Radiology

Three-dimensional Deep Convolutional Neural Networks for Automated Myocardial Scar Quantification in Hypertrophic Cardiomyopathy: A Multicenter Multivendor Study

Ahmed S. Fahmy, PhD • Ulf Neisius, MD, PhD • Raymond H. Chan, MD, MPH • Ethan J. Rowin, MD • Warren J. Manning, MD • Martin S. Maron, MD • Reza Nezafat, PhD

From the Departments of Medicine (Cardiovascular Division) (A.S.F., U.N., W.J.M., R.N.) and Radiology (W.J.M.), Beth Israel Deaconess Medical Center and Harvard Medical School, 330 Brookline Ave, Boston, MA 02215; Toronto General Hospital, University Health Network, Toronto, Ontario, Canada (R.H.C.); and Hypertrophic Cardiomyopathy Center, Division of Cardiology, Tufts Medical Center, Boston, Mass (E.J.R., M.S.M.). Received April 5, 2019; revision requested May 21; revision received August 25; accepted September 25. **Address correspondence to** R.N. (e-mail: *rnezafat@bidmc.harvard.edu*).

Supported by the National Institutes of Health (1R01HL129157-01A1, 5R01HL129185) and American Heart Association (15EIA22710040, 19AIML34850090).

Conflicts of interest are listed at the end of this article.

Radiology 2020; 294:52–60 • https://doi.org/10.1148/radiol.2019190737 • Content code: CA

Background: Cardiac MRI late gadolinium enhancement (LGE) scar volume is an important marker for outcome prediction in patients with hypertrophic cardiomyopathy (HCM); however, its clinical application is hindered by a lack of measurement standardization.

Purpose: To develop and evaluate a three-dimensional (3D) convolutional neural network (CNN)-based method for automated LGE scar quantification in patients with HCM.

Materials and Methods: We retrospectively identified LGE MRI data in a multicenter (n = 7) and multivendor (n = 3) HCM study obtained between November 2001 and November 2011. A deep 3D CNN based on U-Net architecture was used for LGE scar quantification. Independent CNN training and testing data sets were maintained with a 4:1 ratio. Stacks of short-axis MRI slices were split into overlapping substacks that were segmented and then merged into one volume. The 3D CNN per-site and per-vendor performances were evaluated with respect to manual scar quantification performed in a core laboratory setting using Dice similarity coefficient (DSC), Pearson correlation, and Bland-Altman analyses. Furthermore, the performance of 3D CNN was compared with that of two-dimensional (2D) CNN.

Results: This study included 1073 patients with HCM (733 men; mean age, 49 years ± 17 [standard deviation]). The 3D CNN-based quantification was fast (0.15 second per image) and demonstrated excellent correlation with manual scar volume quantification (r = 0.88, P < .001) and ratio of scar volume to total left ventricle myocardial volume (%LGE) (r = 0.91, P < .001). The 3D CNN-based quantification strongly correlated with manual quantification of scar volume (r = 0.82–0.99, P < .001) and %LGE (r = 0.90–0.97, P < .001) for all sites and vendors. The 3D CNN identified patients with a large scar burden (>15%) with 98% accuracy (202 of 207) (95% confidence interval [CI]: 95%, 99%). When compared with 3D CNN, 2D CNN underestimated scar volume (r = 0.85, P < .001) and %LGE (r = 0.83, P < .001). The DSC of 3D CNN segmentation was comparable among different vendors (P = .07) and higher than that of 2D CNN (DSC, 0.54 ± 0.26 vs 0.48 ± 0.29; P = .02).

Conclusion: In the hypertrophic cardiomyopathy population, a three-dimensional convolutional neural network enables fast and accurate quantification of myocardial scar volume, outperforms a two-dimensional convolutional neural network, and demonstrates comparable performance across different vendors.

Online supplemental material is available for this article.

Hypertrophic cardiomyopathy (HCM) is the most prevalent hereditary cardiomyopathy (1) and is a major contributor to sudden cardiac death in young adults and athletes (2). Quantification of late gadolinium enhancement (LGE) with cardiac MRI is an established tool in outcome prediction (3,4). The presence of LGE has been associated with long-term outcomes in numerous studies (5–8), and the value of LGE quantification in HCM has been established (3,4). Robust and reproducible scar quantification could substantially improve adoption of LGE volume as a prognostic imaging marker in the clinical care of patients with HCM, particularly at low-volume centers or those with less experience in this area. However, the current practice of quantifying LGE scar relies on tedious and time-consuming manual analysis (9) of stacks of two-dimensional (2D) images to identify areas of LGE. Additionally, variation among readers, cardiac MRI centers, and analysis core laboratories (10,11) reduces the reproducibility of scar quantification and hinders its clinical utility. Thus, there is an unmet need for automated LGE scar quantification in patients with HCM.

Recently we demonstrated the potential of 2D fully convolutional neural networks (CNNs) to quantify LGE scar in patients with HCM (12). However, 2D CNNs are generally limited because they do not allow comprehensive processing

Abbreviations

CI = confidence interval, CNN = convolutional neural network, DSC = Dice score coefficient, HCM = hypertrophic cardiomyopathy, LGE = late gadolinium enhancement, LV = left ventricle, %LGE = ratio of scar volume to total LV myocardium volume, 3D = three dimensional, 2D = two dimensional

Summary

Three-dimensional deep convolutional neural networks allowed fast and accurate quantification of myocardial scar in a multicenter MRI study of patients with hypertrophic cardiomyopathy.

Key Results

- Three-dimensional (3D) convolutional neural networks (CNNs) determined the extent of hypertrophic cardiomyopathy scar, with a performance that was comparable to that of manual analysis by an expert reader ($r = 0.89$, $P < .001$).
- The 3D CNN identified patients with a large scar burden (>15% left ventricle volume) (accuracy, 98%).
- The 3D CNN performed better than did a two-dimensional CNN (Dice similarity coefficient, 0.54 vs 0.48; $P = .02$).

of volumetric data (eg, stacks of 2D slices). This limitation has been addressed with several approaches, including recurrent CNN (13), ad hoc modifications of 2D CNN models (14), and three-dimensional (3D) CNN. Use of 3D CNN enables optimal processing of interslice information and demonstrates strong promise in several medical image applications, including segmentation of bony structures (15,16), abdominal anatomy (17), and lung nodules (18). However, 3D CNN models can be limited due to their reduced set size (number of volumes instead of number of images) and extensive demands on computational resources. Also, 3D CNN models are designed and trained assuming there will be a fixed number of input slices, which can lead to a substantial performance decrease when the number of input slices varies during testing. The latter is particularly important for multicenter LGE data, where images are acquired with mixed 2D and 3D imaging protocols. We hypothesize that 3D CNN would outperform 2D CNN in producing accurate and reliable LGE segmentation. We developed and evaluated a 3D CNN model with a fixed number of input slices and used a sliding window scheme to process inputs of arbitrary sizes. The proposed model was trained and evaluated by using a multicenter data set with images from different imaging protocols and vendors.

