Human-level CMR image analysis with deep fully convolutional networks

Wenjia Bai[1*] Matthew Sinclair[1] Giacomo Tarroni[1] Ozan Oktay[1] Martin Rajchl[1]
Ghislain Vaillant[1] Aaron M. Lee[2] Nay Aung[2] Elena Lukaschuk[3] Mihir M. Sanghvi[2]
Filip Zemrak[2] Kenneth Fung[2] Jose Miguel Paiva[2] Valentina Carapella[3]
Young Jin Kim[3] Hideaki Suzuki[4] Bernhard Kainz[1] Paul M. Matthews[4]
Steffen E. Petersen[2] Stefan K. Piechnik[3] Stefan Neubauer[3] Ben Glocker[1]
Daniel Rueckert[1]

arXiv:1710.09289v1 [cs.CV] 25 Oct 2017

Cardiovascular magnetic resonance (CMR) imaging is a standard imaging modality for assessing cardiovascular diseases (CVDs), the leading cause of death globally. CMR enables accurate quantification of the cardiac chamber volume, ejection fraction and myocardial mass, providing a wealth of information for sensitive and specific diagnosis and monitoring of CVDs. However, for years, clinicians have been relying on manual approaches for CMR image analysis, which is time consuming and prone to subjective errors. It is a major clinical challenge to automatically derive quantitative and clinically relevant information from CMR images. Deep neural networks have shown a great potential in image pattern recognition and segmentation for a variety of tasks. Here we demonstrate an automated analysis method for CMR images, which is based on a fully convolutional network (FCN). The network is trained and evaluated on a dataset of unprecedented size, consisting of 4,875 subjects with 93,500 pixelwise annotated images, which is by far the largest annotated CMR dataset. By combining FCN with a large-scale annotated dataset, we show for the first time that an automated method achieves a performance on par with human experts in analysing CMR images and deriving clinical measures. We anticipate this to be a starting point for automated and comprehensive CMR analysis with human-level performance, facilitated by machine learning. It is an important advance on the pathway towards computer-assisted CVD assessment.

An estimated 17.7 million people died from cardiovascular diseases (CVDs) in 2015, representing 31% of all global deaths [1]. More people die annually from CVDs than any other cause. Technological advances in medical imaging have led to a number of options for non-invasive investigation of CVDs, including echocardiography, computed tomography (CT), cardiovascular magnetic resonance (CMR) etc., each having its own advantages and disadvantages. Due to its good image quality, excellent soft tissue contrast and absence of ionising radiation, CMR has established itself as the gold standard for assessing cardiac chamber volume and mass for a wide range of CVDs [2–4]. To derive quantitative measures such as volume and mass, clinicians have been relying on manual approaches to trace the cardiac chamber contours. It typically takes a trained expert 20 minutes to analyse images of a single subject at two time points of the cardiac cycle, end-diastole (ED) and end-systole (ES). This is time consuming, tedious and prone to subjective errors.

Here we propose a computational method which can automatically analyse images at all time points across the cardiac cycle and derive clinical measures within seconds. The accuracy for clinical measures is comparable to human expert performance. The method would assist clinicians in CMR image analysis and diagnosis with an automated and objective way for deriving clinical measures, therefore reducing cost and improving work efficiency. It would also facilitate large-population imaging studies, such as the UK Biobank project, which aims to conduct imaging scans of vital organs for 100,000 subjects [5]. An automated method is crucial for analysing such a large amount of images and extracting clinically relevant information for subsequent clinical studies.

Machine learning algorithms, especially deep neural networks, have demonstrated their great potential, achieving or surpassing human performance in a number of visual tasks including object recognition in natural images [6], Go game playing [7], skin cancer classification [8] and ocular image analysis [9]. Previously, neural networks have been explored for CMR image analysis [10–12]. These studies either use relatively shallow network architectures or are limited by the size of the dataset (45 subjects). None of them has demonstrated a human-level performance. In 2016, Kaggle organised the second Data Science Bowl for left ventricular volume assessment [13]. Images from 700 subjects were provided with the left ventricular volumes, however, none of the images were annotated. In 2017, MICCAI organised the ACDC challenge [14], where a training set of 100 subjects were provided with manual annotation. In this paper, we utilise a large dataset of 4,875 subjects with 93,500 images, one or two orders of magnitude larger than previous datasets, and for which all the images have been pixelwise annotated by clinical experts. We train fully convolutional networks for both short-axis and long-axis CMR image analysis. By combining the power of deep learning and a large annotated dataset for training and evaluation, this is the first paper that demonstrates an automated method that matches human-level performance on CMR image analysis.

