Generalizable Framework for Atrial Volume Estimation for Cardiac CT Images Using Deep Learning With Quality Control Assessment

Musa Abdulkareem¹,²,³, Mark S. Brahier⁴, Fengwei Zou⁵, Alexandra Taylor⁶, Athanasios Thomaides⁷, Peter J. Bergquist⁷, Monvadi B. Srichai⁷, Aaron M. Lee¹,², Jose D. Vargas⁸,⁹ and Steffen E. Petersen¹,²,³,¹⁰

¹Barts Heart Centre, Barts Health National Health Service Trust, London, United Kingdom, ²National Institute for Health Research Barts Biomedical Research Centre, William Harvey Research Institute, Queen Mary University of London, London, United Kingdom, ³Health Data Research UK, London, United Kingdom, ⁴Georgetown University School of Medicine, Washington, DC, United States, ⁵Montefiore Medical Centre, Bronx, NY, United States, ⁶Northeastern University, Boston, MA, United States, ⁷MedStar Heart and Vascular Institute, Washington, DC, United States, ⁸Veterans Affairs Medical Center, Washington, DC, United States, ⁹Georgetown University, Washington, DC, United States, ¹⁰The Alan Turing Institute, London, United Kingdom

Objectives: Cardiac computed tomography (CCT) is a common pre-operative imaging modality to evaluate pulmonary vein anatomy and left atrial appendage thrombus in patients undergoing catheter ablation (CA) for atrial fibrillation (AF). These images also allow for full volumetric left atrium (LA) measurement for recurrence risk stratification, as larger LA volume (LAV) is associated with higher recurrence rates. Our objective is to apply deep learning (DL) techniques to fully automate the computation of LAV and assess the quality of the computed LAV values.

Methods: Using a dataset of 85,477 CCT images from 337 patients, we proposed a framework that consists of several processes that perform a combination of tasks including the selection of images with LA from all other images using a ResNet50 classification model, the segmentation of images with LA using a U-Net image segmentation model, the assessment of the quality of the image segmentation task, the estimation of LAV, and quality control (QC) assessment.

Results: Overall, the proposed LAV estimation framework achieved accuracies of 98% (precision, recall, and F1 score metrics) in the image classification task, 88.5% (mean dice score) in the image segmentation task, 82% (mean dice score) in the segmentation quality prediction task, and $R^2$ (the coefficient of determination) value of 0.968 in the volume estimation task. It correctly identified 9 out of 10 poor LAV estimations from a total of 337 patients as poor-quality estimates.

Conclusions: We proposed a generalizable framework that consists of DL models and computational methods for LAV estimation. The framework provides an efficient and robust strategy for QC assessment of the accuracy for DL-based image segmentation and volume estimation tasks, allowing high-throughput extraction of reproducible LAV measurements to be possible.

Keywords: CT, left atrium, left atrial volume, quality control, deep learning
INTRODUCTION

Cardiac computed tomography (CCT) is commonly used to evaluate pulmonary vein (PV) anatomy and left atrial (LA) appendage thrombus prior to catheter ablation (CA). These images can also be used for full volumetric LA measurement, as there is strong evidence of the association of LA volumes (LAV) with atrial fibrillation (AF) recurrence after CA (1, 2). Evidence also exists that suggests an association between LAV and successful PV isolation in CA of AF (3). The anatomic structure of LA is varied and complex, limiting precision of echocardiographic LAV measurements and making manual contouring and volume measurements of CCT time-consuming and user-dependent.