Materials and Methods

Training and Testing Data Set

A set of short-axis LGE MRI scans from 1073 patients with HCM (Table 1) acquired as part of a multicenter ($n = 7$) multivendor ($n = 3$) study from November 2001 to November 2011 (3) was retrospectively used for network training and testing. Of the 1073 patients, 1041 were previously used for 2D CNN analysis, and findings were reported by Fahmy et al (12). In this study, we used 3D CNN analysis, compared 3D CNN analysis with 2D CNN analysis, and studied the performance across different sites and vendors. All patients signed statements approved by the internal review boards of participating institutions agreeing to the use of their medical information for research.

Cardiac MRI was performed with a 1.5-T scanner (GE Medical Systems, Waukesha, Wis; Philips Medical Systems, Best, the Netherlands; Siemens Healthineers, Erlangen, Germany) using electrocardiographic-gated acquisition of short-axis slices from the atrioventricular ring to the apex (3). LGE images were acquired 10–20 minutes after intravenous administration of 0.2 mmol per kilogram of body weight gadopentetic acid with breath-hold 2D or navigator-gated 3D segmented inversion recovery sequences. Inversion time was optimized to null normal myocardial signal. Image acquisitions included magnitude- or phase-sensitive inversion recovery sequences. For phase-sensitive sequences, uncorrected magnitude images were used.

Image analysis was performed in a core laboratory setting, where an expert with cardiac MRI level III training manually delineated myocardium boundaries and adjusted an intensity threshold to segment areas of visually identified LGE MRI scar (Appendix E1 [online]) (3). This resulted in a set of images with pixels labeled *normal myocardium*, *scar*, *blood pool*, and *background*. All MRI scans were normalized to an intensity range from 0 to 1 and were cropped to 128 × 128 around the heart to reduce computational requirements and potential interslice misalignment due to patient motion. To avoid training biases, patient data sets from each site were classified into scar (>1% of LV volume) or no-scar subgroups that were then randomly split into training and testing data sets (4:1 ratio) (12) (Fig 1).

Convolutional Neural Network

We developed a 3D CNN based on U-Net architecture (19) with four multiresolution processing levels and 3 × 3 × 3 convolutional kernels with in-plane pooling (Appendix E1 [online]). The network was designed assuming input volumes of 128 × 128 × 3, where the number of slices ($n = 3$) corresponded to the minimum anticipated number of LGE slices in a protocol. Segmentation of larger volumes was achieved with a sliding window approach (described later in this article). The network output for each input volume was four probability maps with a size of 128 × 128 × 3, representing the probability of each pixel to belong to scar, normal myocardium, blood, or background regions. Image augmentation (using in-plane translation, rotation, and mirroring) was used to increase the training data set size and avoid overfitting (20). Similar parameters (eg, rotation angles) were used to transform all slices in the input volume to avoid disrupting interslice anatomic correlations. We used cross entropy between the CNN output and manual segmentation as a training loss function. The loss of each volume was weighted by log(10 + scar volume) to compensate for the low percentage of cases with scar in the training set. To compare 3D CNN with 2D CNN, we implemented a 2D CNN with the exact same architecture as the 3D CNN but with 2D convolutional kernels with a size of 3 × 3 (Appendix E1 [online]). The networks were implemented by using Python (version 3.6; Python Software Foundation, *www.python.org*) and Tensorflow (version 1.12; Google, Mountain View, Calif) software on a DGX-1 workstation (Nvidia, Santa Clara, Calif). Network imple-

Table 1: Demographics of Patients with Hypertrophic Cardiomyopathy

Characteristic	Site 1 (*n* = 404)	Site 2 (*n* = 231)	Site 3 (*n* = 58)	Site 4 (*n* = 14)	Site 5 (*n* = 158)	Site 6 (*n* = 55)	Site 7 (*n* = 153)	Total* (*n* = 1073)
Male sex	247	142	0	8	81	32	110	620
Age (y)†	44 ± 18 (7–84)	47 ± 17 (9–84)	43 ± 14 (14–70)	51 ± 16 (13–73)	45 ± 17 (9–81)	53 ± 16 (10–80)	51 ± 14 (13–87)	46 ± 17 (7–87)
Body surface area (m^2)†	2.0 ± 0.3 (0.9–2.8)	1.9 ± 0.3 (1.0–0.3)	1.9 ± 0.2 (1.5–2.3)	1.8 ± 0.2 (1.3–2.0)	1.9 ± 0.2 (0.9–2.4)	1.9 ± 0.2 (1.5–2.2)	2.0 ± 0.2 (1.5–2.5)	1.9 ± 0.3 (0.9–2.8)
LVOTO	84	69	16	2	33	17	31	252
New York Heart Association classification								
I	240	94	32	9	77	30	74	556
II	78	64	20	3	49	21	58	293
III	36	46	3	2	23	4	20	134
IV	3	0	1	0	1	0	0	5
Coronary artery disease	19	11	2	9	4	4	9	51
Atrial fibrillation	43	35	11	5	21	5	13	133
Risk factors								
Nonsustained ventricular tachycardia	47	23	13	3	23	23	25	157
Unexplained syncope	45	15	2	2	12	2	22	100
Family history of sudden cardiac death	60	20	3	4	35	10	28	160

Note.—Unless otherwise indicated, data are numbers of patients. LVOTO = left ventricular outflow tract obstruction, defined as a basal left ventricular outflow tract gradient of 30 mm Hg or greater.

* Data records were not available for 81 patients.

† Data are mean ± standard deviation, with the range in parentheses.

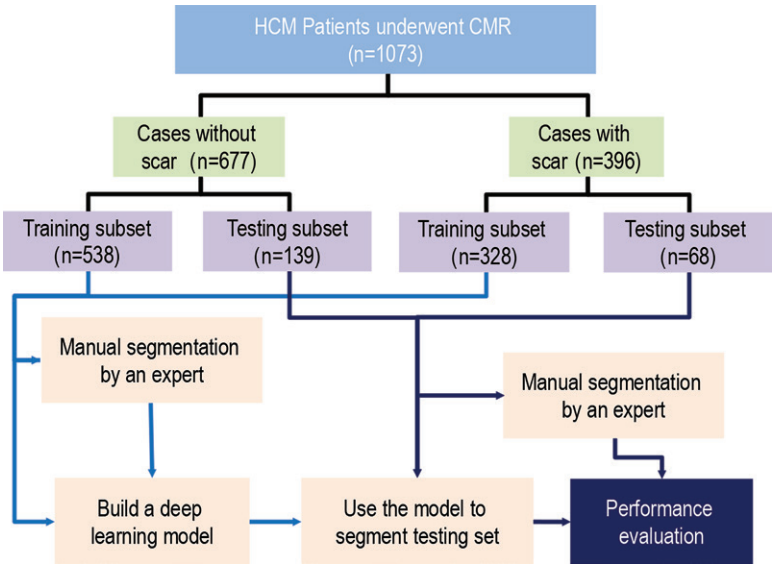

Figure 1: Flowchart shows a workflow overview of the proposed method. The data set (1073 patients) is stratified according to presence of scar. Each group is then split into training (approximately 80%) and testing (approximately 20%) subsets. All images are manually segmented by an expert reader. Late gadolinium enhancement images and their corresponding manually segmented images are used to build a deep learning model that, in turn, segments the testing images. Automatically segmented and manually segmented images are compared with each other to evaluate the performance of the deep learning model. CMR = cardiac MRI, HCM = hypertrophic cardiomyopathy.

mentation is available online (version 1.0; *https://github.com/cmr-bidmc/3D-Scar-Quantification*).