[1]Biomedical Image Analysis Group, Department of Computing, Imperial College London, UK
[2]NIHR Biomedical Research Centre at Barts, Queen Mary University of London, UK
[3]Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, UK
[4]Division of Brain Sciences, Department of Medicine, Imperial College London, UK

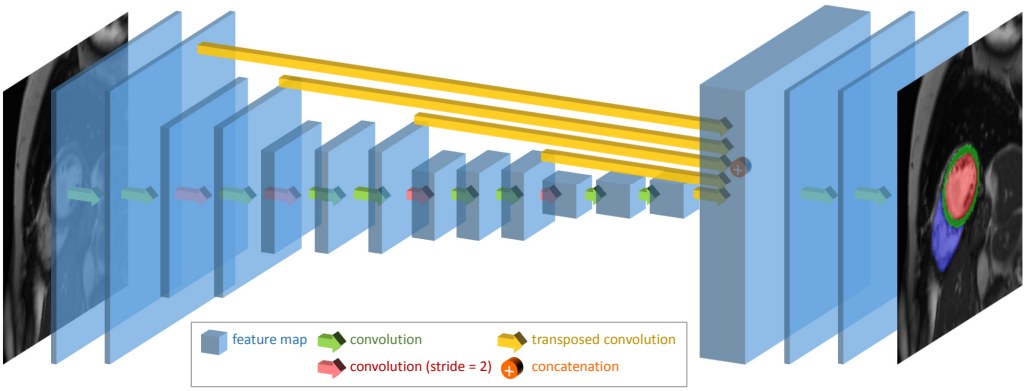

Figure 1: **The network architecture.** A fully convolutional network is used, which takes the CMR image as input, learns image features from fine to coarse scales through a series of convolutions, concatenates multi-scale features and finally predicts a pixelwise image segmentation.

Results

To achieve this, we utilise a fully convolutional network (FCN) architecture, which is a type of neural network that can predict a pixelwise image segmentation by applying a number of convolutional filters onto an input image [15]. The network architecture is illustrated in Figure 1. The FCN learns image features from fine to coarse scales using convolutions and combines multi-scale features for predicting the label class at each pixel (refer to the Methods section for details). Figure 2a illustrates the predicted segmentation of the left ventricle (LV) and the right ventricle (RV) on short-axis images. It shows that automated segmentation agrees well with manual segmentation by a clinical expert at both ED and ES time frames.

For quantitative assessment, we evaluate the performance of the automated method in two ways, respectively using commonly used metrics for segmentation accuracy assessment, including the Dice metric, mean contour distance and Hausdorff distance, and using clinical measures derived from segmentations, including ventricular volume and mass. Figure 3 illustrates the definitions of the Dice metric and contour distance metrics. The Dice metric evaluates the overlap between automated segmentation A and manual segmentation B and it is defined as,

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|}.$$

It is a value between 0 and 1, with 0 denoting no overlap and 1 denoting perfect agreement. The higher the Dice metric, the better the agreement. The mean contour distance and Hausdorff distance evaluate the mean and the maximum distance respectively between the segmentation contours ∂A and ∂B. They are defined as,

$$\text{mean dist.} = \frac{1}{2|\partial A|} \sum_{p \in \partial A} d(p, \partial B) + \frac{1}{2|\partial B|} \sum_{q \in \partial B} d(q, \partial A)$$

$$\text{Haus. dist.} = \max \left(\max_{p \in \partial A} d(p, \partial B), \max_{q \in \partial B} d(q, \partial A) \right)$$

where $d(p, \partial)$ denotes the minimal distance from point p to contour ∂. The lower the distance metric, the better the agreement.

Supplementary Table 1(a) reports the Dice metric, mean contour distance and Hausdorff distance between automated and manual segmentations, evaluated on a test set of 600 subjects, which the network has never seen before. The table shows a mean Dice value of 0.94 for the LV cavity, 0.88 for the LV myocardium and 0.90 for the RV cavity, demonstrating a good agreement between automated and manual segmentations. The mean contour distance is 1.04 mm for the LV cavity, 1.14 mm for the LV myocardium and 1.78 mm for the RV cavity, all of which are smaller than the in-plane pixel spacing of 1.8 mm. The Hausdorff distance ranges from 3.16 mm to 7.25 mm for each class. Of the 600 test subjects, 39 are with CVDs (refer to the Methods section for details). Supplementary Table 1(b) reports the Dice and distance metrics on these pathological cases. It shows a consistent segmentation performance as on the full test set for the Dice metric and just slightly larger errors for the contour distance metrics.

For quantitative evaluation of human performance, we assess the inter-observer variability between segmentations by different clinical experts. A set of 50 subjects was randomly selected and each subject was analysed by three expert observers (O1, O2, O3) independently. Supplementary Table 2 compares the Dice and distance metrics between automated segmentation and manual segmentation, as well as between segmentations by different human observers. It demonstrates that the computer-human difference is close to or even smaller than the human-human difference for all the metrics.

As an additional qualitative assessment, an image analyst with over ten years experience in cardiovascular image analysis visually compared automated segmentation to manual segmentation for 250 test subjects. Extended Table 3 shows that for mid-ventricular slices, automated segmentation agrees well with manual segmentation for 84.8% of the cases by visual inspection. In addition, when there is discrepancy between automated and manual segmentations, the analyst scores that for 12.2% of the cases, automated segmentation performs better, whereas for 2.4% of the cases, manual segmentation performs better. For basal slices where the ventricular contours are more complex and thus more difficult to segment, automated segmentation agrees well with manual segmentation for 40.0% of the cases and performs better for 26.2% of the cases. On the contrary, manual segmentation performs better for 20.6% of the cases.