Recent popularity of the use of artificial intelligence (AI), a concept of developing and using computer algorithms with human-like intelligence to solve specific tasks, in healthcare is driven by the advances and accessibility to high-performance scalable computing equipment, prompting an expansion of research into new AI methods and algorithms. AI algorithms have the potential to interpret biomedical and healthcare data. These algorithms can be used to optimize care pathways, standardize clinical diagnosis and develop predictive models (4). The choice of the method, data quality and context associated with the way data is used are important for the success of an AI approach for a given problem. For example, deep learning (DL) algorithms, which are a set of techniques based on neural networks designed to achieve AI, are particularly suitable for medical imaging segmentation problems (5, 6). Medical image segmentation helps to quantify anatomical structures and produce quantities that are used to diagnose, monitor or prognosticate diseases. AI-based image segmentation tools are particularly useful in that they encourage standardization and consistency, reducing variability that characterizes manual human segmentation. They also permit high-throughput extraction of reproducible measurements that can be used for further analyses.

It is costly if image-derived variables that support clinical decisions in diagnosis and treatments are obtained from low quality or corrupted images, e.g., due to artifacts or noise introduced during image acquisition. Importantly, with the increasing use of DL models for medical image processing and analysis, it is pertinent that quality control (QC) mechanisms are in place to ensure cases where such models fail in deducing quantitative measures are flagged or identified. In clinical research, such failed cases can be processed manually or discarded altogether in subsequent statistical analysis. However, to the best of our knowledge, to date, QC has not been applied to LAV estimation obtained via DL approaches. Of course, LAV estimation machine learning (ML) methods have been proposed for some modalities (7–9) other than CT, but even these have only been developed with very limited amount of data (e.g., 56–58 subjects), meaning that, the generalizability of these methods have not been established. Importantly, in tasks like LAV estimation using DL in which image segmentation is only part of the process, the errors at the segmentation level can add up cumulatively in the computation of the volumes. QC assessment via visual inspection of each segmentation mask is time consuming and infeasible, particularly for large datasets. Thus, there is the need for a QC framework for DL-based LAV estimations that is generalizable to large datasets and can identify cases that are difficult for DL models to read or where the predictions of such models are inaccurate.

In this study, we present a novel fully automated framework for estimating LAV from CCT images. The framework consists of several processes that perform a combination of tasks including the selection of images with LA from all other images using a DL classification model, the segmentation of images with LA using a DL image segmentation model, the assessment of the quality of the image segmentation task, the estimation of LAV by integrating the interslice volumes computed using segmentation areas across neighboring slices and QC assessment. Existing automated DL methods only combine some of these essential processes and not all, making their practical utility and general adoption of such algorithms limited, particularly for large cohort studies. Using a dataset of 85,477 contrast-enhanced CCT images from 337 patients, we show that our approach provides an efficient and robust methodology for fully automated LAV estimation. The proposed approach is also equipped with a QC mechanism that allows LAV estimations to be flagged for rejection if it determines that the prediction is of poor quality. Thus, the framework allows high-throughput extraction of reproducible LAV measurements.

METHODS

The presentations related to DL in this paper follow the relevant aspects of the recommendation of the Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME) guidance (10). The overview of the proposed framework for estimating the volume of LA is shown in Figure 1 and consists of four main processes. The image selection process performs an image classification task for selecting images containing the LA from the set of all CT images acquired for a given patient. The image segmentation process obtains the segmentation mask from the selected images of the preceding process. The segmentation quality assessment process quantifies the quality of the segmentation task while the quality-controlled volume estimation process computes the LAV and quantifies the accuracy of that computation. Each of these processes is described in detail in subsequent subsections.

Data Acquisition and Analysis Tools

Study Population

The study included 337 patients who underwent catheter ablation for symptomatic, anti-arrhythmic medication-refractory atrial fibrillation. This retrospective observational study was performed based on analysis of CCT within a database from MedStar
TABLE 1 | Baseline characteristics of the patients.

| Baseline characteristics | Total (N = 337) |
|--------------------------|----------------|
| Age, years               | 63.2 (± 10.2)  |
| Men, n (%)               | 223 (66)       |
| BMI (kg/m²)              | 31.1 (± 7.0)   |
| LAV (ml)                 | 137.1 (± 46.4) |
| Paroxysmal AF, n (%)     | 219 (65)       |

