Journal of Cardiovascular Magnetic Resonance 26 (2024) 101040

Contents lists available at ScienceDirect

Journal of Cardiovascular Magnetic Resonance

journal homepage: www.sciencedirect.com/journal/jocmr

Original research

Quality assurance of late gadolinium enhancement cardiac magnetic resonance images: a deep learning classifier for confidence in the presence or absence of abnormality with potential to prompt real-time image optimization

Sameer Zaman[a,b,c,1], Kavitha Vimalesvaran[a,c,1], Digby Chappell[c], Marta Varela[a], Nicholas S. Peters[a,b], Hunain Shiwani[e,f], Kristopher D. Knott[e,g], Rhodri H. Davies[e,f], James C. Moon[e,f], Anil A. Bharath[d], Nick WF Linton[b,d,*], Darrel P. Francis[a,b], Graham D. Cole[a,b,2], James P. Howard[a,b,2]

[a] National Heart and Lung Institute, Imperial College London, London SW7 2AZ, UK
[b] Imperial College Healthcare NHS Trust, London W12 0HS, UK
[c] AI for Healthcare Centre for Doctoral Training, Imperial College London, London SW7 2AZ, UK
[d] Department of Bioengineering, Imperial College London, London SW7 2AZ, UK
[e] Institute of Cardiovascular Science, University College London, London WC1E 6DD, UK
[f] Barts Health Centre, St. Bartholomew's Hospital, London EC1A 7BE, UK
[g] St. George's University Hospitals NHS Foundation Trust, London SW17 0QT, UK

ARTICLE INFO

Keywords:
Late gadolinium enhancement
Artificial intelligence
Deep learning
Neural networks
Efficiency

ABSTRACT

Background: Late gadolinium enhancement (LGE) of the myocardium has significant diagnostic and prognostic implications, with even small areas of enhancement being important. Distinguishing between definitely normal and definitely abnormal LGE images is usually straightforward, but diagnostic uncertainty arises when reporters are not sure whether the observed LGE is genuine or not. This uncertainty might be resolved by repetition (to remove artifact) or further acquisition of intersecting images, but this must take place before the scan finishes. Real-time quality assurance by humans is a complex task requiring training and experience, so being able to identify which images have an intermediate likelihood of LGE while the scan is ongoing, without the presence of an expert is of high value. This decision-support could prompt immediate image optimization or acquisition of supplementary images to confirm or refute the presence of genuine LGE. This could reduce ambiguity in reports.
Methods: Short-axis, phase-sensitive inversion recovery late gadolinium images were extracted from our clinical cardiac magnetic resonance (CMR) database and shuffled. Two, independent, blinded experts scored each individual slice for "LGE likelihood" on a visual analog scale, from 0 (absolute certainty of no LGE) to 100 (absolute certainty of LGE), with 50 representing clinical equipoise. The scored images were split into two classes—either "high certainty" of whether LGE was present or not, or "low certainty." The dataset was split into training, validation, and test sets (70:15:15). A deep learning binary classifier based on the EfficientNetV2 convolutional neural network architecture was trained to distinguish between these categories. Classifier performance on the test set was evaluated by calculating the accuracy, precision, recall, F1-score, and area under the receiver operating characteristics curve (ROC AUC). Performance was also evaluated on an external test set of images from a different center.
Results: One thousand six hundred and forty-five images (from 272 patients) were labeled and split at the patient level into training (1151 images), validation (247 images), and test (247 images) sets for the deep learning binary classifier. Of these, 1208 images were "high certainty" (255 for LGE, 953 for no LGE), and 437 were "low

Abbreviations: CMR, cardiovascular magnetic resonance; LGE, late gadolinium enhancement; ROC, receiver operating characteristic curve; AUC, area under the curve; LV, left ventricular; AI, artificial intelligence; PSIR, phase-sensitive inversion recovery; GPU, graphical processing unit; 2D, two dimensional

* Corresponding author at: Department of Bioengineering, Imperial College London, London SW7 2AZ, UK.
E-mail address: n.linton@imperial.ac.uk (N.W. Linton).

[1] S.Z. and K.V. contributed equally to this work (joint lead authors).
[2] J.P.H. and G.D.C. contributed equally to this work (joint senior authors).

https://doi.org/10.1016/j.jocmr.2024.101040
Received 6 December 2023; Received in revised form 10 March 2024; Accepted 19 March 2024

certainty". An external test comprising 247 images from 41 patients from another center was also employed. After 100 epochs, the performance on the internal test set was accuracy = 0.94, recall = 0.80, precision = 0.97, F1-score = 0.87, and ROC AUC = 0.94. The classifier also performed robustly on the external test set (accuracy = 0.91, recall = 0.73, precision = 0.93, F1-score = 0.82, and ROC AUC = 0.91). These results were benchmarked against a reference inter-expert accuracy of 0.86.

Conclusion: Deep learning shows potential to automate quality control of late gadolinium imaging in CMR. The ability to identify short-axis images with intermediate LGE likelihood in real-time may serve as a useful decision-support tool. This approach has the potential to guide immediate further imaging while the patient is still in the scanner, thereby reducing the frequency of recalls and inconclusive reports due to diagnostic indecision.

1. Background

Cardiac magnetic resonance (CMR) imaging is the gold-standard investigation for a wide range of clinical conditions because of its unique ability to detect myocardial fibrosis through late gadolinium enhancement (LGE) [1–4]. LGE's role in prognosis is well-established but its assessment remains primarily qualitative [5], with even small amounts appearing to be prognostic [4,6]. There have been extensive efforts to develop tools to optimize and standardize LGE quantification from already-acquired images [7,8]. However, these approaches cannot tackle a problem with the upstream process, particularly for images in which the presence or absence of LGE is equivocal. Factors contributing to LGE equipoise include poor image quality and artifact (Fig. 1). If these issues are not identified during the scan, the resulting "intermediate LGE likelihood" images create a post-scan challenge for clinicians reporting scans to determine whether the signal seen is genuine or not. The ideal radiology report arising from a clinician's review of the image data should reduce diagnostic uncertainty [9], but many reports contain ambiguous elements that create misunderstanding between reporting radiologists and referring clinicians [10].