Next, we evaluate the accuracy of clinical measures,

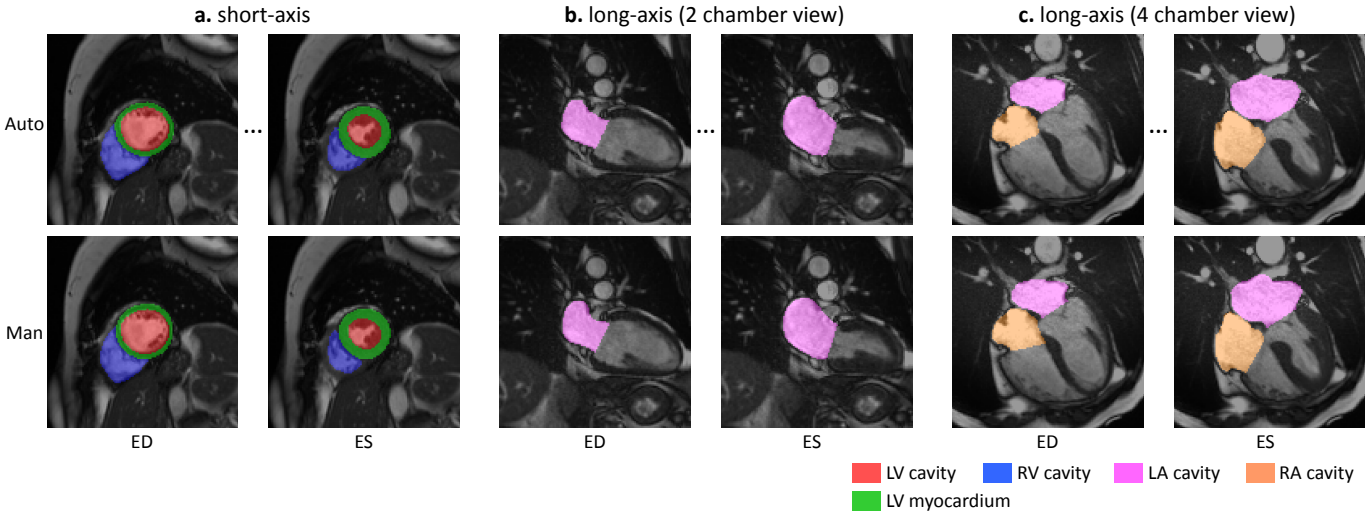

Figure 2: **Illustration of the segmentation results for short-axis and long-axis images.** The top row shows the automated segmentation, whereas the bottom row shows the manual segmentation. The automated method segments all the time frames. However, only ED and ES frames are shown, as manual analysis only annotates ED and ES frames. The cardiac chambers are represented by different colours.

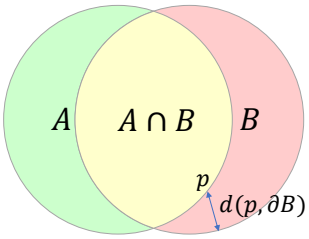

Figure 3: **Illustration of the Dice metric and contour distance metrics.** A and B are two sets representing automated segmentation and manual segmentation. The Dice metric calculates the ratio of the intersection $|A \cap B|$ over the average area of the two sets $(|A| + |B|)/2$. The mean contour distance first calculates, for each point p on one contour, its distance to the other contour $d(p, \partial)$, then calculates the mean across all the points p. The Hausdorff distance calculates the maximum distance between the two contours.

which are derived from image segmentations. We calculate the LV end-diastolic volume (LVEDV) and end-systolic volume (LVESV), LV myocardial mass (LVM), RV end-diastolic volume (RVEDV) and end-systolic volume (RVESV) from automated segmentation and compare them to measurements from manual segmentation. Supplementary Table 4 reports the mean absolute difference and relative difference between automated and manual measurements and between measurements by different expert observers. It shows that for the clinical measures, the computer-human difference is on par with the human-human difference.

Figure 4 shows the Bland-Altman plots of the clinical measures. The Bland-Altman plot is commonly used for analysing agreement and bias between two measurements. The first column of the figure compares automated measurements to manual measurements on 600 test subjects. These subjects were annotated by a group of eight observers and each subject was annotated only once by one observer. The first column shows that the mean difference is centred close to zero, which suggests that the automated measurement is almost unbiased relative to the group of observers. Also, there is no evidence of bias over hearts of difference sizes or volumes. By contrast, the bias between different pairs of human observers (second to fourth columns) is often larger than that, especially for RVEDV and RVESV. This indicates that individual observers may be biased. As the automated method is trained with annotations from multiple observers, it learns a consensus estimate across the group of observers and it is thus less susceptible to biases.

We further demonstrate the performance of the method on long-axis CMR images, which are commonly used for assessing the cardiac chambers from a different angle. Figures 2b and 2c illustrate the segmentations of the left atrium (LA) and the right atrium (RA) for the 2-chamber view (2Ch) and 4-chamber view (4Ch) images respectively. We evaluate the Dice metric and the contour distances on a test set of 600 subjects, as reported in Supplementary Table 5. The mean Dice metric is 0.93 for the LA (2Ch), 0.95 for the LA (4Ch), 0.96 for the RA (4Ch), whereas the mean contour distance is smaller than the in-plane pixel spacing of 1.8 mm, demonstrating a good segmentation accuracy on long-axis images. Supplementary Table 6 demonstrates that for long-axis images, the computer-human difference is also on par with or smaller than the human-human difference.