*Values are numbers and percentage (%) of the variables (± standard deviation). AF: atrial fibrillation; BMI: body mass index; LAV: left atrial volume.*

Georgetown University Hospital in Washington, D.C. All patients underwent CCT for pre-operative assessment and gave written informed consent. Images were de-identified prior to analysis. This study was approved by the Georgetown University Institutional Review Board (STUDY-0400, approved 7/20/2017). A total of 85,477 images for all the 337 patients were available for the study. This dataset and its subsets are used for all the experiments (analysis, modeling, and evaluation) in this paper. The baseline characteristics of the studied patients are given in Table 1.

CT Acquisition

Pre-operative CCT was acquired using a high resolution 256-slice scanner (Brilliance iCT; Royal Philips; Amsterdam, Netherlands) with 0.625 mm detector collimation. Images were acquired using prospectively ECG-triggered CCT scanning at 40% of the cardiac cycle, capturing the chest from carina to diaphragm during the angiodynamic administration of contrast medium. The contrast protocol included administration of 60 mL iohexol (Omnipaque; GE Healthcare; Chicago, Illinois) intravenously at a rate of 5 mL/s with a kV range between 80 and 120.

Manual Segmentation of LA as Ground Truth Dataset

CCT scans were analyzed using the semi-automated post-processing program 3D Slicer (Version 4.11.0, http://www.slicer.org). The LAV was calculated based on manual segmentation using axial views: the LA was encircled on each 2D slice of the CCT scan superiorly to inferiorly by a reader with several years of experience. Care was taken to exclude the left atrial appendage and pulmonary veins. 3D Slicer calculates the area of each corresponding LA section and estimates the LAV based on the distance between adjacent planes.

Machine Learning Software Tools

All experiments were conducted in Python programming language with TensorFlow 2.0 Python API machine learning framework, available from https://www.tensorflow.org/ on a Nvidia Tesla M40 machine. The source code for the experiments can be found on the following GitHub repository: https://github.com/mabdulkareem/lav_volume_with_qc.

Image Classification Model

We consider three DL architectures that can be used for the image classification task of selecting images containing the LA from the set of all CT images acquired for a given patient. These architectures are based on the VGG (VGG16 and VGG19) (11) and the ResNet (ResNet50) (12) architectures. The configuration of the VGG16 and VGG19 architectures are given in Table 2.
TABLE 2 | Configuration of the VGG16 and VGG19 DL Architectures.

| VGG16 | VGG19 |
|-------|-------|
| **Input (224 × 224 one channel image)** | **conv2-64** |
| **conv2-64** | **conv2-64** |
| **maxpool (kernel size = 2, stride = 2)** | **maxpool (kernel size = 2, stride = 2)** |
| **conv2-128** | **conv2-128** |
| **conv2-128** | **conv2-128** |
| **maxpool (kernel size = 2, stride = 2)** | **maxpool (kernel size = 2, stride = 2)** |
| **conv2-256** | **conv2-256** |
| **conv2-256** | **conv2-256** |
| **maxpool (kernel size = 2, stride = 2)** | **maxpool (kernel size = 2, stride = 2)** |
| **conv2-512** | **conv2-512** |
| **conv2-512** | **conv2-512** |
| **conv2-512** | **conv2-512** |
| **maxpool (kernel size = 2, stride = 2)** | **maxpool (kernel size = 2, stride = 2)** |
| **global average pooling** | **global average pooling** |
| **fully connected layer (1 unit, sigmoid activation function)** | **fully connected layer (1 unit, sigmoid activation function)** |

The convolutional layer parameters are denoted as “conv2-(number of filters),” where “conv2” denotes 2D convolution operation and the height and width of the 2D convolution window is 2×2. For brevity, the ReLU activation function is not shown.

