nature portfolio

Peer Review File

A Generalizable Deep Learning System for Cardiac MRI

Corresponding Author: Dr Rohan Shad

Version 0:

Decision Letter:

Dear Dr. Shad,

Thank you again for submitting to *Nature Biomedical Engineering* your manuscript, "A Generalizable Deep Learning System for Cardiac MRI". The manuscript has been seen by 2 experts, whose reports you will find at the end of this message.

You will see that the reviewers appreciate aspects of the work. However, they articulate concerns about the degree of support for some of the claims and about the advance that the work represents over relevant published studies, and provide useful suggestions for improvement. We would ask that you address all of these concerns - they are mostly technical and should be feasible. Reviewer 2 especially has some comments related to the clinical feasibility and application of the method, and we would like to see more of this in the paper, if possible.

When you are ready to resubmit your manuscript, please upload the revised files, a point-by-point rebuttal to the comments from all reviewers, the reporting summary, and a cover letter that explains the main improvements included in the revision and responds to any points highlighted in this decision.

Please follow the following recommendations:

*Please submit your revised manuscript in a word doc or LaTeX format.

* Clearly highlight any amendments to the text and figures to help the reviewers and editors find and understand the changes (yet keep in mind that excessive marking can hinder readability).

* If you and your co-authors disagree with a criticism, provide the arguments to the reviewer (optionally, indicate the relevant points in the cover letter).

* If a criticism or suggestion is not addressed, please indicate so in the rebuttal to the reviewer comments and explain the reason(s).

* Consider including responses to any criticisms raised by more than one reviewer at the beginning of the rebuttal, in a section addressed to all reviewers.

* The rebuttal should include the reviewer comments in point-by-point format (please note that we provide all reviewers will the reports as they appear at the end of this message).

* Provide the rebuttal to the reviewer comments and the cover letter as separate files.

We expect that you will be able to resubmit the manuscript within 25 weeks of receiving this message. If this is the case, you will be protected against potential scooping. Otherwise, we will be happy to consider a revised manuscript as long as the significance of the work is not compromised by work published elsewhere or accepted for publication at *Nature Biomedical Engineering*.

We hope that you will find the referee reports helpful when revising the work. Please do not hesitate to contact me should you have any questions.

Best wishes,

Barbara Cheifet

Editor
Nature Biomedical Engineering

_____
Reviewer #1 (Report for the authors (Required)):

I had the pleasure of reviewing the manuscript by Shad et al. This manuscript introduces a foundational deep learning system for cardiac MRI, leveraging self-supervised contrastive learning to integrate visual data from cine-sequence cardiac MRI scans with the corresponding radiology reports. Trained on data from multiple academic institutions and validated using diverse datasets, the system achieves good performance in estimating left ventricular ejection fraction and diagnosing 35 cardiovascular conditions, including hypertrophic cardiomyopathy and cardiac amyloidosis. By reducing reliance on extensive manual labeling and training data, the model demonstrates robust generalizability and efficiency, allowing diverse diagnostic applications and phenotyping for cardiovascular MRI . Similar work was recently shown for echocardiography and pathology applications.

The authors are commended for their detailed methodology, use of multisite data, external validation, and commitment to open science by providing code under an MIT license. As the first major application of its kind for cardiovascular MRI, this work represents significant advancement. Notwithstanding, more comparisons and perspective should be provided to discuss this advance in the context of similar work in echocardiography and other fields.

The limitations which could be improved involve use of only basics cine sequences, image only use without usual demographic data, and lack of clarity about what fine tuning -which pollutes the data hygiene of otherwise well separated datasets.

Some specific comments are given below

Demographic data for UPenn is missing -this is the most important set to look into this as this is the only complete clinical external dataset Kaggle set is also missing demographic information.
This work is overly reliant on basic cine sequences. This basic information Cine MRI, while useful, only captures a subset of cardiac pathologies. The exclusion of other data types, such as LGE or T1/T2 mapping, limits the model's ability to handle more complex diagnoses that require these inputs.
Unlike clinicians, the system does not seem to consider demographic data for reports but rather learns them from the images. This seems redundant. Why not provide additional demographics data routinely available to the clinicians along the images to further improve diagnosis.
There is no evidence or discussion of prospective validation in clinical practice to show real-world utility.
Was center cropping performed fully automatically and was there any quality control for images going outside of the field-of-view?
Why was ChatGPT turbo 3.5 used to parse the free text to diagnosis rather than later systems ?
The bland Altman 95% LOA -12 +13% for LVEF seem rather large -similar to what is achieve in 2D echo. MRI is considered to be a more precise technique. Can outliers be shown and discussed. references 32 and 33, 34 are rather old and newer approaches shoud be discussed
Finetuning is performed in Biobank set, raises concerns about the separation between training and testing datasets, which could lead to inflated performance metrics. Authors should either perform fine tuning in one of the 3 sets used for development or shoud not consider biobank as part of external testing
Several figures in the appendix contain undefined abbreviations
AUCS for detection of some clinically important conditions seems very low -borderline significant or non-significant for example cardiomyopathy, bicuspid valve, myocarditis aortic stenosis. The authors do not provide sufficient analysis of why this is the case or how the system could be improved for these critical conditions. The authors should consider these accuracies in the clinical context. What would clinical accuracy be ?
Certain figures and tables lack clarity or are not sufficiently detailed (e.g., the number of cases with specific diagnoses in training and testing datasets is missing). N for each occurrence shoud be given in Supplementary table 5 -7 and figures 9 and 10

Supplementary table 8. Bland Altman plot should be provided

The self-attention maps provided are not intuitive or clinically meaningful. The paper does not demonstrate how these maps could aid clinicians in interpreting model decisions or improve trust in the system. Can some validation and proposed use be shown for these.
In addition, clinical examples of attention maps should be shown for specific disease patterns. Otherwise, this material is not very helpful. Results without any validation performed should not be shown.

The augmentation scheme seems simplistic: only resize, color jitter, shear, translate and rotate videos in spatial dimensions. These may not adequately reflect real-world variations in cardiac imaging, such as motion artifacts, poor image quality, or varying contrast noise levels. Can more details be given and potential improvements to achieve more realistic augmentation schemes be discussed? Unclear what color jitter represent in terms of cardiac imaging

Were Delong comparisons performed as paired? it is unclear from the methods.

Reviewer #1 (Remarks on code availability):

While I did not run or install the code the MIT license is given and code seems complete on Github

There is a README file with basic information

Few key datasets shoud be provided with he code as well as expected results for these cases should be provided.

Reviewer #2 (Report for the authors (Required)):

This paper describes a foundational vision system for CMR using deep learning. The system is based on self-supervised contrastive learning on CMR scans and accompanying reports. The model was trained and evaluated using data from four large academic institutions in the US, with additional validation on the UK Biobank and other public datasets.
The authors use a transformer-based vision system trained on 19,041 CMR scans employing a large language model to vectorize the radiology reports.
The model seems generalizable and data-efficient and allows basic grouping of patients into major disease categroies without supervision.
In general, the approach is exciting, novel and potentially game changing. Using a self-trained vision model instead of requiring contouring would change the workflow and possibly reduce human interaction and thus improve scalability.

However, so far, I see a significant number of shortcomings and further work that mandatorily needs to be done to convince me.

Main issues:

1. The core data provided demonstrates, that the pretrained model performs better, than the model trained on a kinetic dataset. While it shows the value of pretraining to learn cardiac-specific features, the untrained model focusses on the distinction of short- and long-axis views. Again, this is not surprizing, as this is the core difference of the datasets provided. It would be much more meaningful to see how the extensively self-trained model performs versus a model trained just to recognize the different anatomical views, so it can then – without further training – start to assess the differences within the views.

2. The ROC curves and AUCs provided are helpful to see the generally superior performance of the model versus an ill-trained model (see above), but they do not provide any information on clinical value. For some clinical issues it is highly important to have very high sensitivity even if the specificity is low (e.g. in rare cases with amyloidosis, which MUST be detected, as there is early, lifesaving therapy). While this is discussed in the outlook, it is not assessed in the data-analysis. Given the low prevalence, the NIR (no information rate) would be very high (and potentially even better than the model), but would miss these highly relevant cases. The model must be able to flag these cases. The ROC curves also show – different to how the data is presented in the main text – that the accuracy of the models is terrible. While the AUCs were reasonable for Tetralogy (which is easy to spot due to massively abnormal cardiac chamber sizes and position) and amyloid (with the caveats mentioned above), the majority of AUCs is not convincing. In addition, the main indications for a CMR study, such as myocardial infarction, viability and ischemia were not assessed at all.

3. Another major issue with the data presented is that the majority of CMR scans are not done to generate a diagnosis in a patient with no previous knowledge, but rather the opposite. CMR scans are ordered to make differential diagnoses, e.g. such as working out the underlying cause of heart failure or left ventricular hypertrophy. It would be important to test the ability of the algorithm to perform differential diagnoses in borderline cases.

4. The presentation of results lacks sufficient detail to fully assess the model's performance. For continuous variables like ejection fraction, detailed plots showing the distribution of errors and outliers are needed. More clinically relevant metrics and visualizations are required to understand the practical implications of the model's performance. The ROC curves provide most information, and I have explained my issues with them above.

A few smaller comments:
5. Please provide data about the severity of the diseases? Early recognition or end-stage disease? Please provide data on prevalence of the various categories in each dataset?

6. Do you know how accurate the vectorization of the reports was (do you have a metric for this and did you assess it against this metric).
7. Do you know how accurate the reports themselves were, as there can be large differences in reporting between different radiologists. Was this validated?
8. How would a physician looking after the patient analysed with your algorithm understand how likely the patient was correctly classified, how can they reassess if there are no numbers for volumes, wall thickness, muscle mass, etc. as this is the classical pathway for physicians to make decisions.
9. Cine imaging – while important – provides only a small fraction of the important data for the final diagnosis and especially the differential diagnoses as discussed earlier. Most differential diagnoses are based on scar images (LGE), perfusion and maps.
10. Generalizability remains an issue. The UK Biobank data is highly homogenous due to identical scanners and protocols used.
11. The paper repeatedly claims the model "understands" cardiovascular disease. This is an overstatement. The model learns statistical associations but doesn't possess genuine understanding in the way a clinician does.

I am not an expert in models themselves and can not assess the technical performance, requirements or novelty of the vision model or the LMMs used for fine-tuning.

Version 1:

Decision Letter:

Dear Rohan,

Thank you for submitting your revised manuscript "A Generalizable Deep Learning System for Cardiac MRI" (NBME-24-3442A). Having consulted with the original reviewers, I am pleased to write that we shall be happy to publish the manuscript in Nature Biomedical Engineering, pending minor revisions to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW
Nature Biomedical Engineering offers a transparent peer review option for new original research manuscripts submitted from 17th February 2021. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't.** Failure to state your preference will result in delays in accepting your manuscript for publication.

Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Author names using non-Roman characters
Nature Portfolio journals can support presentation of author names using non-Roman characters in the HTML version of the article. If you wish to, please include author names in parentheses after the Roman-character spelling; see example online here. Currently supported scripts are: Arabic, Chinese, Cyrillic, Devanagari, Greek, Hebrew, Hangul, Japanese and Persian. You will be asked to verify the rendering is correct at proof stage.

Sincerely,
Rita

Rita Strack, Ph.D.
Chief Editor

Nature Biomedical Engineering

Reviewer #2 (Report for the authors (Required)):

Thank you for the thorough response to my questions and criticism. I have no further issues with the manuscript. Well done.

Version 2:

Decision Letter:

Dear Dr Shad,

I am happy to inform you that your manuscript, "A Generalizable Deep Learning System for Cardiac MRI", has now been accepted for publication in *Nature Biomedical Engineering*.

Over the next few weeks, the figures will be checked for production quality, the text edited to ensure that it conforms to house style, and the manuscript typeset.

Our Articles are published about 40 days after the acceptance date (we recommend that you inform your institutional press office of this timeframe), and you will be notified of the actual publication date a few days in advance. Articles can be published any working day of the week, and are pushed live shortly after 10 am London time.

Publishing agreement. You will be asked to digitally sign a publishing agreement (grant of rights). After the signed publishing agreement has been received, the proofs of the article will be sent to you for review. If you have any queries during the production process, or you cannot meet the requested deadline for returning the proofs, please contact rjsproduction@springernature.com.

Nature Biomedical Engineering is a Transformative Journal. Authors may publish their research with us through the traditional subscription access route, or make their paper immediately open access through payment of an article-processing charge. More information about publication options is available.

You may need to take specific actions to comply with funder and institutional open-access mandates. If the work described in the accepted manuscript is supported by a funder that requires immediate open access (as outlined, for example, by Plan S) and your manuscript was originally submitted on or after January 1st 2021, then you should select the gold OA route. Authors selecting subscription publication will need to accept our standard licensing terms (including our self-archiving policies), and these will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Acceptance of your manuscript is conditional on agreement, by all authors, with both our media embargo and confidentiality and pre-publicity policies. In particular, you may arrange your own publicity of the Article (for instance, through your institutional press office), as long as you ensure that journalists strictly adhere to the media embargo.

To assist you in disseminating the work, as soon as the Article is published you will be able to take advantage of the Springer Nature SharedIt initiative to generate a unique shareable link to the Article that will allow anyone (with or without a subscription) to read it. Recipients of the link who are subscribers will also be able to download and print the PDF.

Thank you for having submitted this work to *Nature Biomedical Engineering*.

Best wishes,
Rita

Rita Strack, Ph.D.
Chief Editor
Nature Biomedical Engineering

We thank the reviewers for their insightful feedback and for highlighting the strengths of our work, including the detailed methodology, openly available codebase and model weights, and validation across multiple external datasets. We have added two authors to this revision—**Mrudang Mathur** and **Joseph Cho**—who contributed substantially to the new experiments and analyses presented here. The updated list and order of authors is reflected in our manuscript and supplementary appendix.

In response to the reviewers' comments, we have performed several additional experiments and analyses that we believe further strengthen the rigor and generalizability of our study. These include revised classification experiments with an updated labeling strategy and improved results, expanded analyses of our regression models, and a detailed case-by-case discussion of borderline examples. A comprehensive summary of these updates is provided below. We have attached our revised manuscript with markup, and supplementary appendix (with and without markup for ease of reading).

Each reviewer comment is addressed in depth in blue text, with specific amendments to the content of the manuscript or supplementary appendix clearly *demarcated and italicized*.