The proposed automated method enables us to perform clinical studies on large-scale datasets. Using automatically derived clinical measures, we investigate the association between cardiac function and anthropometric covariates in an exemplar application here. Supplementary Table 7 compares the ventricular volume and mass for two groups of subjects, the normal weight group and the obese group. The two groups were matched for sex, age, height and blood pressure (refer to the Methods section for details). The table shows that obesity is associated with increased ventricular volume and mass with statistical significance. This is consistent with a previous finding in [16], which was performed on a dataset of 54 subjects with manual segmentation. Now we can confirm the find-

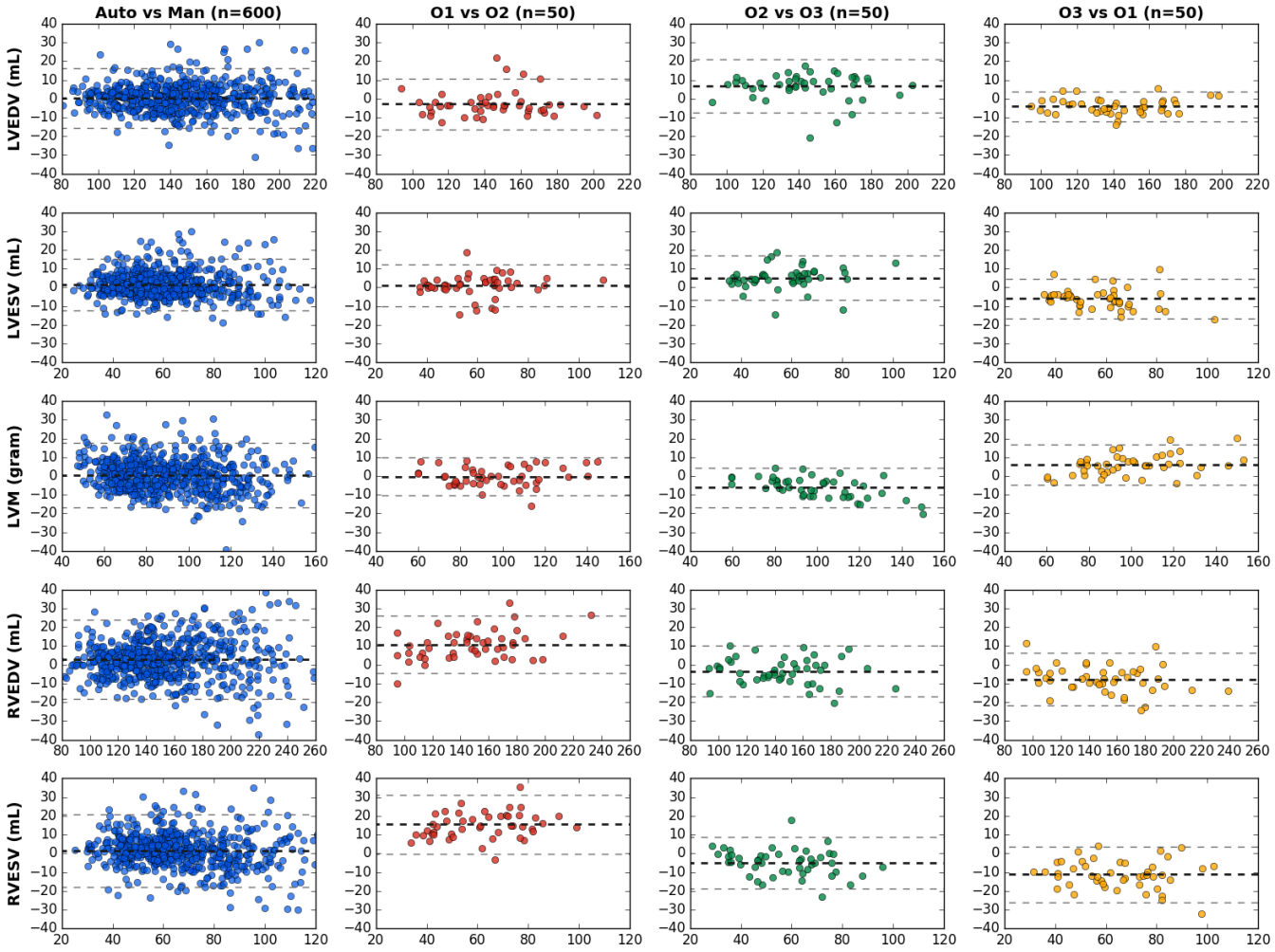

Figure 4: **Bland-Altman plots of clinical measures between automated measurement and manual measurement, as well between measurements by different human observers.** The first column shows the agreement between automated and manual measurements on a test set of 600 subjects. The second to fourth columns show the inter-observer variability evaluated on the randomly selected set of 50 subjects. In each Bland-Altman plot, the x-axis denotes the average of two measurements and the y-axis denotes the difference between them. The dark dashed line denotes the mean difference (bias) and the two light dashed lines denote ± 1.96 standard deviations from the mean.

ing with automated analysis on a much larger dataset with thousands of subjects.

