ESC
European Society
of Cardiology

European Heart Journal - Imaging Methods and Practice (2024) **2**, qyae094
https://doi.org/10.1093/ehjimp/qyae094

ORIGINAL ARTICLE

Quality control of cardiac magnetic resonance imaging segmentation, feature tracking, aortic flow, and native T1 analysis using automated batch processing in the UK Biobank study

Sucharitha Chadalavada [1,2,†], Elisa Rauseo[1,2,3,†], Ahmed Salih[1,4], Hafiz Naderi [1,2], Mohammed Khanji[1,2], Jose D. Vargas[5], Aaron M. Lee [1], Alborz Amir-Kalili[6], Lisette Lockhart[6], Ben Graham[6], Mihaela Chirvasa[6], Kenneth Fung [1,2], Jose Paiva[1], Mihir M. Sanghvi [1,2], Gregory G. Slabaugh[3,7,8], Magnus T. Jensen[1,9], Nay Aung [1,2,3,‡], and Steffen E. Petersen[1,2,6,8,10,11,*‡]

[1]William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University of London, Charterhouse Square, London, UK
[2]Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, West Smithfield, London, UK
[3]Digital Environment Research Institute, Queen Mary University of London, London, UK
[4]Department of Computer Science, University of Zakho, Zakho, Kurdistan of Iraq, Iraq
[5]Department of Cardiology, US Department of Veterans Affair Medical Center, Washington, D.C., USA
[6]Circle Cardiovascular Imaging Inc., Calgary, Canada
[7]School of Electronic Eng. & Computer Science, Queen Mary University of London, UK
[8]Alan Turing Institute, The British Library, John Dodson House, London, UK
[9]Steno Diabetes Center Copenhagen, Cardiometabolic Disease Department, Borgmester Ib Juuls Vej 83, 2730 Herlev, Denmark
[10]Health Data Research UK, London, UK
[11]National Institute for Health and Care Research, UK

Received 7 June 2024; accepted after revision 30 August 2024; online publish-ahead-of-print 16 September 2024

Abstract

Aims Automated algorithms are regularly used to analyse cardiac magnetic resonance (CMR) images. Validating data output reliability from this method is crucial for enabling widespread adoption. We outline a visual quality control (VQC) process for image analysis using automated batch processing. We assess the performance of automated analysis and the reliability of replacing visual checks with statistical outlier (SO) removal approach in UK Biobank CMR scans.

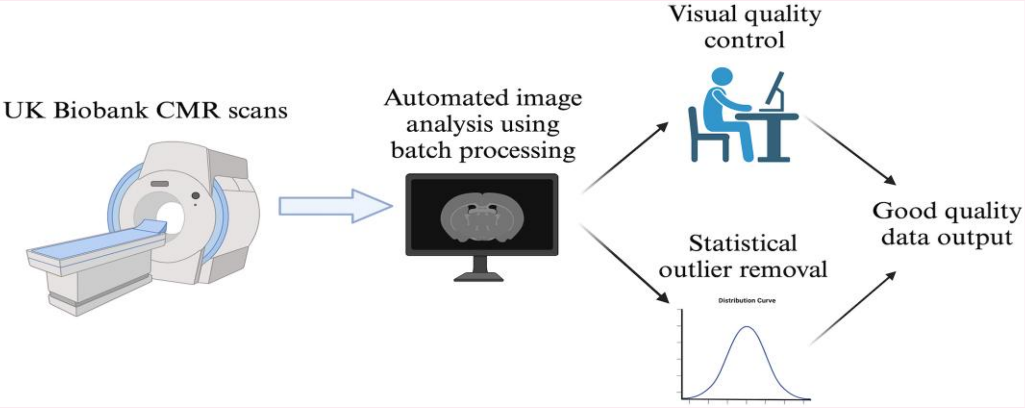
Methods and results We included 1987 CMR scans from the UK Biobank COVID-19 imaging study. We used batch processing software (Circle Cardiovascular Imaging Inc.—CVI42) to automatically extract chamber volumetric data, strain, native T1, and aortic flow data. The automated analysis outputs (~62 000 videos and 2000 images) were visually checked by six experienced clinicians using a standardized approach and a custom-built R Shiny app. Inter-observer variability was assessed. Data from scans passing VQC were compared with a SO removal QC method in a subset of healthy individuals ($n = 1069$). Automated segmentation was highly rated, with over 95% of scans passing VQC. Overall inter-observer agreement was very good (Gwet's AC2 0.91; 95% confidence interval 0.84, 0.94). No difference in overall data derived from VQC or SO removal in healthy individuals was observed.

* Corresponding author. E-mail: s.e.petersen@qmul.ac.uk
† These authors contributed equally and share first authorship.
‡ These authors contributed equally and share senior authorship.

Conclusion Automated image analysis using CVI42 prototypes for UK Biobank CMR scans demonstrated high quality. Larger UK Biobank data sets analysed using these automated algorithms do not require in-depth VQC. SO removal is sufficient as a QC measure, with operator discretion for visual checks based on population or research objectives.

Lay summary Cardiac magnetic resonance (CMR) imaging is often considered the gold standard in imaging the heart and is being used more clinically and especially for research. However, the analysis can be time-consuming. The image analysis can be automated, but the quality of this has often been questioned. In this study, we assess the quality of automated techniques used to analyse CMR scans from the UK Biobank study with experts visually checking the automated analysis. We have shown that automated image analysis can be effective and produce good-quality data for large-scale research. We have also shown that, for UK Biobank scans, visual check may not be necessary, which saves time and resources and improves cardiovascular research.

Graphical Abstract



Created using Biorendr.com.

Keywords cardiac magnetic resonance imaging • automated image analysis • quality control • Shiny app • machine learning

Introduction

Cardiac magnetic resonance (CMR) imaging is considered the reference standard for assessing cardiac structure and function.[1,2] Manual image segmentation is essential in deriving several quantitative measurements and is time-consuming and prone to error.[3] Automated segmentation based on machine learning (ML) algorithms can improve the clinical workflow and increase the reproducibility of measurements by reducing intra- and inter-observer variability.[4,5] However, quality assessment of the outputs generated by ML algorithms remains an essential step before they are deployed in clinical practice.

The quality of automated segmentation is generally assessed against a ground-truth reference, typically represented by manual expert annotations not always available in large data sets. Furthermore, the quantitative metrics used to evaluate the segmentation algorithms do not necessarily correlate with clinical acceptability.[6,7] A quality assessment involving clinicians is thus essential to ensure the reliability and trustworthiness of ML outputs in clinical workflows. However, visual quality control (VQC) is a tedious, time-consuming, and often subjective process that relies on clinicians inspecting images and assigning scores. Furthermore, it requires multiple well-trained people to perform VQC, which is not always practical, especially for large data sets like the UK Biobank. Alternative automated QC techniques that do not require large, fully annotated data sets have been proposed in the literature, but these approaches do not guarantee the generalizability of the algorithm in unseen data.[8,9]

As current automated methods have become more accurate and successfully used for assessing several heart diseases,[10–12] we hypothesized that it may be possible to visually QC only a subset of output data to assess the clinical acceptability and reliability of the ML algorithm. Here, we describe a QC process applied to a subset of UK Biobank participants enrolled in the COVID-19 study in which CMR images were analysed using automated batch processing algorithms. First, a team of six clinicians with extensive experience in interpreting and analysing CMR images visually assessed the automated outputs by assigning a quality score. Then, the automated measurements of cardiac function and structure related to the images that passed the QC were compared against those that were selected using a statistical outlier (SO) removal, a common method used in the UK Biobank imaging studies. That was done to assess to what extent the ML algorithm could generate reliable data without incurring time and resource-consuming VQC.