Once a patient exits the scanner, opportunities for additional imaging within the same session are lost, leaving clinicians with limited options to resolve uncertainties in LGE interpretation. Interrogating alternative image planes may provide some clarification, but these approaches do not help if the reason for equipoise is poor image quality or a lack of intersecting images. There are strategies to acquire more images which may remove equipoise, which in most cases, could turn a "low certainty" situation into a "high certainty" one, but applying these strategies indiscriminately to *all* cases is inefficient and unnecessary, when the uncertainty affects only a *minority* of cases.

In the current workflow, the detection of equivocal LGE may occur only when a clinician reviews the images at the end of the scan, after which an opportunity to acquire further images may have been lost (Fig. 2). Even if image optimization or additional imaging (which may add only minutes to the total scan time) are all that is needed to overcome the diagnostic uncertainty, realizing this after the scan has finished is too late because the patient has already left the scanner. The only choice then is to either accept the acquired data and report within the limits of uncertainty, or to recall the patient for another scan with resulting cost, inconvenience, and inefficiency.

The optimal moment for detecting intermediate LGE likelihood (and possible diagnostic uncertainty) is during the scan itself. This quality assurance could be provided by an expert human reviewing images in real time, but since qualitative LGE assessment is a complex task requiring extensive training and experience, this form of quality assurance is neither practical nor universally applicable. Computational approaches to automatically improve image quality are primarily directed at post-processing of images after they have already been acquired rather than flagging uncertainty during the scan while there is still opportunity to acquire more data [11,12]. We hypothesized that an artificial intelligence (AI) algorithm could provide automatic, accurate, and potentially real-time quality control of late gadolinium short-axis images based on LGE likelihood to guide further imaging.

In this study, we present a deep learning model that classifies short-axis left ventricular (LV) late gadolinium images into one of two classes: "high certainty" or "low certainty," based on the likelihood of LGE being present. We propose that this system could provide decision support to clinical users, while the scan is ongoing, to guide image optimization or the acquisition of additional sequences (such as intersecting images). This could reduce clinical uncertainty and enhance diagnostic accuracy by helping confirm or refute the presence of genuine LGE.

2. Methods

2.1. Data extraction and characteristics

Anonymized short-axis phase-sensitive inversion recovery (PSIR) late gadolinium images for 300 scans (3152 image slices) performed at our center in London, UK, between 2018 and 2021, were randomly extracted from our local CMR database. Additionally, for an external test set, the same type of images were extracted from 41 randomly selected scans of all-

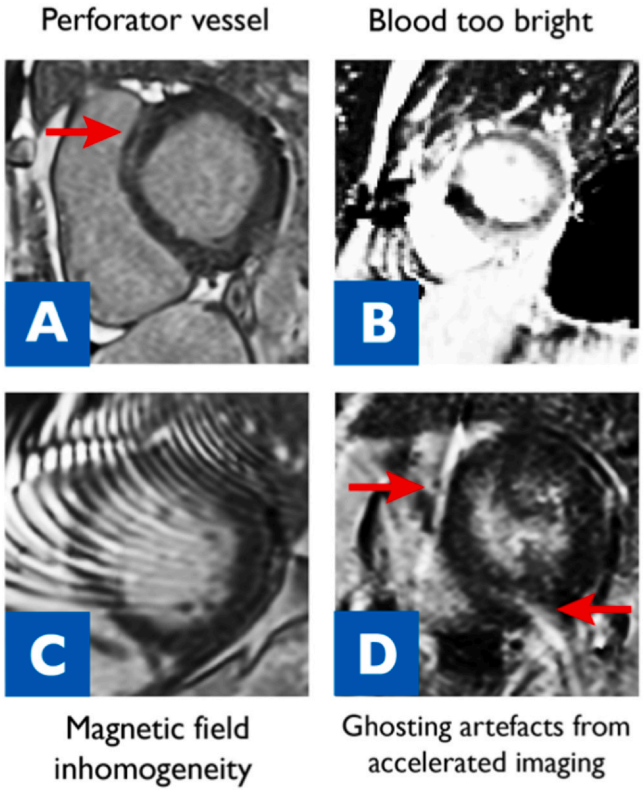

Fig. 1. Four examples of short-axis late gadolinium images with an intermediate (uncertain) likelihood of LGE. (A) Midwall signal in the basal anteroseptum (red arrow), likely artifact from a myocardial perforating vessel. (B) Signal from blood pool is too bright, preventing reliable assessment of subendocardial LGE. There is also signal dropout in the septum, artifact from a pacemaker. (C) Moire fringes artifact from magnetic field inhomogeneity, which obscures half of the myocardial circumference. (D) Two localized ghosting artifacts (red arrows) from accelerated (parallel) imaging, obscuring part of the anteroseptum and inferior walls. LGE, late gadolinium enhancement

