and operator technique across the six participating centres may introduce systematic biases. Furthermore, temporal variability in cardiac function due to treatment

timing, patient hydration status, and heart rate at the time of acquisition may affect LVEF measurements independently of true dysfunction status. Site-specific differences in image quality and patient positioning may also contribute to performance variation across centres.

## ASSESSMENT METHODS

### Metric(s)

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

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

The primary ranking metric is the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which provides a threshold-independent measure of discriminative performance. In addition, the following secondary metrics will be reported at a clinically defined operating point (LVEF < 50%): Accuracy, Precision (Positive Predictive Value), Recall (Sensitivity), F1-Score, and Specificity. These secondary metrics are reported for descriptive and clinical interpretability purposes only and do not contribute to the final ranking. In the event of a tie in AUC-ROC (to three decimal places), the algorithm with higher sensitivity at 90% specificity will be ranked higher. If a tie persists, the algorithm with higher overall accuracy will take precedence.

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

AUC-ROC was selected as the primary ranking metric because it provides a threshold-independent assessment of how well an algorithm discriminates between patients with and without left ventricular dysfunction across all possible operating points. This is essential in a challenge setting where different clinical scenarios may favour different sensitivity–specificity trade-offs. Secondary metrics (Accuracy, Precision, Recall, F1-Score, Specificity) are reported at the clinically defined LVEF < 50% threshold to provide interpretable, actionable performance summaries. Recall (Sensitivity) is emphasised because missed cases of LV dysfunction carry greater clinical risk than false positives in a cardio-oncology screening context. The expected prevalence of LV dysfunction in the dataset is approximately 15–25%, reflecting the real-world class imbalance in this population; AUC-ROC is robust to such imbalance, further justifying its use as the ranking metric.

### 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. Ideally, provide the ranking scheme as a concrete pseudo code.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve) is the sole metric used for final ranking. For each participating algorithm, AUC-ROC is computed on the held-out test set. Algorithms are ranked in descending order of AUC-ROC. In the event of a tie (to three decimal places), the algorithm achieving higher sensitivity at 90% specificity is ranked higher; if still tied, overall accuracy at the LVEF < 50% threshold is used. Secondary metrics (Accuracy, Precision, Recall, F1-Score, Specificity) are reported but do not affect ranking.

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



Participating teams are required to submit Docker containers that automatically process all test cases on the challenge evaluation server. If a submission fails to produce a valid output for one or more test cases, those cases will be assigned the worst possible metric value for the corresponding task and metric. This policy ensures fair comparison across submissions and discourages partial or non-robust solutions.

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

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was selected as the sole evaluation metric for ranking participating algorithms in this challenge. This choice is motivated by several key considerations. First, AUC-ROC provides a threshold-independent measure of discriminative performance, evaluating how well an algorithm distinguishes between patients with and without left ventricular dysfunction across all possible decision thresholds. This is particularly important in a challenge setting, as different clinical scenarios may require different operating points depending on the desired balance between sensitivity and specificity. Second, AUC-ROC offers robustness to class imbalance, which is a common characteristic of clinical datasets where the prevalence of dysfunction may vary. Third, this metric is widely established and accepted within medical imaging and clinical AI communities, ensuring comparability with existing literature and facilitating the interpretation of results. Finally, AUC-ROC captures the overall ability of an algorithm to correctly rank positive cases higher than negative cases, providing a comprehensive and clinically meaningful assessment of model performance. For these reasons, AUC-ROC represents an objective, fair, and standardized basis for algorithm comparison in this task. A clinically useful algorithm should target an AUC-ROC  $\geq 0.85$ , with particular emphasis on high sensitivity ( $\geq 0.90$ ) to minimise missed cases of LV dysfunction, as false negatives carry greater clinical risk in a cardio-oncology screening context. To further contextualise results, we will report sensitivity and specificity at the clinically recommended LVEF  $< 50\%$  decision threshold, as well as the negative predictive value, given its direct relevance to ruling out dysfunction in a screening population. To anchor interpretation, two organiser-provided reference baselines (a standard video classification network and a fine-tuned EchoNet-Dynamic model) will be evaluated under the same protocol and published alongside the training data (see Further Analyses).

## Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

Statistical analysis is designed to ensure robust and unbiased evaluation of submitted algorithms for left ventricular cardiac dysfunction assessment. Performance is assessed using case-level classification metrics that are aggregated across the test set, and final rankings are computed using a rank-based aggregation scheme to avoid bias from differing measurement scales. Missing or invalid outputs are handled consistently by assigning worst-case performance values. No distributional assumptions are required, as the evaluation relies on non-parametric, rank-based methods.

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

To assess the statistical reliability and precision of the performance estimates for individual algorithms, we will compute 95% Confidence Intervals (CIs) for the primary metric (AUC-ROC) using non-parametric stratified bootstrapping with 1,000 resampling iterations. Stratification will preserve the class ratio (normal vs. dysfunction)

in each bootstrap sample. In addition, 95% CIs will be computed for all secondary metrics (Accuracy, Precision, Recall, F1-Score, Specificity) at the LVEF < 50% operating point using the same bootstrapping procedure.

Provide a description of how variability of the performance of individual algorithms across test cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

Since AUC-ROC is a "set-based" metric (calculated using the whole dataset at once), a standard deviation of the AUC across individual cases will not be calculated.

Provide a description of how variability of rankings is assessed.

The stability and reliability of algorithm rankings will be assessed using bootstrap-based ranking analysis. Specifically, the test set will be resampled with replacement across a large number of bootstrap iterations (e.g., 1,000 samples). For each bootstrap sample, the AUC-ROC will be computed for all participating algorithms, and a ranking will be derived based on these performance scores. This process generates an empirical distribution of ranks for each algorithm across all iterations.

Several complementary approaches will be employed to characterize ranking variability. First, the mean and standard deviation of ranks across bootstrap samples will be reported for each algorithm, providing a quantitative measure of ranking stability. Second, rank frequency matrices or heatmaps will be constructed to visualize how often each algorithm achieves each possible rank, offering insights into the consistency of relative performance.

