nature portfolio

Corresponding author(s): Rohan Shad

Last updated by author(s): 12/12/2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Repository for this project containing model checkpoints, raw table and figure data, and additional scripts to preprocess and analyze data are available at https://github.com/rohanshad/cmr_transformer. Final contrastive pretrained checkpoint for our models are available for academic use: https://huggingface.co/rohanshad/cmr_c0.1. Please contact the corresponding and first authors for commercial licenses. Additional scripts and preprocessing code for handling Cardiac MRI data are available free of cost at https://github.com/rohanshad/cmr_toolkit.
Data analysis	We use the pytorch deep learning library (v1.11.0) and the pytorch lightning framework (v1.8.6). Major python packages used in this work include numpy (v1.21.2), pydicom (v2.0.0), transformers (v4.4.2), and stanza (v1.5.0). Statistical analyses were performed, and graphs were plotted using R (v4.1.0); major packages used include pROC (v 1.17.0), ggplot2 (3.3.5), blandr (0.5.1). The online test-set leaderboard webapp was created using shiny (1.8.1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Cardiac MRI Data from Stanford Medicine, UCSF, and Medstar were secured via agreements with Bunkerhill Health. Applications for data access can be found at https://www.bunkerhillhealth.com. Cardiac MRI data from the University of Pennsylvania are available to researchers under appropriate data use agreements. Please contact Rohan Shad, MD for additional information. Applications for data access from the UK Bio-Bank can be found at https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access. The ACDC dataset is publicly available at https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html. The Kaggle MRI dataset is publicly available at https://www.kaggle.com/competitions/second-annual-data-science-bowl/data. Leaderboard for classification and regression test datasets are available at https://rohanshad.shinyapps.io/cmr_leaderboard/. We are happy to accept model weights as submissions to evaluate on our existing datasets, please contact Rohan Shad, MD for additional information.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We include Cardiac MRI scans from all self-reported genders. No gender based analyses were performed for this study given adequate representation or small sample sizes (for certain disease classification experiments).
Reporting on race, ethnicity, or other socially relevant groupings	We report demographic data from all our clinical sites including age, sex, and self reported race and ethnicity in this study. We do not use any socially relevant groupings to draw conclusions from our results.
Population characteristics	Population characteristics including demographics, gender, age distributions along with scanner manufacturer and field strength are detailed in the Supplementary Appendix. Additionally, we provide prevalences and severity of various disease conditions in our datasets stratified by clinical site in the Supplementary Appendix.
Recruitment	All available cardiac MRI studies were included from all sites given the retrospective nature of the study
Ethics oversight	Data collection and research began following approval and waiver of consent by the Institutional review board of Stanford University (Protocol #60342; March 2021). Additional data from the University of Pennsylvania was sourced after retrospective collection was deemed IRB exempt by the University of Pennsylvania Health System (Protocol #852332; November 2022).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were performed. We used all available Cardiac MRI scans from public and private sources (65544 unique individuals; 71080 unique scans) across the Stanford, MedStar, UCSF, UPenn, Kaggle Data Science Bowl, UK BioBank, and ACDC datasets. For some experiments we show the effect of fine-tuning our deep learning models on subsets of available data (i.e. 1% and 10% of the available data). For certain rare disease labels, sample size was limited by the number of cases available. Details of source data are available in the Methods and Supplementary Appendix
Data exclusions	We excluded MRI scans with no cine-sequences or those with corrupted and unrecoverable DICOM files across all datasets. For the pre-training and downstream clinical evaluation experiments we excluded 98 scans that did not have associated reports, or those with fewer than 3 sentences.
Replication	All attempts at replication of our experiments across large-scale pre-training, zero shot evaluations on the UK BioBank and ACDC dataset, and finetuning experiments on the LVEF% regression and disease diagnosis problems were successful. In Fig 3a. we explicitly show replicated results across different randomized seeds and random subsets of the UK BioBank data as boxplots with interquartile ranges described where appropriate. We replicated final experiments three times on separate hardware each time with a fixed random seed yielding identical results.

Randomization	Cardiac MRI scans were randomly split to training, validation, and testing arms wherever applicable at the patient level (ensuring scans of the same patient / individual were not present across arms). Specific details are available in the Methods section of our manuscript. Detailed experiment level descriptions of splits along with proportions are described in the Methods and Supplementary Appendix. No attempt was made to control for covariates as we only assess the diagnostic abilities of our deep learning system.
Blinding	Blinding was not relevant to our study given the abscence of intervention arms.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A