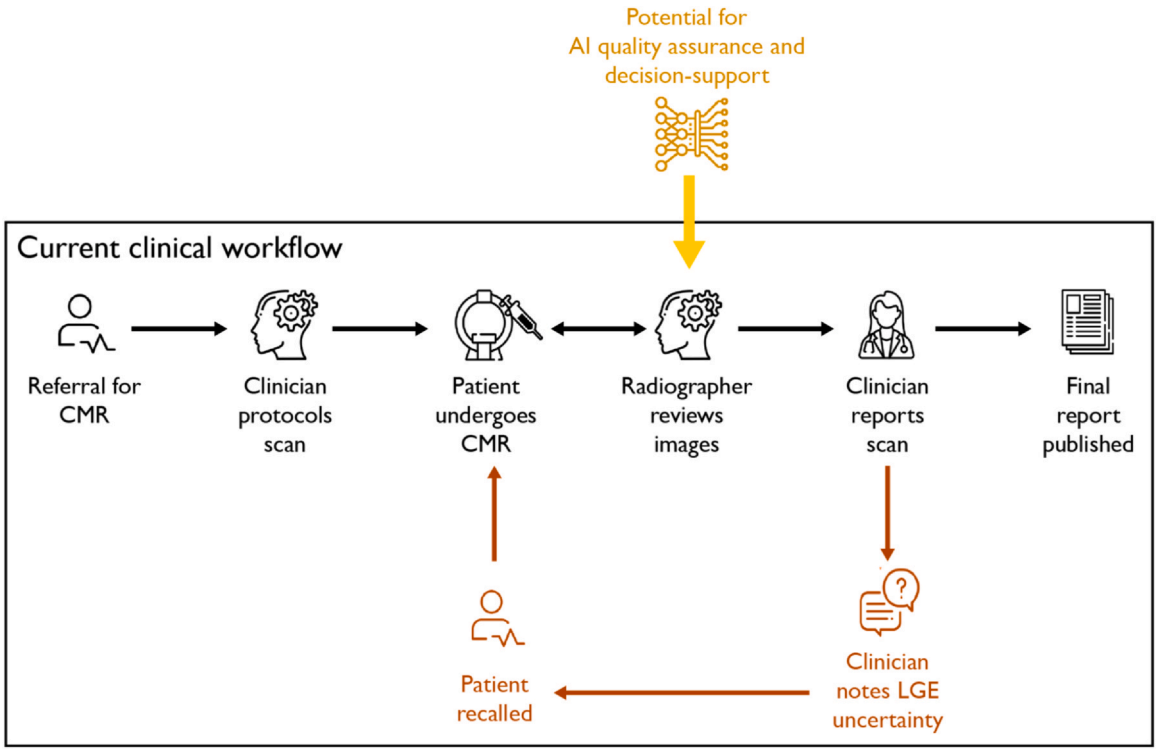

Fig. 2. The current clinical CMR workflow. Black text and arrows denote usual workflow. Red text and arrows denote additional workflow in situations with intermediate LGE likelihood (i.e., uncertainty). Yellow text and arrow denote potential application of an artificial intelligence-based quality assurance algorithm with real-time decision support. CMR: cardiovascular magnetic resonance, LGE: late gadolinium enhancement.

comers to another center in London, UK, between 2016 and 2018 (totaling 496 image slices). Ethical approval was gained from the Health Regulatory Agency (Integrated Research Application System identifier 243023).

Images were from a 1.5T or 3.0T scanner (Siemens Aera, Skyra or Prisma, Siemens Healthineers, Erlangen, Germany). Acquisition parameters for the included sequences are shown in Appendix A.

2.2. Data curation

Each short-axis stack was split into its constituent individual images and shuffled. Two experienced CMR clinicians reviewed the extracted images and excluded them if:

(i) The file was corrupted during transfer or extraction.
(ii) Incomplete imaging due to hardware or early termination of the scan.
(iii) The slice did not have at least 50% of the LV myocardium visible around the blood pool, according to the Society of Cardiovascular Magnetic Resonance guidelines [5], due to being "too basal" or "too apical."

Slices that had sub-optimal imaging (e.g., incomplete myocardial nulling or artifact obscuring part of the myocardium) were not excluded. The remaining images were uploaded onto the Unity image labeling platform [13] (Fig. 3).

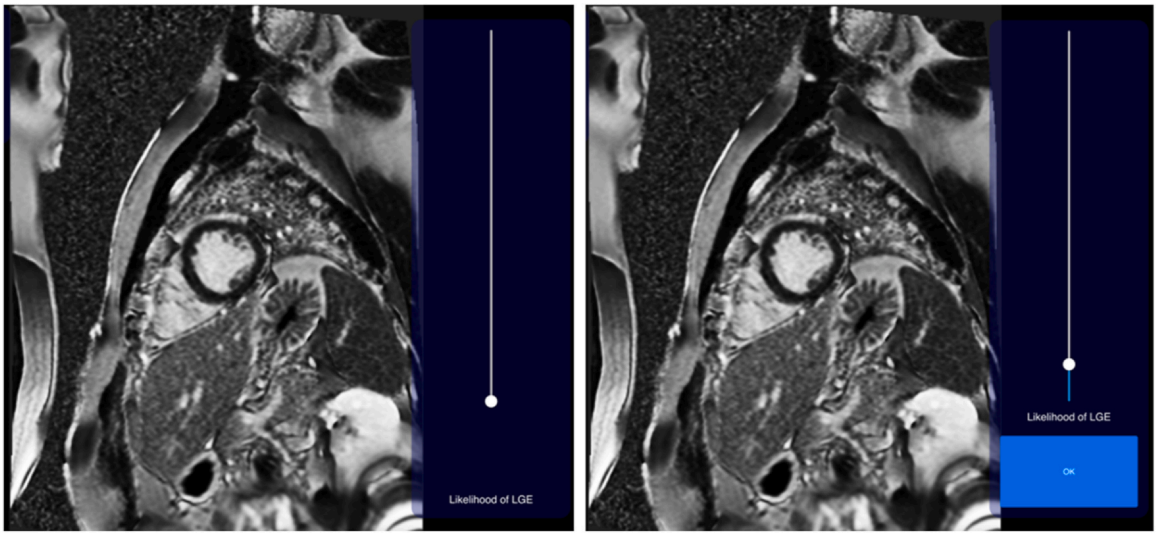

Fig. 3. Screenshots of LGE likelihood labeling on the Unity imaging platform. LGE: late gadolinium enhancement.

Table 1
Key training parameters of deep learning classifier. (Data are descriptions, ratios, counts or settings).

Parameter	Value
Architecture	EfficientNetV2 + feed forward neural network
Train:validation:test	70:15:15
Trainable parameters	23,590,049
Optimizer	Adam with weight decay
Learning rate	1e-4

2.3. Data labeling

Two independent experienced CMR clinicians (S.Z. and K.V., board accredited with > 3 years CMR reporting experience each) performed labeling on the same set of short-axis PSIR late gadolinium images. All images were presented to both labelers, who worked on the task independently and were blinded from each other's ratings.

Individual images were presented one at a time on the platform, accompanied by a single sliding visual analog scale (the values on the scale were not shown to labelers). The scale was titled "Likelihood of LGE." Labelers were able to zoom, pan, and adjust the windowing for each image. After the labeler submitted a rating for a given image, the label was stored on a cloud-based data frame and that image was not shown again to that labeler.

Human operators were asked to follow the following guidelines:

(i) 0% LGE likelihood: if you are sure there is no genuine LGE in this image
(ii) 100% LGE likelihood: if you are sure that genuine LGE is present (no matter the amount of LGE) somewhere in this image

Apart from these broad guidelines, labelers were free to rate images based on their personal opinion of LGE likelihood for each image.