Reviewer #1:

I had the pleasure of reviewing the manuscript by Shad et al. This manuscript introduces a foundational deep learning system for cardiac MRI, leveraging self-supervised contrastive learning to integrate visual data from cine-sequence cardiac MRI scans with the corresponding radiology reports. Trained on data from multiple academic institutions and validated using diverse datasets, the system achieves good performance in estimating left ventricular ejection fraction and diagnosing 35 cardiovascular conditions, including hypertrophic cardiomyopathy and cardiac amyloidosis. By reducing reliance on extensive manual labeling and training data, the model demonstrates robust generalizability and efficiency, allowing diverse diagnostic applications and phenotyping for cardiovascular MRI. Similar work was recently shown for echocardiography and pathology applications.

The authors are commended for their detailed methodology, use of multisite data, external validation, and commitment to open science by providing code under an MIT license. As the first major application of its kind for cardiovascular MRI, this work represents significant advancement. Notwithstanding, more comparisons and perspective should be provided to discuss this advance in the context of similar work in echocardiography and other fields.

We thank reviewer #1's comments on the significant technical advance of our work. Our preprint preceded many similar but independent areas of research including those employing contrastive learning in the field of echocardiography and CT. We are happy to expand the discussion section to summarize some of the more recent developments in the interim period, we have added citations for these papers in the main section of our manuscript ("Pretraining framework and evaluation of low-dimensional representations"; Paragraph 1):

"Recent advances in self-supervised deep learning have shown promise in reducing this reliance on vast quantities of expert labelled data across a wide range of modalities ranging from pathology slides to chest X-rays.[1–8] *Yuhao et. al describe a framework wherein unstructured text could be used as a method of self-supervision for a Chest X-Ray deep learning system.*[3] *In this method of contrastive learning, two neural networks are used to produce a pair of low-dimensional representations for each pair of contextually related inputs from two separate modalities.*[9] *We extend these concepts to the spatiotemporal and multi-view problem of Cardiac MRI. During the pre-training process, the networks are trained to match true pairs of text report and MRI scans by optimizing for a contrastive objective.*[9] *Since*

the release of our preprint, others have reported promising results using similar contrastive-learning methods to train foundation models for computed tomography (CT) and echocardiography[10–12]*."*

The limitations which could be improved involve use of only basics cine sequences, image only use without usual demographic data, and lack of clarity about what fine tuning -which pollutes the data hygiene of otherwise well separated datasets.

Some specific comments are given below

Demographic data for UPenn is missing -this is the most important set to look into this as this is the only complete clinical external dataset Kaggle set is also missing demographic information.

We apologize for this accidental omission, UPenn demographic data are now detailed in the supplementary appendix (Supplementary Table 1). Demographic information from the Kaggle dataset was never made publicly available. We contacted the original authors of the dataset, and while we can confirm that the patients were recruited by the NIH from the DC metro area, we have no demographic summaries available to report. The original authors have since left their roles at the NIH and moved to industry.

This work is overly reliant on basic cine sequences. This basic information Cine MRI, while useful, only captures a subset of cardiac pathologies. The exclusion of other data types, such as LGE or T1/T2 mapping, limits the model's ability to handle more complex diagnoses that require these inputs.

We fully agree with the reviewers assessment regarding our reliance on cine-sequence images and this unfortunately remains a limitation of our work. The primary reason for this is that our work is reliant on video-input, and the image-only LGE sequences are incompatible with our overall framework. A complete system that can take image and video inputs would require a entirely novel vision encoder that is capable of multimodal input. There are some examples of multimodal input models now available in contemporary literature however these models are **significantly** larger and thus require significant additional engineering efforts beyond simple scaling of methods within the budgetary constraints of an academic environment. Our implementation of mViT contains around 36 million parameters and takes a full 16 frame video as input, whereas PEcore, for example - a new multimodal vision encoder released April of 2025 - **has 90 million parameters** for the smallest model for a **single frame input** – a near 100x increase in computational expense. Multi-modality remains an active area of research for our multi-center consortium.

Our thesis for working with cine-sequences to start with was made in consultation with our CMR imaging advisors given that that the algorithms used for reconstruction of cine-sequences (specifically bSSFP) involves balancing both T1 and T2 signal acquired by the scanner. As pointed out by reviewer #1, clinical practice relies on contrast enhanced or other dedicated sequences (T1 / T2) that better show these tissue differences. We argue that this does not necessarily imply that useful diagnostic information doesn't exist within the cine-sequences. While humans struggle with quantifying subtle pixel differences in a moving video in a reliable reproducible fashion, our efforts were to see if a large deep learning system could extract such information. We show that in certain situations, computers are rather good at identifying some of these features. In fact, part of the premise of our work to begin with, is to show we could do all these incredible things with cine sequences alone, without relying on explicit measurements of chamber size and function! Our Discussion section transparently acknowledges these limitations (Discussion; Paragraph 4):

"Another limitation of our current work is the reliance on cine-SSFP sequences alone. While the reports describe findings from gadolinium enhanced imaging, the networks do not make use of LGE (Late Gadolinium Enhanced) sequences as inputs. This is largely because LGE sequences are often captured as images rather than dynamic-motion videos, necessitating a separate and parallel image-based neural network to be built into the current framework. ***Future research will incorporate non-standard view-planes along with T1, T2, LGE, and perfusion scans via separate image encoders or dimensionally agnostic vision encoders capable of handling both image and video data…*** *"*

Additionally, we make mention of disease labels where our cine-only system performs remarkably well despite the absence of the above-mentioned sequences (Discussion; Paragraph 4):

"… Despite this, our models consistently perform well on tasks with just cine-SSFP sequences in situations where clinicians typically require additional imaging data. LGE, for instance, is routinely used by clinicians for diagnosing Amyloidosis; similarly LGE along with T1, and ECV (extracellular volume) maps are often used in diagnosing Hypertrophic cardiomyopathy.[13–15]*"*

We concede that the absence of LGE / T1 / T2 may have hindered the performance of our sub-classifier modules at tasks such as detection of myocarditis or the detection of cardiac thrombus vs tumor. This may have also contributed to the imperfect performance for these labels ("Diagnostic performance strengths and limitations"; Paragraph 1):

"… However, performance on tasks involving the detection of intracardiac thrombi (AUC 0.744; 95% CI 0.653 – 0.836) or certain cardiac tumors (AUC 0.649; 95% CI 0.535 – 0.764) remained marginal. Differentiating tumor from thrombus, or diagnosing Myocarditis can be particularly challenging for clinicians without contrast enhancement or T2 imaging [16–18]*. The Lake Louise imaging criteria used for the diagnosis of Myocarditis in clinical practice, for instance, relies on T1, T2, or LGE sequences.*[19] *Our results indicate that cine-sequences alone are unlikely to provide enough signal to reliably diagnose such conditions despite the presence of textual information in the reports."*

Unlike clinicians, the system does not seem to consider demographic data for reports but rather learns them from the images. This seems redundant. Why not provide additional demographics data routinely available to the clinicians along the images to further improve diagnosis.

We thank reviewer #1 for this observation. Our description of results on the UK biobank dataset re: demographic information and the ability of our model to learn these features was not clear and we have amended this to read *("Validation on the UK BioBank"; Paragraph 1):*

"… Key to driving invariance based on view-planes was to ensure all view planes from the same study are aligned with a text-embedding from the same text-report during pre-training. This allows for attention to be placed instead on features of importance highlighted in the text. We see a relatively dense cluster of patients with ejection fractions < 35%. ***Surprisingly, without any explicit instruction****, we find that contrastive pre-training allowed for sharp delineation of sex, and separation by age…"*

Briefly, we did not explicitly instruct the models to learn these demographic features nor consider this as a goal for this project. We stumbled upon this fact. We agree – this is redundant, it would be much more straightforward for us to append this information into the classifier directly to make use of this data. Furthermore, we have also not found a way to **completely prevent** the model from learning these demographic features.

In a separate body of work being performed by the lab that is currently in review at another journal, our team attempted to study algorithmic fairness and biases of ML systems including our pre-trained model – specifically the finetuned classifier for LVEF% prediction in the UK Biobank. We found that our LVEF% predicting model outputs had

slightly higher mean absolute errors in women with LVEF < 55% compared to men. We employed multiple approaches in an attempt to address this, including: 1. Pretraining the network by explicitly masking demographic features (age / sex) in the report text, 2. Sex-specific pre-trained networks, 3. Providing additional input to the networks for age / sex as variables, similar to how reviewer #1 has suggested.

Our results showed that pre-training our network with all the demographic textual features omitted resulted in models that **continued to separate the features** in t-SNE plots similar to that of Fig 2a of the main manuscript, with no reduction in downstream errors in women. We do not yet have a good explanation for why this is the case. Pre-training with data belonging to only one sex vs another (ie. sex specific foundation models) made performance worse, likely due to the overall reduction in training data each model was seeing given the sex stratification.

As reviewer #1 alluded to, we appended a numeric value for age and gender to the 512-dimensional video embedding, essentially feeding these demographics as additional input variables to our models. We found with this approach the models began to struggle with inconsistent performance (even worse for men, slightly better for women).

In summary: First, we found no improvement in performance with explicit addition of demographics – often times this reduced performance. And second, we have not yet identified a method to prevent the networks from learning features indicative of demographic differences on their own.

There is no evidence or discussion of prospective validation in clinical practice to show real-world utility.

We agree that we have not provided prospective validation in clinical practice to showcase clinical utility. However, in the time since original submission of our preprint and submission for peer-review we worked on uncovering cases of undiagnosed HCM in the large UK Biobank Dataset. The prevalence of hypertrophic cardiomyopathy (HCM) in the UK Biobank based on ICD-10 codes (.07%) is lower than global estimates of disease prevalence (0.2 - 0.5%). Prior studies on the genetic background of individuals diagnosed with HCM in the UK BioBank have remarked on this likely underdiagnosis problem. The key issue stems from the fact that that Cardiac MRI scans performed at the UK Biobank research centers are not actually read by a certified imaging cardiologist.[20] One of our HCM classifiers achieving an AUC of 0.94 was deployed on the UK Biobank data in an effort to re-label the dataset accurately for this disease. Cases exceeding a threshold of 95% – correlating to the top 0.5% of cases (expected specificity of 97% and sensitivity of 60%) – were screened in for manual reading. In a blinded fashion, a board-certified radiologist was tasked with diagnosing HCM on a sample of cases composed of high and low predicted probabilities.

Our goal was to use our ML system to whittle down an untenable volume of CMR data (>40,000 patients) to a subset of patients with a very high likelihood of disease for manual physician over-read. Of the 43,017 patients with cardiac MRIs, only 9 (.02%) had an ICD diagnosis of HCM. 266 cardiac MRIs were screened for manual review: 216 had greater than 95% predicted probability of HCM; 50 negative controls were randomly selected amongst cases with predicted probability less than 10%. **In a blinded assessment,** a board certified cardiac imaging radiologist concurred with an HCM diagnosis for 115 cases (sensitivity 53%, specificity 98%), 112 of which were previously undiagnosed. The prevalence of hypertension and aortic stenosis did not significantly differ between the cohort of true positives (69.2%) and false positives (76.6%). Our corrected prevalence of HCM in the UK Biobank MRI cohort was 0.28% - a figure that is closer to what epidemiological estimates would expect.

While this is not prospective clinical deployment in practice, this is a significant real world application of our technology, and is the subject matter of a recent abstract and presentation at the American Heart Association conference (Nov 11, 2024). The result of our AI + human screening run is being relayed back to the UK Biobank in the hopes to integrate our labels as official diagnoses. This research remains work in progress, but we briefly touch on these findings in the manuscript (“Clinical applications and future directions”; Paragraph 3):

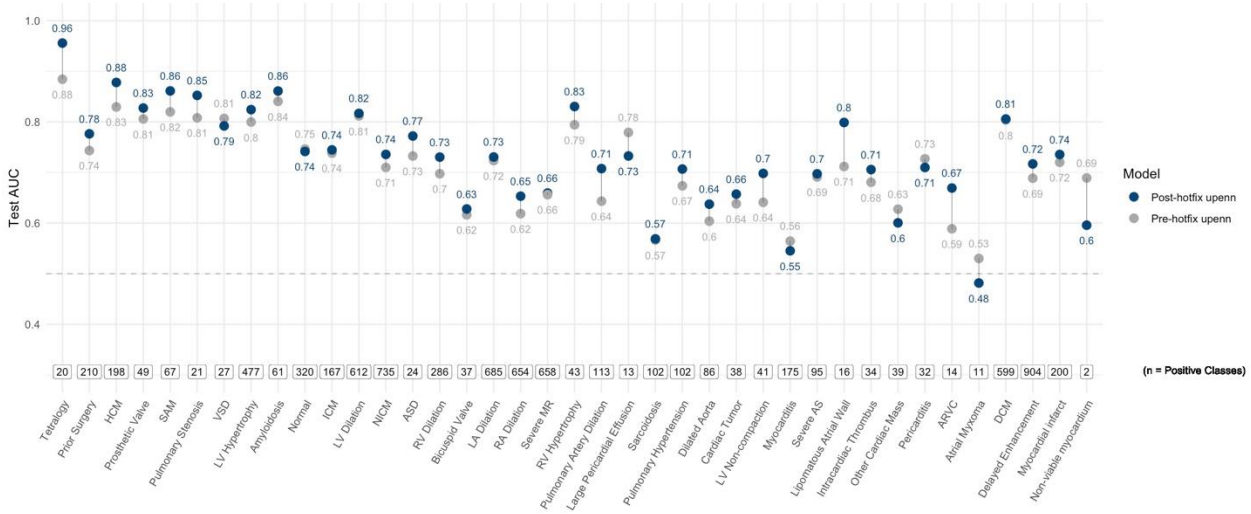
"We recently presented work on our pre-trained models in screening the UK BioBank population for Hypertrophic Cardiomyopathy.[20] CMR studies acquired as part of the UK BioBank are usually not reviewed by an imaging cardiologist unless glaring abnormalities are noted, likely contributing to the 5-10x lower reported prevalence compared to the general population.[21] Given the overwhelming majority of normal studies, cases exceeding a probability threshold of 95% – correlating to the top 0.5% of cases (expected specificity of 97% and sensitivity of 60%) – were screened in for manual reading. With this approach, our group screened through over 40,000 CMR scans, and identified 200 scans for manual physician over-read. From this high probability subpopulation, we confirmed 112 new individuals with imaging features of HCM. Our models thus have significant potential for clinical expert-grade phenotypic refinement of large research datasets."

Finally, we are in the process of integrating our work prospectively at the University of Pennsylvania as suggested by reviewer #1. Our collaborators and co-authors on this work Dr. Walter Witschey has recently showcased work for a cloud-based automated system that serves AI model predictions in a live clinical environment at Penn.[22] This is a multi-year effort given the administrative and regulatory requirements for prospective deployment.

Was center cropping performed fully automatically and was there any quality control for images going outside of the field-of-view?