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

Pairwise comparisons of AUC-ROC between the top-ranked algorithm and all other submissions will be performed using the DeLong test for comparing correlated AUCs. To account for multiple comparisons, Benjamini-Hochberg correction will be applied. A significance level of  $\alpha = 0.05$  will be used throughout. Additionally, McNemar's test will be applied to assess whether differences in binary classification outcomes (at the LVEF < 50% threshold) between pairs of algorithms are statistically significant. These tests will be reported alongside confidence intervals to provide a comprehensive picture of whether observed ranking differences reflect meaningful performance gaps or fall within the margin of statistical uncertainty.

Provide a description of the missing data handling.

No missing data will be provided to participants.

Indicate any software product that is used for all data analysis methods.

Participants are free to use any software product available.

## 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.

Task 2 is evaluated independently of Tasks 1 and 3. Participating teams may enter any combination of tasks. Rankings and awards are determined separately per task; there is no requirement or incentive to participate in all

three tasks.

## TASK 3: Early Prediction of Therapy-Induced Cardiotoxicity from Baseline Echocardiography

### 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.

Therapy-induced cardiotoxicity is a major cause of morbidity and treatment interruption in breast cancer patients receiving potentially cardiotoxic therapies. Although echocardiography is routinely used for longitudinal cardiac monitoring, current clinical practice primarily detects cardiac dysfunction after it has already developed, limiting opportunities for early intervention. There is a critical unmet need for methods that can predict cardiotoxicity risk before therapy initiation, using baseline imaging alone. This task challenges participants to develop AI models that predict future cardiotoxicity from baseline echocardiography videos. The task is based on data from 254 patients (249 baseline echocardiograms), reflecting a realistic clinical scenario in which only pre-treatment imaging is available at inference time. Participants may leverage models or imaging-derived parameters developed in Tasks 1 and 2, enabling integrated and multi-stage predictive approaches. From a technical standpoint, the task focuses on future outcome prediction from subtle and potentially subclinical imaging patterns, requiring robust spatiotemporal representation learning and risk modeling from heterogeneous ultrasound data. Methodological approaches may include video-based deep learning, multi-task or transfer learning strategies, and outcome-oriented risk prediction frameworks. The envisioned impact is twofold. Clinically, early prediction of cardiotoxicity could enable proactive risk stratification, closer monitoring, and timely cardioprotective interventions, improving patient outcomes and reducing unnecessary discontinuation of effective cancer therapies. Technically, this task promotes a shift from retrospective detection to prospective risk prediction from echocardiography, advancing the development of clinically actionable AI for personalized cardio-oncology care. Task 3 deliberately focuses on imaging-only predictive modeling using baseline echocardiography in order to isolate and quantify the prognostic contribution of imaging features. These models are not intended to replace comprehensive clinical risk assessment. In real-world practice, predicted risk scores would be interpreted alongside patient-specific factors such as age, comorbidities, therapy regimen, and longitudinal biomarker trends. By standardizing inputs to imaging alone, the challenge enables transparent benchmarking of echocardiographic risk prediction while preserving compatibility with multimodal clinical decision-making frameworks.

#### Keywords

List the primary keywords that characterize the task.

Echocardiography, therapy-induced cardiotoxicity, early risk stratification, predictive modeling

### ORGANIZATION

#### Organizers

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

Kostas Marias  
Foundation for Research and Technology - Hellas

Manolis Tsiknakis  
Hellenic Mediterranean University

Grigorios Kalliataakis  
Foundation for Research and Technology

Georgios Manikis  
Foundation for Research and Technology - Hellas

Georgia Karanasiou  
University of Ioannina

Eleni Georga  
University of Ioannina

Katerina Naka  
University Hospital of Ioannina

Dorothea Tsekoura  
National and Kapodistrian University of Athens

Gerasimos Filippatos  
National and Kapodistrian University of Athens

Anastasia Constantinidou  
Bank of Cyprus Oncology Centre

Andri Papakonstantinou  
Karolinska University Hospital

Ketti Mazzocco  
Istituto Europeo di Oncologia IRCCS Milano

Domen Ribnikar  
Institute of Oncology Ljubljana

Dimitrios Fotiadis  
University of Ioannina

b) Provide information on the primary contact person.

Kostas Marias, kmarias@ics.forth.gr  
Foundation for Research and Technology - Hellas

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

Yes, clinicians are integral members of the organizing team and play a central role in the clinical design, validation, and interpretation of the challenge. Their involvement ensures that the challenge addresses clinically relevant questions, reflects real-world clinical practice, and that the chosen evaluation metrics and outcomes are meaningful, interpretable, and impactful for patient care.

Katerina Naka (University Hospital of Ioannina) – Clinical lead for cardiac imaging data curation and annotation. She oversees the clinical validation of echocardiographic datasets, contributes to defining clinically relevant endpoints, and supports interpretation of challenge outcomes from a cardiology perspective.

Dorothea Tsekoura (National and Kapodistrian University of Athens) – Clinical advisor for cardio-oncology pathways and patient stratification. She contributes to the definition of inclusion/exclusion criteria, clinical metadata harmonization, and ensures alignment with current cardio-oncology guidelines.

Gerasimos Filippatos (National and Kapodistrian University of Athens) – Senior clinical advisor and scientific oversight. He provides high-level clinical guidance on heart failure and therapy-related cardiac dysfunction, ensuring the clinical credibility and translational relevance of the challenge.

Anastasia Constantinidou (Bank of Cyprus Oncology Centre) – Oncology clinical advisor. She supports the integration of oncological treatment context, therapy exposure variables, and clinically meaningful outcomes related to cancer therapy-related cardiotoxicity.

Andri Papakonstantinou (Karolinska University Hospital) – Clinical expert for international validation and benchmarking. She supports interpretation of results in a broader clinical context.

Ketti Mazzocco (Istituto Europeo di Oncologia IRCCS, Milan) – Clinical advisor for patient-centered outcomes. She provides input on clinically relevant outcome measures and supports alignment with patient-reported and longitudinal assessment perspectives.

Domen Ribnikar (Institute of Oncology Ljubljana) – Clinical advisor for real-world oncology practice. He advises on translational applicability of the developed methods.

### 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).

29th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2026)

b) Report the platform used to run the challenge.