Segmentation of Variable Number of Slices

Given a stack of n slices (ie, size 128 × 128 × n), a sliding window size of three was used to create a subvolume from every three consecutive slices, resulting in n–2 volumes of size 128 × 128 × 3 (Fig 2a). Because of subvolume overlapping, each slice could be processed a certain number of times ($m \leq 3$). To generate a single segmentation image per slice, the resulting m probability maps of each tissue type, excluding scar, were merged by calculating their pixelwise average (Fig 2b). Scar probability maps were merged by taking their pixelwise maximum to increase sensitivity for detecting scarred regions. Isolated scar regions of four or fewer pixels were considered noise and were removed from the resulting segmentation. The final segmentation output was computed by assigning each pixel a label corresponding to the tissue type with maximum probability.

Image Analysis and Scar Quantification

The scar and healthy myocardium volumes were computed by multiplying the voxel volume (in cubic centimeters) by the number of pixels la-

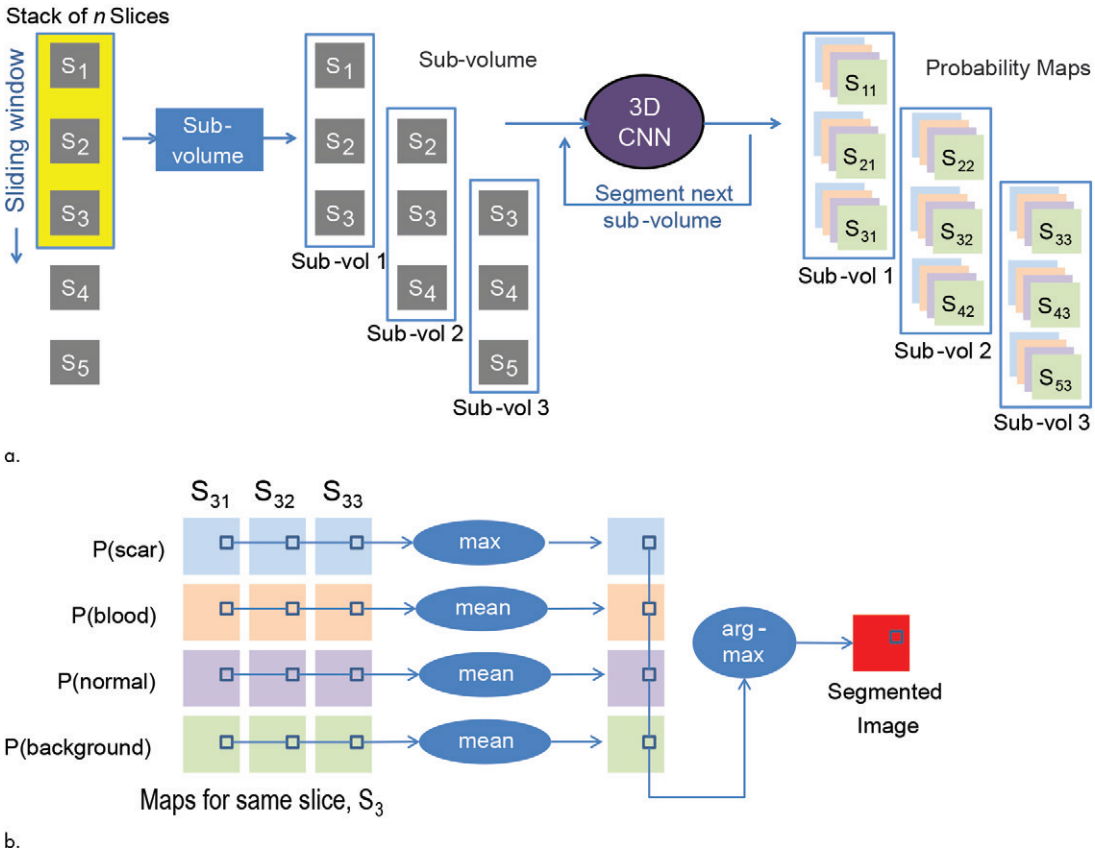

Figure 2: Three-dimensional (3D) convolutional neural network (CNN)-based segmentation of an arbitrary number of slices. **(a)** An input stack of *n* slices, Sj (j = 1 to 5) is divided into (n−2) three-slice subvolumes (Sub-vol), and the 3D CNN is used to segment each subvolume. One slice can belong to more than one subvolume; thus, it can be processed several times. Sjm represents the resulting probability maps of slice Sj in subvolume m (≤3). **(b)** Example of merging the probability maps for slice S3, where the pixelwise maximum (for scar) or average (for all other tissue types) is calculated to create one probability map for each tissue. The final segmentation is computed by using the arg-max function that assigns a label to each pixel corresponding to the tissue type with maximum probability.

Table 2: Size of Training and Testing Sets Grouped by Imaging Site and Scar Presence

A: Size Grouped by Imaging Site

Site No.	Vendor	Training Set	Testing Set	Total
1	3	323	81	404 (38)
2	2	183	48	231 (22)
3	1	46	12	58 (5)
4	3	11	3	14 (1)
5	2	133	25	158 (15)
6	1	47	8	55 (5)
7	1	123	30	153 (14)

B: Size Grouped by Scar Presence

Presence of Scar	Training Set	Testing Set	Total
No scar	538	139	677 (63)
With scar	328	68	396 (37)
Total	866	207	1073 (100)

Note.—Data are numbers of patients; data in parentheses are percentages and are relative to total number of patients in the data set (*n* = 1073). Vendor 1 = GE Medical Systems (Waukesha, Wis), vendor 2 = Philips Medical Systems (Best, the Netherlands), vendor 3 = Siemens Healthineers (Erlangen, Germany).

beled *scar* or *healthy myocardium*, respectively (21). Summation of segmented healthy and scarred myocardial volumes was used to calculate the LV volume. The ratio of scar volume to total LV myocardium volume (%LGE) was computed for automatic and manual segmentations. Overlap between the automatically segmented volume and the manually segmented volume in each patient data set was measured by using the Dice score coefficient (DSC) for both the LV scar and the myocardium (22). The DSC values were computed over the subject's volume of the LV scar (22). Cases without scar in both manual and automatic segmentation (ie, where DSC is undefined) were excluded from DSC calculations, which were performed only for cases with manually identified scars. Image processing and performance evaluation were performed by using Matlab software (R2017a; Mathworks, Natick, Mass).