Methods

Study population/data set

UK Biobank was a prospective study of half a million volunteers from the general population aged between 40 and 69, who were recruited between 2006 and 2010[13] (www.ukbiobank.ac.uk). The UK Biobank imaging study aims to collect brain, heart, and abdomen scans from 100 000 participants, with 70 000 participants scanned so far.[14] The CMR scanning protocol using a 1.5 T scanner (MAGNETOM Aera, Syngo Platform VD13A, Siemens

Healthcare, Erlangen, Germany) has been described in detail elsewhere.[15] Approximately 50 000 participants were scanned prior to the COVID-19 pandemic.[16] Around 2000 of these participants, some of whom were infected with COVID-19, and others who were not, were invited back for repeat imaging.[16,17]

Our research group analysed the CMR scans of this COVID-19 UK Biobank sub-study to compare the cardiovascular changes between the baseline imaging visit and repeat imaging visit for those that were infected with COVID-19 and the control group that were not. The characteristics of the 1069 participants whose scans were included in this study are shown in *Table 1*, with their definitions detailed in Supplementary data online, *Table S1*.

This paper outlines the automated methods used for large-scale image analysis and the QC process to ensure the accuracy and reliability of these automated methods. *Figure 1* outlines in detail the timeline of image analysis and QC steps.

CMR image analysis: batch processing

CVI42 software (Circle Cardiovascular Imaging, Inc., prototype 5.14.1.2875) was used to perform an automated batch ML analysis of CMR scans included in the COVID-19 study to derive volumes of cardiac chambers, aortic flow analysis, and native T1. CVI42 (Circle Cardiovascular Imaging, Inc., prototype 5.13.7I) was used to perform an automated batch feature tracking analysis of the same cases to derive myocardial strain. The settings used for each image analysis are detailed below.

Batch processing is a semi-automated process involved preparing ∼300 scans (the maximum the software could execute each time), to be analysed by the automated algorithm imbedded in the new CVI 42 prototypes for the respective type of image analysis. The correct settings had to be selected (detailed below), which was discussed and agreed with experts within the team. The correct destination file paths had to be selected to import each 'batch' and for the data output to be safely recorded. All of the automated image analysis generated videos or images of the automated analysis that were used to assess quality.

The resolution of the videos and images was limited to 1024 × 1024 pixels due to storage space, given the large data set. A total of 1987 scans (1069 baseline scans, 918 repeat scans) were analysed and included in the VQC process detailed in this paper.

Volumes of cardiac chambers

The whole of the short axis (SAX) stack was chosen to be segmented, where, in the basal slices, only closed contours (i.e. complete circles of both endocardium and epicardium are visible) were included for volume calculation. Simplified endocardial contours without papillary muscle detection [papillary muscles are not identified and therefore not included in the left ventricular (LV) mass calculation] was the protocol for this analysis.

LV endocardial and epicardial contours were placed both in SAX and long axis (LAX) cine images (two-chamber, three-chamber, and four-chamber). Right ventricle (RV) endocardial and epicardial were traced from the SAX and four-chamber cine CMR views using the automated segmentation tool. Left and right atrial (LA and RA) contours were automatically drawn in four-chamber views, and the LA area was also segmented in two-chamber cine CMR view.

Table 1 Baseline participants' characteristics

	Participants ($n = 1064$)
Age, years (mean ± SD)	59.37 ± 7.23
Males (*n*, %)	478 (44.92%)
Ethnicity—Caucasian (*n*, %)	1028 (96.61%)
Ethnicity—other (*n*, %)	36 (3.38%)
Current smoking (*n*, %)	44 (4.12%)
BMI (median + IQR)	26.3 (23.5, 29.3)
Prevalent diabetes (*n*, %)	44 (4.13%)
Prevalent hypertension (*n*, %)	224 (21.05%)

Aortic flow

The aortic flow was automatically measured using the aortic valve phase contrast image after having defined the inner wall of the aorta in the systolic phase (ROI). A ML algorithm was used to segment the cross-sectional image of the aorta for all images in the sequence independently. Background correction and anti-aliasing background correction was not required due to the high quality of images in UK Biobank.

Native T1

CMR native T1 maps were acquired at the mid-ventricular SAX view using the ShMOLLI T1 mapping sequence. A ML algorithm was used to segment the endocardial and epicardial borders into six segments of the native T1 image, which included a 10% offset on both endocardium and epicardium contours. The contoured results are saved as a greyscale image to aid VQC of the algorithm-generated epicardial and endocardial borders, offsets, and myocardial segments.

Myocardial strain using feature tracking

The protocol applied temporal smoothing and used LV diastole and systole to guide tracking. LV diastole was set as the reference phase. Any SAX images where there are open LV contours (missing endocardium or epicardium contours) were excluded from the feature tracking. Contours were first drawn with SAX stack segmented only in LV end-diastole and LV end-systole. LAX images were also segmented.

Visualization of CMR images and videos using a Shiny app

We built a custom Shiny[18] app to perform the visual evaluation of segmentation videos and still images produced by the automated batch processing pipeline. Shiny is an R package designed for the development of interactive web applications and data visualization dashboards. A typical Shiny app consists of a user interface script that controls the layout and appearance of the app and a server script that details the information to build the objects shown in the user interface. The Shiny framework also allows customization by using HTML, CSS, or JavaScript. In this project, we leveraged the ShinyProxy tool[19] which permits user authentication, wrapped in a Docker[20] container to deploy our Shiny app for multiple readers. *Figure 2* shows the user-friendly interface of the custom-built Shiny app.

VQC process

A detailed description of the criteria used to assess the quality and assigning a score of each image analysis modality is provided in *Table 2*.

The quality of the automatically derived CMR measurements was assessed by a group of six experienced clinicians in reading CMR (E.R., H.N., J.V., M.Y.K., S.C., and S.E.P.) including three consultant cardiologists with level 3 certification in CMR (M.Y.K., J.V., and S.E.P.) and three well-trained clinical research fellows (E.R., H.N., and S.C.) with >3 years' experience in analysing CMR. The evaluators were asked to visualize the images and assign quality scores ranging from 1 to 3 ('good', 'satisfactory', 'poor').