The challenge will be run on the Synapse platform (<https://www.synapse.org>). Synapse provides a robust and widely used infrastructure for running biomedical data challenges, supporting secure data hosting, participant registration, submission management, leaderboard evaluation, and transparent benchmarking. The platform is particularly well suited for challenges involving sensitive clinical data, offering fine-grained access control, governance mechanisms, and compliance with ethical and data-use requirements. The use of Synapse ensures a secure, transparent, and reproducible challenge execution, aligned with MICCAI policies and best practices for clinical AI benchmarking.

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

Yes

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

The challenge will be hosted on Synapse (Sage Bionetworks), which will serve as the central platform for data access, documentation, submissions, and leaderboards. A public placeholder Synapse project has been created and will be populated with all challenge materials upon acceptance: [Synapse Project URL <https://www.synapse.org/Synapse:syn72001386>]

### Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

Fully Automatic.

All submitted methods must operate in a fully automatic manner at inference time, with no user interaction or manual intervention on validation or test data. This includes no manual selection of frames, regions of interest, contours, or case-specific parameter tuning. During training, participants are allowed to perform standard data curation, preprocessing, and augmentation on the provided training set, as well as pre-train models using external public datasets or foundation models. Participants using external public datasets or foundation models for pretraining or model development will be required to disclose them in their submission description. External retrieval, API calls, or internet access during inference will not be permitted.



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 may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

#### **Publicly available data is 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.

#### **May participate but not eligible for awards and not listed in leaderboard**

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

The challenge will recognize outstanding contributions based on the official final leaderboard ranking.

Top-performing teams will be invited to present their methods and results during the associated MICCAI 2026 challenge session or workshop, and selected teams will be invited to contribute to a joint challenge summary paper to be submitted to a peer-reviewed journal. Official certificates of achievement will be awarded to top-ranked teams. To further acknowledge excellence, monetary awards totaling €900 will be granted and distributed as follows: €450 for 1st place, €300 for 2nd place, and €150 for 3rd place. Rankings are computed exclusively based on performance on the hidden test set for Task 3. This award policy emphasizes excellence while ensuring a transparent and fair evaluation process in line with MICCAI challenge best practices.

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.

**One official leaderboard will be generated per task based on performance on the hidden test set. Test-set results will be released at the official challenge result announcement.**

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).

Upon completion of the challenge, the organizers will coordinate a comprehensive peer-reviewed publication summarizing the challenge design, key analyses, and results. The paper will be a collaborative effort involving the challenge organizing team, clinical and technical collaborators, and the top-performing participating teams. Each eligible team will be invited to nominate up to three co-authors, provided that the team has made a valid submission and complied with all challenge rules.

Participating teams remain free to publish their own methods and results independently; however, to preserve the integrity of the challenge outcomes, an embargo period of three months following the release of the official challenge results will apply. The organizers reserve the right to exclude teams from co-authorship in the challenge paper in cases of rule violations or non-compliant submissions.

## **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.

Submissions will be accepted in the form of Docker containers via the Synapse platform. Detailed submission instructions and technical requirements will be provided at the time of the official challenge announcement. All submitted algorithms must operate fully automatically within the provided containerized evaluation environment and must not access external resources during test-time execution. Participants are permitted to use publicly available external datasets and pretrained models during model development; however, all external data sources and pretrained models must be explicitly disclosed in the method description submitted with the final results. Any use of private or non-public datasets must be declared and must comply with the challenge data usage policy. Participants must ensure that no direct or indirect information from the hidden test set is used during training or model selection. Any form of metadata-based case identification, retrieval, or linkage to external databases is prohibited.

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.

Participants may submit up to three development runs evaluated on the validation set to verify correctness and obtain performance feedback. Validation metrics are returned privately to the submitting team and are not used to determine official rankings. Final rankings are based exclusively on performance on the hidden test set. Only the final valid submission before the deadline will be evaluated on the test set and used for official leaderboard placement.

### 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

April 1, 2026: Challenge website opens for registration; release of training and validation data

April 10, 2026: Submission system opens for validation submissions

July 15, 2026: Submission system opens for test submissions

August 20, 2026: Registration and Docker submission deadline

October 7, 2026: Release of final results during the MICCAI Annual Meeting

January 31, 2027: Publication of the challenge summary

### Ethics approval

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

The dataset used in the EchoRisk-MICCAI Challenge originates from the CARDIOCARE project (Horizon 2020, Grant Agreement No. 945175), a prospective, multicenter clinical study conducted across six European hospital sites in five countries. The CARDIOCARE study received ethical approval from the relevant Institutional Review Boards (IRBs) / Ethics Committees at each participating clinical center, in accordance with national regulations, the Declaration of Helsinki, and the EU General Data Protection Regulation (GDPR). Ethics approvals were obtained locally at each recruiting institution prior to patient enrollment and data collection, covering: (i) prospective acquisition of echocardiography imaging, (ii) collection of clinical and biomarker data, and (iii) longitudinal follow-up of breast cancer patients undergoing potentially cardiotoxic therapies. Imaging and associated metadata were fully de-identified following established DICOM anonymization standards prior to any secondary processing. For the purposes of the MICCAI 2026 Challenge, formal secondary-use authorization for release of the fully anonymized dataset is being processed across participating institutions. Approvals have already been received by multiple centers, while the remaining institutions are in the final stages of institutional confirmation following completion of the required review procedures. The dataset will be released exclusively in anonymized form and strictly for non-commercial scientific research. Secondary-use authorization is treated as a key prerequisite in the release process, and the current center-level approval status is reported transparently in the "Further Comments" section of this proposal.

### 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)

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

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

### 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 related to the EchoRisk-MICCAI training dataset, evaluation procedures, and challenge scoring metrics will be made publicly available through the EchoRisk-MICCAI GitHub repository. The evaluation code will be released

prior to the official start of the challenge, enabling participants to test their algorithms locally and to clearly understand the evaluation and ranking process.

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

All participants are required to submit their algorithms in the form of a self-contained Docker container for evaluation. The submitted code must be able to run without requiring additional manual setup or external dependencies beyond those specified by the organizers. For the purpose of reproducibility and transparency, participants are strongly encouraged to provide links to their source code on a public repository (e.g., GitHub, GitLab, or a similar platform); however, public code release is not a condition for participation during the challenge.