Statistical Analysis

Continuous variables were expressed as mean ± standard deviation. Pearson correlation coefficient, *r*, and Bland-Altman analysis (23) were used to as-

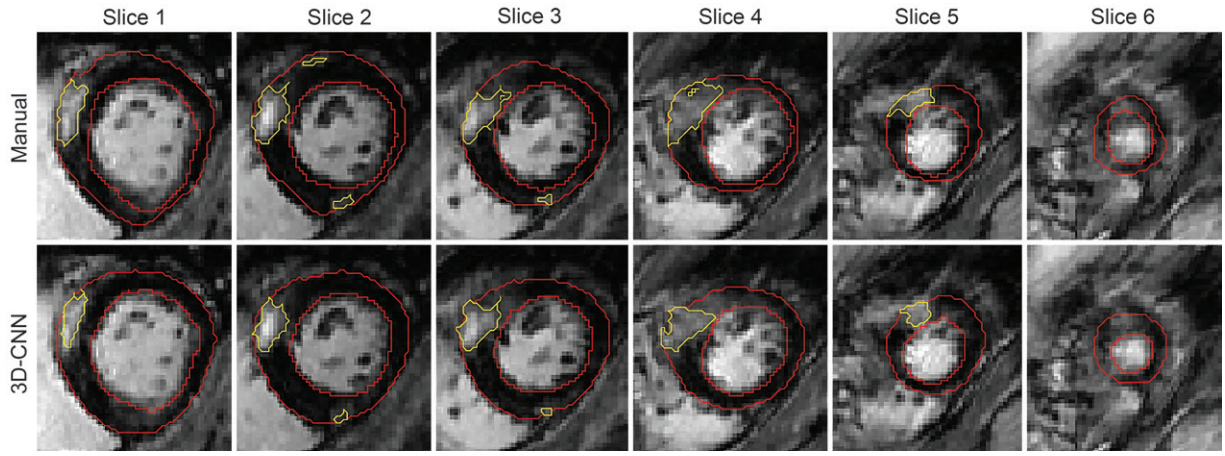

Figure 3: Images in a 60-year-old man diagnosed with hypertrophic cardiomyopathy. Segmentation results for all cardiac MRI short-axial slices are shown. Contours resulting from manual (top row) and three-dimensional (3D) convolutional neural network (CNN) (bottom row) segmentations for the myocardium (red) and scar (yellow) are overlaid on late gadolinium enhancement images. Manual and 3D CNN quantification of scar volume in this patient were 15.3 cm^3 and 12.3 cm^3, respectively, with a Dice similarity coefficient of 0.77.

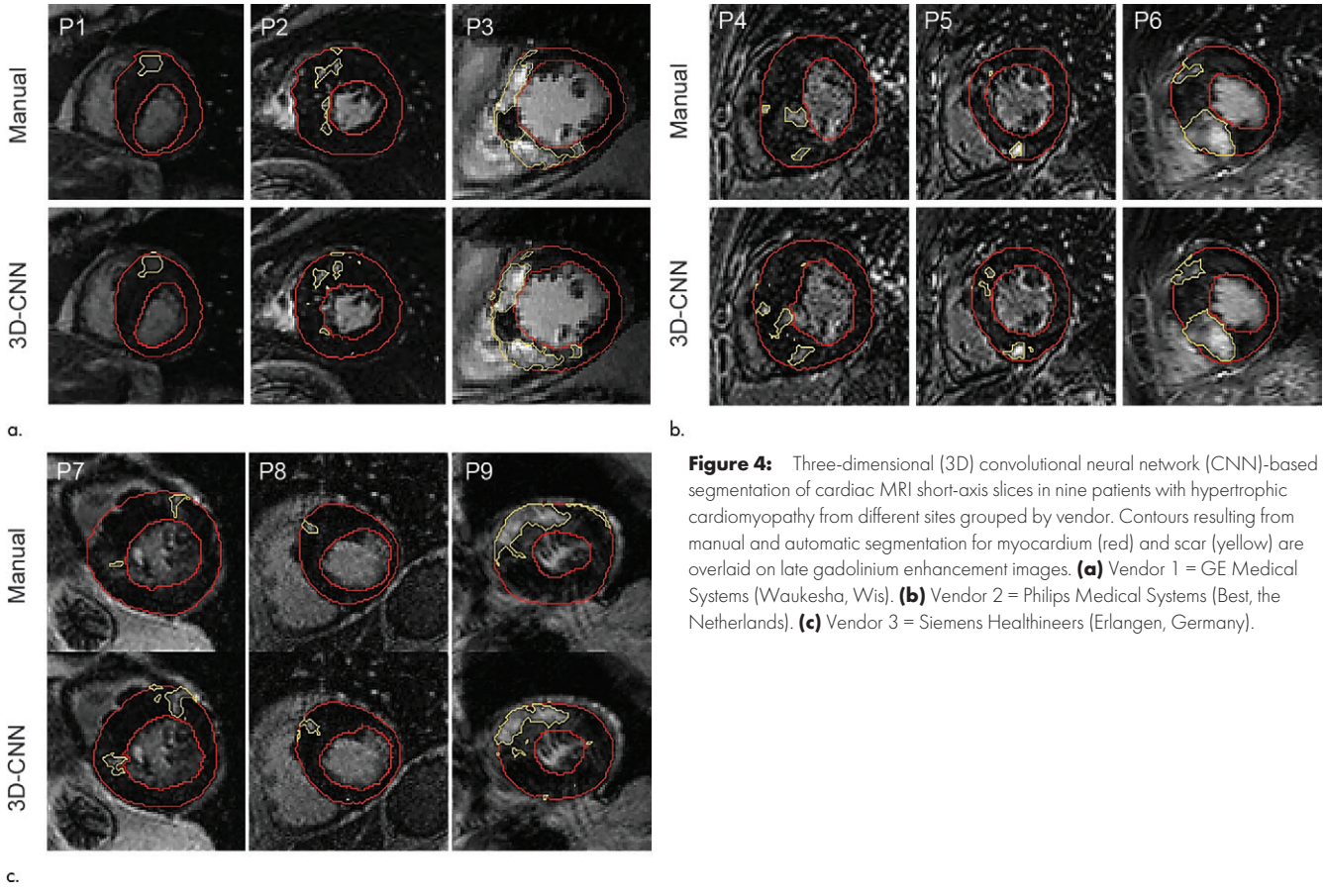

Figure 4: Three-dimensional (3D) convolutional neural network (CNN)-based segmentation of cardiac MRI short-axis slices in nine patients with hypertrophic cardiomyopathy from different sites grouped by vendor. Contours resulting from manual and automatic segmentation for myocardium (red) and scar (yellow) are overlaid on late gadolinium enhancement images. **(a)** Vendor 1 = GE Medical Systems (Waukesha, Wis). **(b)** Vendor 2 = Philips Medical Systems (Best, the Netherlands). **(c)** Vendor 3 = Siemens Healthineers (Erlangen, Germany).