2.4. Evaluation of labeler agreement and determination of class labels

Agreement between the two labelers' ratings of LGE likelihood was evaluated by calculating the Spearman's rho correlation coefficient and the inter-rater agreement (accuracy). The mean of the two labelers' ratings was taken as the average LGE likelihood for each image. The mean difference between the two labelers' ratings and its 95% confidence interval was used to determine the thresholds of LGE likelihood between classes. Initially, two thresholds were applied, resulting in three classes:

(i) Threshold 1: Between "low LGE likelihood" and "intermediate LGE likelihood"
(ii) Threshold 2: Between "intermediate LGE likelihood" and "high LGE likelihood"

Preliminary experiments showed that a three-class classification network did not perform well. Therefore, the final two classes were determined by combining the "low LGE likelihood" and "high LGE likelihood" images into a single class: "high certainty." The "intermediate LGE likelihood" images were renamed: "low certainty." These two classes were the final labels for training, validation, and testing of the deep learning network.

2.5. Neural network design and training

The labeled data were split into training, validation, and test sets (70:15:15), maintaining the relative class distributions. The neural

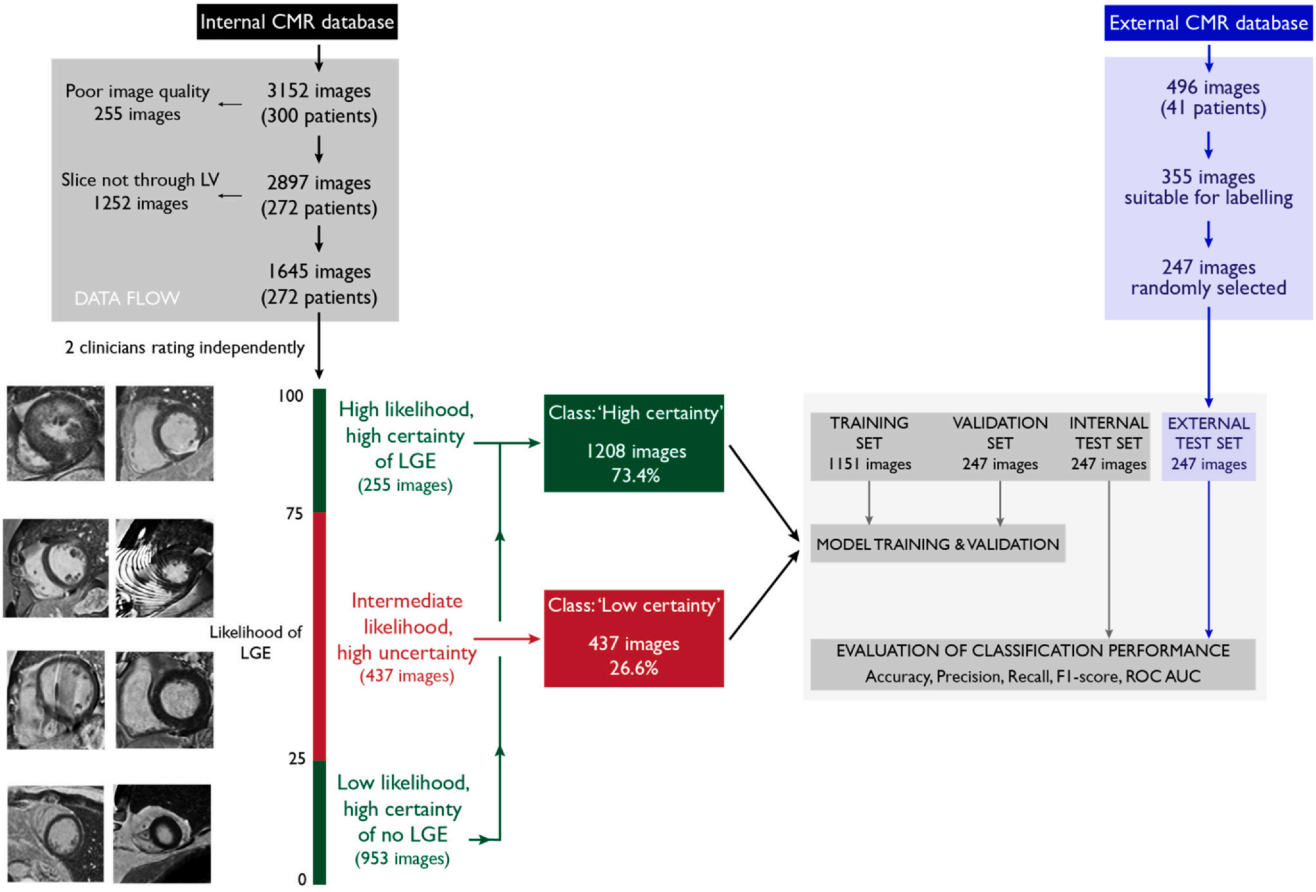

Fig. 4. Study design and data flow. CMR: cardiovascular magnetic resonance, LV: left ventricle, LGE: late gadolinium enhancement, ROC AUC: receiver operating characteristics area under the curve.

Table 2
Two labelers' ratings of LGE likelihood for 1645 images. (Data are counts or proportions).

LGE likelihood (%)	Rater 1	Rater 2
0-25	955	968
26-74	438	401
75-100	252	276
Spearman's rho (vs rater 1)	-	0.82

LGE late gadolinium enhancement.

network architecture chosen was an adapted version of EfficientNetV2 [14], modified by replacing the final layer with a feed-forward neural network. The weights of the features from EfficientNetV2 were reused but not the weights of its classifier. The weights of the feed-forward neural network were randomly initialized according to the He Normal method [15]. The final layer in the feed-forward network was programmed to output the probability of the input image being "high certainty" or "low certainty." The class with the higher output probability was assigned as the prediction for that image.

To enhance the model's generalizability and prevent overfitting, we applied specific pre-processing steps in the training dataset. Images were first cropped 40% top/bottom and 30% left/right to remove non-cardiac structures from the image periphery. This resulted in rectangular images that were made square by applying a 256 × 256 pixel center crop. To prevent the network from memorizing specific images, the training set was then augmented by random crops to a 224 × 224 pixel size, flips and rotations. These augmentation operations were exclusive to the training dataset; images in the validation and test sets were only resized to 224 × 224 pixels to match the size of the training images.

Our training process underwent 10 training runs starting with different random weight initializations. The model was trained on a Tesla P100-PCIE-16 GB graphical processing unit. Model parameters are shown in Table 1.