### **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.

CARDIO CARE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 945175. No other conflicts of interest. Test images will only be accessible to the challenge organizers.

## **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

Decision support, Research, CAD, Assistance

### **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.

The target cohort for this challenge comprises elderly and multimorbid women diagnosed with breast cancer who are undergoing potentially cardiotoxic cancer therapies, including anthracyclines and HER2-targeted agents. These patients represent a clinically vulnerable population with an elevated risk of therapy-induced cardiotoxicity due to age-related cardiovascular changes and the high prevalence of pre-existing comorbidities. The target cohort is characterized by a need for frequent, longitudinal cardiac assessment to enable early detection of subclinical cardiac dysfunction, risk stratification, and timely intervention. Given that older breast cancer patients are underrepresented in clinical trials but highly prevalent in real-world clinical practice, this cohort represents a critical target population for deployable, scalable, and equitable AI-based decision support tools in cardio-oncology.

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

The proposed challenge will release 2D echocardiography video data acquired from 421 breast cancer patients enrolled in the CARDIOCARE prospective clinical study. The dataset originates from six European cancer and cardiology centers across five countries (Italy, Cyprus, Sweden, Greece, and Slovenia), ensuring diversity in patient populations, acquisition protocols, and ultrasound equipment. All data represent real-world clinical scenarios, including patients undergoing potentially cardiotoxic cancer therapies and exhibiting a spectrum of normal cardiac function, subclinical changes, and therapy-induced cardiac dysfunction. The imaging dataset consists of grayscale DICOM echocardiography videos capturing at least one representative cardiac cycle, acquired in standard apical 4-chamber and 2-chamber views, and collected at multiple longitudinal timepoints, including baseline and follow-up examinations. To reflect routine clinical practice, the data include variability in image

quality, framing, patient anatomy, and operator technique, as well as occasional missing or suboptimal views. Echocardiography data were acquired using multi-vendor ultrasound systems from major manufacturers commonly used in clinical practice. In addition to imaging, the challenge cohort includes cardiotoxicity-related clinical labels, such as left ventricular ejection fraction, global longitudinal strain, and blood biomarkers (e.g., troponin and NT-proBNP), enabling clinically grounded evaluation of automated analysis and prediction methods. Ethical approvals for data acquisition were obtained at all participating centers, and all data are fully de-identified in compliance with GDPR and institutional regulations.

### Imaging modality(ies)

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

The challenge is based on Cardiac ultrasound (echocardiography)

### Context information

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

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

Additional information provided alongside the imaging data includes clinically relevant quantitative and biomarker measurements, specifically: (i) Left ventricular ejection fraction (LVEF), (ii) Cardiac troponin levels, (iii) N-terminal pro-B-type natriuretic peptide (NT-proBNP); these parameters are directly related to cardiac function and myocardial injury and are provided to support clinically meaningful model development and evaluation.

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

No additional patient-level context information (e.g., demographics, medical history, or clinical background) will be provided beyond the imaging data and the specified cardiac measurements.

### Target entity(ies)

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

The data for this challenge originate from cardiac ultrasound (echocardiography) imaging of the heart, with a primary focus on the left ventricle and associated cardiac chambers as visualized in standard apical 4-chamber and apical 2-chamber views. These views provide essential anatomical and functional information for assessing left ventricular size, systolic function, myocardial deformation, and early signs of therapy-induced cardiotoxicity. In the target biomedical application, echocardiography would be acquired from breast cancer patients undergoing potentially cardiotoxic therapies as part of routine cardiac monitoring before, during, and after treatment. The imaging focuses on longitudinal assessment of cardiac structure and function to enable early detection of subclinical dysfunction and risk stratification in real-world cardio-oncology practice.

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

For Task 3, participating algorithms are designed to analyze baseline 2D echocardiography videos acquired from standard clinical views and to predict the future development of therapy-induced cardiotoxicity at subsequent

follow-up time points. The algorithms leverage spatiotemporal information from baseline cardiac imaging, capturing structural and functional patterns across the cardiac cycle that may be indicative of increased cardiotoxicity risk. Approaches may implicitly or explicitly rely on learned video representations, imaging-derived functional parameters, or predictive models developed in Tasks 1 and 2. The emphasis of this task is on prospective risk prediction from pre-treatment echocardiography, rather than retrospective detection of established cardiac dysfunction.

### 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)).

Sensitivity, Specificity, Reliability, Robustness

## 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).

#### Ultrasound System

- Manufacturer: GE HealthCare
- Model: Vivid iq portable cardiovascular ultrasound system

#### System Configurations

- Standard Vivid iq
- Vivid iq Premium configuration
- Vivid iq Ultra Edition (latest release with enhanced AI and workflow tools)

#### Typical Hardware / Software Revisions in Clinical Use

- Vivid iq v204
- Vivid iq v206 (These correspond to commonly deployed hardware and software revision levels in routine clinical practice.)

#### Key Technical Specifications

- 15.6-inch touchscreen display
- Approximate weight: 5.2 kg (with battery)
- Portable laptop-style form factor
- Imaging modalities: B-mode, Color Doppler, Pulsed-Wave Doppler, Continuous-Wave Doppler, 2D and limited 4D



## imaging

- Integrated AI-based quantification tools (Ultra Edition), including automated ejection fraction and strain analysis

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

## Patient Preparation and Setup

Patient positioned supine or in the left lateral decubitus position

Continuous ECG tracing displayed during acquisition

Image optimization performed for gain, depth, and sector width

## Standard Imaging Views Acquired

Parasternal long-axis view

Parasternal short-axis views at basal, mid, and apical levels

Apical four-chamber view (cine loops and end-diastolic/end-systolic frames)

Apical two-chamber view (cine loops and end-diastolic/end-systolic frames)

Apical three-chamber view (cine loops and end-diastolic/end-systolic frames)

## Left Ventricular Systolic Function Assessment

Visual assessment of global and regional function across all views

Left ventricular ejection fraction (LVEF) measured using the biplane Simpson method from apical four- and two-chamber views