Center cropping was performed in a fully automated fashion after an initial period of manual review to confirm images were not excluded from the field of view for each data source. Our center-cropping operation is quite mild (224px crop from a 260px resized image; original 360px) – only done to trim off excess soft tissue and pixels outside the body, this is not a center crop performed to perfectly encompass the heart for example. This was to ensure our models would identify structures such as sternal wires, or the thoracic aorta.

During manual re-evaluation of our cropping methods prompted by reviewer #1, we did however uncover a bug with how UPenn scans were being pixel normalized. Test-set results now are reported from the external UPenn classification results with normalization corrected. The impact of addressing the normalization issue (independent of the relabeled dataset) is shown below for transparency of review (dark-blue: final results post-normalization fix; grey: results on data without normalization error rectified). Ultimately this narrows the gap between our internal and external test results, bolstering our claims of generalizability. Raw UPenn test set results are uploaded publicly to the project GitHub for full transparency.



Why was ChatGPT turbo 3.5 used to parse the free text to diagnosis rather than later systems?

This is an important point raised by reviewer #1, given significant time has elapsed since we first began parsing our text reports with LLM based approaches. When we began working on this in 2022, earlier iterations of models such as OpenAI's davinci-003 were plagued by crippling hallucinations. LLAMA released by meta research early 2023 that was similarly unusable at least for zero-shot labelling tasks. ChatGPT turbo 3.5 was released in March 2023 and was the first usable solution in a zero-shot fashion, and was the best model available for this task. The remainder of our research has proceeded with these labels since.

The landscape as of the writing of this response has changed dramatically. Newer models – both open source and proprietary are far more reliable and less prone to hallucinogenic behavior at 'straightforward' tasks such as free text parsing of reports into structured formats. We have explored use of other models and our current approach in the lab is local large language model (LLM) deployment with llama.cpp or ollama, where each GPU hosts a copy of a quantized LLM. We now routinely use a 27 billion parameter medgemma model for tasks of this nature, and have made some adjustments to how we grade the severity of certain disease labels for classification. We have provided our LLM labelling script (llm_labeller.py) free for academic use in our main project repository, and detail the exact prompting strategy in a dedicated section within the Supplementary Appendix with some examples.

Keeping in mind the prevalences of certain labels are different from the original chatgpt-turbo3.5 labelled data, some of this is due to medgemma aligning better with human labellers than earlier models, some are conscious decisions to clearly delineate what should be labelled as 'diseased' – eg: we now explicitly label only severe mitral regurgitation as 'severe', whereas previously 'moderate and above' was classified as such. Our new medgemma3:27B labelling system achieves a median IRR of 83% which is conventionally considered to be "excellent agreement".

We **have repeated all classification experiments with this new labelling scheme.** Fig 4 in the main manuscript and the results have been updated to reflect this change. Additionally, all Supplementary figures and tables detailing performance across various datasets (internal – test / validation / training; and UPenn datasets) have been updated. Finally, figures detailing prevalences of various labels have been updated. Our conclusions and discussion however remain the same, and performance overall is slightly improved. Updated figures include: Fig 4, Supplementary Fig 12, 13. Supplementary tables 6, 7, and 8 have also been updated.

The bland Altman 95% LOA -12 +13% for LVEF seem rather large -similar to what is achieve in 2D echo. MRI is considered to be a more precise technique. Can outliers be shown and discussed. references 32 and 33, 34 are rather old and newer approaches should be discussed

We thank the reviewer for this observation – we must emphasize that the 95% LOA of +/- 12% number cited are the results reported from prospective clinical trials.[23] Our baseline model (non-pretrained) performs at this level with a mae of 4.603 (BA limits -12.15% to +12.8%) Our final results with our contrastive pre-training paradigm are superior with BA limits 9.91% to +9.61% and mae of 3.34%. The presentation of this material has been edited to avoid confusion. The section on Automated Estimation of Left Ventricular Ejection Fraction now reads:

"Prospective studies have shown that there is considerable variability in institutional and dataset specific protocols for calculation of certain metrics, clinicians can typically be expected to make estimates of LVEF within Bland-Altman limits of -12% to +12%.[23,24] *With our pretrained vision encoder frozen barring the last linear layer, we finetune the network on 34488 unique CMR scans from the UK BioBank for the task of predicting LVEF. The vision encoder can no longer learn dataset specific features with this approach, and must rely on learned representational abilities from prior pre-training. To incorporate information from each available view to produce a 'patient level' prediction of LVEF we use a multi-instance self-attention regression head as described above, and input cine-CMR sequences from all*

available views without any additional quality control steps (Methods). On a hold-out test subset of UK BioBank participants using our approach, we report a mean absolute error (mae) of 3.344 (sd 3.615), with Bland Altman limits of agreement -9.91% to +9.61%. At baseline, using a traditionally trained deep learning system of the same model architecture as our contrastive pre-trained models, we report mae of 4.603 (sd 4.409), with Bland Altman limits of agreement -12.15% to +12.8%."

We highlight recent advances that have shown promise in the field of ML in Cardiac MRI. CineMA – a self-supervised model for Cardiac MRIs with a masked auto-encoder using 74,916 CMR studies from the UK Biobank was recently described.[25] Despite the network being trained entirely on the UK Biobank data on **twice as much data**, the mean absolute errors (mae) described for LVEF% estimation is 3.33%. This furthermore, is achieved only with the addition of segmentation supervision. Direct regression of LVEF% achieved a mean absolute error of 5.21%. Segmentation supervision was key for CineMA.

Our model in contrast, does not rely on segmentation. The vision encoder is furthermore **entirely frozen** for finetuning tasks (such as estimation of LVEF%), the visual features learned entirely from our clinical datasets. When finetuning on the UK BioBank, at most we use data from 31,693 scan. Despite this, our model essentially performs the same with a mean absolute error of 3.34%, and dramatically outperforms CineMA without segmentation supervision. Direct comparisons are of course challenging, but notably the authors of CineMA show that when using 50% of the data (a more comparable volume of finetuning data to our work – even though the key difference is that our frozen vision encoder cannot learn more features from the UK BioBank for this task), the performance even with segmentation dips to a mean absolute error of 3.72%. Finally, our work evaluated on the Kaggle dataset (once bias corrected) shows a mean absolute error for LVEF of 4.86% (Supplementary Appendix; Fig 7). This again is comparable to results of CineMA at 4.83% .

A second recent paper by Zhang *et al.* showcases a Cardiac MRI model that also makes use of cine-sequences in the UK BioBank, with a staged masked auto-encoder pre-training stage, followed by a contrastive objective applied to tabular data with ICD diagnostic codes from the UK Biobank.[26] As with CineMA, the model is pre-trained and finetuned exclusively on the UK Biobank data, but tests were performed on a 1000 participant set, making it challenging to compare results directly. The cardiac MRI model developed by this team was able to return a mean absolute error of 2.95%, but this again was without a frozen vision encoder – essentially creating a fully supervised model.

Our discussion is not to claim our models are *better* in a direct head-to head comparison – that would require blinded assessment on the same external datasets. The methods in the two papers are innovative and provide thought provoking perspective on the architectural choices we make for future work. However, we claim that the performance of our **generalist** models rivals those of contemporary efforts at CMR deep learning **models designed from the ground up for specific tasks on the same specific datasets**. We argue that our claims of generalizability are bolstered with these new results: our pre-trained vision encoders are not just as capable at tasks showcased by CineMA and that of Zhang *et al*, but can be immediately and readily applied to new tasks with a fraction of the finetuning data showcased by other modern approaches. This data efficient transferability shown by our work is not a feature of these two cutting edge models. We incorporate these studies and a brief discussion of the same that now reads ("Automated Estimation of Left Ventricular Ejection Fraction"; Paragraph 1):

"… Nonetheless, these networks perform exceptionally well in the task of LVEF estimation as they replicate the measurement workflow of routine clinical practice. Bai et. al for instance, report a mean absolute error (mae) of 3.2 for the task of predicting LVEF in a subset of the UK Biobank participants using short-axis sequences.[27] Two recent papers released while our manuscript was in review are notable for their technical advance: Fu et al describe CineMA, a masked-autoencoder system utilizing a hybrid-segmentation supervision; and Zhang et al who describe masked

autoencoders combined with a contrastive objective.[25,26] *While both achieve competitive performance on LVEF estimation in the UK BioBank (CineMA with a mae of 3.34% and Zhang et al. with a mae of 2.95%), critically, these systems were both pre-trained and finetuned on the UK BioBank itself making direct comparisons of generalizability challenging."*

Finetuning is performed in Biobank set, raises concerns about the separation between training and testing datasets, which could lead to inflated performance metrics. Authors should either perform fine tuning in one of the 3 sets used for development or shoud not consider biobank as part of external testing

We understand these concerns. Data from the UK BioBank was used for finetuning for the LVEF diagnosis challenge as numerous studies have trained on this dataset in the past, allowing for reasonable comparisons to be drawn. A key difference in our approach is that we freeze the entire vision encoder, finetuning happens only at the level of the fully-connected layers past the encoder itself.

Our goal was to see how well a clinical-dataset trained contrastive-pretrained model would translate to an entirely **new task** for LVEF regression on data that is **distinct from the** underlying pre-training dataset population of Stanford / MedStar / UCSF. This was to ensure we could fairly compare our model with the kinetics pre-trained baseline. We argue that our approach reduces the likelihood of inflated performance (vs say if we were to evaluate the task of LVEF estimation on the clinical dataset). If we had fine-tuned on Stanford data, our contrastive pre-trained models would have an unfair advantage over baseline methods. We use this as a proof of concept that our models can be generalized and finetuned with great efficiency to data outside of what we have collected for pre-training. Furthermore, we demonstrate our results on yet another dataset, the Kaggle LVEF dataset to ensure that even the LVEF challenge has it's own nested "external test set". The term 'external validation' has been removed to avoid confusion, and we replace the subheading of this paragraph as now simply "**Validation on the UK BioBank".**

AUCS for detection of some clinically important conditions seems very low -borderline significant or non-significant for example cardiomyopathy, bicuspid valve, myocarditis aortic stenosis. The authors do not provide sufficient analysis of why this is the case or how the system could be improved for these critical conditions. The authors should consider these accuracies in the clinical context. What would clinical accuracy be ?

Reviewer #1 raises the limitations with the performance of our system across some of our listed tasks. Of interest are ARVC (arrhythmogenic right ventricular cardiomyopathy), myocarditis, and some of the valvular conditions such as bicuspid valve disease.

The Lake Louise imaging criteria is used for the diagnosis of Myocarditis in clinical practice, which relies on T1, T2, extracellular volume assessments, or LGE sequences. Diagnostic performance with this approach is typically of AUCs > 0.90 when benchmarked against histopathological gold standard. Our model performance for this task is 0.698 on the internal test set. This likely stems from our reliance on cine-sequences, a limitation of the technology in its current form that we fully acknowledge. Similarly, for bicuspid valve disease, it is helpful to have phase-contrast imaging centered at the plane of the aortic valve. A standard short-axis cine-stack may miss the aortic valve structures by a few millimeters. This again would be something we could improve by incorporation of sequences beyond the cine-sequences that we utilize.

And ("Diagnostic performance strengths and limitations"; Paragraph 2):

"… Detecting bicuspid valves on the other hand requires acquisition of images in the plane of the aortic valve to truly define the location and orientation of fused aortic valve leaflets. While these features are sometimes visible on cine

sequences, accurately quantifying regurgitant volumes and the severity of valvular heart disease requires additional post-processing of phase contrast sequences in specific anatomical planes or 4D-flow scans that are seldom performed as part of standard CMR scanning protocols.[28]"

The substandard performance on other conditions potentially stems from how conditions such as ARVC are diagnosed to begin with. For instance, diagnosis of ARVC relies on a Clinical Task Force Criteria (TFC) scoring system of major and minor criteria: a combination of clinical symptoms, medical and family history, EKG findings, and imaging findings. Our training paradigm while capable of learning some demographic features (as we show in Fig 3), it is incapable of directly learning features beyond images and the CMR reports. Real world evaluations of the TFC scoring system have shown that using just the CMR structural features yield Sensitivity & Specificity of 69% and 88% for a Youden's index of 0.57 (roughly translating to an AUC of at least 0.78) for the diagnosis of ARVC. The full TFC scoring system yields sensitivities and specificities in the 95% range based on most estimates. Our model performance returns an AUC of 0.65 for ARVC, indicating the challenges of diagnosing conditions - that in routine clinical practice - rely on a comprehensive evaluation of the patient beyond what images alone can provide.

("Diagnostic performance strengths and limitations"; Paragraph 1):

"… However, performance on tasks involving the detection of intracardiac thrombi (AUC 0.744; 95% CI 0.653 – 0.836) or certain cardiac tumors (AUC 0.649; 95% CI 0.535 – 0.764) remained marginal. Differentiating tumor from thrombus, or diagnosing Myocarditis can be particularly challenging for clinicians without contrast enhancement or T2 imaging [16–18]*. The Lake Louise imaging criteria used for the diagnosis of Myocarditis in clinical practice, for instance, relies on T1, T2, or LGE sequences.*[19] *Our results indicate that cine-sequences alone are unlikely to provide enough signal to reliably diagnose such conditions despite the presence of textual information in the reports. Similarly, the diagnosis of Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC) relies on a Clinical Task Force Criteria (TFC) scoring system: a combination of clinical symptoms, medical and family history, EKG findings, and imaging findings.*[29] *Our model performance returns an AUC of 0.652 (95% CI 0.487 – 0.818) for ARVC, indicating the challenges of diagnosing conditions - that in routine clinical practice - rely on a comprehensive evaluation of the patient beyond what images alone can provide."*

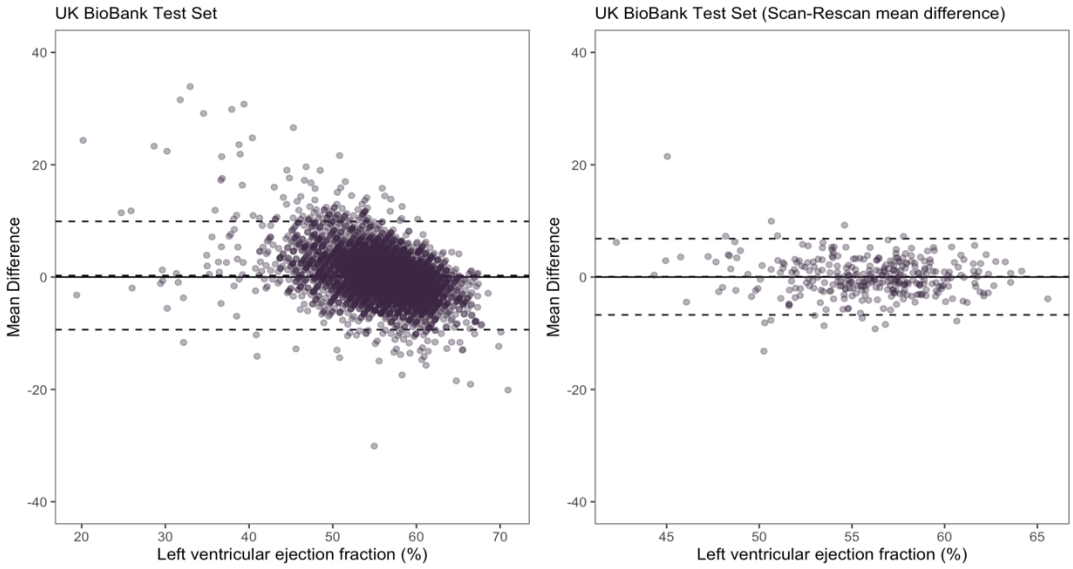
Certain figures and tables lack clarity or are not sufficiently detailed (e.g., the number of cases with specific diagnoses in training and testing datasets is missing). N for each occurrence shoud be given in Supplementary table 5 -7 and figures 9 and 10. Several figures in the appendix contain undefined abbreviations

We have amended figures In the supplement to address concerns re: abbreviations. We have also now featured disease positive class prevalences across the various datasets and data splits, featured prominently in Supplementary Fig 10, 11, 12, 13 and Supplementary Tables 6, 7, and 8.