Discussion

Here we propose an automated method using deep FCN for short-axis and long-axis CMR image analysis. By training and evaluating on a large-scale annotated dataset, we demonstrate that the method matches human expert performance on image segmentation accuracy and clinical measurement accuracy. In terms of speed, it can analyse the short-axis and long-axis images for one subject in two seconds. The method is fast and scalable, overcoming limitations associated with current clinical CMR image analysis routine, which is manual, time-consuming and prone to subjective errors. Our method has a great potential for improving work efficiency and assisting clinicians in diagnosis and performing large-scale clinical research. The method has been evaluated on the UK Biobank dataset. But it is generic and can be adapted to other CMR datasets using the technique of transfer learning [17]. Future research will explore evaluating the method performance on a wider range of CMR images, such as multi-site images acquired from different machines and using different imaging protocols, and integrating automated segmentation results into diagnostic reports. The current method trains networks for short-axis images and long-axis images separately. It would be interesting to combine the two views for image analysis, which can provide complementary information about the anatomy of the heart. Finally, we believe that a benchmark platform based on this annotated dataset is needed, which would benefit the whole community and greatly advance the development of CMR image analysis algorithms.

Methods

Dataset The dataset consists of short-axis and long-axis cine CMR images acquired from the UK Biobank study. For short-axis images, the in-plane image resolution is 1.8×1.8 mm^2 with slice thickness of 8.0 mm and slice gap of 2 mm. A short-axis image stack typically consists of 10 image slices. For long-axis images, the in-plane image resolution is 1.8×1.8 mm^2 and only 1 image slice is acquired. Each cardiac cycle consists of 50 time frames. Details of the image acquisition protocol can be found in [18]. Manual image annotation was undertaken by a team of eight observers under the guidance of three principal investigators and following a standard operating procedure [19]. For short-axis images, the LV endocardial and epicardial borders and the RV endocardial borders were manually traced at ED and ES time frames using the cvi^{42} software (version 5.1.1, Circle Cardiovascular Imaging Inc., Calgary, Canada). For long-axis 2Ch images, the LA endocardial border was traced. For long-axis 4Ch view images, the LA and RA endocardial borders were traced.

Preprocessing The CMR DICOM images were converted into NIfTI format for analysis. The manual annotations from the cvi^{42} software were exported as XML files and then converted into NIfTI format. The images and annotations were quality controlled to ensure that annotations cover both ED and ES frames and without missing slices or missing anatomical structures. For short-axis images, 4,875 subjects (with 93,500 annotated image slices) were available after quality control, which were randomly split into three sets of 3,975/300/600 for training/validation/test, i.e. 3,975 subjects for training the network, 300 validation subjects for tuning model parameters, and finally 600 test subjects for evaluating performance. For long-axis 2Ch images, 4,723 subjects were available after quality control, which were split into 3,823/300/600. For long-axis 4Ch images, 4,682 subjects were available, which were split into 3,782/300/600.

Pathological cases For evaluation of the method in the context of CVDs, pathological cases were selected using the following criteria: cases with the International Classification of Diseases code, 10th Revision (ICD-10) of I21 (acute myocardial infarction), I22 (subsequent myocardial infarction), I23 (certain current complications following acute myocardial infarction), I25 (chronic ischaemic heart disease), I42 (cardiomyopathy), I50 (heart failure); cases where participants had self-reported heart attack.

Network architecture We use a fully convolutional network architecture adapted from the VGG-16 network [20] and similar to the DeepLab architecture used in [21] for natural image segmentation. It is illustrated in Figure 1. The network consists of a number of convolutional layers for extracting image features. Each convolution uses a 3×3 kernel and it is followed by batch normalisation and ReLu. After every two or three convolutions, the feature map is downsampled by a factor of 2 so as to learn features at a more global scale. Feature maps learnt at different scales are upsampled to the original resolution using transposed convolutions and the multi-scale feature maps are then concatenated. Finally, three convolutional layers of kernel size 1×1, followed by a softmax function, are used to predict a probabilistic label map. The segmentation is determined at each pixel by the label class with highest softmax probability. The mean cross entropy between the probabilistic label map and the manually annotated label map is used as the loss function. Excluding the transposed convolutional layers, this network has in total 16 convolutional layers. Details of the network architecture can be found in Supplementary Table 8.

We also experimented with a deeper network by replacing the convolutional layers from scale 3 to 5 with residual blocks as described in [22] and constructed a residual network which has 33 convolutional layers. In experiments, we found the residual network achieves a similar performance as the VGG-16 network. Thus, we only report the results from the VGG-16 network in the paper.

Network training and testing For training the network, all images are cropped to the same size of 192×192 and intensity normalised to the range of $[0, 1]$. We perform data augmentation on-the-fly, which applies random translation, rotation, scaling and intensity variation to each mini-batch of images before feeding them to the network. Each mini-batch consists of 20 image slices. The Adam method [23] is used for optimising the loss function, with a learning rate of 0.001 and iteration number of 50,000. The method was implemented using Python and TensorFlow. It took about 10 hours to train the VGG-16 network on a Nvidia Tesla K80 GPU. During the testing stage, it took about 2.2 seconds to segment 50 time frames of short-axis

images for one subject and about 0.2 second to segment 50 time frames of long-axis images.

Clinical measures The clinical measures including the LV and RV volumes are calculated by summing up the number of voxels belonging to each class, multiplied by the volume per voxel. The LV mass is calculated by multiplying the LV myocardial volume with the density of 1.05 g/mL [24].

Exemplar clinical study The clinical study was designed similar to [16]. Pathological cases with CVDs were excluded. The normal weight group and the obese group were matched for sex, age, height, diastolic blood pressure (DBP) and systolic blood pressure (SBP) using the nearest neighbour propensity score matching, implemented using the MatchIt package in R. The clinical measures were then compared between the matched groups using two-sided t-tests.