sess agreement between automatic and manual segmentations of scar volume and the ratio of scar volume to total LV myocardium volume. We used a nonparametric Kruskal-Wallis test to compare DSC values, the Welsh *t* test to compare regression slopes, and the Fisher *z* test to compare correlation coefficients. DSC scores were classified as follows: less than 0.25, poor; 0.25–0.49, moderate; 0.5–0.74, good; and 0.75 or greater, excellent. Statistical significance was defined as $P <$ .05. Statistical analyses were performed by using the Statistics Toolbox of Matlab R2017a (Mathworks, Natick, Mass).

Results

The image set for the 1073 patients with HCM (733 men; mean age, 49 years ± 17 [standard deviation]) from all sites (Table 1) was divided into mutually exclusive training and testing subsets. Data splitting resulted in a training set of 866

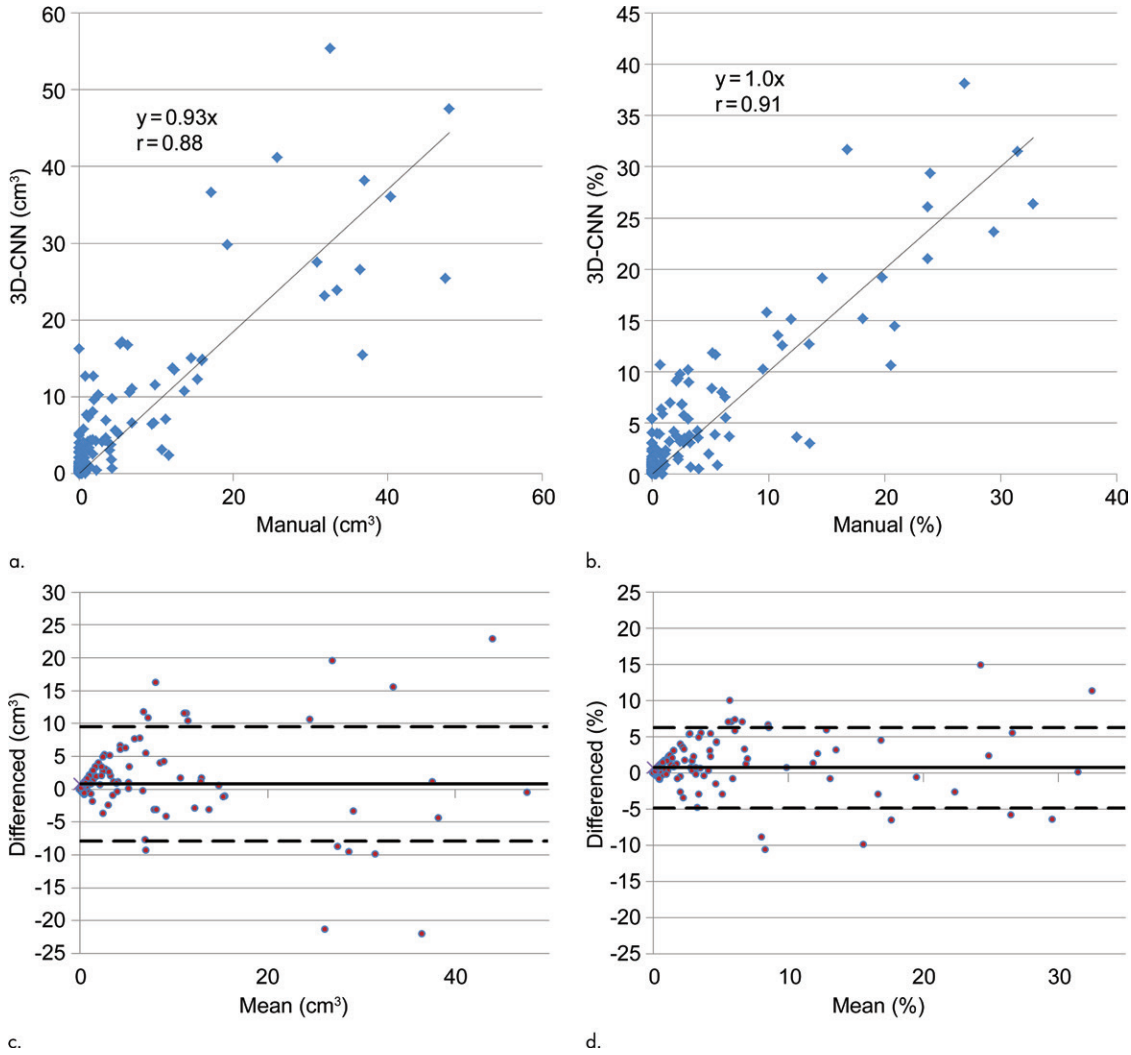

Figure 5: Evaluation of automatic scar segmentation. **(a,b)** Linear regression and **(c,d)** Bland-Altman plots of automatic segmentation versus manual segmentation of the scar volume (left column) and ratio of scar volume to total left ventricular myocardium volume (right column). The solid and dashed horizontal lines in **b** and **d** represent bias and limit of agreement lines, respectively.

patients (328 patients with scar; 38%) and a testing set of 207 patients (68 patients with scar; 33%) (Table 2). Stacks of short axial slices (mean, 11 images per patient ± 4; range, three to 30 images per patient; median, 11 slices) were acquired with 1.5-T MRI from three major vendors with different imaging protocols, including magnitude (n = 979) and phase-sensitive (n = 94) inversion recovery and 2D (n = 839) and 3D (n = 234) acquisitions (Table 2). Only magnitude images of phase-sensitive inversion recovery data were used for CNN training and testing.