Supplementary table 8. Bland Altman plot should be provided

We have now added Bland Altman plots to better illustrate scan-rescan variance in the UK Biobank test-set. Plot-axes are scaled identical to that of the main LVEF% results of the UK Biobank from Fig. 3b. Of note the 95% BA limits here are -6 to +6. This is superior to previously reported clinical expert level scan-rescan variance (95% BA limits of -9 to 10; Bhuva *et al.*) The Supplementary Appendix has been updated to now read:

"… We additionally show Bland-Altman plots for the scan-rescan assessment of LVEF% in the UK BioBank dataset. We find the 95% BA limits are -6 to +6 for the scan-rescan assessment. These findings are superior to previously reported clinical expert level scan-rescan variance (95% BA limits of -9 to 10; Bhuva et al"



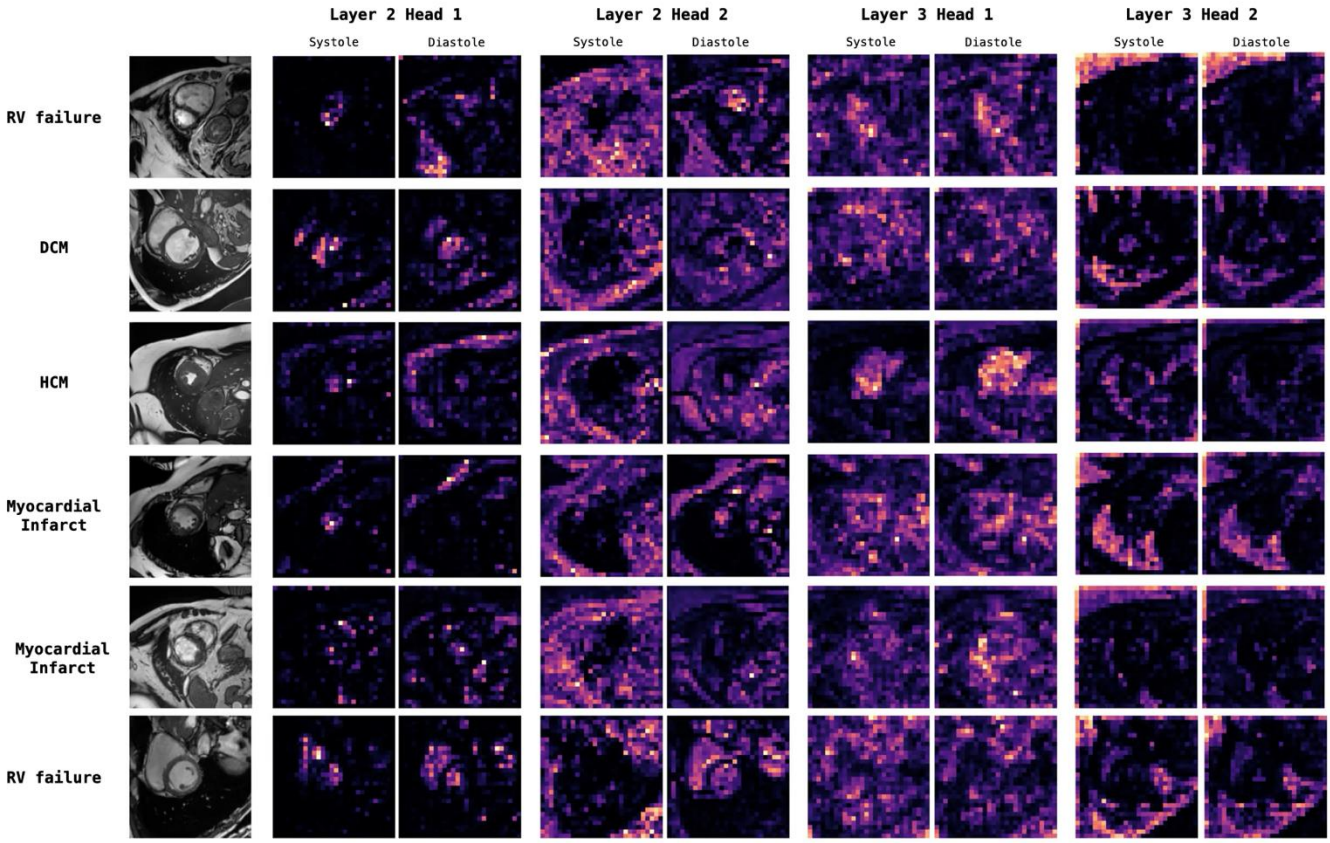
Additionally, the main manuscript has been updated to reflect the same ("Automated Estimation of Left Ventricular Ejection Fraction"; Paragraph 4) now reads:

"Finally, we studied the scan-rescan variability of LVEF predictions generated by our models for participants with more than one available CMR study acquired at two distinct time periods (n = 311; test-set). The mean variance was found to be 5.98% (sd 1.53%), with Bland Altman limits of agreement between scans of -6 to +6%. These figures exceed previous benchmarks of clinical expert level performance in prospective trials (Supplementary Table 9).[23]*"*

The self-attention maps provided are not intuitive or clinically meaningful. The paper does not demonstrate how these maps could aid clinicians in interpreting model decisions or improve trust in the system. Can some validation and proposed use be shown for these. In addition, clinical examples of attention maps should be shown for specific disease patterns. Otherwise, this material is not very helpful. Results without any validation performed should not be shown.

We agree with reviewer #1 re: the attention visualizations that are in our Supplementary Appendix and Fig 4c. We acknowledge that the attention maps are themselves not always very 'explanatory'. In the supplement, we are transparent about these limitations: *"Of note, it is impossible to determine how the network uses this information, we can only infer what structures are attended to. This limits explainability…."*. We have omitted the line that proposed potential uses of these attention maps re: clinical interpretation and trust. We have added an additional figure displaying different disease patterns to showcase differences in the way these inputs are attended to by our pre-trained vision encoder (Supplementary Fig. 18):

Attention visualizations are included in the Supplementary Appendix only for completeness in response to frequent requests to provide a qualitative sense of what features the model attends to in prior submissions and publications, despite the inherent limitations of all these methods. We caution that their purpose here is descriptive rather than inferential, and we are happy to remove these if needed.



Caption for Supplementary Figure 18 also reads: "*Supplementary Figure 1: Examples of attention maps randomly sampled for different disease processes including isolated right ventricular failure (RV failure), Myocardial infarctions, hypertrophic cardiomyopathy (HCM), and dilated cardiomyopathy (DCM) from patients not-seen previously by our vision encoders as part of pre-training or finetuning. Attention maps are displayed for attention heads from layers 2 and 3 for brevity. Systolic and diastolic differences in attention localization are noted for different disease processes. Layer 3; Head 1 for instance attends to myocardium with greater intensity in diastole, and Layer 2; Head 2 shows increased attention localized to the right ventricular chamber cavity in diastole. We caution that understanding how the models utilize this information is infeasible from these visualizations, and their purpose here is descriptive rather than inferential.*"

In Fig 4c. we present self-attention heatmaps for each view plane – these allow us to understand which view-planes most significantly contribute to the final disease classification labels. This does inform concordance of human vs model behavior in assigning importance to various view planes. The line now reads (Methods; Attention Visualizations; Paragraph 1):

"The multi-instance classifier head self-attention scores show that the network learns to differentially prioritize view-planes for different clinical tasks."

The augmentation scheme seems simplistic: only resize, color jitter, shear, translate and rotate videos in spatial dimensions. These may not adequately reflect real-world variations in cardiac imaging, such as motion artifacts, poor image quality, or varying contrast noise levels. Can more details be given and potential improvements to achieve more realistic augmentation schemes be discussed? Unclear what color jitter represent in terms of cardiac imaging

We thank reviewer #1 for raising concerns re: our augmentation strategy. Our description and explanation of our augmentation methods were without sufficient detail. We conducted experiments with two broad approaches to augmentations for all phases of our work: Random augmentations and AugMix.

RandAug chains transformations randomly selected from a pool of available operations that include contrast, solarization, shearing, rotation, and translation. We follow the default RandAug recepie for two randomly selected sequential operations of fixed a scalar magnitude term that dictates the severity of the transformations. AugMix was proposed as an alternative method to improve the robustness of models as traditional augmentations schemes were found to be insufficient in lending strong generalization to minor corruptions.

At a high level, AugMix chains simple augmentation operations in concert with a consistency loss. These augmentation operations are sampled stochastically and layered one on top of another to produce a high diversity of augmented images. Mixing augmentations allows AugMix to generate diverse transformations, which are important for inducing robustness, as a common failure mode of deep models in the arena of corruption robustness is the memorization of fixed augmentations. Pre-training losses were found to dramatically decrease with AugMix compared to RandAug. The layered augmentations simulate motion artifacts, and the complex augmentations (unlike simple gaussian noise) allow for models to be more resilient to corruptions.

We have updated the manuscript to read as follows (Methods; Pre-training framework; paragraph 2):

"We employ randomized sequential data augmentation schemes (AugMix) to stochastically sample and layer a series of chained transformations including but not limited to resizing, solarization, shear, translate and random rotation of videos in the spatial dimensions, all while preserving the same augmentations along the temporal dimension for temporal consistency.[30]*"*

The Supplementary Appendix additionally now has been amended to read as follows under "Augmentations":

"For the vision pipeline we experimented with and without random temporal subsampling, and assessed different data augmentation schemes including simple chained random rotations and shearing operations, RandAug and AugMix.[30,31] *RandAug randomly chains transformations selected from a pool of common operations that include contrast, solarization, shearing, rotation, and translation. We follow the default RandAug recipe for two randomly selected sequential operations of fixed a scalar magnitude term that dictates the severity of the transformations.*

AugMix was proposed as an alternative method to improve the robustness of models as traditional methods were insufficient in lending strong generalizations against minor image corruptions. At a high level, the original description of AugMix stochastically chains simple augmentation operations, layering them to generate a perturbed composite image. This is done in concert with a consistency loss. Mixing augmentations allows AugMix to generate diverse transformations, which are important for inducing robustness, as a common failure mode of deep models in the arena of corruption robustness is the memorization of fixed augmentations. We use AugMix without the consistency enforcement for all pre-training, finetuning, and transfer learning experiments.

AugMix Transform Operations:

{ ShearX. ShearY, TranslateX, TranslateY, Rotate, Posterize, Solarize, AutoContrast, Equalize, Brightness, Saturation, Contrast, Sharpness }

Default Settings:

Severity: Severity of the transformations; default 3
Mixture width: Number of augmentation chains; default 3
Chain Depth: The depth of augmentation chains, sampled stochastically between [1, 3]

We achieved the best pretraining loss with a uniform temporal subsample operation of 16 frames, random crop operation to 224x224 pixels, followed by AugMix for augmentation strategy. We additionally experimented with both RandAug and AugMix for the downstream task of LVEF% prediction. We found the networks converged faster with lower losses with AugMix for both the pre-training phase and the downstream task of LVEF% prediction. We continued to use AugMix for all further experiments. Significant performance gains were achieved by first temporally subsampling the frames followed by the cropping and chained augmentations."

Finally, we review the literature on more advanced methods of realistic augmentations: specifically with generative methods such as diffusion networks and variational autoencoders to generate synthetic cardiac MRI studies to augment both dataset size and the variety of data available for training deep learning models. Specific methods tailored towards echocardiography and Cardiac MRI have been developed that allow for conditional generation of Cardiac MRI sequences based on certain phenotypes (larger or smaller left-ventricular diameters for instance). Our lab has recently presented work that enables generation of synthetic data based on textual prompting. These methods are not viable for the pre-training phase (as one would need to generate correct synthetic reports to match the synthetic images – a major project in its own right and beyond the scope of this manuscript). But they are viable for downstream regression tasks.

Of note, experiments at least with Cardiac MRI specific methods by Li *et al,* demonstrate that synthetically augmenting datasets by 5x what we finetuned with (in a fully-supervised setting) did not yield materially superior results on the task of LVEF% estimation on the UK BioBank compared to what we have demonstrated in this manuscript: MAE 3.27 vs 3.34 (ours). We briefly discuss this immediately below the technical description of our AugMix strategy in the supplementary Appendix:

"Expansion of low-data regime datasets with synthetic images as a method of augmentation are discussed briefly here, but remain beyond the scope of this manuscript. Specific methods tailored towards echocardiography and Cardiac MRI have been developed that allow for conditional generation of synthetic images based on certain phenotypes (larger or smaller left-ventricular diameters for instance).[32,33] *Our group has recently presented work that enables generation of synthetic data by way of text-guided latent diffusion models that could serve to expand small CMR datasets, though currently is only capable of generating images and cannot generate temporally coherent Cine-Sequences.*[34] *Of note, experiments with Cardiac MRI specific methods by Li et al, demonstrate that synthetically augmenting datasets by 5x our finetuning dataset (in a fully-supervised setting) did not yield materially superior results on the task of LVEF% estimation on the UK BioBank compared to what we have demonstrated in this manuscript: MAE 3.27 vs 3.34 (ours)."*

Were Delong comparisons performed as paired? it is unclear from the methods.

Yes, DeLong comparisons were all performed as paired analyses. Statistical methods and supplementary appendix figure captions have been updated to reflect this clearly (Methods; Statistical analyses, paragraph 1):

"To compare the performance of finetuned classifier models (ie Contrastive pre-trained vs Baseline), we calculated non-parametric confidence intervals on the AUROC using DeLong's method (paired),[35] following which p-values were computed for the mean difference between AUROC curves."