For each of these two VGG architectures, the pooling layer after the last convolutional layer of the network is followed by a global average pooling operation and then a dense layer of 1 unit (neuron) with the sigmoid nonlinear activation function applied. The sigmoid function is chosen given that we are attempting to solve a binary (two-class) classification problem, that is, determine the presence or the absence of LA in a given CCT image. Each of the convolutional layers uses rectified linear unit (ReLU) as its activation function.

The ResNet50 architecture consists of five stages. The first stage consists of one convolutional layer and each of the remaining four stages consists of several blocks of convolutional layers with skip connections. These blocks are referred to as identity blocks if they contain no convolutional layer in the skip connection path, otherwise, they are called as convolutional block. Similar to the VGG16 and VGG19 architectures in this work, the last convolutional layer of the ResNet50 architecture is followed by a global average pooling operation and then a dense layer of 1 unit (neuron) with the sigmoid nonlinear activation function applied and the outputs of the model yields a binary prediction (i.e., an image is with/without LA). The configuration of the ResNet50 architecture is given in Table 3. The number of parameters of the three classification models are given in Table 4.

For the training of the VGG16, VGG19, and ResNet50 classification models, the input CCT images were resized to 224×224. For generalizability of the architectures, we rotated the images up to ±10° and their intensities normalized as part of the on-the-fly data augmentation (which ensures a different set of images is used after each complete training epoch). The parameters (weights) of the models were randomly initialized.
and training proceeded for 60 complete epochs using a batch size of 128 images. The binary cross-entropy function was used as the loss function. The optimization method used was the Root Mean Squared Propagation (RMSProp) optimizer, with an initial learning rate of 1e-4, decreasing as shown in Table 5. If 24 epochs elapsed with no decrease in the loss function, training was set to cease and the weights from the best epoch (i.e., one that yielded the lowest loss) is restored as the model’s weights.

The CT image datasets of 150 patients (Dataset 1) of which with 15,961 images contain LA were used to train, validate and test the three models. Dataset 1 consists of 42,296 CT images in total. Of these, 15% of these images were selected and divided into two equal sets representing the validation and test sets. Specifically, the number of the training, validation and test (evaluation) images are 35,952, 3,172, and 3,172, respectively.

Of the 35,952 images that constitute the training set, the number of images with LA present and absent are 13,586 and 22,366, respectively. In order to address this data imbalance, we used weighted loss function. Let \( \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), \ldots, (x_N, y_N)\} \) denote a training set of \( N \) samples where \( x \) is the two-dimensional input image and \( y \in \{0, 1\}^C \) denote a binary one-hot encoded label with \( C = 2 \) in our case, then the weighted loss function is defined as:

\[
E_w(\theta) = -\frac{1}{N} \left[ \sum_{n=1}^{N} \lambda_0 T_0(x_n) y_n \log(\hat{y}_n(x_n, \theta)) + \lambda_1 \sum_{n=1}^{N} T_1(x_n) y_n \log(\hat{y}_n(x_n, \theta)) \right]
\]

where \( \theta \) denotes the trainable parameters of a model; \( \hat{y}_n(x_n, \theta) \) is the posterior probability obtained after the application of sigmoid activation function on the output layer of the model; \( T_0(x_n) \) and \( T_1(x_n) \) are functions that indicate whether image \( x_i \) belongs to class 0 or class 1, respectively (in our case, whether LA is absent or present in image \( x_i \), respectively); and \( \lambda_0 \) and \( \lambda_1 \) are weights that penalize the loss function for false negatives errors \( x_i \) that belongs to class 0 is misclassified as belonging to class 1) and false positive errors \( x_i \) that belongs to class 1 is misclassified as belonging to class 0), respectively. The weights, \( \lambda_0 \) and \( \lambda_1 \), can be computed using the following formula:

\[
\lambda_i = \frac{1}{k_i} \cdot \frac{N}{C}
\]