Endocardial borders traced at end-diastole and end-systole

Foreshortening avoided during acquisition and analysis

LVEF reported as a percentage

## Left Ventricular Volumes and Dimensions

LV end-diastolic volume (LVEDV) and end-systolic volume (LVESV) derived using biplane Simpson method

LV end-diastolic and end-systolic diameters measured in parasternal long-axis view

Septal and posterior wall thickness recorded

Global Longitudinal Strain (GLS)

Acquired from apical four-, two-, and three-chamber views

Frame rates between 50–90 frames per second

Adequate tracking quality required across all myocardial segments

GLS reported as a percentage

Regional Wall Motion Analysis

Segmental assessment using the standardized 17-segment LV model

Wall motion graded for each segment

Left Ventricular Diastolic Function Assessment

Mitral inflow E and A wave velocities

E-wave deceleration time

Tissue Doppler imaging of septal and lateral  $e'$  velocities

E/ $e'$  ratio

Left atrial volume index

Tricuspid regurgitation velocity

Supporting Findings and Additional Assessments

LV mass and geometric pattern

Presence of hypertrophy or dilation

Pericardial effusion

Valvular disease affecting LV loading conditions

Reporting Standards

Quantitative values reported with measurement method specified

Image quality statement included

Comparison with prior studies when available

Integrated clinical interpretation provided for patient management

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

The challenge dataset was collected across multiple clinical sites to ensure diversity, robustness, and generalizability. Participating institutions include: (i) Istituto Europeo di Oncologia (IEO), Italy; (ii) Bank of Cyprus Oncology Centre (BOCOC), Cyprus; (iii) Karolinska University Hospital / Karolinska Institutet (KSBC), Sweden; (iv) National and Kapodistrian University of Athens (NKUA), Greece; (v) University Hospital of Ioannina (UOI), Greece; and (vi) Institute of Oncology Ljubljana (IOL), Slovenia.

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 data were acquired by trained and certified cardiac sonographers with clinical experience in transthoracic echocardiography. Data acquisition was conducted under the oversight of cardiologists and senior clinical advisors, who ensured adherence to standardized imaging protocols, clinical quality control, and consistency across participating sites.

### Training and test case characteristics

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

Examples:

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

A case in this challenge refers to the complete set of data required to produce a single algorithm output that is compared against a corresponding reference outcome. For Task 3, a case corresponds to a single baseline echocardiography examination from one patient, acquired prior to or at the initiation of cancer therapy, with the reference outcome defined within the available follow-up window (up to 12 months). No follow-up imaging or clinical information beyond the baseline examination is provided to participants at inference time; the prediction must rely exclusively on baseline echocardiographic imaging data. The cases used in Task 3 correspond to

patients enrolled in the prospective CARDIO CARE multicenter study for whom sufficient longitudinal follow-up data were available to adjudicate cardiotoxicity outcomes. While the intended deployment population includes elderly and multimorbid women undergoing potentially cardiotoxic breast cancer therapy, the challenge dataset reflects the specific characteristics of the enrolled CARDIO CARE study participants. Accordingly, Task 3 should be interpreted as benchmarking predictive performance within this prospective clinical cohort rather than as a fully representative epidemiological sample of all patients receiving cardiotoxic therapy. External validation in independent cohorts would be required prior to broader clinical deployment.

Each case includes one or more 2D echocardiography video sequences (grayscale DICOM format) acquired from standard clinical views, capturing at least one representative cardiac cycle, together with associated metadata as provided by the organizers. The expected algorithm output for each case is a scalar probability (between 0 and 1) representing predicted risk of therapy-induced cardiotoxicity at follow-up. Participants may additionally leverage imaging-derived parameters or learned representations obtained in Tasks 1 and 2. For training cases, reference outcome labels indicating the presence or absence of cardiotoxicity at follow-up are provided to participants. For validation cases, baseline echocardiography videos are provided together with reference outcome labels to enable model tuning and internal evaluation. For test cases, baseline echocardiography videos are provided, while the corresponding outcome labels are strictly withheld by the organizers and used exclusively for final evaluation and ranking. Validation results are intended solely for model development and do not influence final rankings or awards. Each case is evaluated independently, enabling standardized and fair comparison of algorithm performance across patients.

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

For Task 3, the dataset is split at the patient level into training, validation, and test sets to avoid data leakage. Each case corresponds to a single baseline echocardiography examination from one patient. Of the 254 patients included in this task, 178 patients (70%) are allocated to the training and validation sets, while the remaining 76 patients (30%) are reserved for a held-out test set. The training and validation sets are further divided using an 80/20 split, resulting in 142 training cases and 36 validation cases. Test cases and their corresponding outcome labels remain hidden and inaccessible to participants and are used exclusively by the organizers for final evaluation and ranking.

c) How much of the data are already annotated (stratified by train test in percentage)?

All data included in Task 3 are annotated at the case level with clinically derived reference outcome labels. Cardiotoxicity labels are established by expert clinicians using a standardized adjudication process developed specifically for this project, representing a key methodological contribution of the challenge. This expert-driven labeling framework integrates longitudinal imaging findings and clinically relevant criteria to ensure accurate, consistent, and clinically meaningful outcome definitions.

Training set ( $\approx 56\%$ , 142 cases): All cases are annotated. Reference labels indicate the presence or absence of therapy-induced cardiotoxicity at a predefined follow-up time point. These labels are provided to participants for model training.

Validation set ( $\approx 14\%$ , 36 cases): All cases are annotated. Reference outcome labels are provided to participants and may be used for model development, hyperparameter tuning, and internal validation.

Test set ( $\approx 30\%$ , 76 cases): All cases are annotated. Reference outcome labels are available to the organizers only and remain hidden from participants throughout the challenge. These annotations are used exclusively for final evaluation and ranking.

Overall, 100% of the dataset is annotated, with outcome labels distributed across training, validation, and test

sets according to the predefined patient-level split.