Code:

While I did not run or install the code the MIT license is given and code seems complete on GitHub. There is a README file with basic information. Few key datasets should be provided with the code as well as expected results for these cases should be provided.

We have added some helper scripts (such as the LLM labelling scripts described above) amongst some other minor updates. We are unfortunately prevented by institutional policy and IRB restrictions in sharing our internal clinical datasets, and additionally are prevented from directly providing subsets of the UK Biobank and Kaggle datasets by their respective terms and conditions. We are only able to point researchers to the websites of these datasets. We have however, released checkpoints of our models for researchers on HuggingFace: https://huggingface.co/rohanshad/cmr_c0.1 to further our commitment to open science.

Reviewer #2

This paper describes a foundational vision system for CMR using deep learning. The system is based on self-supervised contrastive learning on CMR scans and accompanying reports. The model was trained and evaluated using data from four large academic institutions in the US, with additional validation on the UK Biobank and other public datasets. The authors use a transformer-based vision system trained on 19,041 CMR scans employing a large language model to vectorize the radiology reports. The model seems generalizable and data-efficient and allows basic grouping of patients into major disease categories without supervision.
In general, the approach is exciting, novel and potentially game changing. Using a self-trained vision model instead of requiring contouring would change the workflow and possibly reduce human interaction and thus improve scalability.

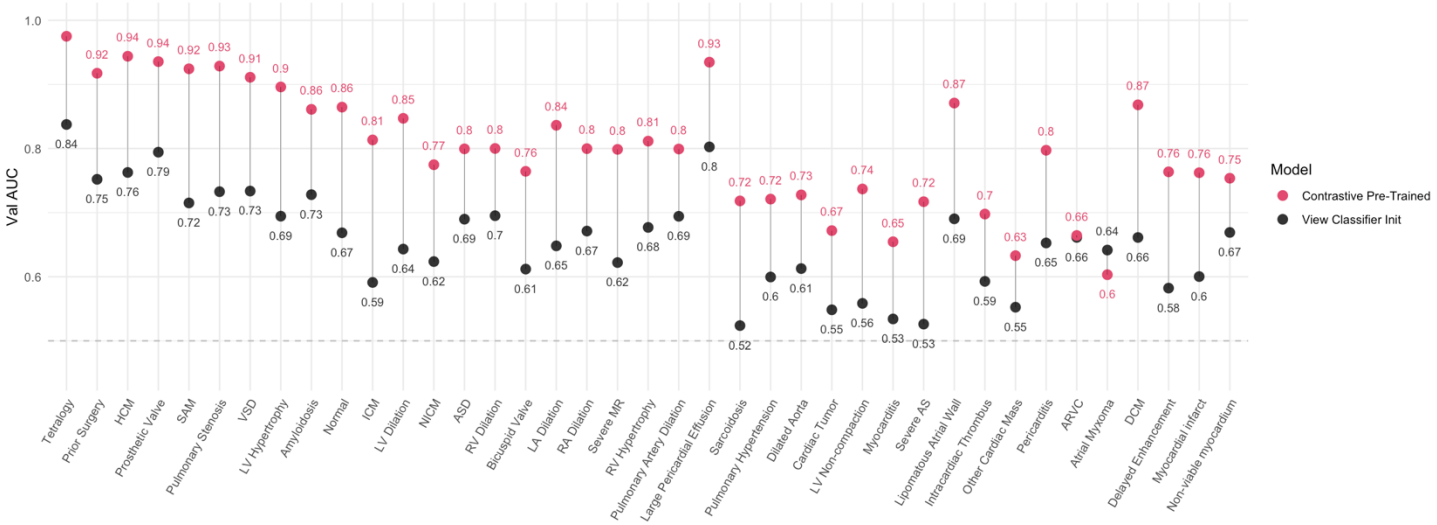
However, so far, I see a significant number of shortcomings and further work that mandatorily needs to be done to convince me.

Main issues:

1. The core data provided demonstrates, that the pretrained model performs better, than the model trained on a kinetic dataset. While it shows the value of pretraining to learn cardiac-specific features, the untrained model focusses on the distinction of short- and long-axis views. Again, this is not surprising, as this is the core difference of the datasets provided. It would be much more meaningful to see how the extensively self-trained model performs versus a model trained just to recognize the different anatomical views, so it can then – without further training – start to assess the differences within the views.

We thank reviewer #2's comments on our work and overall approach, over the past few months we have performed substantial additional experiments. Our hope is that these results and data better showcase the capabilities of our methods and their clinical utility.

We retrain a version of vision system extensively with a supervised signal to perform just the anatomical views (SAX, 4CH, 3CH, 2CH), the model as one can imagine separates these views very well. However, when freezing the network and attempting to finetune towards our disease classification task, the models fail to learn meaningful signal (evaluated on the internal validation set). See figure below: view-selective model run performance in dark grey, our contrastive pre-trained model in pink.



Human experts would likely be able to discern features indicative of disease when exclusively shown certain view planes in a repetitive fashion, and then be able to collect their learned experiences of different views to develop a cohesive diagnosis similar to what reviewer #2 hinted at in their comment. ML systems typically struggle at such adaptation.

Our observations of this dramatically worse performance here are explained by (i) shortcut learning / spurious correlations, (ii) **spectral or simplicity bias** (networks learn "easy" signals first) – that is the network **once it learns something 'easy' fails to unlearn it quickly**, and only sees the world within the 'easy' framework of separating different scans. We suspect this is what happens when we attempt this method of staged learning. Coaxing the network to ignore the 'easy' differences in anatomical views may be essential for our methods to focus on the more challenging tasks of disease classification. We note that our contrastive pre-trained models have no concept of "view" and are unable to separate them, and it may be necessary to ignore these obvious anatomical differences to learn more complex disease specific features. This phenomenon of 'shortcut learning' – where deep learning systems fail at generalizing after learning quick heuristics that explain the data distribution is well studied.

The Supplementary Appendix has now been amended to include this detail in brief ("MRI vision encoder, BERT text-encoder architecture and pretraining parameters"; View vs Study based vision-text alignment; Paragraph 1):

"Our contrastive pre-pretraining routine aligns report text with randomly sampled views of the same patient. Over the course of 600 epochs a comprehensive representation of disease from the vantage of different views are learned by the vision encoder. A byproduct of this decision is view-invariance, that is best visualized in Fig 3 of the main manuscript. Explicitly constraining models to separate views results models to fail at downstream finetuning, our position is that it may be necessary to ignore obvious anatomical differences to learn more complex disease specific features."

2. The ROC curves and AUCs provided are helpful to see the generally superior performance of the model versus an ill-trained model (see above), but they do not provide any information on clinical value. For some clinical issues it is highly important to have very high sensitivity even if the specificity is low (e.g. in rare cases with amyloidosis, which MUST be detected, as there is early, lifesaving therapy). While this is discussed in the outlook, it is not assessed in the data-analysis. Given the low prevalence, the NIR (no information rate) would be very high (and potentially even better than the model), but would miss these highly relevant cases. The model must be able to flag these cases. The ROC curves also show – different to how the data is presented in the main text – that the accuracy of the models is terrible. While the AUCs were reasonable for Tetralogy (which is easy to spot due to massively abnormal cardiac chamber sizes and position) and amyloid (with the caveats mentioned above), the majority of AUCs is not convincing. In addition, the main indications for a CMR study, such as myocardial infarction, viability and ischemia were not assessed at all.

We are happy to expand on reviewer #2's astute observations on our model performance and the importance of clinical context depending on the disease being studied. The issue raised of NIR is one that is very relevant, and spurious null classifiers with high accuracies (e.g. 99% accuracy by always predicting a '0' label for a disease with a 1% prevalence) can be clinically misleading and unhelpful. AUROCs are invariant to class imbalance due to low prevalences and there is an extensive body of research most recently with the seminal work of Richardson *et al* published in Patterns on this topic [36].

To address this, we first report the prevalences of the different disease labels in our dataset are shown as bar graphs with percentage and raw number prevalence in Supplementary Figures 10 and 11. Additionally, we show

the number of diseased individuals for each disease label in Fig 4 along the x-axis. The caption has been updated to reflect this to explicitly label the same. We also now display the number of positive labels for each class in our AUROC plots in Supplementary Figures 13 and 14.

The area under receiver-operating curve (AUC) is 0.5 for each null model (i.e. that only predicts label '0' or no disease). The accuracy of such a model equals the no-information rate (NIR). To illustrate the example provided by reviewer #2 we force a small classifier to always output a class label of '0' for each label yielding impressive accuracies on face value, and we empirically show that for each such classifier the AUC calculates to exactly 0.5.

For each disease label, we calculate the optimal threshold for a certain disease, we calculate the NIR, the p-value for the null hypothesis that the model accuracy = NIR, and the actual model accuracy. For Amyloidosis as an example, our test-set performance AUC was 0.921. We calculate an optimal threshold based on Youden's method to binarize predictions from the continuous probabilities generated by our models (ie. predicted disease = 1 vs predicted normal = 0). From this we calculate our **model accuracy to equal 80%** (95% CI 0.779 - 0.825), the NIR for Amyloid given the low prevalence is 98.7% - meaning a model that consistently predicts a label of '0' is going to be 98.7% accurate, **but** with a calculated AUC of 0.5 representing an entirely useless model. The p-value for the null hypothesis that our model behavior equals NIR was < 0.00001.

Through these calculations we hope to address multiple aspects of reviewer #2's critique – we show that our models never behave merely as null classifiers at NIR for any of the disease labels, we prove this difference to show our models output predictions that are clinically meaningful. We agree with the reviewer here that a model that does nothing but unintelligently output '0' can have very high accuracies at face value. Accuracy as a result, is an unreliable metric to compare classifier performance given the fact that 'higher' may not mean 'better'. For illustrative purposes and completeness, we statistically prove that for each null-classifier mimicking NIR with these high accuracies, the AUC calculates to exactly 0.5 – random chance ability to discriminate between disease and controls.

Finally, our approach allows us to dynamically select appropriate thresholds for the probability output – ie. we can choose to balance sensitivity and specificity, or prioritize one over the other depending on the clinical context without any additional training / finetuning. We use this for our work on identifying underdiagnosed Hypertrophic Cardiomyopathy in the UK Biobank (work described above and presented at AHA 2024). The NIR for HCM based on the prevalence in our test set was 0.924. With a balanced threshold setting, the accuracy for the HCM model is 88% for a sensitivity of 0.837 and specificity of 0.888. For our work on the UK Biobank we pre-specify a threshold at a specificity of 97% and a corresponding sensitivity of 62%, the accuracy here calculates to 94%.

This further supports our decision to primarily report AUC over Accuracy in our initial submission. Accuracy varies with threshold selection and can be misleading: It is quite feasible, for instance, to cherry pick a high accuracy number for the same AUC (generally accuracy will increase if one selects a high specificity threshold for a given AUC in highly imbalanced datasets). The AUC curves and values on the other hand assess the ability of the model to discriminate disease from controls across the entire sensitivity / specificity continuum and **are immutable for each model.** In the supplementary appendix we now report additional data for completeness and context for readers. In Supplementary Table 12 we report the internal test AUCs, the accuracy (for a sensitivity / specificity balanced strategy), NIR and null model AUCs. In Supplementary table 13 we display the same but with a strategy of thresholding at a sensitivity of 0.90, and in Supplementary table 14 we threshold for a specificity of 0.90.

Finally, we update our list of conditions to include detection of LGE, DCM, non-viable myocardium, and myocardial infarction. Test set results are summarized in Supplementary Table 7, but briefly, detection of LGE from pre-contrast cine-sequences is a challenging task and models struggle with AUC of 0.76. Detection of Myocardial infarction

(0.809) and Dilated Cardiomyopathy (0.867) fared better. Non-viable myocardium was a challenging clinical diagnostic label as there is considerable variation both in anatomical location and descriptive practices by radiologists. We find for non-viable myocardium the AUC was 0.83 though with wide confidence intervals (0.637 – 0.991). Future work may consider alternative approaches to label myocardial viability (such as a percentage of total myocardium, or localizing areas of non-viability with a bounding box). These remain beyond the scope of the current model architecture. Data with new labels are included in all major results figures and tables.

3. Another major issue with the data presented is that the majority of CMR scans are not done to generate a diagnosis in a patient with no previous knowledge, but rather the opposite. CMR scans are ordered to make differential diagnoses, e.g. such as working out the underlying cause of heart failure or left ventricular hypertrophy. It would be important to test the ability of the algorithm to perform differential diagnoses in borderline cases.

We thank the reviewers for this insightful critique. Our methods and results demonstrate impressive performance in disease diagnostics, and one of the deliberate features of our approach the ability to predict probabilities for each disease label in a multi-label fashion – that is, our models are able to detect the presence of overlapping conditions that are not-mutually exclusive, and provide ranked probabilities in the setting of borderline cases. All analyses described below were performed on the internal test set with no additional training or finetuning.

We examine the broad label of "non-ischemic cardiomyopathy" or NICM (n = 242) in the internal test set. This broadly includes cardiomyopathies associated with infiltrative diseases, genetic hypertrophic or dilated cardiomyopathies, myocarditis among others. Importantly within this subset of 242 patients, there are none **without** heart failure – a far more challenging task for our models given they were originally optimized for identifying disease features in the background of over 30 different disease conditions. Such subgroup analyses demonstrate conditional model performance in clinically challenging contexts, for instance: Certain non-ischemic cardiomyopathies share overlapping phenotypic features of left-ventricular hypertrophy (sarcoidosis, amyloidosis, and hypertrophic cardiomyopathy). This makes discerning the underlying etiology challenging in both clinical practice and for deep learning systems. Nonetheless, our models demonstrate impressive performance in this subset for Amyloid, DCM, and HCM (Supplementary Table 18)

Disease Label	(N) label	Contrastive Pretrained	
		AUC	95% Confidence Intervals
Amyloid	15	0.87	0.675 – 0.957
Sarcoidosis	21	0.725	0.623 – 0.827
Dilated.Cardiomyopathy	140	0.879	0.836 – 0.923
Myocarditis	37	0.423	0.318 - 0.529
Hypertrophic.Cardiomyopahy	46	0.925	0.873 – 0.977

In clinical practice, there is additional context obtained from chart review or discussions with the primary cardiologists, for our work we are prevented by IRB and data-procurement restrictions in re-identifying individuals for secondary chart review. These factors in addition to the limitations of using cine-sequences alone likely contribute to the poor discriminatory ability of our model for sarcoidosis and myocarditis in this sub-population. We incorporate portions of this discussion in the main manuscript ("Diagnostic performance strengths and limitations"; Paragraph 3):

"In clinical practice, Cardiac MRI is often used to establish differential diagnoses, for example interrogating the underlying etiologies of individuals presenting with heart failure. To illustrate this, we interrogate a subset of patients with the clinical label of "Non-Ischemic Cardiomyopathy" (n = 242) in our internal test dataset. This broadly includes

cardiomyopathies associated with infiltrative diseases, genetic, hypertrophic or dilated cardiomyopathies, and myocarditis among others. Such subgroup analyses demonstrate conditional model performance in clinically challenging contexts, for instance: Certain non-ischemic cardiomyopathies share overlapping phenotypic features of left-ventricular hypertrophy (sarcoidosis, amyloidosis, and hypertrophic cardiomyopathy for instance). This makes discerning the underlying etiology challenging for models finetuned to detect disease in the background of phenotypically very distinct negative classes. Nonetheless, our models demonstrate impressive performance in this subset for Amyloidosis (AUC 0.870; 95% CI 0.675 – 0.957), Dilated Cardiomyopathy (AUC 0.879; 95% CI 0.836 – 0.923), and Hypertrophic cardiomyopathy (AUC 0.925; 95% CI 0.873 – 0.977) (Supplementary Table 18).”