Regarding contouring of cardiac chambers, a score of 1 ('good') was assigned to good-quality images, with contouring considered of excellent/good quality with no manual changes required. A score of 2 ('satisfactory') was given to contouring with minor or negligible inaccuracies that if the scan was being interpreted in a clinical setting would not affect the clinical decisions. A score of 3 ('poor') was assigned to either absent contouring, poor-quality images, or gross inaccuracies of the algorithm to segment the cardiac structures significant enough to be considered unacceptable if the scan was interpreted in a clinical setting.

Similar to previous standard operating procedures for contouring analysis, the quality of T1 maps was defined as 1 ('good' quality, without motion artefacts), 2 ('satisfactory': images are suboptimal but still analysable), and 3 ('poor' image quality that makes it not analysable, no image available).

The quality of aortic flow images was judged as 1 ('good': images acquired at the correct slice, of good quality, and free of artefacts), 2 ('satisfactory': suboptimal images, with minor issues and artefacts but still analysable), and

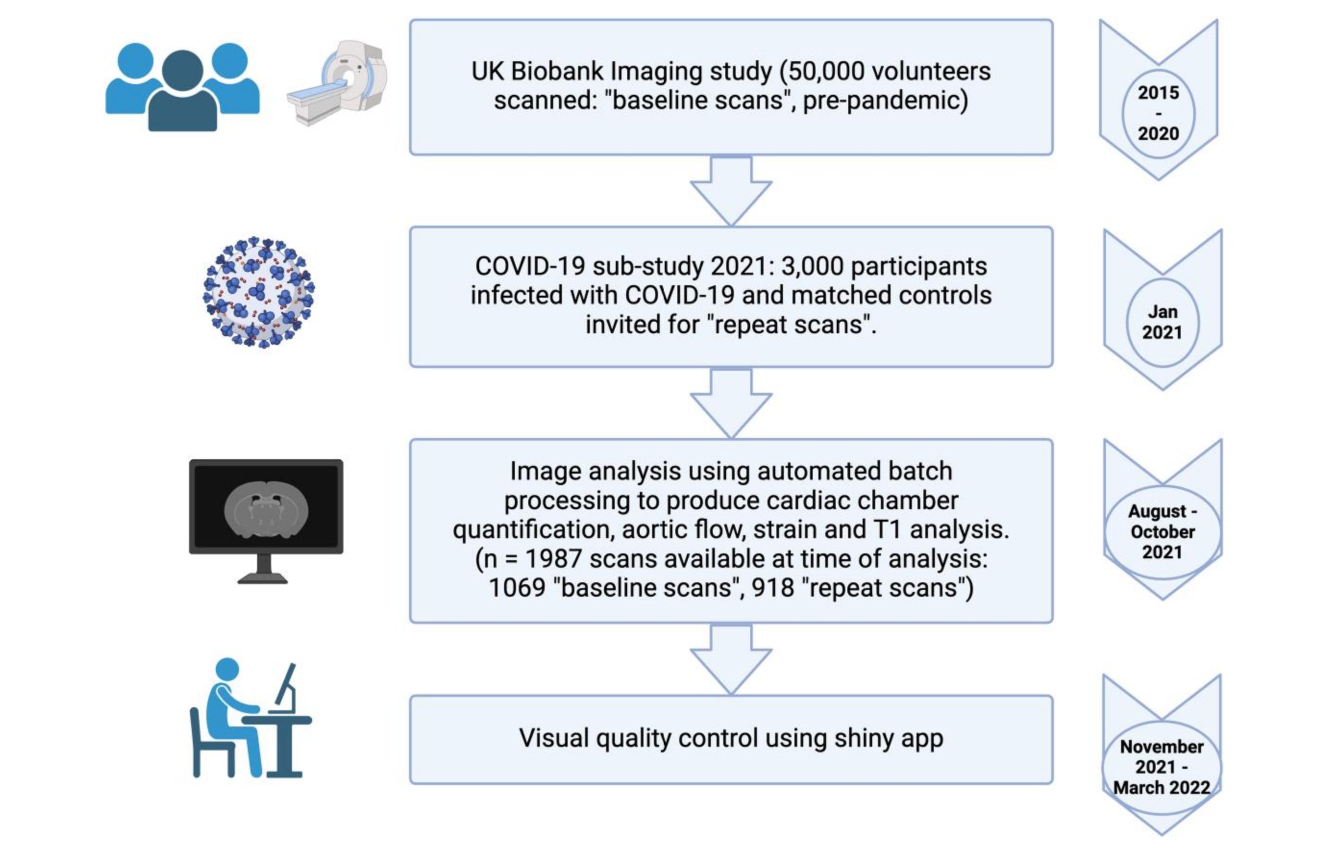

UK Biobank Imaging study (50,000 volunteers scanned: "baseline scans", pre-pandemic)
2015 - 2020
COVID-19 sub-study 2021: 3,000 participants infected with COVID-19 and matched controls invited for "repeat scans".
Jan 2021
Image analysis using automated batch processing to produce cardiac chamber quantification, aortic flow, strain and T1 analysis. (n = 1987 scans available at time of analysis: 1069 "baseline scans", 918 "repeat scans")
August - October 2021
Visual quality control using shiny app
November 2021 - March 2022

Figure 1 Flow chart detailing timeline of CMR image analysis and quality control process. Created using Biorendr.com.

3 ('poor' quality due to incorrect slice location, extreme imaging artefacts, aliasing in the aortic flow, or missing images).

A detailed description of the criteria used to assess the quality and assigning a score of each image analysis modality is provided in *Table 2*.

The QC process comprised three distinct phases. The first QC phase was performed on a sample of 20 CMR cases. Six readers were asked to independently assess the same cases using the quality scoring system previously described. The analysis was followed by a group discussion to assess the inter-observer variability. The discussion also included a review of CMR studies where there was the largest variability in quality scores measured by inter-reader agreement statistics to establish a consensus on the appropriate approach for future such cases.

The second phase of QC included 50 different CMR cases assigned to each observer, followed by another round of group discussion. The observers reviewed challenging cases together, and a consensus was reached to establish a uniform approach to the VQC process.

The third and final phase of the QC was performed on the remaining subset of cases ($n = 1987$) distributed among all the observers, who were asked to score them independently.

Comparing clinical measures generated from VQC vs. SO removal

We assessed the CMR measures of cardiac structure and function, directly calculated from the automated image segmentations, in a subset of participants ($n = 1069$, using only baseline scans) through two QC methods: VQC and SO removal. Specifically, in this analysis, we assessed the following clinical measures: LV and RV volumes, LV mass, LV ejection fraction (LVEF), LV global longitudinal strain (GLS), aortic flow, and native T1.

The output data from the VQC data set, derived from scans rated as either 'good' or 'satisfactory', were compared with the data from the SO data set, which were generated by removing outliers three times the interquartile range below the first quartile and above the third quartile. That was done to assess the impact of the QC methods on the clinical measurements. Participants with documented cardiovascular disease were excluded from this analysis. The UK Biobank field IDs used to identify and define those with pre-existing cardiovascular disease can be found in Supplementary data online, *Table S2*.