Data and code availability The imaging data and manual annotations were provided by the UK Biobank Resource under Application Number 2946. Researchers can apply to use the UK Biobank data resource for health-related research in the public interest [25]. The code will be documented and made available at the time of publication. The code is used for data format conversion, preprocessing, segmentation network training, testing and clinical measure calculation.

References

[1] World Health Organisation. Cardiovascular diseases (CVDs) fact sheet. `http://www.who.int/mediacentre/factsheets/fs317/en/` (accessed on 11 Jul 2017).

[2] Ripley, D. *et al.* Cardiovascular magnetic resonance imaging: what the general cardiologist should know. *Heart* **102**, 1589–1603 (2016).

[3] Fihn, S. *et al.* 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS guideline for the diagnosis and management of patients with stable ischemic heart disease. *Circulation* **60**, e44–e164 (2012).

[4] McMurray, J. *et al.* ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012. *European Journal of Heart Failure* **14**, 803–869 (2012).

[5] UK Biobank Imaging Study. `http://imaging.ukbiobank.ac.uk/` (accessed on 11 Jul 2017).

[6] He, K. *et al.* Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision*, 1026–1034 (2015).

[7] Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).

[8] Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).

[9] Long, E. *et al.* An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nature Biomedical Engineering* **1**, 0024 (2017).

[10] Avendi, M. *et al.* A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Medical Image Analysis* **30**, 108–119 (2016).

[11] Ngo, T. *et al.* Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. *Medical Image Analysis* **35**, 159–171 (2017).

[12] Tran, P. A fully convolutional neural network for cardiac segmentation in short-axis MRI. *arXiv:1604.00494* (2017).

[13] Kaggle Second Annual Data Science Bowl. `https://www.kaggle.com/c/second-annual-data-science-bowl/` (accessed on 11 Jul 2017).

[14] MICCAI 2017 ACDC Challenge. `https://www.creatis.insa-lyon.fr/Challenge/acdc/` (accessed on 25 Oct 2017).

[15] Long, J. *et al.* Fully convolutional networks for semantic segmentation. In *Conference on Computer Vision and Pattern Recognition*, 3431–3440 (2015).

[16] Rider, O. *et al.* Determinants of left ventricular mass in obesity; a cardiovascular magnetic resonance study. *Journal of Cardiovascular Magnetic Resonance* **11**, 9 (2009).

[17] Yosinski, J. *et al.* How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, 3320–3328 (2014).

[18] Petersen, S. *et al.* UK Biobank's cardiovascular magnetic resonance protocol. *Journal of Cardiovascular Magnetic Resonance* **18**, 8 (2016).

[19] Petersen, S. *et al.* Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. *Journal of Cardiovascular Magnetic Resonance* **19**, 18 (2017).

[20] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 1–14 (2015).

[21] Chen, L. *et al.* DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[22] He, K. *et al.* Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).

[23] Kingma, D. & Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations* (2015).

[24] Grothues, F. *et al.* Comparison of interstudy reproducibility of cardiovascular magnetic resonance with two-dimensional echocardiography in normal subjects and in patients with heart failure or left ventricular hypertrophy. *The American Journal of Cardiology* **90**, 29–34 (2002).

[25] UK Biobank Register and Apply. `http://www.ukbiobank.ac.uk/register-apply/` (accessed on 11 Jul 2017).

Acknowledgements

This research has been conducted mainly using the UK Biobank Resource under Application Number 2946. The initial stage of the research was conducted using the UK Biobank Resource under Application Number 18545. The authors wish to thank all UK Biobank participants and staff. This work is supported by the SmartHeart EPSRC Programme Grant (EP/P001009/1). G.T. is supported by a Marie Skłodowska Curie European Fellowship. A.L. and S.E.P. acknowledge support from the NIHR Barts Biomedical Research Centre and from the MRC for the MRC eMedLab Medical Bioinformatics infrastructure (MR/L016311/1), which enables data access. N.A. is supported by a Wellcome Trust Research Training Fellowship (203553/Z/Z). S.N. and S.K.P. acknowledge support from the NIHR Oxford Biomedical Research Centre and the Oxford BHF Centre of Research Excellence. S.E.P., S.K.P. and S.N. acknowledge the British Heart Foundation (BHF) for funding the manual analysis to create a cardiovascular magnetic resonance imaging reference standard for the UK Biobank imaging resource in 5000 CMR scans (PG/14/89/31194). H.S. is supported by a Research Fellowship from the Uehara Memorial Foundation. P.M.M. gratefully acknowledges support from the Edmond J. Safra Foundation and Lily Safra, the Imperial College Healthcare Trust Biomedical Research Centre, the EPSRC Centre for Mathematics in Precision Healthcare and the MRC.

Author contributions

W.B., B.G. and D.R. conceived and designed the study; M.S., G.T., O.O., M.R., and G.V. provided advice and support on computing method aspects; S.N., S.E.P., S.K.P. provided the design of a large data resource to be used for training and testing of artificial intelligence approaches; A.M.L., N.A., S.E.P., S.K.P. and S.N. provided advice and support on clinical aspects; N.A., E.L., M.M.S.,

F.Z., K.F., J.M.P., V.C. and Y.J.K. performed manual image annotation under the senior supervision of S.E.P., S.K.P. and S.N.; E.L. performed qualitative visual assessment of automated segmentation; A.M.L. and V.C. curated the annotation database; H.S. and P.M.M. provided advice and support in the initial stage of model development; W.B., B.K. and A.M.L. performed data pre-processing; W.B. designed the method, performed data analysis and wrote the manuscript. All authors read and approved the manuscript.