With the MRI reports available to us, we are however, able to perform some additional analyses on borderline cases as requested by reviewer #2. We report the probabilities of relevant labels for discussion (whether high or low probability, and whether or not they are correct), but as described in the methods, our models output probabilities for all 39 disease labels for each case discussed below. For each of these discussed cases, the models have no a-priori knowledge of the text in the reports and are purely test-set results. The following discussion is detailed in the Supplementary Appendix in its own dedicated section.

Case #1: Report mentions the following: “Diffuse thickening of the RV and LV compatible with hypertrophic cardiomyopathy. Patchy and diffuse areas of delayed enhancement in the apex, anterior wall and inferior wall which can be seen both with hypertrophic cardiomyopathy or infiltrative processes such as amyloidosis. Mildly depressed LV function with ejection fraction of 54%”

Our model returns probabilities in the range of 0 to 1. Probabilities for Amyloid (0.61, 85.5th percentile) and HCM (0.991, 96.8th percentile) indicating greater overall model confidence in HCM as disease label. True clinical ground truth without biopsy results is unlikely to be conclusive for this case.

Case #2: Report mentions the following: “Redemonstrated extensive multifocal thrombus involving the mid and apical left ventricle. Persistent small thrombus in the left atrium. Severe global ventricular contraction abnormality. … Transmural delayed enhancement involving the mid anterior and apical lateral segments. Additional extensive subepicardial and midwall delayed enhancement involving the mid through apical septal, anterior, and lateral segments… Severe left ventricular enlargement, biventricular contractile dysfunction (LVEF 10% and RVEF 22%), and associated extensive left ventricle myocardial delayed enhancement with predominantly subepicardial involvement and areas of transmural involvement as described above. Findings are compatible with a nonischemic dilated cardiomyopathy, with the clinical course and imaging pattern suggesting an aggressive myocarditis (giant cell versus other viral), versus less likely other etiologies including HIV myocarditis or sarcoidosis.

For this case, our model returns the following probabilities: DCM (0.989), Myocarditis (0.96), LV dilation (0.99), Delayed enhancement (0.961), Sarcoidosis (0.58), Thrombus (0.94). Between the differentials raised of DCM from aggressive myocarditis to sarcoidosis, our model appears to agree with the reporting radiologist that this is more likely DCM / Myocarditis than Sarcoidosis.

Case #3: This is a complex case with numerous atypical overlapping disease features present. Report reads “There is prominent trabeculation of the anterior, lateral, and inferior left ventricle. Compacted myocardium at end diastole … The noncompacted myocardium along the lateral wall at the same level measures 15 mm, giving a noncompacted to compacted myocardial ratio of 3.75. There is a focal area of thinning and rightward bowing of the apical septum… Transmural delayed enhancement is seen in the subapical and mid inferior wall of the left ventricle. Dilated left ventricle with an appearance consistent with partial noncompaction. Reduced biventricular function with LVEF 22% and RVEF 22… Transmural delayed enhancement of the of the inferior mid and apical left

ventricle consistent with prior infarct. Mitral stenosis and regurgitation, with regurgitant fraction of 49% based on volume measurement."

Our model returns the following probabilities: DCM (0.97), LV Non-compaction (0.84), LV dilation (0.93), RV dilation (0.91). Overall, the model detects the presence of dilated cardiomyopathy, but the 'partial non-compaction' is a challenging feature evidenced by the slightly lower predicted probability for LV non-compaction. Our model correctly rejects the diagnosis of either Amyloid (0.03) or HCM (0.06) with very low predicted probabilities. The mitral regurgitant fraction of 49% approaches the "Severe" cut-off of 50% and our model returns an appropriately high probability for Severe MR (0.82).

Case #4: This is another complex case with multiple atypical overlapping and unrelated feature of HCM, pericarditis and an old myocardial infarction, the report reads: "There is left atrial enlargement. The left ventricle is hypertrophied, most prominently involving the basal anterior/anteroseptal wall measuring up to 2.3-cm, the mid septum measuring up to 1.9-cm, and the apical region concentrically. There is mid-cavity LV obliteration, with associated flow acceleration.. Systolic anterior motion of the anterior leaflet of the mitral valve is seen, with resultant mitral regurgitation… Delayed enhancement with focal thinning and akinesis of the basal inferior wall consistent with prior myocardial infarct. On the 3-chamber view, there are punctate foci of delayed enhancement along the anteroseptal wall, which may represent scar related to hypertrophic myocardium. Diffuse circumferential pericardial delayed enhancement, without pericardial thickening or effusion… Findings may still be related to inflammatory process such as pericarditis… Mediastinal lymphadenopathy with a 2.7 x 1.9-cm low right paratracheal lymph node… MRI findings consistent with hypertrophic obstructive cardiomyopathy. Flow acceleration in the left ventricular outflow tract, with associated systolic anterior motion of the anterior leaflet of the mitral valve and resultant mitral regurgitation. Anteroseptal maximal thickening of 2.5-cm, associated with punctate foci of delayed enhancement in that region, consistent with myocardial scar. Evidence of prior myocardial infarction along the basal inferior. Abnormal enhancement of the pericardium, concerning for inflammatory process such as pericarditis"

Our model returns the following probabilities: HCM (0.99), Severe MR (0.98), SAM (0.99), Myocardial infarct (0.84). Our model correctly rejects Ischemic cardiomyopathy (0.21) despite the high probability returned for detecting an old infarct. However, our model incorrectly predicts a high probability of amyloidosis (0.97) and lower probability for pericarditis (0.31) – possibly confused by the hypertrophic ventricle and appearance of the pericardium.

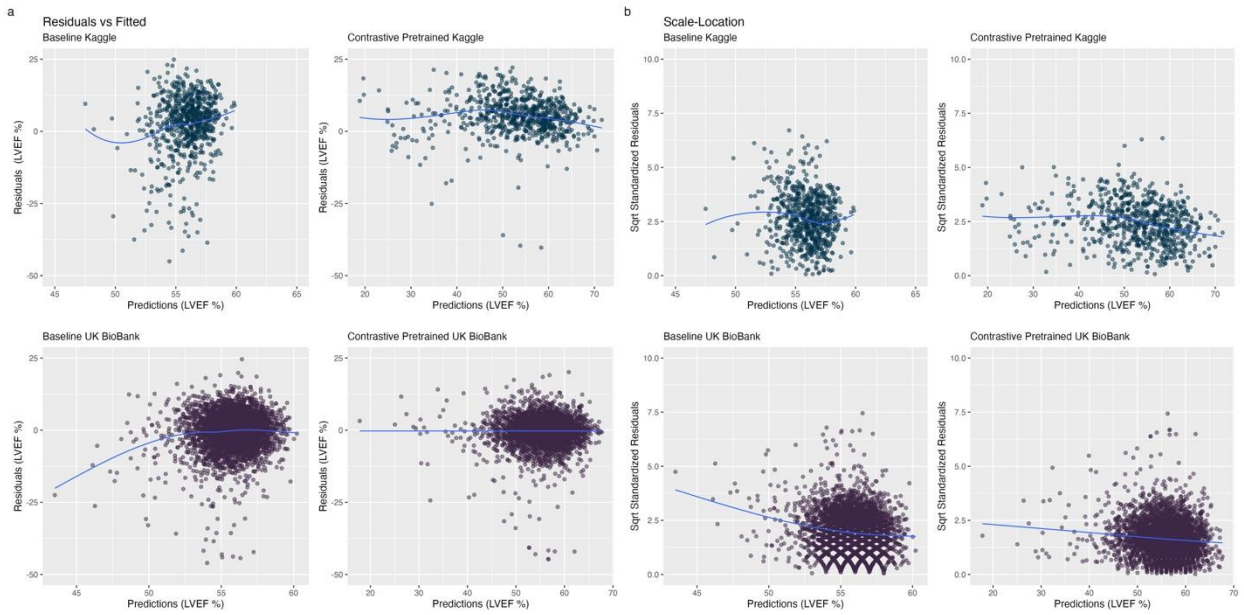
Case #5: Report reads as follows: "The left ventricle is mildly enlarged. Left ventricular wall thickness is upper normal… The pericardium appears of normal thickness and without significant effusion. Function: LV systolic function is mildly reduced. There are no wall motion abnormalities. RV systolic function is normal. Focal, transmural enhancement is noted within the basal to mid inferior wall. No other areas of abnormal late gadolinium enhancement. Mild left ventricular enlargement with mild global systolic dysfunction (EF 45%) … Focal, transmural late gadolinium enhancement noted within the basal to mid inferior wall. Differential considerations include myocardial infarction within the right coronary artery distribution, likely of embolic origin, as well as atypical appearance of a nonischemic etiology such as myocarditis"

The report is rather inconclusive for this case with a broad differential ranging from either an old infarct or non-ischemic disease such as myocarditis. Our model returns the following probabilities: Amyloid (0.94), Dilated cardiomyopathy (0.74), Ischemic cardiomyopathy (0.96), ischemic cardiomyopathy (0.96), Myocardial infarct (0.97), Myocarditis (0.25), Sarcoidosis (0.93). Our model and the report are in concordance given the broad differential and high probability outputs for a variety of different etiologies.

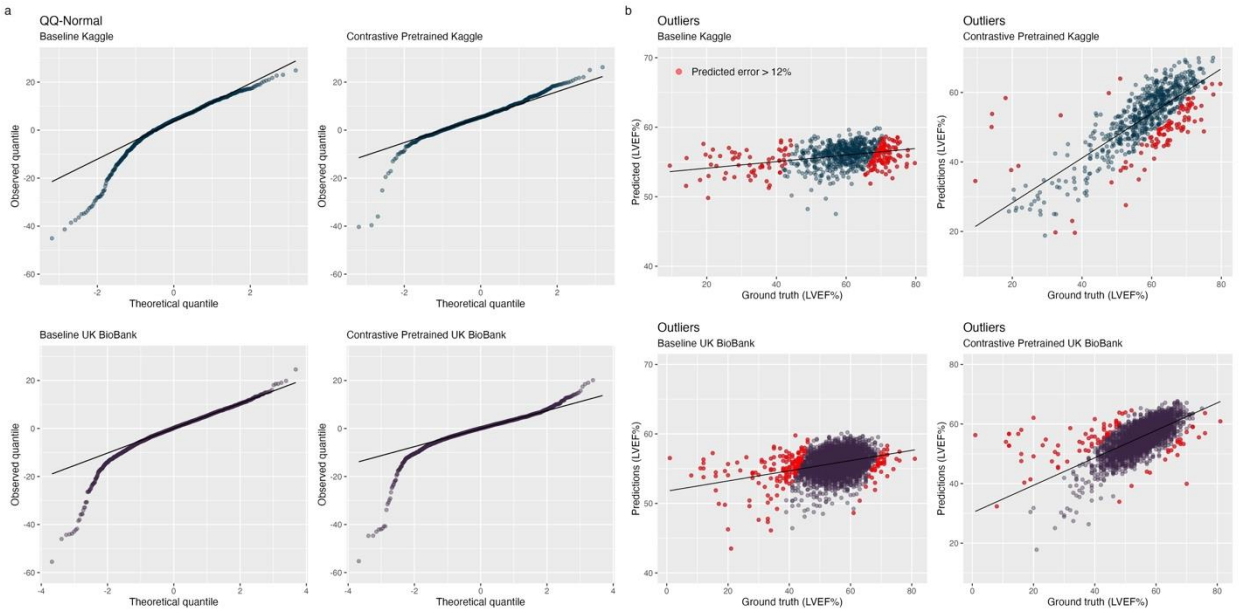
4. The presentation of results lacks sufficient detail to fully assess the model's performance. For continuous

variables like ejection fraction, detailed plots showing the distribution of errors and outliers are needed. More clinically relevant metrics and visualizations are required to understand the practical implications of the model's performance. The ROC curves provide most information, and I have explained my issues with them above.