Statistical analysis

Continuous variables are presented as proportions and means ± standard deviations. Gwet's second-order agreement coefficient with ordinal weighting applied (AC2) values was used to assess inter-observer scoring variability.[21] This method has been validated for use when assessing reliability between multiple operators.[7,21,22] The coefficient value, 95% confidence interval (CI), and associated P values were generated. An agreement coefficient value between 0.61 and 0.80 has been used as an acceptable benchmark in previous literature and by Gwet.[7,23] Therefore, we have adopted the same limits for this paper. Two-sided t-tests were used to determine the significance of the differences in the CMR measurements derived from the VQC and SO data sets for both male and female participants. A $P < 0.05$ was considered statistically significant for all analyses. Statistical analysis was performed using Python Version 3.6.4 (Python Software Foundation, DE, USA).

Results

Inter-observer agreement

The first phase of the QC process was conducted on a subset of studies (20 scans) equally distributed across the six readers to assess the inter-observer quality score variability using Gwet's AC2 agreement coefficient with ordinal weighting. The inter-observer agreement for each

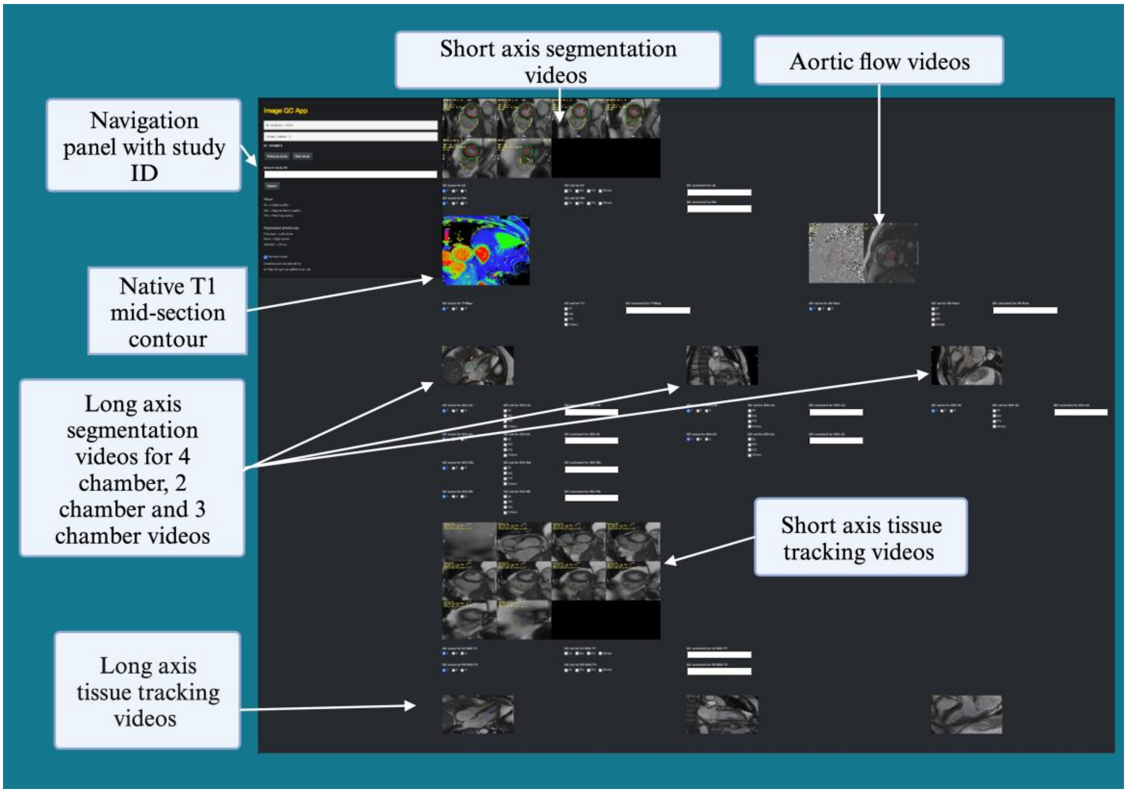

Short axis segmentation videos
Aortic flow videos
Navigation panel with study ID
Image QC App
Native T1 mid-section contour
Long axis segmentation videos for 4 chamber, 2 chamber and 3 chamber videos
Short axis tissue tracking videos
Long axis tissue tracking videos

Figure 2 Custom-built shiny app user-friendly interface and layout. Created using Biorendr.com.

image sequence is shown in *Figure 3* with the more detailed results seen in Supplementary data online, *Table S3*.

There was substantial inter-observer agreement in quality scoring each image sequence with an overall AC2 score of 0.91 (95% CI 0.84, 0.94), which exceeds the benchmark suggested in the literature.[21] However, the score agreement was lower for the RV contours in SAX images (AC2 = 0.53, 95% CI 0.42–0.64), and for RV tissue tracking (AC2 = 0.57, 95% CI 0.44–0.69).

VQC results

Contouring of cardiac chambers, native T1 maps, and aortic flow images

Of the 1987 studies selected, we removed those with missing images or clips across each CMR sequence (the highest number of missingness was observed for native T1 maps and aortic flow sequences), and the remaining studies were left to be visually quality-controlled. Of these, the proportions of exams judged of good/acceptable quality (score 1 or 2) were around 99.7%, 99.8%, 98.3%, and 99.8%, for LV, RV, LA, and RA contouring, respectively. Furthermore, the score assigned to each contouring was consistent across all the image sequences (*Table 3*). Similar results were observed for T1 maps and aortic flow analysis, where the proportions of CMR scans scored 1 or 2 were 98.9% and 99.7%, respectively (*Table 3*).

The CMR image sequence judged as of poor quality (score 3) had either poor imaging quality (e.g. artefacts, aliasing for aortic flow sequence) or plane quality issues (e.g. images foreshortened) or incomplete/failed segmentation (see Supplementary data online, *Table S4*).

Strain analysis

Scans from the COVID-19 cohort ($n = 1987$) were visually quality-checked, 1069 of these scans were baseline scans and 918 were repeat scans. The results can be seen in *Table 4*. The scores given during the visual quality check for the automated image analysis to produce strain data were examined to determine the overall quality of the automated batch image analysis using this software.

LAX image feature tracking and RV SAX were >96% of the time graded as being either excellent or good (grade 1 or 2, respectively). The LV-SAX tissue tracking score was comparatively lower at 90% for being either good or satisfactory.

We examined the score and comments for the 1982 scans where LV SAX tissue tracking was available to further understand why the LV-SAX tissue tracking image sequence had comparatively a lower proportion graded as good or satisfactory or poor. Given the score of ‘good’ or ‘satisfactory’ (1 or 2) were 90.2% (1788 scans) and 9.8% (194 scans) were given the score of poor.[3] The QC comments revealed that there was only partial tracking of LV in some slices in 24 of these scans. The comments also revealed that 170 scans were graded as poor quality[3] due to missing any feature tracking markers on the LV in SAX stack images. Lack of feature tracking on the LV in SAX stack images would result in lack of data output for global circumferential strain (GCS) and global radial strain (GRS) as these are derived from LV myocardial tracking in the SAX images. Therefore, the grade of poor quality[3] in this case would be redundant since there is no clinical output data derived from these images.