Conflict of interest

S.E.P. receives consultancy fees from Circle Cardiovascular Imaging Inc., Calgary, Alberta, Canada.

Supplementary Table 1: **The Dice metric, mean contour distance (MCD) and Hausdorff distance (HD) between automated segmentation and manual segmentation for short-axis images.** The mean and standard deviation (in parenthesis) are reported.

(a) The full test set (n = 600)

	Dice	MCD (mm)	HD (mm)
LV cavity	0.94 (0.04)	1.04 (0.35)	3.16 (0.98)
LV myocardium	0.88 (0.03)	1.14 (0.40)	3.92 (1.37)
RV cavity	0.90 (0.05)	1.78 (0.70)	7.25 (2.70)

(b) Cases with cardiovascular diseases (n = 39)

	Dice	MCD (mm)	HD (mm)
LV cavity	0.94 (0.04)	1.19 (0.41)	3.62 (1.14)
LV myocardium	0.87 (0.04)	1.23 (0.40)	4.28 (1.18)
RV cavity	0.90 (0.04)	2.02 (0.88)	8.19 (2.94)

Supplementary Table 2: **The Dice metric and contour distance metrics between automated segmentation and manual segmentation for short-axis images, as well between segmentations by different human observers.** The first column shows the difference between automated and manual segmentations on a test set of 600 subjects. The second to fourth columns show the inter-observer variability, which is evaluated on a randomly selected set of 50 subjects, each being analysed by three different human observers (O1, O2, O3) independently. The mean and standard deviation (in parenthesis) of the metrics are reported.

(a) Dice metric

	Auto vs Man (n = 600)	O1 vs O2 (n = 50)	O2 vs O3 (n = 50)	O3 vs O1 (n = 50)
LV cavity	0.94 (0.04)	0.94 (0.04)	0.92 (0.04)	0.93 (0.04)
LV myocardium	0.88 (0.03)	0.88 (0.02)	0.87 (0.03)	0.88 (0.02)
RV cavity	0.90 (0.05)	0.87 (0.06)	0.88 (0.05)	0.89 (0.05)

(b) Mean contour distance (mm)

	Auto vs Man (n = 600)	O1 vs O2 (n = 50)	O2 vs O3 (n = 50)	O3 vs O1 (n = 50)
LV cavity	1.04 (0.35)	1.00 (0.25)	1.30 (0.37)	1.21 (0.48)
LV myocardium	1.14 (0.40)	1.16 (0.34)	1.19 (0.25)	1.21 (0.36)
RV cavity	1.78 (0.70)	2.00 (0.79)	1.78 (0.45)	1.87 (0.74)

(c) Hausdorff distance (mm)

	Auto vs Man (n = 600)	O1 vs O2 (n = 50)	O2 vs O3 (n = 50)	O3 vs O1 (n = 50)
LV cavity	3.16 (0.98)	2.84 (0.70)	3.31 (0.90)	3.25 (0.96)
LV myocardium	3.92 (1.37)	3.70 (1.16)	3.82 (1.07)	3.76 (1.21)
RV cavity	7.25 (2.70)	7.56 (2.51)	7.35 (2.19)	7.14 (2.20)

Supplementary Table 3: **Qualitative visual assessment of automated segmentation.** An experienced image analyst visually compared automated segmentation to manual segmentation for 250 test subjects and assessed whether the two segmentations achieved a good agreement (visually close to each other) or not. If there was a disagreement between the two, the analyst would score in three categories: automated segmentation performs better; manual segmentation performs better; not sure which one is better. The visual assessment was performed for basal, mid-ventricular and apical slices. The percentage of each score catetory is reported.

	Agreement (%)	Disagreement (%)		
		Auto. better	Man. better	Not sure
Basal	40.0	26.2	20.6	13.2
Mid-ventricular	84.8	12.2	2.4	0.6
Apical	44.0	29.0	22.0	5.0

Supplementary Table 4: **The difference in clinical measures between automated segmentation and manual segmentation, as well between measurements by different human observers.** The first column shows the difference between automated and manual segmentations on a test set of 600 subjects. The second to fourth columns show the inter-observer variability, which is evaluated on a randomly selected set of 50 subjects, each being analysed by three different human observers (O1, O2, O3) independently. The mean and standard deviation (in parenthesis) of the absolute difference and relative difference are reported.