where \( k_i \) is the number of images in class \( i \). In our case, \( N = 35952 \); class 0 and class 1 are subgroups indicating the collection of samples where LA is absent and present, respectively; then, \( \lambda_0 = \frac{1}{13586} \times \frac{35952}{2} = 0.8037 \) and \( \lambda_1 = \frac{1}{11386} \times \frac{35952}{2} = 1.3231 \). In other words, the images contain LA (class 1) are weighted as being more valuable than those without LA (class 0) since the original Dataset 1 is skewed (i.e., images in class 0 are more than those in class 1).
| Table 5 | Learning rate adopted for training the image classification models with $lr = 1e-4$. |
|---|---|
| Epoch | Learning rate $lr$ |
| 1–10 | $0.5 \times lr$ |
| 11–15 | $0.2 \times lr$ |
| 16–20 | $0.1 \times lr$ |
| 21–25 | $0.05 \times lr$ |
| 26–30 | $0.02 \times lr$ |
| 31–35 | $0.01 \times lr$ |
| 36–40 | $0.006 \times lr$ |
| 41–45 | $0.002 \times lr$ |
| 46–50 | $0.001 \times lr$ |
| 51–60 | $0.001 \times lr$ |

**Image Segmentation Model**

The configuration of the UNet DL architecture (13) that is used for the image segmentation task is given in Table 6. The UNet architecture includes batch normalization following every convolutional layer to enhance robustness of the model [i.e., reduce sensitivity to initial parameters and learning rates (14)]. Also, “dropout” operation (15), dropping out 30% hidden neurons, is performed on the first six consecutive up-sampling convolutional layers of the architecture to avoid problems associated with model overfitting. The output of the model yields an image segmentation mask with the background pixels labeled 0 and the foreground pixels (region of the LA) labeled 1. The total number of parameters of the model is 29,812,034 out of which 29,800,514 parameters are trainable.

For the training of the UNet model for image segmentation, the input CCT images were resized to 224×224. The images were rotated up to ±30° and their intensities normalized as part of the on-the-fly data augmentation. The parameters of the models were randomly initialized and training proceeded for 60 complete epochs using a batch size of 64 images. The sparse categorical cross-entropy function was used as the loss function. The optimization method used was the Adam optimizer, with an initial learning rate of 1e-4. If 24 epochs elapsed with no decrease in the loss function, training was set to cease and the weights from the best epoch is restored as the model’s weights. The optimization method used was the Adam optimizer, with an initial learning rate of 1e-4, decreasing as shown in Table 7.

For the training set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), \ldots, (x_N, y_N)\}$ of $N$ samples where $x$ is the two-dimensional input image and $y$ denote the two-dimensional segmentation mask, then the sparse categorical cross-entropy loss function is defined as:

$$E_w (\theta) = -\frac{1}{N} \sum_{n=1}^{N} y_n \log (\hat{y}_n (x_n, \theta))$$

where $\theta$ denotes the trainable parameters of a model and $\hat{y}_n (x_n, \theta)$ is the posterior probability obtained after the application of “softmax” activation function on the output layer of the model.

The CT image datasets of 150 patients in Dataset 1 with LA present (15,961) were used to train, validate and test the UNet models for the segmentation of LA. Of these, 15% of these images were selected and divided into two equal sets representing the validation and test sets. The training, validation and test images are 13,567, 1,197, and 1,197, respectively.

**Image Segmentation Quality Assessment**

One of the methods that have been proposed to address QC for AI-based segmentation tasks is the so-called Reverse Classification Accuracy (RCA) method (16, 17). The RCA algorithm attempts to predict the dice similarity coefficient (DSC), or dice score, using the segmentation mask in the absence of the ground truth. DSC is a measure of similarity between the label and predicted segmentation masks. Given two sets (two images in this case) $A$ and $B$, DSC score can be expressed as follows:

$$DSC = \frac{2|A \cap B|}{|A| + |B|}$$

where $|A|$ and $|B|$ represent the cardinalities of set $A$ and $B$ (i.e., the number of elements in each set), respectively. The DSC, which has a range of $[0,1]$, is a useful summary measure of spatial overlap that can be applied to quantify the accuracy in image segmentation tasks. Rather than using one test image to validate a segmentation model, it can be used as a statistical validation metric by computing the DSC of several images from a segmentation model and evaluating the mean DSC, allowing reproducibility and comparison of the model with other models.