We examined the data output that revealed that 64 of the scans graded as poor quality and missing LV tracking in SAX stack did not have any output for GCS and GRS, as predicted, making the score of poor quality redundant. Therefore, the total number of scans that is

Table 2 Criteria to assess the quality of each image analysis modality

CMR image sequence	QC score 1 (good)	QC score 2 (satisfactory)	QC score 3 (poor)
Contours of cardiac chambers (~15 videos per scan) Points to consider when assessing quality: • Accuracy of contouring. • Level of operator corrections required to make contours clinically acceptable.			
SAX (RV)	Contours considered of good quality with no operator corrections needed.	Fair or clinically acceptable contours with minor or negligible inaccuracies in the segmentation considered not clinically relevant.	Significantly inaccurate segmentation or requiring significant manual changes to make them clinically acceptable. If an LV or LA severely foreshortened due to planning quality.
SAX (LV)		If an LV or LA somewhat foreshortened due to planning quality.	
4ch (RV)			
4ch (LV)			
2ch (LV)			
3ch (LV)			
4ch (RA)			
4ch (LA)			
Strain (~ 15 videos per scan) Points to consider when assessing quality: • Are the contours for end-systole and end-diastole appropriate? • Are the tracking markers and the 'quivers' (a line which spreads out from the feature dot that is being tracked) appropriately placed in relation to the contours and the 'quivers' aligned in the appropriate direction. • Look at the 'extent' tool—the blue line going from the base to the apex. Is this correct? If the blue line is going outside of the heart, then the data is unlikely to be accurate. • If all three long axis images are missing, then this will produce very unreliable strain data and should be classed as grade '3'. • If >3 points are not tracking appropriately, then the quality of the data will be poor • Between 1 and 3, need to look, and if it is related to papillary muscle, then this is inappropriate tracking from the software but could be graded as quality grade 2 if only one tracking marker associated with papillary muscle. • The protocol that we have selected should exclude any slices where the blood pool is not visible (apical) or where there are open contours (basal slices) • If RA is tracked when tracking the RV then this is a grade 2 and need to enter in free text 'RA tracked'.			
SAX	Perfect or near perfect	Acceptable (clinically we would include)	Completely inaccurate contours and border points
2ch	Perfect or near perfect	Extent tool (blue line) slightly inaccurate but not outside of the heart	Extent tool (blue line) outside the heart
3ch	Perfect or near perfect	Look at extent tool (blue line) in this image also.	Two or more strain markers tracking papillary muscle & extent tool outside.
4ch	Perfect or near perfect	1–3 markers inappropriate tracking. Slightly inaccurate contours.	Two or more strain markers tracking papillary muscle
Native T1 (1 image) Points to consider: • Appropriate contours • Appropriate colour maps • Artefacts in the ROI.			
SAX	Perfect or near perfect	Some irregularities but would be acceptable. If an artefact is in the ROI but can still get an accurate value if manually drawn omitting artefact.	Incorrect contour, colour map/incompatible artefact in ROI.
Aortic valve flow (1 video)			
Aortic valve phase contrast image	Perfect or near perfect contour, no aliasing	Contour slightly inaccurate but acceptable	Incorrect contour and/or aliasing present.

CMR, cardiac magnetic resonance; QC, quality control; SAX, short axis; LV, left ventricle; RV, right ventricle; RA, right atrium; LA, left atrium; 2ch, 2 chamber; 4ch, 4 chamber; ROI, region of interest.

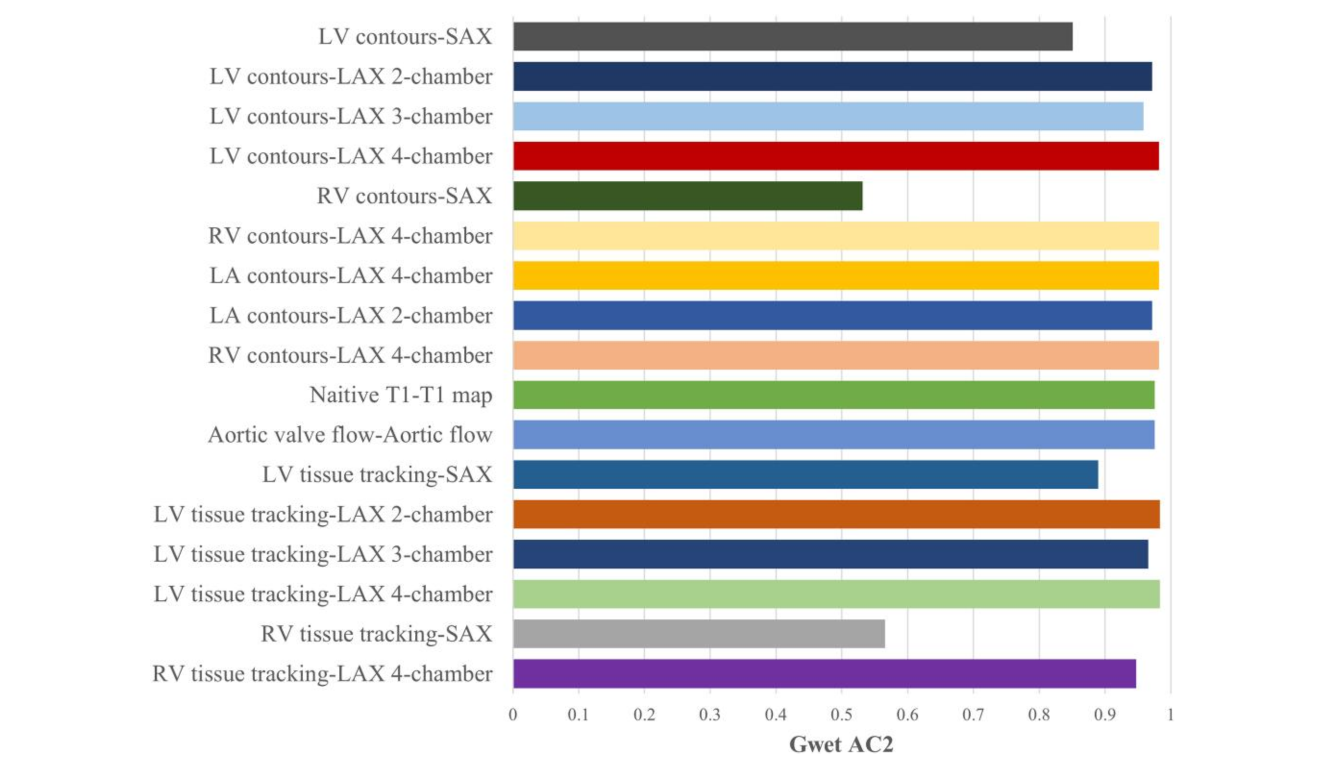


Figure 3 Coefficient (Gwet AC2) by image sequence of scans visually checked by six different operators. The *P* value for all the results shown was <0.05, and the detailed results are shown in Supplementary data online, *Table S3*. Created using Biorendr.com. LV, Left ventricle; SAX, short axis; LAX, long axis; RV, right ventricle; LA, left atrium.

valid to be included in the assessment is 1918 (1982–64). The remaining 106 scans (170–64), which were given a 'poor' score, were visually re-checked and showed that LV tracking was present in all these scans, but were difficult to visualize at a glance (reduced quality image output which was necessary due to limited storage capacity for larger data set). Therefore, these scans correctly had corresponding data output for GCS and GRS. Only 6 of these 106 scans had truly 'poor' LV tracking limited to few LV slices deserving the score of 3 when visually re-checked.