2.6. Evaluation of model performance

Model performance was evaluated on an internal, held-out test set of 247 images (15% of the dataset) that had been curated and labeled by the same clinicians in the same way as the training set, but reserved exclusively for post-training evaluation. Performance was also evaluated on an external test equivalent in size to the internal test set (247 images) obtained from an external center, having been curated and labeled in the same way as the internal test set.

Performance metrics included accuracy, sensitivity (recall), specificity (precision), F1-score (the harmonic mean of precision and recall), and area under the receiver operating characteristics curve (ROC AUC).

An overview of the data flow and study design is shown in Fig. 4.

3. Results

3.1. Data acquisition and image selection

Three thousand one hundred and fifty-two short-axis PSIR late gadolinium images were extracted for 300 patients. Two hundred and fifty-five images (28 patients) were removed after applying the exclusion criteria. One thousand two hundred and fifty-two images were excluded because they did not image > 50% of the LV myocardium due to being "too basal" or "too apical." The remaining 1645 images (from 272 patients) were uploaded to the Unity medical image labeling platform [13]. Using a split of 70:15:15, this resulted in 1151 images for training, 247 images for validation, and 247 images in the internal, held-out test set.

Additionally, 496 short-axis PSIR late gadolinium images (from 41 patients) were randomly extracted from an external CMR database. After the application of the exclusion criteria, 141 images were removed. The remaining 355 images (from 41 patients) were labeled in the same way as the internal set of images. Finally, to match the size of the internal test set, 247 images were randomly selected to make the final external test set.

3.2. Clinician label agreement

There was overall good agreement between the two clinicians' independent blinded ratings of LGE likelihood on the internal dataset (Spearman's Rho correlation coefficient = 0.82) (Table 2). The inter-expert agreement (accuracy) was 86% (Appendix B, Figure B1). The median difference in LGE likelihood between the two clinicians was 2% (−4 to +8%). The mean of the two clinicians' readings was taken as the final LGE likelihood label for each image.

3.3. Categorization of LGE likelihood into certainty classes

To establish class thresholds (cut-offs) for our binary classifier, we employed a statistically-driven approach based on the 95% confidence interval for the mean difference between the two expert's labels. The mean absolute difference between the two clinicians was 1.409% (± 13.660%). This meant that in 95% of cases, the difference in the LGE likelihood between raters lay between −25.364% and +28.183%. The following LGE likelihood cut-offs were applied to categorize the data into two classes (Fig. 4):

(i) "High certainty": images with LGE likelihood scores ranging from 0%-25% and 75%-100%. These scores represented a high degree of confidence in the presence or absence of LGE.
(ii) "Low certainty": images with LGE likelihood scores between 26% and 74%. These scores indicated a moderate or uncertain level of confidence regarding the presence or absence of LGE.

These class labels were used to train, validate, and test the deep binary classifier.

3.4. Performance of binary classifier

After 100 epochs, the trained model showed good performance on both the internal test set of 247 images and the external test set of 247 images (Table 3). The confusion matrices of the classifier's predictions on the internal and external test sets and ROC curves are shown in Figs. 5 and 6, respectively. Overall, the model's performance was better for all evaluation metrics on the internal test set compared to the external test set.

4. Discussion

This study demonstrates proof of concept that a deep learning approach can classify the certainty of LGE being present in a short-axis CMR image with good performance on the internal dataset

Table 3
Performance on the internal and external test sets of 247 images each. (Data are proportions).

Evaluation metric	Internal held-out test set (247 images)	External test set (247 images)
Accuracy	0.94	0.91
Recall (sensitivity)	0.80	0.73
Precision	0.97	0.93
Specificity	0.99	0.98
F1-score	0.87	0.82
ROC AUC	0.94	0.91

ROC AUC area under the receiver operating characteristics curve.

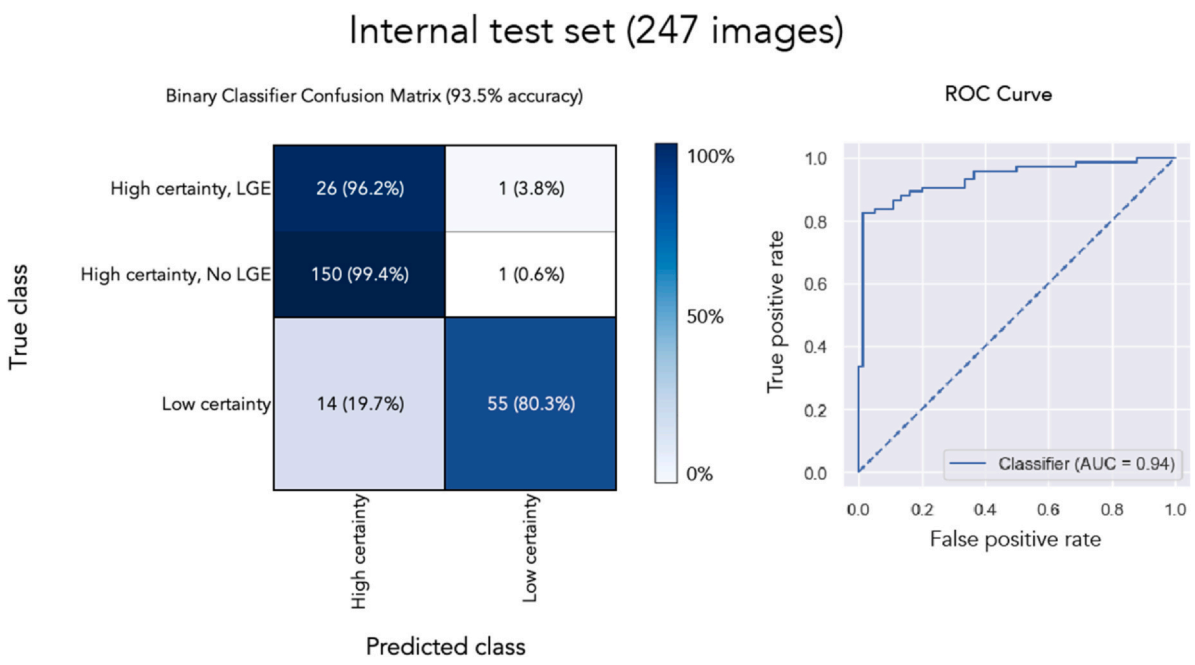

Fig. 5. Confusion matrix and receiver operating characteristics (ROC) curve for the trained model's performance on the internal held-out test set of 247 images. AUC: area under the curve, LGE: late gadolinium enhancement.