The average computation time of 3D CNN–based segmentation was 0.15 second per volume (ie, three slices). Figure 3 shows segmentation results for all short axial slices (n = 6) from one patient, where the 3D CNN segmentation agrees with manual segmentation in slices with scar (slices 1–5) and a slice without scar (slice 6). The 3D CNN and manual segmentations also were comparable across different vendors (Fig 4). Scar volumes quantified by 3D CNN (mean, 4.3 cm^3 ± 8.8) strongly correlated with manually quantified scars (3.5 cm^3 ± 8.8) (y = 0.93×, r = 0.88; P < .001) (Fig 5a). The %LGE quantified by 3D CNN (mean, 3.3% ± 6.4) also strongly correlated with manually quantified %LGE (mean, 2.6% ± 6.1) (y = 1.0×, r = 0.91; P < .001) (Fig 5b). Bland-Altman analysis showed agreement between automated and manually quantified scar volumes (mean, 0.82 cm^3 ± 8.8) and %LGE (mean, 0.73% ± 5.6) (Fig 5c, 5d). In all data sets across different sites and vendors, 3D CNN and manual scar quantifications were strongly correlated (volume, r > 0.82; %LGE, r > 0.90) (Table 3). Furthermore, 3D CNN identified 12 patients with large scars (defined as %LGE ≥15% [3]) with an accuracy of 98% (202 of 207) (95% confidence interval [CI]: 95%, 99%), a sensitivity of 83% (10 of 12) (95% CI: 51%, 97%), and a specificity of 98% (192 of 195) (95% CI: 95%, 99%) (Table E1 [online]).

The 3D CNN segmentation demonstrated good similarity to manual segmentation, with good DSC scores (mean, 0.54 ± 0.26; interquartile range, 0.41–0.69) averaged across all testing data sets, sites, and vendors (Table 3). Exceptions included sites 4 (mean DSC, 0.77 ± 0.00) and 5 (mean DSC, 0.41 ± 0.25), and

Table 3: Performance of 3D CNN versus 2D CNN Categorized by Cardiac MRI Vendor and Site

	Correlation Coefficient Volume*		Correlation Coefficient %LGE*		Dice Similarity Coefficient†	
Vendor or Site	3D CNN	2D CNN	3D CNN	2D CNN	3D CNN	2D CNN
Vendor 1 (n = 50)	0.92 (0.82, 1.0) [0.93]	0.73 (0.67, 0.79) [0.95]	0.93 (0.85, 1.0) [0.94]	0.78 (0.69, 0.86) [0.92]	0.59 ± 0.18	0.53 ± 0.19
Vendor 2 (n = 73)	1.20 (1.0, 1.3) [0.91]	0.90 (0.80, 1.0) [0.90]	1.26 (1.1, 1.4) [0.91]	1.06 (0.93, 1.2) [0.88]	0.49 ± 0.12	0.48 ± 0.2
Vendor 3 (n = 84)	0.74 (0.63, 0.85) [0.82]	0.48 (0.39, 0.54) [0.87]	0.90, (0.81, 0.99) [0.90]	0.65 (0.54, 0.76) [0.78]	0.51 ± 0.21	0.47 ± 0.22
Site 1 (n = 81)	0.74 (0.63, 0.85) [0.82]	0.46 (0.40, 0.53) [0.84]	0.90 (0.81, 0.99) [0.90]	0.65 (0.56, 0.74) [0.83]	0.50 ± 0.18	0.46 ± 0.22
Site 2 (n = 48)	1.01 (0.88, 1.15) [0.91]	0.85 (0.72, 0.97) [0.89]	1.15 (1.0, 1.3) [0.92]	1.05 (0.89, 1.2) [0.88]	0.53 ± 0.15	0.56 ± 0.12
Site 3 (n = 12)	1.04(0.67, 1.4) [0.85]	0.53 (0.47, 0.59) [0.98]	1.00 (0.78, 1.2) [0.93]	0.65 (0.58, 0.72) [0.98]	0.67 ± 0.09	0.59 ± 0.06
Site 4‡ (n = 3)	0.90 (...) [0.95]	0.75 (...) [0.55]	0.90 (...) [0.96]	0.80 (...) [0.55]	0.77 ± 0.00	0.72 ± 0.00
Site 5 (n = 25)	1.60 (1.4, 1.8) [0.95]	1.10 (0.92, 1.24) [0.94]	1.65 (1.4, 1.9) [0.92]	1.01 (0.87, 1.3) [0.88]	0.41 ± 0.25	0.33 ± 0.26
Site 6 (n = 8)	1.02 (0.91, 1.1) [0.99]	0.78 (0.71, 0.86) [0.99]	1.16 (0.94, 1.4) [0.97]	0.88 (0.75, 1.0) [0.98]	0.61 ± 0.27	0.44 ± 0.25
Site 7 (n = 30)	0.82 (0.75, 0.90) [0.97]	0.79 (0.70, 0.87) [0.95]	0.86 (0.76, 0.95) [0.95]	0.78 (0.65, 0.90) [0.90]	0.55 ± 0.23	0.55 ± 0.20
All Sites (n = 207)	0.93 (0.86, 0.99) [0.88]	0.7 (0.64, 0.74) [0.85]	1.0 (0.94, 1.1) [0.91]	0.8 (0.74, 0.87) [0.83]	0.54 ± 0.26	0.48 ± 0.29

Note.—CNN = convolutional neural network. Vendor 1 = GE Medical Systems (Waukesha, Wis), vendor 2 = Philips Medical Systems (Best, the Netherlands), vendor 3 = Siemens Healthineers (Erlangen, Germany).

* Data are the slope, 95% confidence interval of the slope, and Pearson correlation coefficient of the linear regression model between automatic and manual scar quantifications, with all regression models (except site 4) having a $P < .001$.

† Data are mean ± standard deviation computed over cases with scar.

‡ Insufficient data (n = 3) for proper statistical analysis.

vendor 2 (mean DSC, 0.49 ± 0.12). On the other hand, 2D CNN segmentation showed moderate DSC scores (mean, 0.48 ± 0.29) averaged across all testing data sets, sites, and vendors (Table 3). Exceptions included sites 2, 3, 4, and 7 and vendor 1 (Table 3). The DSC was also comparable among image sets from the three vendors (P = .07). Table 3 summarizes 3D CNN performance metrics compared with 2D CNN for all data sets. The 3D CNN showed higher correlation with manually quantified %LGE (r = 0.91 vs r = 0.83, $P < .001$) and higher overall DSC (mean, 0.54 ± 0.26 vs 0.48 ± 0.29, P = .02) than did 2D CNN. Furthermore, 3D CNN versus manual quantification had a linear relationship with a significantly higher slope (ie, less underestimation) when compared with 2D CNN versus manual quantifications (volume, 0.93 vs 0.70; $P < .001$; %LGE: 1.0 vs 0.80, $P < .001$; n = 207). The regression slope for 3D CNN was significantly higher ($P < .001$) than that for 2D CNN for both scar volume and %LGE for all vendors and all sites except site 4.

Discussion

In this study, we present a three-dimensional (3D) convolutional neural network (CNN)-based method for automatic late gadolinium enhancement (LGE) segmentation in patients with hypertrophic cardiomyopathy (HCM). The proposed CNN was successfully trained and tested by using a manually segmented multicenter and multivendor LGE data set. After the offline training phase, fast segmentation of LGE images into background, healthy, and scarred myocardium categories was performed. A strong correlation was observed between 3D CNN and manual quantification of scar volume (r = 0.88, $P < .001$) and %LGE (r = 0.91, $P < .001$). Also, the method enabled accurate identification of patients at high risk of sudden cardiac death, with accuracy of 98% (202 of 207; 95% CI: 95%, 99%). Furthermore, our developed CNN showed accurate quantification in image sets acquired by using imaging protocols from different cardiac MRI centers and vendors.