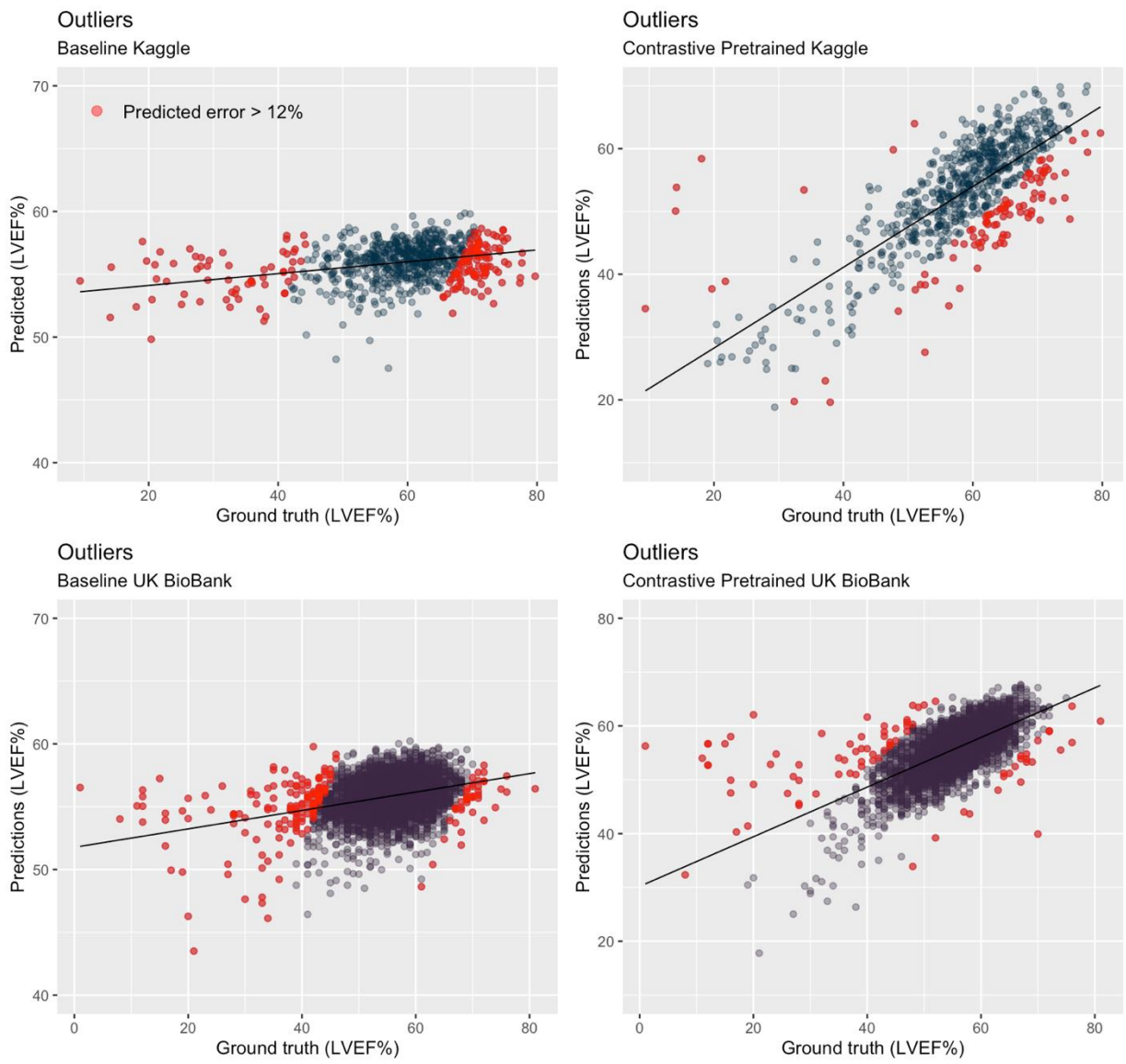
We have revamped the presentation of the regression experiments for ejection fraction. We now display multiple plots in a larger format for review in the Supplementary appendix, accompanied by a detailed explanation and interpretation of the same. We have created new plots for predicted vs ground-truth labels for LVEF that clearly highlight predictions that are incorrect by a margin greater than 12% for LVEF (based on generally accepted Bland-Altman limits in clinical practice) in red for clarity.[23] Each plot for the following well established methods of running model diagnostics on regression analyses have been redone: residuals vs fitted plots, the scale-location plots, and the qq-normal plots, in addition to direct plots of predicted vs ground truth LVEF% values. We stratify each plot by baseline and our contrastive pre-trained method for both the Kaggle and the UK Biobank test sets. Figures and image captions displayed below for reference:



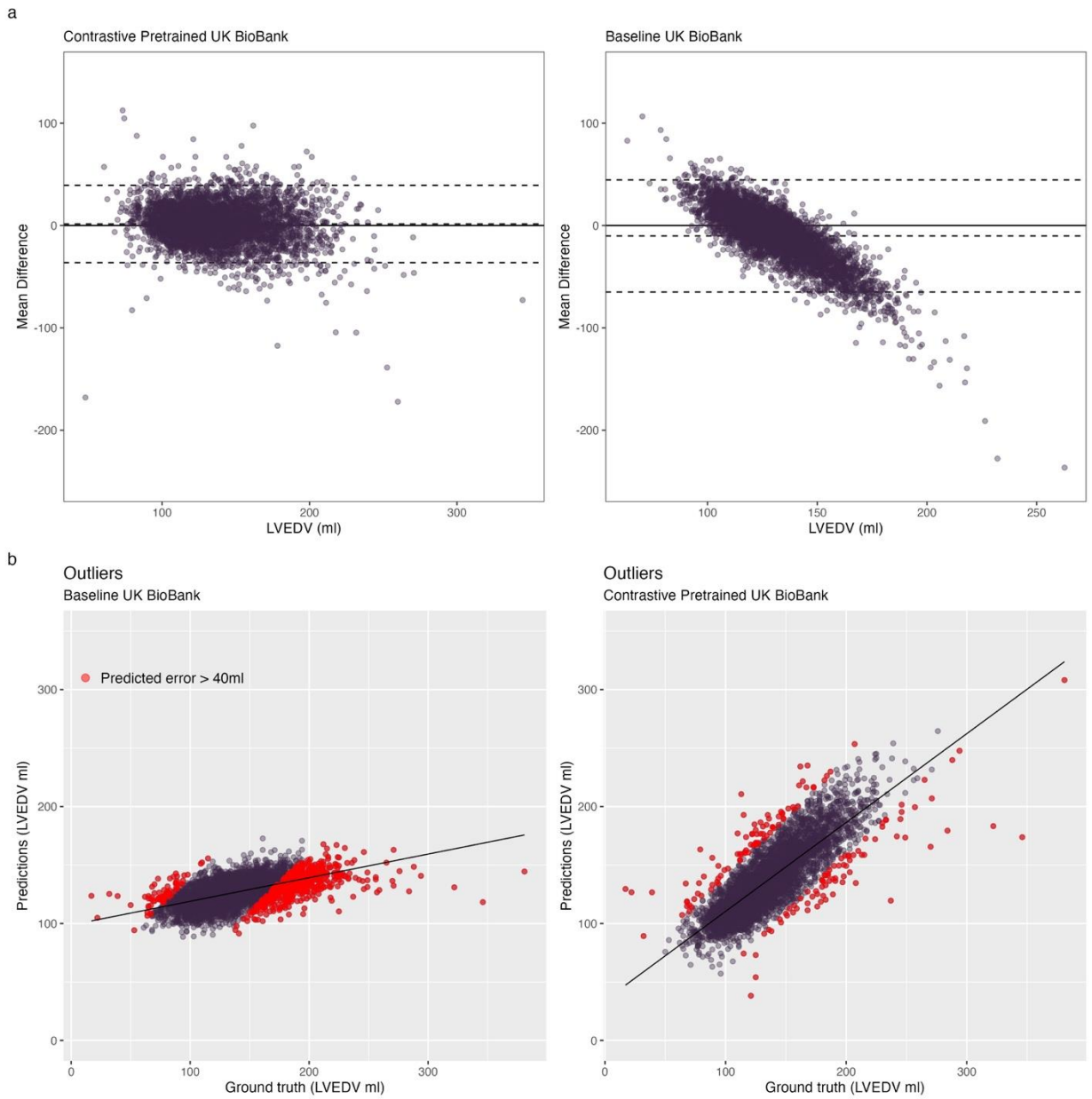
*"Supplementary Figure 7: Regression diagnostic plots **a.** Residuals vs. Fitted and **b.** Scale-Location - were generated for baseline and contrastively pretrained models on both the Kaggle and UK Biobank datasets. Baseline kinetics-pretrained models exhibited noticeable curvature in the residual plots, particularly for UK Biobank, suggesting systematic bias and potential model misspecification. In both datasets, baseline models also showed heteroscedasticity, with residual variance changing across the prediction range. In contrast, the contrastively pretrained models demonstrated flatter residual trends and more uniform variance in the Scale–Location plots, indicating improved calibration, reduced systematic bias, and greater stability of prediction error across the output range."*



"Supplementary Figure 8: Quantile–quantile (QQ) plots and outlier analyses were performed for baseline and contrastively pretrained models on the Kaggle and UK Biobank datasets. ***a****. In the QQ plots, both baseline models showed systematic deviations from normality, with heavier-than-normal tails and curvature at the extremes, particularly for the UK Biobank baseline. Contrastive pretraining dramatically reduced tail deviations and brought the residual distribution closer to normality.* ***b****. Outlier plots (predictions vs. ground truth), with prediction outliers (errors > 12% LVEF) in red, revealed that baseline models exhibited more widespread and systematic outliers, including clusters of under- and overestimation across the target range. In contrast, contrastively pretrained models showed tighter clustering around the identity line and fewer high-leverage outliers, indicating improved agreement with ground truth and reduced prediction error variance."*



We also create outlier plots and Bland Altman plots for the illustrative LVEDV regression example in Supplementary Figure 16 with caption that now reads:

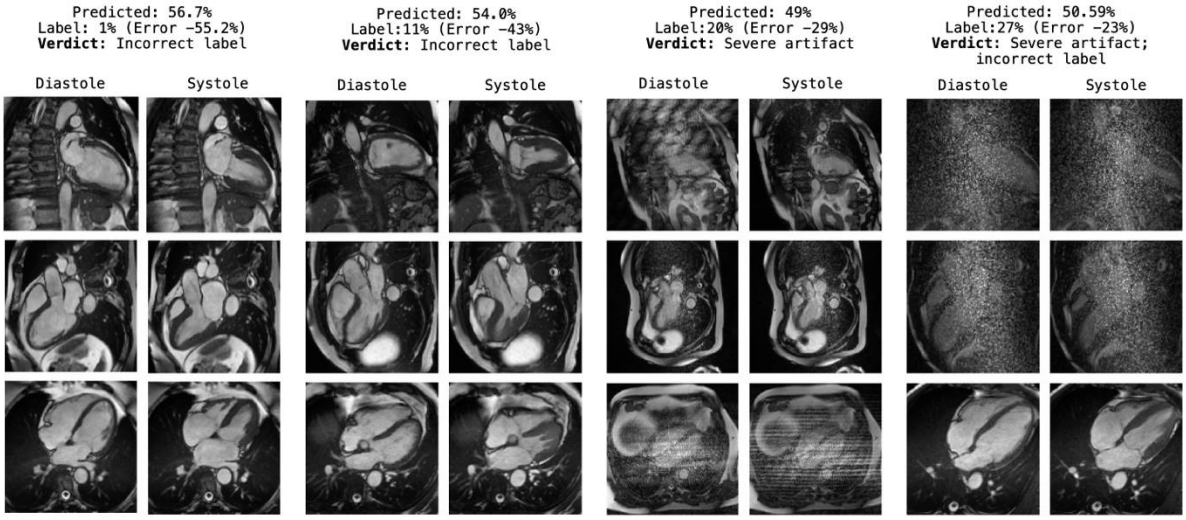


*"Supplementary Figure 16**a**. Bland-Altman plots for test-set results on the UK BioBank), with BA limits of agreement for contrastive pre-trained models resulting -36.26 ml to +32.32ml. For baseline methods, the BA limits of agreement are higher with a -10ml bias on average, with 95% BA limits of agreement at -64.98ml to + 44.59ml.* ***b.*** *Outlier plots (predictions vs. ground truth), with prediction outliers (errors > 40ml LVEDV) in red, revealed that baseline models exhibited more widespread and systematic outliers, including significant under- and overestimation across the target range. Our contrastively pretrained models showed tighter clustering around the identity line and fewer high-leverage outliers, indicating improved agreement with ground truth and reduced prediction error variance"*

Furthermore, to explore this point raised by reviewer #2 specifically dealing with understanding the practical implications of the model, we separately review each scan where the predictions of our model are egregiously incorrect (>20% LVEF error) in its predictions to understand the source of these errors – this directly informs readiness for prospective clinical deployment.

We found 25 cases in our test set where our model predictions were incorrect by a margin greater than 20%. For one case, we noted severe apical hypokinesia likely overestimation of LVEF by our model, and another case we noted a more hypertrophic LV where our model underpredicted LVEF. The remainder however: 16 cases had obviously incorrect ground truth values for LVEF. Exemplar studies are shown in Supplementary Figure 9 with systolic and diastolic phases captured displayed. The remaining 8 studies were either unusable because of significant artefactual degradation or poor image quality. Our observations are detailed in Supplementary Table 4.

If one were to exclude these 25 cases from the analysis the performance of our model on the UK Biobank test set in terms of mean absolute error (mae) would be **3.17!** (down from 3.33). We have decided against reporting this improvement in performance in our manuscript and supplement, given this would make our results challenging to compare against with contemporary literature that also likely using the same incorrect labels.



"Supplementary Figure 9: Examples of scans where our model predictions for LVEF% were incorrect by an absolute margin of more than 20% in the UK Biobank test set. Scans for each failure were manually reviewed. Examples of clearly incorrect labels (from left to right) shown where the LVEF% has erroneously been labelled at 1% or 11%, where in fact the label should be 50-60% based on independent review. Other examples where our model generated large errors is with severe artifacts in the scans where some or all views were impacted."

Filenames	Preds	Labels	residuals	Comments
ukbiobank_highresid_scan_1	52.7222	12	-40.7222	Incorrect ground truth label
ukbiobank_highresid_scan_2	60.88862	81	20.11138	Likely incorrect ground truth
ukbiobank_highresid_scan_3	58.594143	32	-26.594143	Incorrect ground truth label
ukbiobank_highresid_scan_4	56.683796	15	-41.683796	Incorrect ground truth label
ukbiobank_highresid_scan_5	56.647926	12	-44.647926	Incorrect ground truth label
ukbiobank_highresid_scan_6	52.857643	23	-29.857643	Incorrect ground truth label
ukbiobank_highresid_scan_7	56.265575	1	-55.265575	Incorrect ground truth label
ukbiobank_highresid_scan_8	56.709087	12	-44.709087	Incorrect ground truth label

ukbiobank_highresid_scan_9	47.456722	26	-21.456722	artifact; unusable
ukbiobank_highresid_scan_10	49.133	20	-29.133	severe artifact; unusable
ukbiobank_highresid_scan_11	54.786976	24	-30.786976	poor 3CH slice
ukbiobank_highresid_scan_12	54.000492	11	-43.000492	Incorrect ground truth label
ukbiobank_highresid_scan_13	50.592274	27	-23.592274	artifact and incorrect ground truth
ukbiobank_highresid_scan_14	49.881943	28	-21.881943	poor 4CH and 3CH slices, motion artifact
ukbiobank_highresid_scan_15	58.011852	16	-42.011852	poor 4CH slice, incorrect ground truth
ukbiobank_highresid_scan_16	32.343876	8	-24.343876	artifact; unusable
ukbiobank_highresid_scan_17	40.306206	17	-23.306206	artifact; unusable
ukbiobank_highresid_scan_18	52.7838	28	-24.7838	Incorrect ground truth label
ukbiobank_highresid_scan_19	61.638092	40	-21.638092	Incorrect ground truth label
ukbiobank_highresid_scan_20	62.076603	20	-42.076603	Incorrect ground truth label
ukbiobank_highresid_scan_21	41.40514	19	-22.40514	Apical dyskinesia; incorrect prediction
ukbiobank_highresid_scan_22	39.91553	70	30.08447	incorrect prediction
ukbiobank_highresid_scan_23	52.708202	12	-40.708202	Incorrect ground truth label
ukbiobank_highresid_scan_24	47.55127	16	-31.55127	Incorrect ground truth label
ukbiobank_highresid_scan_25	49.92222	16	-33.92222	Incorrect ground truth label

"Supplementary Table 4: Manual adjudication of all cases in the UK BioBank test set where our model predictions were incorrect by an absolute margin greater than 20%. For one case, we noted severe apical hypokinesia likely overestimation of LVEF by our model, and another we noted a more hypertrophic LV where our model underpredicted LVEF. The remainder however: 16 cases had obviously incorrect ground truth values for LVEF. The remaining 8 studies were either unusable because of significant artefactual degradation or poor image quality."