For the purpose of this study, we will consider the segmentation model’s performance as good if the mean DSC is $>0.8$ on evaluation dataset.

The RCA algorithm consists of 3 main steps: Step 1 involves the image registration (rigid registration) of some $N$ number of (moving) images against the $k^{th}$ (fixed) image to obtain a set of $N$ transformation matrices; Step 2 involves using the transformation matrices from Step 1 to warp the segmentation masks of the $N$ images—thereby obtaining a set of $N$ warped segmentation masks; and Step 3 involves the prediction of the dice score. Thus, for a given mask, say mask $m$, from the output of the DL segmentation model, a set of warped masks are computed using the transformation matrices obtained in Step 1; these warped masks of mask $m$ are then compared with the all warped segmentation masks obtained in Step 2 as ground truth in order to obtain the set of dice scores. The maximum dice score in the set of dice scores is assumed to be the predicted dice score—which is often an estimate that is good.

In our segmentation quality assessment approach, we modified the RCA method to improve its prediction accuracy—we call this the modified RCA (mRCA) method. Specifically, rather than using one fixed image in Step 1, we used several fixed images—thereby, building an atlas-like knowledge base. From $M = 704$ images from $p = 5$ patients (with patients selected at random) containing LA and their segmentation masks, starting from the first image of these stacked layers of images, we selected images at 25 slice intervals—a set of $K = 29$ images in total—for obtaining the transformation matrices and warped segmentation masks. In particular, the transformation matrices $\{T_1, T_2, \ldots, T_N\}$ and warped segmentation masks $\{w_1, w_2, \ldots, w_N\}$ were computed by selecting each of these 29 images as the $k^{th}$ image used as the...
fixed image in the rigid registration process of Step 1 and the remaining 28 images as the moving images—iterating through all the 29 images in the set. In total, we have $N = 812$ (i.e., $29 \times 28$) transformation matrix and warped segmentation mask pairs in Step 2. Thus, in Step 3, in order to predict the dice score of mask $m$, the 812 transformation matrices are used to warp $m$ and each of these warped masks is compared with its corresponding warped segmentation mask to obtain a list of 812 dice scores $D = \{d_1, d_2, \ldots, d_{812}\}$. The maximum dice score in this list, that is $\max(D)$, is the predicted dice score for mask $m$.

The $M = 704$ images from $p = 5$ patients is a subset of the entirety of images (1,315). This subset is the complete selection of images in which LA is present. We find using several images across several patients and across several slices in the mRCA method built using 29 images superior in dice score prediction accuracy than using 100 images built with the RCA method in
the original paper. Figure 2 shows a schematic description of the mRCA method.

**Volume Estimation and Quality Control**

The estimation of LAV involves the summation of the interslice volumes where each interslice volume is estimated by computing the approximate volume between two consecutive slices using the following formula:

\[ V = \sum_{i}^{N} v_i \]  

(5)

where \( V \) is the estimated LAV, \( N \) is the number of slices, and \( v_i \) is the \( i \)th interslice volume and is defined as follows:

\[ v_i = \frac{A_i + A_{i+1}}{2} z_{k_i} \]  