(accuracy 94%; F1-score 0.87; ROC AUC 0.94). Performance on external data was also robust (ROC AUC 0.91), indicating a degree of transferability between images from different hospitals. This model's capacity to identify images that are neither clearly normal nor abnormal, and flag them for real-time review during scanning represents a significant step forward in AI-guided decision support for clinical CMR workflows.

The timely provision of this information could facilitate further imaging with optimization or the targeted acquisition of additional (e.g., intersecting) images to reduce the chances that clinicians are left to make a judgment call on equivocal images with an intermediate likelihood of LGE after the patient has left the scanner.

4.1. Impact on diagnostic certainty and comparison with other approaches

Existing systems capable of recognizing ambiguity and hedging from text reports are limited in their application to the post-scan phase of the workflow, rendering them ineffective in generating unequivocal images for reporting [16,17]. When faced with suboptimal images, clinicians will usually try to steer to a diagnosis, but they cannot always be entirely unambiguous in their language when the source of the uncertainty is the images [18,19]. Our approach tackles the upstream problem and demonstrates that, in the majority of cases, uncertainty in LGE likelihood can be detected during the scan while there is still chance to acquire more data that could reduce ambiguity in the

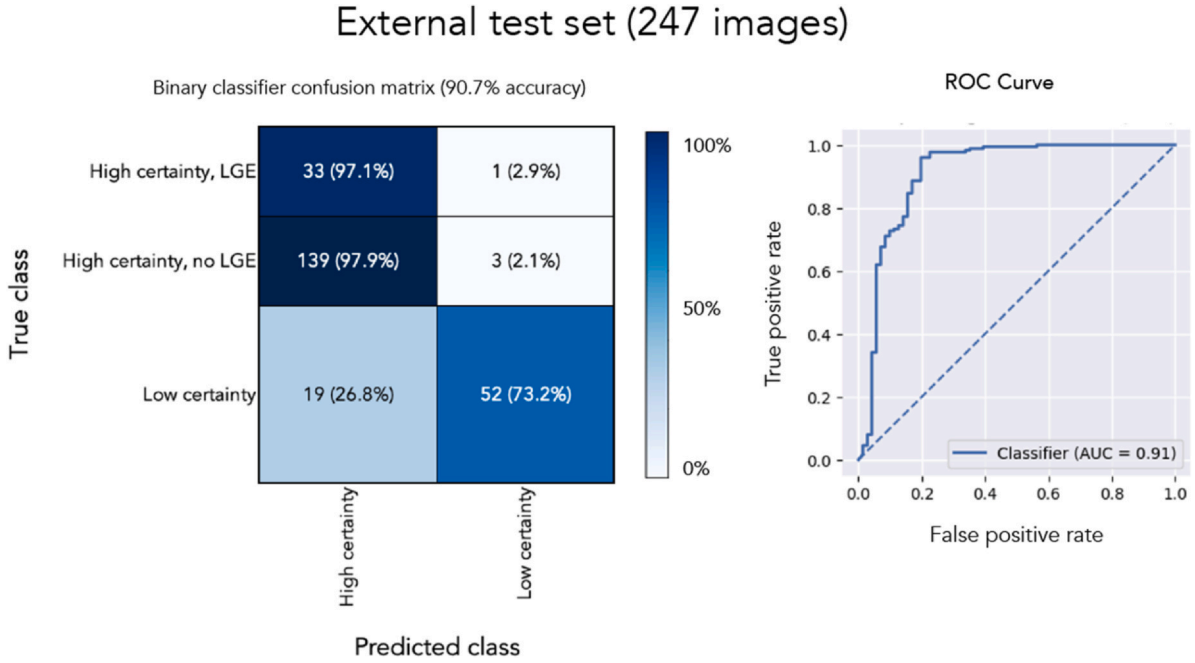

Fig. 6. Confusion matrix and receiver operating characteristics (ROC) curve for the trained model's performance on the external test set of 247 images. AUC: area under the curve, LGE: late gadolinium enhancement.

subsequent report. Interestingly, our model's overall accuracy (94% on internal test set; 91% on external test set) surpassed the inter-rater accuracy (86%). This shows that the AI agreed with the human consensus better than the humans agreed with each other.

Previous work has shown that deep learning can be used to automatically plan CMR standard view images [20]. Deep learning algorithms can also be implemented to analyze CMR images while the scan is ongoing, and their output is useful to guide selection of scan protocols based on initial findings [21]. However, leveraging AI technology for real-time CMR quality assurance has so far been limited to tissue segmentation and automatic measurements rather than assessing LGE likelihood [22,23]. Our algorithm is novel in that it offers the potential to provide real-time quality assurance and decision-support by flagging images in which there is equipoise over whether LGE is present or not, while there is still a window of opportunity to take more images.

4.2. Impact on clinical workflow

Balancing scan duration against the risk of needing to recall the patient for further imaging presents a significant challenge. On one hand, CMR scans are often extensive, frequently producing hundreds of individual images [5] and it is not practical to extend every patient's scan exhaustively. On the other hand, an excessively lean and truncated protocol is likely to risk images in which artifact has not been resolved and risks clinicians having to consider recalling the patient for further images. Both scenarios have impact on efficiency, logistics, and patient experience. An adaptive strategy would potentially deal with this by identifying images with low certainty of LGE or no LGE.

Our model identified the "low certainty" images with 80% sensitivity. Sensitivity was lower on the external images (73%), but the model still detected the majority of "low certainty" cases which would otherwise have passed through to the reporting stage of the clinical workflow. Based on the results of the internal test set, assuming that the patients who had a "low certainty" of LGE could be improved by further imaging on-the-fly, our system has the potential to reduce the proportion of cases with equipoise down to 5%, with 22% of patients receiving further on-the-fly imaging.

We randomly selected 40 studies from our training set, for which at least 1 slice was labeled as "low certainty." For these 40 cases, we reviewed the entire short-axis "stack," the long-axis late gadolinium images, and the cine images. Thirty-one of 40 (77%) remained "low certainty" even after reviewing the other images, while 9/40 (23%) were upgraded to "high certainty". This demonstrates that although reviewing the other routinely protocolled images can resolve a minority of LGE ambiguities; for most cases, additional LGE images are recommended to try to overcome the uncertainty.