Therefore, there were only 30 scans (24 with only partial LV tracking and 6 from the repeat review round of visual QC) where the score of 'poor' quality[3] was appropriate, with the other 64 scan scores no longer applicable to be included in the total, as there was appropriately no corresponding data output when the LV is not tracked in the short axis images. This leads us to reclassify the scores and 98.4% (1888/1918) of the feature tracking for LV SAX tissue tracking images are of either good or satisfactory quality. The score was incorrectly reduced due to human discrepancy in the visual quality checking process.

Comparing clinical measures: SO removal vs. VQC data sets

When comparing the CMR measures between the VQC data set and the SO data set, we found no significant differences in any of the clinical measures related to cardiac structure and function, including LV and RV volumes, LV mass, LVEF, LV GLS, aortic flow, and native T1 (*Table 5*). Additionally, our results showed no significant differences between the two data sets when analysing the measures separately by sex. This suggests that the use of either VQC or SO methods in selecting data for analysis did not have a significant impact on the results obtained.

Discussion

This paper is the first, to our knowledge, that outlines the QC of automated batch processing performed on a large scale for research purposes. Across all images reviewed, there were <5% that were graded 'poor' quality. The results demonstrate the reliability of automated image analysis using the batch processing prototypes developed by Circle Cardiovascular Imaging, Inc. (CVI42 prototypes). Other studies have also demonstrated the superior accuracy and reproducibility of automated image analysis methods using ML algorithms.[4,11,24] Our findings illustrate the impact of human error that can lead to the scan quality being inappropriately downgraded especially with the tissue tracking analysis for LV-SAX where tracking markers were difficult to visualize.

During the first phase of the QC process the results show there was, overall, substantial inter-operator agreement, except for the contouring of RV SAX images. This reflects the reality of CMR image analysis where there is still much debate in both research and clinical settings about the best way to contour/track the right ventricle.[25] These images were the focus of the subsequent group huddle session to establish a consensus between all the operators for the next and final phases of QC. There was agreement with regard to RV tracking, and this did not require extensive discussion to reach a consensus.

The comparison of CMR measurements generated from the VQC data set and the SO data set is also encouraging. There were no statistically significant results when investigating whether the clinical data output differed between the two data sets. Therefore, our findings suggest a statistical approach to remove outliers of image-derived phenotypes is sufficient in UK Biobank image analysis and may save time of experts performing visual quality checks on segmentations.

Table 3 Distribution of CMR scans and segmentations judged of good or acceptable quality (score 1 or 2) according to the CMR image sequence

CMR image sequence	Proportion of studies with good/acceptable quality
LV contours	
Short axis (SAX) stack	99.8% ($n = 1976/1980$)
Long axis 2-chamber	99.6% ($n = 1974/1982$)
Long axis 3-chamber	99.7% ($n = 1979/1984$)
Long axis 4-chamber	99.8% ($n = 1981/1985$)
RV contours	
Short axis (SAX) stack	99.8% ($n = 1977/1980$)
Long axis 4-chamber	99.9% ($n = 1983/1985$)
LA contours	
Long axis 4-chamber	99.6% ($n = 1978/1985$)
Long axis 2-chamber	97.0% ($n = 1965/1982$)
RA contours	
Long axis 4-chamber	99.8% ($n = 1981/1985$)
Native T1 maps	98.9% ($n = 1909/1930$)
Aortic valve flow	99.7% ($n = 1923/1928$)

The proportion of studies deemed of good/acceptable quality was calculated per each image sequence after removing the studies with missing images or clips, which were due to that image sequence not being performed for those studies or corrupted files not being recognized by software.
LV, left ventricle; RV, right ventricle; LA, left atrium; RA, right atrium.

Table 4 The percentage of 1987 scans and segmentations visually scored to be good or satisfactory (grade 1 or 2, respectively) for each image sequence used for strain analysis

CMR image sequence	Proportion of studies with good/acceptable quality
LV tissue tracking	
Short axis (SAX) stack	90.2% ($n = 1788/1982$)
Long axis 2-chamber	97.3% ($n = 1912/1966$)
Long axis 3-chamber	97.6% ($n = 1917/1964$)
Long axis 4-chamber	97.4% ($n = 1921/1973$)
RV tissue tracking	
Short axis (SAX) stack	99.5% ($n = 1972/1982$)
Long axis 4-chamber	98.4% ($n = 1956/1987$)

The proportion of studies deemed of good/acceptable quality was calculated per each image sequence after removing the studies with missing images or clips, which were due to that image sequence not being performed for those studies or corrupted files not recognized by the software.
LV, left ventricle; RV, right ventricle.

The QC processes can be time- consuming and labour- intensive. However, it is currently necessary to safeguard the quality of the data output until the reliability of these automated methods is established and widely accepted. The use of online platforms such as the one developed in-house for this study can greatly improve the efficiency and ease of the QC process.[26] The Shiny app allowed multiple operators to contribute to the process without being bound to a physical space or restricted timings. Therefore, the QC of ∼2000 scans which meant visually reviewing ∼62 000 videos and ∼2000 images, was completed efficiently (median time spent per scan overall was 32 s ranging between operators from 20–56 s), whilst also being convenient for all the operators. The use of the Shiny app also facilitated intense huddle discussions that could be reviewed remotely at the beginning to ensure consistency in scoring. Overall, the time and expertise spent on this QC process is relatively small compared to the time it would have taken for the large number of scans included in this study to be manually analysed and may still need to go through a quality check process. Validating the use of automated image analysis methods is an essential step to allow progress to be made using large data sets that include imaging (for example the UK Biobank imaging data set with ∼50 000 scans to date), which could lead to interesting and previously unknown knowledge in research.

Automated image analysis will also allow greater efficiency in clinical settings where automated analysis can be relied upon for the more mundane task of producing the quantitative data and the expert clinician can then spend more time on the interpretation. This has already been reported in other studies.[4,24] The findings of this paper would suggest that even for more complex image analysis such as aortic flow, native T1, and feature tracking strain, we can rely on automated processes to produce reliable data. The clinicians will always need to be aware of the quality checking parameters such as those described in *Table 2* for each imaging modality when it comes to interpretation, particularly if there are discrepancies in the expected data output and the actual data output. However, if the prototypes described in this paper are used for clinical or research purposes, there can be overall high confidence in the results that are produced based on the results described above.