(6)

where \( z_{k_i} \) is the distance between two consecutive slices in the \( z \) direction and is given by \( z_{k_i} = z_i \hat{k} \), where \( \hat{k} \) is the unit vector in the \( z \) direction and \( z_i \) is the perpendicular distance between two consecutive parallel slices; \( A_i \) and \( A_{i+1} \) are the areas of the two consecutive slices. That is, \( v_i \) is the arithmetic mean of the areas of two consecutive slices multiplied by the distance between them in the direction of the \( z \)-axis. Also, \( A_i \) can be obtained as follows:

\[ A_i = n_i s_x s_y \]  

(7)

where \( n_i \) is the number of pixels that constitutes the LA area on the \( i \)th slice; \( s_x \) and \( s_y \) are the pixel dimensions in the \( x \) and \( y \) directions, respectively; and \( A_i = n_i s_{xz} \) where \( s = s_x = s_y \).

In order to assess the quality of volume estimation, it is tempting to try to check the quality of the segmentation mask of each slice using the mRCA technique described in the preceding section but computational time increases significantly in doing so. Because neighboring slices are often similar, we propose to assess the quality of slices at specific slice intervals. For example, in order to assess the quality of LAV estimation and given that neighboring slices are often similar, we propose to assess the quality of slice at 8 slice intervals using the following two steps:

**Step 1** Estimate the segmentation quality by computing the set of dice scores, \( D \), at 8 slice intervals using mRCA technique; that is, \( D = \{d_1, d_2, \ldots, d_{8}\} \).

**Step 2** Compute the following:

- **QC Score Mean**, \( q_{\text{mean}} \): the arithmetic mean of \( D \).
The scores \( q_{\text{mean}} \), \( q_{\text{perc}} \), and \( q_{\text{no}} \) can then be used as a QC mechanism by the imposition certain conditions on them in order to determine whether to accept or reject an estimated LAV computation.

In addition, in order to compare the predictions of the RCA and mRCA methods against the actual DSC scores, the mean absolute error (MAE) and mean squared error (MSE) may be used. The MAE and MSE are defined as follows:

\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |e_i| \\
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} e_i^2
\]

where error \( e_i = y_i - \hat{y}_i \), and \( y_i \) and \( \hat{y}_i \) represent the label and predicted values, respectively.

**RESULTS**

**Image Segmentation Model With Quality Assessment**

The evaluation of the predicted results of the UNet model was performed using the DSC score, a measure of similarity between the label and predicted segmentation masks, as the performance metric. For the 1,197 evaluation set, the mean DSC is 0.885 (± 0.12 standard deviation; 25, 50, and 75% percentiles are 0.88, 0.93, and 0.95, respectively; the maximum is 0.98). The distribution of the dice score is given in the histogram and boxplot of Figure 4. Figure 5 displays some examples of the predictions of the segmentation model. Overall, the performance of the segmentation model was good (mean DSC > 0.8).

Figure 6 shows some examples of image segmentation quality assessment. For an evaluation dataset of 2,000 test images randomly selected from 15,961 images with LA of Dataset 1, Table 10 shows the mean dice scores computed for the actual DSC and the predicted DSC with RCA and mRCA methods. The table also shows the MAE and MSE values comparing the predictions of these methods against the actual DSC scores. These metrics indicate that mRCA has smaller errors than the RCA method. The image classification process, the image segmentation process and the prediction of the dice score for an image takes 805 ms with RCA method and 4,385 ms with mRCA method.

**Volume Estimation With Quality Assessment**

The regression and the Bland-Altman plots of the label volumes vs. the predicted volumes computed for the study population of 337 patients using the proposed framework are shown in Figure 7. Figure 7A shows the regression plot with volume computed using Equations (5–7) before QC assessment was implemented. In our case, \( s_x = s_y \) and \( z_i \) is approximately between 2.2 and 4.5 mm. Thus, without QC, the mean volume of the population is 137.38 ± 45.58 ml. In order to determine whether an LAV estimate is “poor” quality for a given patient using the scores \( q_{\text{mean}} \), \( q_{\text{perc}} \), and \( q_{\text{no}} \), we express the condition to be satisfied as follows:

\[
((q_{\text{mean}} < 0.6) \ \text{AND} \ (q_{\text{perc}} < 60) \ \text{AND} \ (q_{\text{no}} < 4)) \ \text{AND}
((q_{\text{mean}} < 0.65) \ \text{OR} \ (q_{\text{perc}} < 20) \ \text{OR} \ (q_{\text{no}} < 3))
\]

where AND and OR are logical operators. If this condition is not satisfied, the estimated LAV value is of “good” quality.
Abdulkareem et al. Framework for Atrial Volume Estimation