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

The dataset for Task 3 includes 254 cases, corresponding to baseline echocardiography examinations from patients enrolled in the multicenter CARDIOCARE study. This cohort provides sufficient diversity for the development and evaluation of early cardiotoxicity prediction models. Data are split at the patient level into 142 training cases ( $\approx 56\%$ ), 36 validation cases ( $\approx 14\%$ ), and 76 test cases ( $\approx 30\%$ ) to prevent information leakage. The combined training and validation sets ( $\approx 70\%$ ) support effective model development and tuning across heterogeneous clinical conditions, while the held-out test set ( $\approx 30\%$ ) enables rigorous and unbiased assessment of model generalization in a real-world, multicenter, and multi-vendor setting.

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

For Task 3, cardiotoxicity outcome labels are derived from longitudinal clinical follow-up. While the majority of cases are annotated with therapy-induced cardiotoxicity status, a subset of cases does not carry this specific outcome annotation due to differences in follow-up duration or clinical data availability. Importantly, this does not affect the core challenge task, as all cases included in the evaluation splits are fully supported by the required outcome annotations. The presence of partially annotated data reflects real-world clinical data collection scenarios and introduces methodological flexibility and novelty, enabling the exploration of approaches that leverage incomplete labels, weak supervision, or semi-supervised learning strategies.

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

The challenge will include a substantial proportion of unseen and previously unpublished data derived from the ongoing CARDIOCARE prospective multicenter clinical study. All cases used for Task 3 correspond to baseline echocardiography examinations collected specifically within this study and have not been released in any prior public dataset or benchmark. In particular, all test cases (76 cases,  $\approx 30\%$  of the dataset) consist of unseen and unpublished data, which are withheld entirely from participants during the challenge and used exclusively for final evaluation and ranking. In addition, portions of the training and validation data comprise newly curated, real-world echocardiography acquisitions that have not been previously disseminated.

### 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 annotations for Task 3 correspond to clinically derived outcomes routinely used in cardio-oncology practice, rather than synthetic or in silico ground truth. Cardiotoxicity outcome labels are established based on longitudinal clinical follow-up and expert clinical assessment, integrating standard echocardiographic measurements and complementary clinical information. Left ventricular ejection fraction (LVEF) and global longitudinal strain (GLS), when available, were measured as part of routine clinical echocardiography workflows at the participating centers using vendor-provided or clinically validated software in accordance with local and international guidelines. Blood biomarkers related to cardiotoxicity, including troponin and NT-proBNP, were obtained through routine clinical laboratory testing at follow-up timepoints following standard hospital protocols.

These imaging- and biomarker-derived findings were jointly considered by expert clinicians to derive clinically meaningful cardiotoxicity outcome labels. Cardiotoxicity labels were derived from longitudinal clinical follow-up and structured expert adjudication based on established criteria (including LVEF decline and/or clinically significant GLS deterioration). No additional centralized re-annotation was performed specifically for the challenge. As such, outcome labels reflect real-world clinical practice across participating centers. It is important to note that LVEF and GLS measurements obtained in routine clinical practice are subject to known inter- and intra-observer variability. LVEF variability of approximately  $\pm 5$  percentage points and modest GLS variability across vendors and operators are well documented in echocardiographic practice. In addition, variability in follow-up timing and clinical decision thresholds may introduce heterogeneity in outcome assignment. These factors reflect inherent real-world clinical uncertainty rather than annotation error. Consequently, model performance should be interpreted within the context of this inherent variability, which defines a realistic upper bound for achievable predictive accuracy.

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.

Although the underlying clinical measurements were acquired as part of routine patient care, the cardiotoxicity outcome labels were not routinely recorded and required dedicated expert adjudication. Cardiotoxicity labels were established through a structured manual review process conducted by domain experts, who integrated longitudinal echocardiography findings and cardiotoxicity-related blood biomarkers (e.g., troponin and NT-proBNP) to derive clinically meaningful outcome definitions. This expert-curated labeling framework represents a unique contribution to the field of cardio-oncology, addressing a well-recognized gap in publicly available datasets where therapy-induced cardiotoxicity outcomes are rarely available in a standardized and adjudicated form. While no challenge-specific annotation software was required, the resulting reference labels reflect a deliberate, labor-intensive expert curation process rather than routine clinical reporting or automated extraction from electronic health records.

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.

Clinical measurements underlying the reference annotations were acquired during routine patient care by medically trained professionals, including cardiac sonographers and cardiologists, using vendor-provided or clinically validated echocardiography analysis software. Cardiotoxicity-related blood biomarkers were measured by certified clinical laboratories in accordance with standard hospital protocols. Importantly, while these measurements were obtained as part of routine clinical workflows, the cardiotoxicity outcome labels were subsequently derived through dedicated expert adjudication, applying the same standardized process consistently across the training, validation, and test cases.

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.

No challenge-specific pre-processing is applied to the raw data prior to release. Training and validation data are provided in their original DICOM format, preserving the native image quality, acquisition characteristics, and metadata as acquired in clinical practice. The test data are similarly provided without pre-processing, with reference labels retained by the organizers for evaluation. Participants are free to apply their own pre-processing, normalization, or data augmentation strategies as part of their algorithm development, provided that all processing at test time is fully automatic and compliant with the challenge rules.

### 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.

Reference outcome labels in Task 3 are derived through expert adjudication based on available longitudinal clinical follow-up data, integrating echocardiography findings and cardiotoxicity-related blood biomarkers. While this process ensures clinically meaningful and carefully curated labels, several sources of uncertainty inherent to real-world clinical practice must be acknowledged. First, cardiotoxicity determination relies on echocardiographic measurements such as left ventricular ejection fraction (LVEF) and global longitudinal strain (GLS), which are subject to known inter- and intra-observer variability. In routine practice, LVEF variability of approximately  $\pm 5$  percentage points is common, and GLS measurements may vary across vendors and operators. These measurement variations may influence classification of borderline cases. Second, variability in follow-up duration across patients at the time of data curation introduces temporal uncertainty. Specifically, cardiotoxicity labels reflect the presence or absence of therapy-induced cardiotoxicity within the observed follow-up period, rather than across the full 12-month study duration for all patients. As a result, a subset of patients labeled as non-cardiotoxic at the time of annotation may develop cardiotoxicity at later follow-up timepoints within the study period that were not yet available during data curation. This limitation reflects inherent constraints of prospective longitudinal clinical studies rather than annotation error, and mirrors real-world clinical practice, where risk assessment and clinical decisions are based on the information available at the time of evaluation. Importantly, all reference labels used for training, validation, and testing were derived using the same standardized expert adjudication process, ensuring internal consistency across the dataset. Model performance should therefore be interpreted within the context of inherent clinical measurement variability and follow-up uncertainty, which define a realistic upper bound for achievable predictive accuracy.