Strengths and limitations

This study outlines a robust visual QC process, which can be easily replicated by others, using an online platform which makes the process efficient and convenient even with multiple operators. The results are promising and promote the use of automated image analysis methods as reliable and accurate with the overwhelming majority (>95%) of the analysis deemed to be of 'good' or 'satisfactory' quality. This allows us to recommend that we continue to use automated batch processing image analysis methods on a larger scale without the need for constant, time- and resource-consuming visual quality checks, and instead, the removal of statistical outliers is sufficient. The comparison made between CMR measures generated after the two types of QC demonstrates this recommendation in practice, where there is no difference in the overall data output from both QC methods. Future work could also compare visual quality checks with automated QC algorithms described in the literature.[9,27] This would also be an efficient method for quality-checking the automated image analysis.

However, one limitation is that the encouraging results seen in this paper are limited to scans performed in the UK Biobank imaging study where standardized protocols are implemented, which ensure a relatively high standard of images to begin with. In addition, the comparison made between the visually quality-checked data set and SO data set is limited to those without cardiovascular disease only, after those with overt cardiovascular pathology were excluded. Therefore, variation in data output may exist when those with conditions such as severe heart failure and cardiomyopathies are included, which would lead to more extreme data values being generated. These may then be erroneously removed using the SO method, whereas visual quality checks would note that the extreme values are a reflection of the disease rather than a true erroneous result. This is a key limitation of this secondary analysis in this study.

Finally, the automated batch processing software was provided by Circle and is, therefore, not an open-source resource available to

Table 5 CMR measures at baseline imaging visit of healthy men and women in the COVID-19 imaging study

CMR measure	Male (SO data set) (mean ± sd)	Male (VQC data set) (mean ± sd)	*P* value	Female (SO data set) (mean ± sd)	Female (VQC data set) (mean ± sd)	*P* value
LV end-diastolic volume (mL)	167.7 ± 30.1	167.5 ± 29.8	0.91	130.6 ± 22.3	130.9 ± 22.6	0.83
LV end-systolic volume (mL)	68.9 ± 16.2	69.0 ± 16.5	0.91	49.5 ± 11.4	49.6 ± 11.5	0.86
LV mass (g)	107.9 ± 17.8	108.0 ± 18.0	0.93	74.2 ± 12.7	74.5 ± 12.8	0.77
LV ejection fraction (%)	59.1 ± 5.4	58.9 ± 5.9	0.62	62.2 ± 5.2	62.2 ± 5.3	0.9
LV global longitudinal strain (%)	−17.5 ± 2.0	−17.4 ± 2.3	0.62	−18.8 ± 2.0	−18.8 ± 2.1	0.67
RV end-diastolic volume (mL)	177.9 ± 32.0	177.3 ± 31.7	0.78	132.5 ± 25.0	132.7 ± 25.1	0.92
RV systolic volume	104.1 ± 21.0	103.8 ± 20.8	0.85	82.9 ± 15.8	83.0 ± 15.8	0.93
Aortic total forward volume (mL)	59.7 ± 38.4	59.5 ± 38.2	0.28	52.1 ± 30.7	52.5 ± 30.6	0.84
Native T1	912.4 ± 33.4	920.2 ± 143.3	0.95	943.1 ± 41.2	942.3 ± 47.1	0.76

CMR measures from one data set were generated after removing statistical outliers (SO data set) and the other data set for the same participants were generated from scans that were scored 'good' or 'satisfactory' as part of the VQC data set process. The sample number included varies according to CMR measure due to varied missingness of CMR image sequences from which they were generated. In the statistical outlier data set, the number of men included varied between 409 and 450, and number of women included varied between 518 and 574. In the visually quality-checked data set, the number of men varied between 439 and 478, and the number of women varied between 532 and 585. *P* values based on paired *t*-tests.
CMR, cardiac magnetic resonance; SO, statistical outlier; VQC, visual quality control; LV, left ventricle; RV, right ventricle.

anyone. However, the findings are still encouraging regarding the potential for automated image analysis to be a viable time-saving tool with reliable output.

Conclusion and recommendations

This detailed and comprehensive QC process of the COVID-19 study scans has shown that >95% of the UK Biobank scans analysed using CVI42 batch processing methods are of good or satisfactory quality. Therefore, for the larger imaging cohort of up to ~100 000 studies at baseline and 60 000 at a follow-up, which will be analysed using the same CVI42 batch processing tools, extensive visual quality checks are not required. The recommendation would be that removal of extreme statistical outliers and non-sensical data is sufficient to ensure a good quality for the data outputs. However, when the research question involves looking at a more specific disease affected cohort, additional visual quality check of studies identified as being statistical outliers may be beneficial so that data reflective of true pathology is not inadvertently removed.

Further validation of automated batch processing image analysis will need to be performed for other cohorts and clinical scans before these recommendations can be applied on a wider basis. The robust visual quality check process and the online platform (Shiny app) described in this paper could be beneficial for future work in this field.

Supplementary data

Supplementary data are available at *European Heart Journal - Imaging Methods and Practice* online.

Conflict of interest: S.E.P. provides consultancy to Circle Cardiovascular Imaging, Inc., Calgary, Alberta, Canada. The remaining authors have nothing to disclose.

Funding

Our research group was awarded funding from British Heart Foundation for the COVID-19 repeat imaging study—BHF grant PG/21/10619. S.C. was funded by European Union's Horizon 2020, research and innovation programme under grant agreement no. 825903 (euCanSHare project). E.R. is supported by the mini-Centre for Doctoral Training (CDT) award through the Faculty of Science and Engineering, Queen Mary University of London, UK. A.S. was supported by a British_Heart Foundation project grant (PG/21/10619). H.N. was supported by the British Heart Foundation Pat Merriman Clinical Research Training Fellowship (FS/20/22/34640). A.M.L. and S.E.P. acknowledge support by the London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare (AI4VBH), which is funded from the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed and delivered by Innovate UK on behalf of UK Research and Innovation (UKRI). A.M.L. and S.E.P. also acknowledge support from the 'SmartHeart' Engineering and Physical Sciences and Research Centre programme grant (www.nihr.ac.uk; EP/P001009/1). S.E.P. also acknowledges support from the National Institute for Health and Care Research (NIHR) Biomedical Research Centre at Barts Heart Centre. S.E.P. acknowledges the British Heart Foundation for funding the manual analysis to create a cardiovascular magnetic resonance imaging reference standard for the UK Biobank imaging resource in 5000 CMR scans (www.bhf.org.uk; PG/14/89/31194). M.M.S. is supported by a British Heart Foundation Clinical Research Training Fellowship (FS/CRTF/22/24353). N.A. acknowledges the NIHR Integrated Academic Training program which supports his Academic Clinical Lectureship post. N.A. is supported by the Academy of Medical Sciences Starter Grant for Clinical Lecturers (SGL024\1024).

Data availability

This research was conducted using the UKB resource under access application 2964. UK Biobank will make the data available to all researchers for all types of health-related research that is in the public interest, without preferential or exclusive access for any persons. All researchers are subject to the same application process and approval criteria as specified by UK Biobank. For more details on the access procedure, see the UK Biobank website: http://www.ukbiobank.ac.uk/register-apply/.