Our 3D CNN consists of architecture with only three input slices and an overlapping sliding window to split large input stacks. This design has several advantages over 3D CNN with inputs of a patient's entire image stack. First, it reduces the memory requirements of the CNN model and allows maintenance of a large data set for training and testing. Furthermore, the segmentation reliability in our network potentially improves when several segmentations for the same slice are merged. Although this redundancy increases processing time, there is no tangible delay on our computational platform, where total segmentation time for each slice including the three repetitions is only 0.15 second. On the other hand,

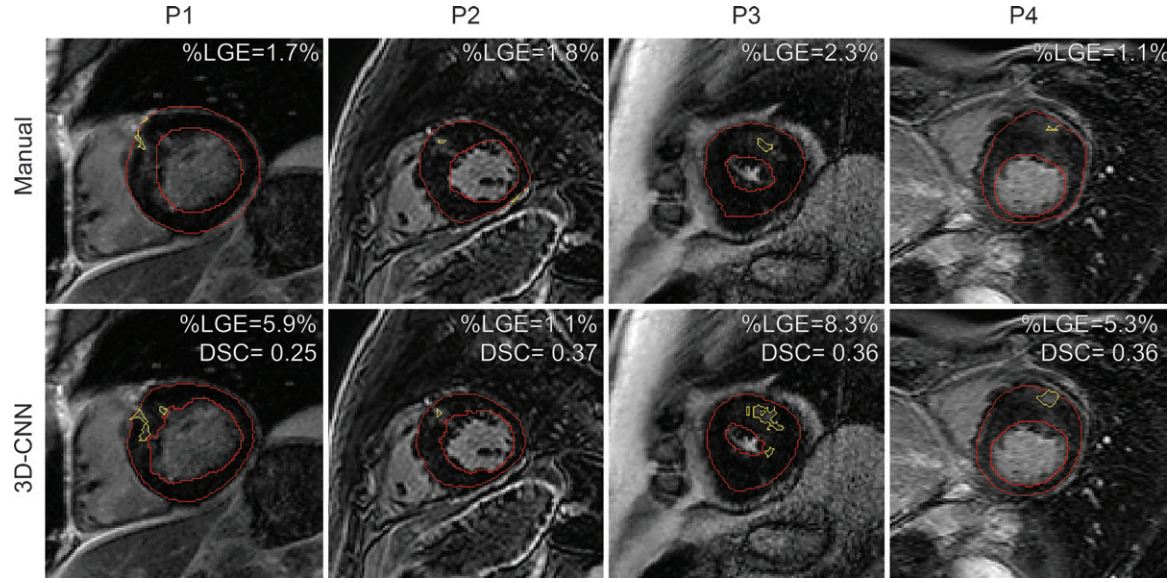

Figure 6: Three-dimensional (3D) convolutional neural network (CNN)-based segmentation (bottom row) and manual segmentation (top row) of cardiac MRI short-axis slices in four different patients (columns 1–4) with hypertrophic cardiomyopathy and with Dice score coefficient (DSC) values within the first quartile of DSC range (DSC < 0.41). Manual and automatic segmentation contours of the myocardium (red) and scar (yellow) regions are overlaid on late gadolinium enhancement images. %LGE = ratio of scar volume to total left ventricular myocardium volume.

3D CNNs trained using large volumes (eg, an entire stack of slices) enable processing of global contextual information for the whole heart. However, this advantage may not be effective in scar segmentation due to the vanishing spatial correlation between slices with increasing distance (eg, scarring in basal slices is more likely to have weak or no spatial correlation with scarring in the apical slices).

Deep 3D CNN shows a high capacity for learning complex HCM scar patterns from different imaging protocols and thus may fulfill the clinical need for accurate and standardized LGE analysis. Currently, there are no fully automated methods for HCM scar quantification. Several semiautomatic segmentation techniques have been proposed for LGE segmentation (24–27), where intensity threshold is combined with image feature analysis (24), spatial constraints (25), or region growing (26,27). Also, probabilistic modeling of myocardial and scar intensities and their spatial distribution have been proposed for LGE segmentation (28,29). These techniques are limited to 2D image analysis and require considerable effort to manually delineate the endo- and epicardium borders. Furthermore, current LGE analysis methods have mainly been developed for myocardial infarction segmentation, which may limit their utility in HCM (26) due to the differences in distribution, shape, and intensity (30,31).

Like all learning-based segmentation methods, the performance of our method is inherently biased toward the most abundant scar patterns in the training set. Thus, segmentation performance for large scars or apical HCM cases is expected to be suboptimal due to the low prevalence of these cases in the HCM population. Additionally, it might be necessary to use proper transfer learning (32) for generalizability of the utility of our network to non-HCM scars (eg, myocardial infarction).

The regression analysis showed that 3D CNN–based quantification of the scar burden strongly agrees with manual segmentation and outperforms 2D CNN. Although the DSC scores show that 3D CNN significantly outperforms 2D CNN, the majority of DSC scores for the network models had mean values less than 0.77. However, we note that low DSC scores may be present in cases where automated and manual scar segmentations are visually comparable (Fig 6). Small and irregular HCM scar patterns lead to substantial reductions in DSC in cases with slight shifts between automated and manual segmentations. Additionally, incorrect identification of clinically unimportant scars (eg, %LGE < 5%) further contributes to lower DSC averages.

Our study had several limitations. First, our training image sets were acquired over 10 years from 2001 to 2011. Thus, extending the utility of the developed method to more recent data sets may require further investigation to account for the change in image quality or newer LGE sequences. Also, we used manual scar segmentation from one expert reader as the ground truth, which is potentially biasing CNN training and assessment toward a specific grading style. Performance may be improved by training the network with multiple sets of contours from different experts. However, inconsistent multiple ground truths of the same image may also negatively affect training and lead to suboptimal performance. Further investigation of this approach is needed. We did not perform statistical tests to compare 3D CNN performance in terms of DSC among different cardiac MRI centers due to the relatively small number of cases with scar from each center. However, we did compare the segmentation accuracy among the different vendors, for which a relatively large number of cases with scar were available. Also, we did not study how performance could be affected by changing the CNN network architecture or

training parameters. Development of new network architectures and optimization of training parameters could potentially improve segmentation performance.

In conclusion, the presented three-dimensional convolutional neural network (CNN)-based MRI late gadolinium enhancement segmentation of the myocardium allows fast and accurate quantification of myocardial scarring in patients with hypertrophic cardiomyopathy from a multicenter image data set. Our three-dimensional method outperforms two-dimensional CNN–based quantification and demonstrates consistent performance across data sets from different vendors.