A few smaller comments:

5. Please provide data about the severity of the diseases? Early recognition or end-stage disease? Please provide data on prevalence of the various categories in each dataset?

Severity of different disease labels are now listed for the internal dataset (Stanford, MedStar, UCSF) and external dataset (UPenn) in Supplementary Tables 16 and 17. Please note that text indicative of severity is not always mentioned in the radiologists' report and these tables offer only an estimate of overall spread of disease severity. Our models detect all instances (mild disease onwards) for labels that don't explicitly have a prefix (i.e. *Severe* Mitral Regurgitation or *Large* Pericardial Effusion)

6. Do you know how accurate the vectorization of the reports was (do you have a metric for this and did you assess it against this metric).

As with reviewer #1's questions on our choice of LLM system for free text-report extraction, this is an important aspect of our work that we are happy to expand on. Please see the above discussion on our updated results and repeated experiments with medgemma3:27B labelling. At the outset of our work, we had manually parsed 537 reports from Stanford, MedStar, and UCSF for 35 disease conditions and a number of volumetric / functional measurements. We benchmark the LLM systems being used for extracting data via inter-rater agreement via Kohen's Kappa (IRR; higher is better) against human labels. The first model we could reliably use for data extraction was ChatGPT-turbo 3.5, which consistently showed median IRR of 74%. Our new medgemma3:27B labelling system achieves a median IRR of 83% which is conventionally considered to be "excellent agreement".

7. Do you know how accurate the reports themselves were, as there can be large differences in reporting between different radiologists. Was this validated?

Data was sourced from multiple large academic centers from scans performed and reported as part of routine clinical practice. Given the wealth of experience at each center in terms of imaging faculty and staff, and that we routinely rely on these reports for clinical use, we assume these are sufficiently accurate for model pretraining. Furthermore, each scan is only ever reported by one physician, with duplicate overreading in clinical practice exceedingly rare that entirely precludes calculations of interrater reliability or 'consensus reporting'. For the purposes of our large-scale project, we accept a degree of variability between radiologists as the reality of clinical practice. Furthermore, we argue that this heterogeneity is good for model training, as the models see scans of multiple variations of similar disease conditions described in slightly unique ways. Variability in reporting practices actually helps serve as a form of natural 'text-augmentation' that stabilizes pre-training.

However, we, accept that for certain cardiomyopathies and future clinical tasks of interest, it would be helpful to have adjudicated datasets of reasonable scale (200-500 patients) from multiple imaging experts for consensus labels. These smaller scale datasets would enable us to finetune models more consistent with consensus guidelines. We have amended the section discussing the strengths and limitations of our work to now read: (Discussion, paragraph 2):

"By pre-training and subsequently finetuning towards specific problems, we show that with remarkable consistency, our networks perform superior to baseline methods. We show that across numerous unrelated tasks, superior performance can be achieved with up to two orders of magnitude less data. This data efficiency is a critical advance for the development of finetuned models for the diagnosis and characterization of complex inherited cardiomyopathies, where expert adjudicated registries with paired imaging data remain scarce.[37,38]"

8. How would a physician looking after the patient analysed with your algorithm understand how likely the patient was correctly classified, how can they reassess if there are no numbers for volumes, wall thickness, muscle mass, etc. as this is the classical pathway for physicians to make decisions.

We thank reviewer #2's comment re: production deployment of our models. Our current architecture and design allow us to finetune and output as many variables and targets as needed. For example, just as we regressed LVEF% directly without segmentation, we can learn and output diameters, mass, and wall thicknesses. We demonstrate a new target - LVEDV on the UK BioBank for example, with experiments detailed in the Supplementary Appendix. For our clinicians, we agree with reviewer #2: our models must output both the probabilities of different disease labels and display traditional metrics to allow clinicians to make decisions. We argue this is a major strength of our approach, as the computational burden of all disease probability labels, traditional volumetrics, and numeric values for muscle mass etc all arise from a single large pre-trained model.

9. Cine imaging – while important – provides only a small fraction of the important data for the final diagnosis and especially the differential diagnoses as discussed earlier. Most differential diagnoses are based on scar images (LGE), perfusion and maps.

This remains a limitation of our methods that we have transparently discussed. (Discussion, Paragraph 4):

"*Another limitation of our current work is the reliance on cine-SSFP sequences alone. While the reports describe*

findings from gadolinium enhanced imaging, the networks do not make use of LGE (Late Gadolinium Enhanced) sequences as inputs. This is largely because LGE sequences are often captured as images rather than dynamic-motion videos, necessitating a separate and parallel image-based neural network to be built into the current framework. Future research will incorporate non-standard view-planes along with T1, T2, LGE, and perfusion scans via separate image encoders or dimensionally agnostic vision encoders capable of handling both image and video data. Despite this, our models consistently perform well on tasks with just cine-SSFP sequences in situations where clinicians typically require additional imaging data. LGE, for instance, is routinely used by clinicians for diagnosing Amyloidosis; similarly LGE along with T1, and ECV (extracellular volume) maps are often used in diagnosing Hypertrophic cardiomyopathy."

While we have demonstrated results that show we are able to detect signal that allows for impressive diagnostic ability for disease conditions such as Amyloidosis, however we accept that the absence of other sequences likely explains the performance limitations for labels such as that of myocarditis and potentially in the ability of our model to differentiate thrombus from certain tumors.

(Diagnostic performance strength and limitations, Paragraph 1):

"Differentiating tumor from thrombus, or diagnosing Myocarditis can be particularly challenging for clinicians without contrast enhancement or T2 imaging [16–18]. The Lake Louise imaging criteria used for the diagnosis of Myocarditis in clinical practice, for instance, relies on T1, T2, or LGE sequences.[19] Our results indicate that cine-sequences alone are unlikely to provide enough signal to reliably diagnose such conditions despite the presence of textual information in the reports"

10. Generalizability remains an issue. The UK Biobank data is highly homogenous due to identical scanners and protocols used.

Our core vision models have learned what they have learned from **our clinical pre-training dataset** (Stanford, MedStar, UCSF) – one that contains a variety of scanners, scanning protocols, and field strengths (Supplementary Table 2).

We only use the UK biobank data to finetune a regression system for LVEF% (ie. the core vision encoder of the model **does not learn** any features from the UK biobank), and thereafter we show through our experiments that this LVEF% prediction model generalizes well externally to the Kaggle dataset. We agree with reviewer #2 regarding the homogenous nature of the UK Biobank, and this limitation is far more relevant for some of the recent papers that **primarily use** the UK Biobank to train and evaluate their models with no other source of data.

11. The paper repeatedly claims the model "understands" cardiovascular disease. This is an overstatement. The model learns statistical associations but doesn't possess genuine understanding in the way a clinician does.

We have amended the language to address this point, instead favoring the use of 'representing' or 'contextualizing' cardiovascular disease wherever applicable.

I am not an expert in models themselves and cannot assess the technical performance, requirements or novelty of the vision model or the LMMs used for fine-tuning.

References:

1. Tiu, E. *et al.* Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng* https://doi.org/10.1038/s41551-022-00936-9 (2022) doi:10.1038/s41551-022-00936-9.
2. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat Med* **30**, 850–862 (2024).
3. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive Learning of Medical Visual Representations from Paired Images and Text. *arXiv:2010.00747 [cs]* http://arxiv.org/abs/2010.00747 (2020).
4. Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nat Med* **30**, 863–874 (2024).
5. Taleb, A. *et al.* 3D Self-Supervised Methods for Medical Imaging. *arXiv:2006.03829v3 [cs.CV]* 19.
6. Azizi, S. *et al.* Big Self-Supervised Models Advance Medical Image Classification. Preprint at https://doi.org/10.48550/arXiv.2101.05224 (2021).
7. Krishnan, R., Rajpurkar, P. & Topol, E. J. Self-supervised learning in medicine and healthcare. *Nat. Biomed. Eng* **6**, 1346–1352 (2022).
8. Zhou, Y. *et al.* A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
9. Oord, A. van den, Li, Y. & Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]* http://arxiv.org/abs/1807.03748 (2019).
10. Vukadinovic, M. *et al.* EchoPrime: A Multi-Video View-Informed Vision-Language Model for Comprehensive Echocardiography Interpretation. Preprint at https://doi.org/10.48550/arXiv.2410.09704 (2024).
11. Blankemeier, L. *et al.* Merlin: A Vision Language Foundation Model for 3D Computed Tomography. Preprint at https://doi.org/10.48550/arXiv.2406.06512 (2024).
12. Beeche, C. *et al.* A Pan-Organ Vision-Language Model for Generalizable 3D CT Representations. Preprint at https://doi.org/10.1101/2025.07.03.25330654 (2025).
13. Martini, N. *et al.* Deep learning to diagnose cardiac amyloidosis from cardiovascular magnetic resonance. *J Cardiovasc Magn Reson* **22**, 84 (2020).

14. Hinojar, R. *et al.* T1 Mapping in Discrimination of Hypertrophic Phenotypes: Hypertensive Heart Disease and Hypertrophic Cardiomyopathy: Findings From the International T1 Multicenter Cardiovascular Magnetic Resonance Study. *Circ: Cardiovascular Imaging* **8**, e003285 (2015).
15. Ommen, S. R. *et al.* 2024 AHA/ACC/AMSSM/HRS/PACES/SCMR Guideline for the Management of Hypertrophic Cardiomyopathy: A Report of the American Heart Association/American College of Cardiology Joint Committee on Clinical Practice Guidelines. *Circulation* CIR.0000000000001250 (2024) doi:10.1161/CIR.0000000000001250.
16. Weinsaft, J. W. *et al.* Contrast-Enhanced Anatomic Imaging as Compared to Contrast-Enhanced Tissue Characterization for Detection of Left Ventricular Thrombus. *JACC: Cardiovascular Imaging* **2**, 969–979 (2009).
17. O'Donnell, D. H. *et al.* Cardiac Tumors: Optimal Cardiac MR Sequences and Spectrum of Imaging Appearances. *American Journal of Roentgenology* **193**, 377–387 (2009).
18. Wang, T. K. M. *et al.* Cardiac Magnetic Resonance Imaging Techniques and Applications for Pericardial Diseases. *Circ: Cardiovascular Imaging* **15**, (2022).
19. Eichhorn, C. *et al.* Multiparametric Cardiovascular Magnetic Resonance Approach in Diagnosing, Monitoring, and Prognostication of Myocarditis. *JACC: Cardiovascular Imaging* **15**, 1325–1338 (2022).
20. Kaur, D. *et al.* Abstract 4124675: Deep Learning Screening of Cardiac MRIs Uncovers Undiagnosed Hypertrophic Cardiomyopathy in the UK BioBank. *Circulation* **150**, (2024).
21. Petersen, S. E. *et al.* UK Biobank's cardiovascular magnetic resonance protocol. *J Cardiovasc Magn Reson* **18**, 8 (2015).
22. Witschey, W. R. T. *et al.* Medical image analysis platform and associated methods. (2024).
23. Bhuva, A. N. *et al.* A Multicenter, Scan-Rescan, Human and Machine Learning CMR Study to Test Generalizability and Precision in Imaging Biomarker Analysis. *Circ: Cardiovascular Imaging* **12**, e009214 (2019).
24. Suinesiaputra, A. *et al.* Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. *J Cardiovasc Magn Reson* **17**, 63 (2015).
25. Fu, Y. *et al.* A versatile foundation model for cine cardiac magnetic resonance image analysis tasks. Preprint at https://doi.org/10.48550/arXiv.2506.00679 (2025).

26. Zhang, Y. *et al.* Towards Cardiac MRI Foundation Models: Comprehensive Visual-Tabular Representations for Whole-Heart Assessment and Beyond. *Medical Image Analysis* **106**, 103756 (2025).
27. Bai, W. *et al.* Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson* **20**, 65 (2018).
28. Mathew, R. C., Löffler, A. I. & Salerno, M. Role of Cardiac Magnetic Resonance Imaging in Valvular Heart Disease: Diagnosis, Assessment, and Management. *Curr Cardiol Rep* **20**, 119 (2018).
29. Bosman, L. P. *et al.* Diagnosing arrhythmogenic right ventricular cardiomyopathy by 2010 Task Force Criteria: clinical performance and simplified practical implementation. *EP Europace* **22**, 787–796 (2020).
30. Hendrycks, D. *et al.* AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. Preprint at http://arxiv.org/abs/1912.02781 (2020).
31. Cubuk, E. D., Zoph, B., Shlens, J. & Le, Q. V. RandAugment: Practical automated data augmentation with a reduced search space. *arXiv:1909.13719 [cs]* http://arxiv.org/abs/1909.13719 (2019).
32. Li, Z. *et al.* Phenotype-Guided Generative Model for High-Fidelity Cardiac MRI Synthesis: Advancing Pretraining and Clinical Applications. Preprint at https://doi.org/10.48550/arXiv.2505.03426 (2025).
33. Reynaud, H. *et al.* Feature-Conditioned Cascaded Video Diffusion Models for Precise Echocardiogram Synthesis. in vol. 14229 142–152 (2023).
34. Cho, J. *et al.* MediSyn: A Generalist Text-Guided Latent Diffusion Model For Diverse Medical Image Synthesis. Preprint at https://doi.org/10.48550/arXiv.2405.09806 (2025).
35. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
36. Richardson, E. *et al.* The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns* **5**, 100994 (2024).
37. Ho, C. Y. *et al.* Genotype and Lifetime Burden of Disease in Hypertrophic Cardiomyopathy: Insights From the Sarcomeric Human Cardiomyopathy Registry (SHaRe). *Circulation* **138**, 1387–1398 (2018).
38. Charron, P. *et al.* The Cardiomyopathy Registry of the EURObservational Research Programme of the European Society of Cardiology: baseline data and contemporary management of adult patients with cardiomyopathies. *European Heart Journal* **39**, 1784–1793 (2018).