Author contributions

S.E.P., N.A., S.C., and E.R. conceived and designed the study. A.M.L., S.C., E.R., N.A., and A.S. performed data analysis. S.C. and E.R. wrote the manuscript, with contribution from N.A. and A.S. The automated segmentation was performed by S.C. using Circle Cardiovascular Imaging

Inc. automated batch processing prototypes. E.R., H.N., J.V., M.K., M.S., S.C., and S.E.P. assessed the quality of automated contours. All authors read and approved the final manuscript.

Lead author biography

Steffen Petersen is a Professor of Cardiovascular Medicine at Queen Mary University of London and a Consultant Cardiologist at Barts Health NHS Trust. He is the British Heart Foundation Data Science Centre Deputy Director and President of the European Society of Cardiology's European Association of Cardiovascular Imaging. Petersen holds medical degrees from Germany and the UK, as well as a Master's in Public Health from Harvard. He is level 3 certified in cardiovascular magnetic resonance and has over 20 years of experience in this field. His research focuses on cardiac imaging, population studies, and cost-effectiveness analysis in prevention.

Dr Nay Aung is a Consultant Cardiologist at Barts Heart Centre, St Bartholomew's Hospital in London, and a Senior Clinical Lecture and MRC Clinician Scientist at Queen Mary University London. He has clinical subspecialty interest in heart failure and cardiac imaging. His research focuses on deep phenotyping of cardiac images using artificial intelligence techniques and the application genomics to understand the determinants of adverse cardiac remodelling. He is an invited speaker for international conferences and serves on multiple committees in the European Society of Cardiology (ESC).

References

1. Arnold JR, McCann GP. Cardiovascular magnetic resonance: applications and practical considerations for the general cardiologist. *Heart* 2020;**106**:174–81.
2. Baritussio A, Scatteia A, Bucciarelli-Ducci C. Role of cardiovascular magnetic resonance in acute and chronic ischemic heart disease. *Int J Cardiovasc Imaging* 2018;**34**:67–80.
3. Bailly A, Lipiecki J, Chabrot P, Alwdja A, Garcier JM, Ughetto S *et al.* Assessment of left ventricular volumes and function by cine-MR imaging depending on the investigator's experience. *Surg Radiol Anat* 2009;**31**:113–20.
4. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA *et al.* Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans Med Imaging* 2018;**37**:2514–25.
5. Leiner T, Rueckert D, Suinesiaputra A, Baeßler B, Nezafat R, Išgum I *et al.* Machine learning in cardiovascular magnetic resonance: basic concepts and applications. *J Cardiovasc Magn Reson* 2019;**21**:1–14.
6. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging* 2015;**15**:1–28.
7. Rauseo E, Omer M, Amir-Khalili A, Sojoudi A, Le TT, Cook SA *et al.* A systematic quality scoring analysis to assess automated cardiovascular magnetic resonance segmentation algorithms. *Front Cardiovasc Med* 2022;**8**:2187.
8. Bhaskaruni D, Moss FP, Lan C. Estimating prediction qualities without ground truth: a revisit of the reverse testing framework. In: *2018 24th International Conference on Pattern Recognition (ICPR)*, Beijing, China. IEEE; 2018. p. 49–54.
9. Robinson R, Valindria VV, Bai W, Oktay O, Kainz B, Suzuki H *et al.* Automated quality control in image segmentation: application to the UK biobank cardiovascular magnetic resonance imaging study. *J Cardiovasc Magn Reson* 2019;**21**:1–14.
10. Avendi MR, Kheradvar A, Jafarkhani H. Automatic segmentation of the right ventricle from cardiac MRI using a learning-based approach. *Magn Reson Med* 2017;**78**:2439–48.
11. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G *et al.* Automated cardiovascular magnetic resonance image analysis with fully convolutional networks 08 information and computing sciences 0801 artificial intelligence and image processing. *J Cardiovasc Magn Reson* 2018;**20**:65.
12. Backhaus SJ, Staab W, Steinmetz M, Ritter CO, Lotz J, Hasenfuß G *et al.* Fully automated quantification of biventricular volumes and function in cardiovascular magnetic resonance: applicability to clinical routine settings. *J Cardiovasc Magn Reson* 2019;**21**:1–13.
13. UK Biobank: Protocol for a large-scale prospective epidemiological resource (AMENDMENT ONE FINAL). 2007. https://www.ukbiobank.ac.uk/media/gnkeyh2q/study-rationale.pdf
14. Raisi-Estabragh Z, Harvey NC, Neubauer S, Petersen SE. Cardiovascular magnetic resonance imaging in the UK biobank: a major international health research resource. *Eur Heart J Cardiovasc Imaging* 2021;**22**:251.
15. Petersen SE, Matthews PM, Francis JM, Robson MD, Zemrak F, Boubertakh R *et al.* UK biobank's cardiovascular magnetic resonance protocol. *J Cardiovasc Magn Reson* 2016;**18**:8.
16. Khanji MY, Aung N, Chahal CAA, Petersen SE. COVID-19 and the UK Biobank—opportunities and challenges for research and collaboration with other large population studies. *Front Cardiovasc Med* 2020;**7**:156.
17. *COVID-19 repeat imaging study [Internet]*. https://www.ukbiobank.ac.uk/explore-your-participation/contribute-further/covid-19-repeat-imaging-study (23 November 2022, date last accessed).
18. Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y *et al. Shiny: Web Application Framework for R. R package version 1.7.4.9002 [Internet]*. https://shiny.rstudio.com/ (6 February 2023, date last accessed).
19. ShinyProxy [Internet]. https://www.shinyproxy.io/ (6 February 2023, cited)
20. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal* 2014;**2014**:2.
21. Li Gwet K. *Handbook of Inter-Rater Reliability, Fourth Edition: the Definitive Guide to Measuring the Extent of Agreement Among Raters*. London: Advanced Analytics, LLC; 2014.
22. Blood E, Spratt KF. *Disagreement on Agreement: Two Alternative Agreement Coefficients*. Orlando, Florida: SAS Global Forum; 2007. https://support.sas.com/resources/papers/proceedings/proceedings/forum2007/186-2007.pdf
23. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977;**33**:363–74.
24. Bhuva AN, Bai W, Lau C, Davies RH, Ye Y, Bulluck H *et al.* A multicenter, scan-rescan, human and machine learning CMR study to test generalizability and precision in imaging biomarker analysis. *Circ Cardiovasc Imaging* 2019;**12**:10.
25. Geva T. Is MRI is the preferred method for evaluating right ventricular size and function in patients with congenital heart disease? *Circ Cardiovasc Imaging* 2014;**7**:190–7.
26. Muller C, Serrano-Solano B, Sun Y, Tischer C, Hériche JK. The image data explorer: interactive exploration of image-derived data. *PLoS One* 2022;**17**:e0273698.
27. 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* 2020;**13**:684–95.