Acknowledgment: The authors thank Jennifer Rodriguez, BA, for the editorial corrections.

Author contributions: Guarantor of integrity of entire study, R.N.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, A.S.F., R.H.C., E.J.R., R.N.; clinical studies, E.J.R., W.J.M., R.N.; statistical analysis, A.S.F., R.H.C.; and manuscript editing, all authors

Disclosures of Conflicts of Interest: A.S.F. disclosed no relevant relationships. U.N. disclosed no relevant relationships. R.H.C. disclosed no relevant relationships. E.J.R. disclosed no relevant relationships. W.J.M. disclosed no relevant relationships. M.S.M. disclosed no relevant relationships. R.N. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: institution holds patents with and receives royalties from Philips and Samsung. Other relationships: disclosed no relevant relationships.

References

1. Maron BJ, Gardin JM, Flack JM, Gidding SS, Kurosaki TT, Bild DE. Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA Study. Coronary Artery Risk Development in (Young) Adults. Circulation 1995;92(4):785–789.
2. Finocchiaro G, Papadakis M, Robertus JL, et al. Etiology of sudden death in sports: insights from a United Kingdom regional registry. J Am Coll Cardiol 2016;67(18):2108–2115.
3. Chan RH, Maron BJ, Olivotto I, et al. Prognostic value of quantitative contrast-enhanced cardiovascular magnetic resonance for the evaluation of sudden death risk in patients with hypertrophic cardiomyopathy. Circulation 2014;130(6):484–495https://doi.org/10.1161/CIRCULATIONAHA.113.007094.
4. Weng Z, Yao J, Chan RH, et al. Prognostic value of LGE-CMR in HCM: a meta-analysis. JACC Cardiovasc Imaging 2016;9(12):1392–1402.
5. O'Hanlon R, Grasso A, Roughton M, et al. Prognostic significance of myocardial fibrosis in hypertrophic cardiomyopathy. J Am Coll Cardiol 2010;56(11):867–874.
6. Adabag AS, Maron BJ, Appelbaum E, et al. Occurrence and frequency of arrhythmias in hypertrophic cardiomyopathy in relation to delayed enhancement on cardiovascular magnetic resonance. J Am Coll Cardiol 2008;51(14):1369–1374.
7. Bruder O, Wagner A, Jensen CJ, et al. Myocardial scar visualized by cardiovascular magnetic resonance imaging predicts major adverse events in patients with hypertrophic cardiomyopathy. J Am Coll Cardiol 2010;56(11):875–887.
8. Rubinshtein R, Glockner JF, Ommen SR, et al. Characteristics and clinical significance of late gadolinium enhancement by contrast-enhanced magnetic resonance imaging in patients with hypertrophic cardiomyopathy. Circ Heart Fail 2010;3(1):51–58.
9. White SK, Flett AS, Moon JC. Automated scar quantification by CMR: a step in the right direction. J Thorac Dis 2013;5(4):381–382.
10. Klem I, Heiberg E, Van Assche L, et al. Sources of variability in quantification of cardiovascular magnetic resonance infarct size—reproducibility among three core laboratories. J Cardiovasc Magn Reson 2017;19(1):62.
11. Suinesiaputra A, Bluemke DA, Cowan BR, et al. Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. J Cardiovasc Magn Reson 2015;17(1):63.
12. Fahmy AS, Rausch J, Neisius U, et al. Automated cardiac MR scar quantification in hypertrophic cardiomyopathy using deep convolutional neural networks. JACC Cardiovasc Imaging 2018;11(12):1917–1918.
13. Poudel RPK, Lamata P, Montana G. Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. Cham, Switzerland: Springer, 2017; 83–94.
14. Zheng Q, Delingette H, Duchateau N, Ayache N. 3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation. IEEE Trans Med Imaging 2018;37(9):2137–2148.
15. Deniz CM, Xiang S, Hallyburton RS, et al. Segmentation of the proximal femur from MR images using deep convolutional neural networks. Sci Rep 2018;8(1):16485.
16. Nie D, Wang L, Trullo R, et al. Segmentation of craniomaxillofacial bony structures from MRI with a 3D deep-learning based cascade framework. Mach Learn Med Imaging 2017;10541:266–273.
17. Gibson E, Giganti F, Hu Y, et al. Automatic multi-organ segmentation on abdominal CT with dense v-networks. IEEE Trans Med Imaging 2018;37(8):1822–1834.
18. Pezeshk A, Hamidian S, Petrick N, Sahiner B. 3-D convolutional neural networks for automatic detection of pulmonary nodules in chest CT. IEEE J Biomed Health Inform 2018;23(5):2080–2090.
19. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III. Cham, Switzerland: Springer International; 2015; 234–241.
20. Hussain Z, Gimenez F, Yi D, Rubin D. Differential data augmentation techniques for medical imaging classification tasks. AMIA Annu Symp Proc 2018;2017:979–984.
21. Pennell DJ. Ventricular volume and mass by CMR. J Cardiovasc Magn Reson 2002;4(4):507–513.
22. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 2004;11(2):178–189.
23. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1(8476):307–310.
24. Hsu LY, Natanzon A, Kellman P, Hirsch GA, Aletras AH, Arai AE. Quantitative myocardial infarction on delayed enhancement MRI. Part I: Animal validation of an automated feature analysis and combined thresholding infarct sizing algorithm. J Magn Reson Imaging 2006;23(3):298–308.
25. Tao Q, Milles J, Zeppenfeld K, et al. Automated segmentation of myocardial scar in late enhancement MRI using combined intensity and spatial information. Magn Reson Med 2010;64(2):586–594.
26. Flett AS, Hasleton J, Cook 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):150–156.
27. Karim R, Bhagirath P, Claus P, et al. Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late Gadolinium enhancement MR images. Med Image Anal 2016;30:95–107.
28. Lu Y, Yang Y, Connelly KA, Wright GA, Radau PE. Automated quantification of myocardial infarction using graph cuts on contrast delayed enhanced magnetic resonance images. Quant Imaging Med Surg 2012;2(2):81–86.
29. Ukwatta E, Arevalo H, Li K, et al. Myocardial infarct segmentation from magnetic resonance images for personalized modeling of cardiac electrophysiology. IEEE Trans Med Imaging 2016;35(6):1408–1419.
30. Choudhury L, Mahrholdt H, Wagner A, et al. Myocardial scarring in asymptomatic or mildly symptomatic patients with hypertrophic cardiomyopathy. J Am Coll Cardiol 2002;40(12):2156–2164.
31. Noureldin RA, Liu S, Nacif MS, et al. The diagnosis of hypertrophic cardiomyopathy by cardiovascular magnetic resonance. J Cardiovasc Magn Reson 2012;14(1):17.
32. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22(10):1345–1359.