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

Multi-vendor and multi-site acquisition variability constitutes the primary additional source of error. Differences in ultrasound equipment settings, image resolution, frame rate, and operator technique across the six participating centres may introduce systematic biases. Furthermore, temporal variability in cardiac function due to treatment timing, patient hydration status, and heart rate at the time of acquisition may affect LVEF measurements independently of true dysfunction status. Site-specific differences in image quality and patient positioning may also contribute to performance variation across centres.

## ASSESSMENT METHODS

### Metric(s)

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



any).

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

For Task 3, algorithm performance will be evaluated based on the ability to predict future therapy-induced cardiotoxicity from baseline echocardiography. The primary ranking metric will be the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), computed on the hidden test set. AUC is selected because it evaluates discrimination independently of any specific decision threshold and is less sensitive to class imbalance than accuracy-based metrics, which may be misleading in settings with unequal outcome prevalence. Given the clinical objective of early risk stratification, false negatives are considered more consequential than moderate false positives, consistent with contemporary cardio-oncology guidelines emphasizing early identification of high-risk patients to enable intensified surveillance and cardioprotective intervention (Lyon et al., 2022). Accordingly, particular emphasis is placed on sensitivity within clinically meaningful operating regions. In addition to AUC, the following secondary metrics will be reported: sensitivity at fixed false positive rates of 10% and 20%, balanced accuracy, positive predictive value (PPV) and negative predictive value (NPV) at a predefined threshold reflecting observed prevalence, and calibration performance assessed using Brier score, calibration curves, and calibration slope and intercept. Calibration is a critical component of clinically reliable prediction models, reflecting agreement between predicted and observed risk (Van Calster et al., 2019; Steyerberg, 2019). The prespecified decision threshold will be set to the observed event prevalence in the training set (i.e., threshold = prevalence), to reflect a pragmatic screening operating point. The intended clinical application of Task 3 is early risk stratification at therapy initiation. In this context, patients predicted to be at high risk would undergo intensified cardiac monitoring, more frequent echocardiographic follow-up, or early initiation of cardioprotective therapy. Therefore, acceptable model performance must support safe screening. A clinically meaningful operating region corresponds to sensitivity  $\geq 80\%$  at false positive rates within 10–20%, reflecting a realistic tolerance for increased surveillance in cardio-oncology practice. Performance outside these operating regions, even if associated with high AUC, may not translate into clinically actionable benefit. In the event of tied AUC values (to three decimal places), ties will be resolved sequentially by: (1) higher sensitivity at a fixed false positive rate of 20%, (2) higher balanced accuracy, and (3) lower Brier score. Higher AUC, strong sensitivity within predefined operating regions, and reliable calibration indicate better predictive performance and greater clinical utility.

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

The area under the receiver operating characteristic curve (AUC) is used as the primary evaluation metric because it provides a threshold-independent measure of discriminative performance, which is essential for early cardiotoxicity risk prediction where optimal decision thresholds may vary across clinical settings. Sensitivity and specificity are reported as complementary metrics to assess clinical utility, reflecting the ability of models to correctly identify patients at high risk of cardiotoxicity while avoiding unnecessary false-positive alerts. Balanced accuracy is additionally reported to account for potential class imbalance, ensuring a fair and clinically meaningful comparison of algorithm performance across participants.

### 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. Ideally, provide the ranking scheme as a concrete pseudo code.



The final ranking for Task 3 is based on performance on the hidden test set using the area under the receiver operating characteristic curve (AUC) as the primary metric. For each submitted algorithm, AUC is computed by comparing the predicted risk probabilities of cardiotoxicity occurrence within the available follow-up window (up to 12 months) against the expert-adjudicated reference outcome labels across all test cases. Algorithms are ranked in descending order of AUC values, with higher AUC indicating better discriminative performance. Secondary metrics are reported for clinical interpretation and are used only for tie-breaking as specified above. In the event of tied AUC values (to three decimal places), ties will be resolved sequentially by: (1) higher sensitivity at a fixed false positive rate of 20%, (2) higher balanced accuracy, and (3) lower Brier score. Task 3 is evaluated independently from Tasks 1 and 2, and participation in this task does not require submission to other tasks. In addition to participant submissions, a predefined clinical reference baseline will be evaluated for benchmarking purposes. This baseline is derived from established cardio-oncology risk stratification frameworks (e.g., HFA-ICOS) and will be computed internally by the organizing team using available clinical variables within the CARDIOCARE dataset. Clinical variables used for baseline computation will not be accessible to participants. Baseline performance will be evaluated under the same test-set partitions and outcome definitions as participant submissions and will not influence participant ranking. The baseline is included solely to contextualize imaging-based predictive performance relative to current guideline-aligned clinical practice. This clinical baseline is reported solely as contextual reference and is not comparable in input modality (multimodal vs imaging-only). Participant ranking remains strictly imaging-only.

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

Participating teams are required to submit Docker containers that automatically process all test cases on the challenge evaluation server. If a submission fails to produce a valid output for one or more test cases, those cases will be assigned the worst possible metric value for the corresponding task and metric. This policy ensures fair comparison across submissions and discourages partial or non-robust solutions.

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

The proposed ranking scheme is designed to reflect clinically meaningful performance in the early prediction of therapy-induced cardiotoxicity. By using a threshold-independent primary metric and evaluating predictions on a fully held-out test set, the ranking prioritizes the ability of algorithms to correctly discriminate between patients who will and will not develop cardiotoxicity within the defined follow-up period. This approach aligns with real-world cardio-oncology practice, where early and reliable identification of patients at elevated risk is essential for guiding monitoring strategies and preventive interventions. The ranking framework provides a fair, transparent, and robust comparison of algorithm performance across heterogeneous, multicenter echocardiography data, supporting the development of clinically actionable predictive models. Final leaderboard positions are determined exclusively based on performance on the hidden test set; validation results are provided for development purposes only and do not influence ranking or awards.

## Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

Statistical analysis is designed to ensure robust and unbiased evaluation of submitted algorithms for early cardiotoxicity prediction.

Performance is assessed at the case level using probabilistic classification metrics, with discrimination (AUC-ROC) as the primary ranking criterion and threshold-based operating characteristics and calibration performance evaluated secondarily. Missing or invalid outputs are handled consistently by assigning worst-case performance values to ensure fairness across submissions. No distributional assumptions are required, as the evaluation relies on non-parametric performance metrics appropriate for binary outcome prediction.

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

For Task 3, statistical analysis is designed to provide reliable estimates of model performance for early prediction of therapy-induced cardiotoxicity from baseline ultrasound videos. Performance metrics are computed at the case level and summarized across the test set using standard descriptive statistics, providing an overall view of predictive performance.

Provide a description of how variability of the performance of individual algorithms across tests cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

Performance variability across cases is assessed by examining case-level predictions across the test cohort. This analysis captures variability related to differences in patient characteristics, imaging quality, and clinical outcomes, and provides insight into model robustness under real-world conditions.

Provide a description of how variability of rankings is assessed.

Variability in rankings is assessed by analyzing the consistency of model performance across evaluation metrics and test cases. Rankings are derived using a consistent evaluation scheme applied uniformly to all submissions, ensuring that overall rankings reflect stable and comparable performance rather than isolated results.

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

To formally assess whether observed differences in predictive performance between submitted algorithms are statistically significant, case-level paired statistical analyses will be conducted on the hidden test set. Differences in AUC between algorithms will be evaluated using DeLong's test for correlated ROC curves, which accounts for the paired nature of predictions on the same test cases. This method is appropriate for comparing discriminative performance of classification models. For threshold-based performance metrics (e.g., sensitivity at fixed false positive rates), paired comparisons will be conducted using McNemar's test for binary outcomes. This ensures that differences in clinically meaningful operating performance are formally evaluated. When multiple pairwise comparisons are performed across submitted algorithms, p-values will be adjusted using the Holm-Bonferroni procedure to control the family-wise error rate. In addition to hypothesis testing, 95% confidence intervals for AUC and sensitivity will be computed using non-parametric bootstrap resampling across test cases. Effect sizes, including absolute differences in AUC and sensitivity, will be reported alongside adjusted p-values to support clinical interpretation.

Provide a description of the missing data handling.

All test cases used in Task 3 undergo quality control prior to evaluation. In cases where a submitted algorithm fails to produce a valid prediction for a test case, predefined handling rules are applied consistently across all submissions to ensure fair comparison.

Indicate any software product that is used for all data analysis methods.

Statistical analyses, performance evaluation, and result visualization are performed using standard scientific computing tools, primarily Python-based libraries commonly used in medical imaging research.

### 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.

N/A

## 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.

Manikis, G., et al. "Association of echocardiographic radiomics-based features with cardiotoxicity effect in breast cancer patients from the CARDIOCARE project." *European Heart Journal-Cardiovascular Imaging* 26.Supplement\_1 (2025): jeae333-028.

Karanasiou, G., et al. "A multimodal approach for the management of co-morbid cardiotoxicity in the elderly breast cancer patients." *European Journal of Cancer* 175 (2022): S40.

Karanasiou, G., et al. "CARDIOCARE: An integrated platform for the management of elderly multimorbid patients with breast cancer therapy induced cardiac toxicity." *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2023.

Georga, Eleni I., et al. "Utilizing Machine Learning for the Identification of Pre-Treatment Prognostic Non-Imaging Biomarkers of Cancer Therapy-Related Cardiac Dysfunction in Female Patients with Breast Cancer." *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*. Vol. 2025. 2025.

Ouyang, David, et al. "Video-based AI for beat-to-beat assessment of cardiac function." *Nature* 580.7802 (2020): 252-256.

Lang, Roberto M., et al. "Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging." *European Heart Journal-Cardiovascular Imaging* 16.3 (2015): 233-271.

Thavendiranathan, Paaladinesh, et al. "Reproducibility of echocardiographic techniques for sequential

assessment of left ventricular ejection fraction and volumes: application to patients undergoing cancer chemotherapy." *Journal of the American College of Cardiology* 61.1 (2013): 77-84.

Lyon AR et al. 2022 ESC Guidelines on cardio-oncology. *Eur Heart J*. 2022.

Plana, Juan Carlos, et al. "Expert consensus for multimodality imaging evaluation of adult patients during and after cancer therapy: a report from the American Society of Echocardiography and the European Association of Cardiovascular Imaging." *European Heart Journal–Cardiovascular Imaging* 15.10 (2014): 1063-1093.

Steyerberg, Ewout W. *Clinical prediction models*. Vol. 201. No. 9. Cham: Springer International Publishing, 2019.

Van Calster, Ben, et al. "Calibration: the Achilles heel of predictive analytics." *BMC medicine* 17.1 (2019): 230.

Vickers, Andrew J., and Elena B. Elkin. "Decision curve analysis: a novel method for evaluating prediction models." *Medical Decision Making* 26.6 (2006): 565-574.

### Further comments

Further comments from the organizers.

Center-level authorization details and current status:

#### 1. BOCOC.

Formal secondary-use approval granted by the local Ethics Committee.

Protocol Number: EEBK/EP/2022/58.

Date of approval: December 18, 2025

#### 2. IOL.

Reviewed by the Scientific Committee for Protocol Assessment.

Legal department confirmation obtained.

Submission to the Ethics Committee pending.

#### 3. NKUA

Formal secondary-use approval granted by the local Ethics Committee.

Protocol Number: 456-3/12-02-2026.

Date of approval: February 12, 2026.

#### 4. IEO.

Submission completed.

Currently under institutional review.

Decision pending.

#### 5. KSBC.

Secondary-use documentation submitted and reviewed.

Amendment requested by the Institutional Review Board and being submitted.

Final approval pending completion of the amendment process.

#### 6. UOI

Submitted to local Ethics Committee.

Currently under review.

Decision anticipated at the end of February 2025.