(a) Absolute difference

	Auto vs Man (n = 600)	O1 vs O2 (n = 50)	O2 vs O3 (n = 50)	O3 vs O1 (n = 50)
LVEDV (mL)	6.1 (5.3)	6.1 (4.4)	8.8 (4.8)	4.8 (3.1)
LVESV (mL)	5.3 (4.9)	4.1 (4.2)	6.7 (4.2)	7.1 (3.8)
LVM (gram)	6.9 (5.5)	4.2 (3.2)	6.6 (4.9)	6.5 (4.8)
RVEDV (mL)	8.5 (7.1)	11.1 (7.2)	6.2 (4.6)	8.7 (5.8)
RVESV (mL)	7.2 (6.8)	15.6 (7.8)	6.6 (5.5)	11.7 (6.9)

(b) Relative difference

	Auto vs Man (n = 600)	O1 vs O2 (n = 50)	O2 vs O3 (n = 50)	O3 vs O1 (n = 50)
LVEDV (%)	4.1 (3.5)	4.2 (3.1)	6.3 (3.3)	3.4 (2.2)
LVESV (%)	9.5 (9.5)	6.8 (7.5)	12.5 (8.5)	11.7 (5.1)
LVM (%)	8.3 (7.6)	4.4 (3.3)	6.0 (3.7)	6.7 (4.6)
RVEDV (%)	5.6 (4.6)	8.0 (5.0)	4.2 (3.1)	5.7 (3.6)
RVESV (%)	11.8 (12.2)	30.6 (15.5)	10.9 (8.3)	16.9 (9.2)

Supplementary Table 5: **The Dice metric, mean contour distance (MCD) and Hausdorff distance (HD) between automated segmentation and manual segmentation for long-axis images.** The mean and standard deviation (in parenthesis) are reported on a test set of 600 subjects.

	Dice	MCD (mm)	HD (mm)
LA cavity (2Ch)	0.93 (0.05)	1.46 (1.06)	5.76 (5.85)
LA cavity (4Ch)	0.95 (0.02)	1.04 (0.38)	4.03 (2.26)
RA cavity (4Ch)	0.96 (0.02)	0.99 (0.43)	3.89 (2.39)

Supplementary Table 6: **The Dice metric and contour distance metrics between automated segmentation and manual segmentation for long-axis images, as well between segmentations by different human observers.** The first column shows the difference between automated and manual segmentations on a test set of 600 subjects. The second to fourth columns show the inter-observer variability, which is evaluated on a randomly selected set of 50 subjects, each being analysed by three different human observers (O1, O2, O3) independently. The mean and standard deviation (in parenthesis) of the metrics are reported.

(a) Dice metric

	Auto vs Man (n = 600)	O1 vs O2 (n = 50)	O2 vs O3 (n = 50)	O3 vs O1 (n = 50)
LA cavity (2Ch)	0.93 (0.05)	0.92 (0.02)	0.90 (0.04)	0.90 (0.04)
LA cavity (4Ch)	0.95 (0.02)	0.95 (0.03)	0.94 (0.02)	0.94 (0.03)
RA cavity (4Ch)	0.96 (0.02)	0.95 (0.02)	0.95 (0.02)	0.95 (0.02)

(b) Mean contour distance (mm)

	Auto vs Man (n = 600)	O1 vs O2 (n = 50)	O2 vs O3 (n = 50)	O3 vs O1 (n = 50)
LA cavity (2Ch)	1.46 (1.06)	1.57 (0.39)	1.94 (0.68)	1.95 (0.57)
LA cavity (4Ch)	1.04 (0.38)	1.08 (0.40)	1.21 (0.33)	1.23 (0.35)
RA cavity (4Ch)	0.99 (0.43)	1.13 (0.35)	1.22 (0.37)	1.16 (0.37)

(c) Hausdorff distance (mm)

	Auto vs Man (n = 600)	O1 vs O2 (n = 50)	O2 vs O3 (n = 50)	O3 vs O1 (n = 50)
LA cavity (2Ch)	5.76 (5.85)	5.66 (1.97)	7.16 (3.12)	6.78 (2.53)
LA cavity (4Ch)	4.03 (2.26)	3.89 (1.85)	4.29 (1.97)	4.06 (1.44)
RA cavity (4Ch)	3.89 (2.39)	4.31 (2.20)	4.20 (2.16)	4.08 (2.06)

Supplementary Table 7: **An exemplar study of cardiac function on large-scale datasets using automatically derived clinical measures.** It compares the normal weight group ($18.5 \leq \text{BMI} < 25$) to the obese group ($\text{BMI} \geq 30$). The mean and standard deviation (in parenthesis) are reported.

	Normal (n = 867)	Obese (n = 867)	p-value
LVEDV (mL)	143 (31)	158 (34)	<0.001
LVESV (mL)	60 (19)	67 (20)	<0.001
LVM (gram)	85 (20)	103 (26)	<0.001
RVEDV (mL)	152 (36)	167 (38)	<0.001
RVESV (mL)	67 (20)	75 (22)	<0.001

Supplementary Table 8: **Architecture of the VGG-16 adapted network.** The first two columns list the resolution scale and feature map size. The third column lists the convolutional layer parameters, with "$3 \times 3, 16$" denoting 3×3 kernel and 16 output features. The last convolutional layer outputs K features, with K denoting the number of label classes.

scale	size	convolution
1	192×192	$3 \times 3, 16$
		$3 \times 3, 16$
2	96×96	$3 \times 3, 32$
		$3 \times 3, 32$
3	48×48	$3 \times 3, 64$
		$3 \times 3, 64$
		$3 \times 3, 64$
4	24×24	$3 \times 3, 128$
		$3 \times 3, 128$
		$3 \times 3, 128$
5	12×12	$3 \times 3, 256$
		$3 \times 3, 256$
		$3 \times 3, 256$
upsample and concatenate scale 1 to 5 features		
predict	192×192	$1 \times 1, 64$
		$1 \times 1, 64$
		$1 \times 1, K$