From the same random sample of "low-certainty" cases, 5/40 (13%) had an ischemic LGE pattern (compared to 21% in the dataset overall). Twenty-eight of 40 (70%) had a non-ischemic LGE pattern (compared to 46% in the dataset overall). This suggests that experts are usually certain when dealing with infarctions, but the ambiguous situations disproportionally feature non-ischemic LGE patterns.

Entirely eliminating diagnostic uncertainty is not possible. In some cases, no amount of additional imaging will resolve image quality issues, and in this situation, the repeated flags of low certainty could be used to avoid radiographers persevering with low likelihood of success. It would also avoid further rescan attempts where it is unclear whether the images could be improved or not, enabling progression of patient care or exploration of other diagnostic tests without delays caused by repeated attempts at CMR.

From a workflow and efficiency point of view, we favored a model that had high precision to one with high recall. Since most situations are "high certainty," misclassifying these as "low certainty" (potentially extending the scan time with unnecessary further imaging) would have a large cumulative resource burden on the system. This also means that the identification of "low certainty" does not come at the cost of needlessly increasing scan time for the "high certainty" group, which makes up the majority (73%) of cases—the chance of a recall decision being made on a high certainty case is only 1%.

5. Limitations

First, the model is not 100% accurate. As with many clinical classifications, determining the likelihood of LGE depends at least on some subjective judgment. Although the overall inter-observer agreement was good in this study (Spearman's rho = 0.82), this was mainly driven by the two raters having good agreement toward the extremes of the spectrum (low LGE likelihood and high LGE likelihood). Clinicians usually agree when an image is definitely normal (and to a lesser extent) definitely abnormal, but there is a wide spectrum of opinion in between these extremes. This is important because it highlights the extent of disagreement in assessing image quality which has been seen in other modalities [18] and the importance of developing tools to identify and flag suboptimal images objectively.

Second, our model analyzes individual short-axis images, rather than the entire stack. We chose to split short-axis stacks into constituent images to aid the human labeling task. In clinical practice, reporters typically look at the entire stack, which may help reduce uncertainty, and intersecting images, if available. This would be very challenging for labeling data to train a neural network, because the same "stack" can simultaneously have "high certainty" and "low certainty" LGE. Since deep learning requires many labeled data, we chose to split and shuffle individual images to make the human workflow amenable to high-volume labeling. Our future work will focus on analyzing the short-axis stack in its entirety, in combination and other intersecting views from the same scan in concert. For example, a model that can pool the confidence of multiple slices, such as a 2D convolutional neural network, which uses neighboring slices and/or long-axis images as inputs would be the next step for our concept.

Third, our work identifies images leading to equivocal reporting regardless of the reason for the poor images and does not suggest to users how to rectify it. The solution for improving the images is likely to depend on the cause of the problem. Since expert clinician data labeling is a finite resource, and network performance is broadly proportional to the amount of labeled data, in this proof-of-concept study we chose not to ask expert labelers additionally for the "reason for uncertainty." Instead, we opted for a lean labeling workflow by distilling the complex, multi-factorial qualitative assessment of LGE into a single metric: "LGE likelihood." Future work should aim toward training a system to identify different causes of poor image quality, with decision-support for possible rectifying actions, but this requires larger amounts of data with specific artifact types.

Finally, as with all deep learning models, it is possible that our findings may not generalize to other data due to "overfitting" [24]. To mitigate this, we report performance on a held-out test set that was not shown to the model until after training and validation were complete. We also demonstrate that 97% of the model's accuracy on the internal images is sustained on an external set of images obtained from a different scanner at a different center performed at a different time. This indicates the robustness of our results and their potential to be transferrable between different clinical settings.

6. Conclusions

Deep learning has potential to automate quality assurance of late gadolinium imaging in CMR. Real-time identification of short-axis images with an intermediate likelihood of LGE (low diagnostic certainty) could provide decision support to trigger further imaging before the patient has left the scanner. Our model shows good performance on detecting "low certainty" situations, while preserving a high specificity for "high certainty" situations so that only the cases that will benefit from additional imaging are flagged. This performance was largely

retained on external images from a different center. This innovation could help clinicians overcome the challenge of reporting equivocal scans and reduce ambiguous communication between reporters and referrers.

Funding

S.Z., K.V., and D.C. are supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1]. J.P.H. is supported by The British Heart Foundation [FS/ICRF/22/26039]. M.V. is supported by The British Heart Foundation Centre of Research Excellence at Imperial College London [RE/18/4/34215]. R.H.D. is funded by the British Heart Foundation (BHF) Accelerator Award (AA/18/6/34223).

Author contributions

Anil A. Bharath: Writing – original draft, Validation, Supervision, Project administration, Writing – review and editing, Formal analysis, Investigation, Conceptualization, Methodology. **Kristopher D. Knott:** Writing – review and editing, Investigation, Data curation. **Hunain Shiwani:** Writing – review and editing, Methodology, Data curation. **Nicholas S. Peters:** Writing – review and editing, Investigation, Methodology, Visualization. **Marta Varela:** Formal analysis, Methodology, Validation, Writing – review and editing. **James P. Howard:** Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization, Visualization, Writing – original draft, Writing – review and editing. **Digby Chappell:** Writing – review and editing, Visualization, Validation, Software, Methodology, Formal analysis. **Graham D. Cole:** Writing – review and editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kavitha Vimalesvaran:** Writing – review and editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Darrel P. Francis:** Writing – review and editing, Supervision, Conceptualization, Data curation, Investigation, Methodology, Project administration, Resources, Software. **Sameer Zaman:** Writing – review and editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Nick W.F. Linton:** Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing – review and editing. **James C. Moon:** Methodology, Writing – review and editing, Data curation. **Rhodri H. Davies:** Data curation, Methodology, Writing – review and editing.

Ethics approval and consent

Ethical approval was gained from the Health Regulatory Agency (Integrated Research Application System identifier 243023).

Consent for publication

NA.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

NA.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jocmr.2024.101040.

References

[1] Becker MAJ, Cornel JH, van de Ven PM, van Rossum AC, Allaart CP, Germans T. The prognostic value of late gadolinium-enhanced cardiac magnetic resonance imaging in nonischemic dilated cardiomyopathy: a review and meta-analysis. JACC Cardiovasc Imaging 2018;11(9):1274–84.
[2] Green JJ, Berger JS, Kramer CM, Salerno M. Prognostic value of late gadolinium enhancement in clinical outcomes for hypertrophic cardiomyopathy. JACC Cardiovasc Imaging 2012;5(4):370–7.
[3] Aquaro GD, Ghebru HY, Camastra G, Monti L, Dellegrottaglie S, Moro C, et al. Prognostic value of repeating cardiac magnetic resonance in patients with acute myocarditis. J Am Coll Cardiol 2019;74(20):2439–48.
[4] Chopra H, Arangalage D, Bouleti C, Zarka S, Fayard F, Chillon S, et al. Prognostic value of the infarct- and non-infarct like patterns and cardiovascular magnetic resonance parameters on long-term outcome of patients after acute myocarditis. Int J Cardiol 2016;212:63–9.
[5] Schulz-Menger J, Bluemke DA, Bremerich J, Flamm SD, Fogel MA, Friedrich MG, et al. Standardized image interpretation and post processing in cardiovascular magnetic resonance: Society for Cardiovascular Magnetic Resonance (SCMR) board of trustees task force on standardized post processing. J Cardiovasc Magn Reson 2013;15(1):35.
[6] Georgiopoulos G, Figliozzi S, Sanguineti F, Aquaro GD, di Bella G, Stamatelopoulos K, et al. Prognostic impact of late gadolinium enhancement by cardiovascular magnetic resonance in myocarditis: a systematic review and meta-analysis. Circ Cardiovasc Imaging 2021;14(1):e011492.
[7] Flett AS, Hasleton J, Cook C, Hausenloy D, Quarta G, Ariti C, et al. Evaluation of techniques for the quantification of myocardial scar of differing etiology using cardiac magnetic resonance. JACC Cardiovasc Imaging 2011;4(2):150–6.
[8] Gräni C, Eichhorn C, Bière L, Kaneko K, Murthy VL, Agarwal V, et al. Comparison of myocardial fibrosis quantification methods by cardiovascular magnetic resonance imaging for risk stratification of patients with suspected myocarditis. J Cardiovasc Magn Reson 2019;21(1):14.
[9] Callen AL, Dupont SM, Price A, Laguna B, McCoy D, Do B, et al. Between always and never: evaluating uncertainty in radiology reports using natural language processing. J Digit Imaging 2020;33(5):1194–201.
[10] Makhnevich A, Sinvani L, Cohen SL, Feldhamer KH, Zhang M, Lesser ML, et al. The clinical utility of chest radiography for identifying pneumonia: accounting for diagnostic uncertainty in radiology reports. AJR Am J Roentgenol 2019;213(6):1207–12.
[11] van der Velde N, Hassing HC, Bakker BJ, Wielopolski PA, Lebel RM, Janich MA, et al. Improvement of late gadolinium enhancement image quality using a deep learning-based reconstruction algorithm and its influence on myocardial scar quantification. Eur Radio 2021;31(6):3846–55.
[12] Oksuz I, Clough J, Bustin A, Cruz G, Prieto C, Botnar R, et al. Cardiac MR motion artefact correction from k-space using deep learning-based reconstruction. In: Knoll F, Maier A, Rueckert D, editors. Machine Learning for Medical Image Reconstruction. Lecture Notes in Computer Science Cham: Springer International Publishing; 2018. p. 21–9.
[13] Howard JP, Stowell CC, Cole GD, Ananthan K, Demetrescu CD, Pearce K, et al. Automated left ventricular dimension assessment using artificial intelligence developed and validated by a UK-wide collaborative. Circ: Cardiovasc Imaging 2021;14(5):e011951.
[14] Tan M, Le Q. EfficientNetV2: smaller models and faster training. In: Proceedings of the 38th International Conference on Machine Learning, PMLR; 2021. p. 10096–10106 (cited October 24, 2022). Accessed: October 1, 2022. https://proceedings.mlr.press/v139/tan21a.html.
[15] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016. p. 770–778 (cited October 24, 2022). Accessed: October 1, 2022 https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.
[16] Lacson R, Odigie E, Wang A, Kapoor N, Shinagare A, Boland G, et al. Multivariate analysis of radiologists' usage of phrases that convey diagnostic certainty. Acad Radio 2019;26(9):1229–34.
[17] Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Jt Summits Transl Sci Proc 2018;2017:188–96.
[18] Cole GD, Dhutia NM, Shun-Shin MJ, Willson K, Harrison J, Raphael CE, et al. Defining the real-world reproducibility of visual grading of left ventricular function and visual estimation of left ventricular ejection fraction: impact of image quality, experience and accreditation. Int J Cardiovasc Imaging 2015;31(7):1303–14.
[19] Mabotuwana T, Bhandarkar VS, Hall CS, Gunn ML. Detecting technical image quality in radiology reports. AMIA Annu Symp Proc 2018;2018:780–8.
[20] Alansary A, Folgoc LL, Vaillant G, Oktay O, Li Y, Bai W, et al. Automatic view planning with multi-scale deep reinforcement learning agents. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Cham: Springer International Publishing; 2018. p. 277–85.

[21] Howard JP, Zaman S, Ragavan A, Hall K, Leonard G, Sutanto S, et al. Automated analysis and detection of abnormalities in transaxial anatomical cardiovascular magnetic resonance images: a proof of concept study with potential to optimize image acquisition. Int J Cardiovasc Imaging 2020;37(3):1033–42. https://doi.org/10.1007/s10554-020-02050-w.

[22] Ruijsink B, Puyol-Antón E, Oksuz I, Sinclair M, Bai W, Schnabel JA, et al. Fully automated, quality-controlled cardiac analysis from CMR: validation and large-scale application to characterize cardiac function. JACC Cardiovasc Imaging 2019;13(3):684–95.

[23] Puyol-Antón E, Ruijsink B, Baumgartner CF, Masci PG, Sinclair M, Konukoglu E, et al. Automated quantification of myocardial tissue characteristics from native T1 mapping using neural networks with uncertainty-based quality-control. J Cardiovasc Magn Reson 2020;22(1):60.

[24] Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. npj Digit Med 2019;2(1):1–3.