**FIGURE 3** Some examples of the prediction of the classification models. The top, middle, and bottom rows correspond to the predictions of the VGG16, VGG19, and ResNet50 models, respectively. The red arrows point to the LA. The middle column shows images with LA that the classification models got wrong.

**FIGURE 4** The histogram (A) and the boxplot (B) of the dice score of the evaluation dataset. The 25, 50, and 75% percentiles are 0.88, 0.93, and 0.95, respectively.

**Figure 7B** shows the regression plot after the imposition of this condition. With QC, the mean volume of the population is 138.38 ± 45.58 ml. The degree of linear correlation between the label and predicted quality-controlled volumes, $R^2$, is 0.864 for the study population of 337 patients without QC. With QC, $R^2$ is 0.968. Bland-Altman plots without and with QC are shown in **Figures 7C** and **7D**, respectively. Also, 95% of the difference between the label and predicted volume values will be between $-34.15$ and $+35.09$ of the mean volume difference value of 0.47 for **Figure 7C**, and between $-16.89$ and $+15.71$ of the mean volume difference value of $-0.59$ for **Figure 7D**. In addition, **Figure 8** shows (a) the regression plot and (b) the kernel density estimate plot along the label and predicted LAV marginal distributions. It also shows (c) the regression plot along the two variables' boxplots and (d) the kernel density and histogram plots of the variables with the dashed vertical
Abdulkareem et al. Framework for Atrial Volume Estimation

FIGURE 5 | Some examples of the variable prediction of the segmentation model (the red and blue contours represent the label and prediction and the corresponding dice scores are shown on the top of each image).

lines representing the arithmetic mean of the distributions. The plots of Figures 8A–D correspond to the label volumes vs. the predicted volumes without quality control (QC) and the plots (Figures 8E–H) correspond to these variables using the proposed QC framework. The estimation of LAV with and without QC took ∼74.25 s per patient and 23.32 s per patient, respectively. Table 11 shows five examples of the predicted volumes using the proposed framework.

DISCUSSION

Summary of Findings

In our methodology, we developed three DL classification models namely, VGG16, VGG19, and ResNet50. We selected ResNet50 model as the basic classification model for our framework as it outperforms the other two models on all the metrics.

For the image segmentation model, the overall performance for the 1,197 evaluation set is 88.5% based on the average DSC score which is close to accuracy of 87.2% reported in Dormer et al. (18) which involved the segmentation of the four cardiac chambers from CT images. The median (50% percentile) dice score of the proposed 2-dimensional UNet image segmentation model on the evaluation set is 0.93. Similarly, a recent effort in Chen et al. (19) that uses a more complex 3-dimensional UNet DL architecture for the segmentation of LA has also been shown to have a performance of 0.93 median dice score.

The RCA method (17) uses 100 images for obtaining the transformation matrix and warped segmentation mask pairs whose number increases with the number of images. In comparison, the mRCA method we proposed uses 29 images selected from a stack of 704 images by selecting images at 25 slice intervals. The choice of $p = 5$ patients and the selection of images at 25 slice intervals was to keep the sizes of the transformation matrices and the warped segmentation masks as small as possible (which reduces the time for computing the predicted DSC score) and at the same time ensure that the selected images are representative of the complex geometry of LA. This slice interval of 25 used for the mRCA method could be reduced further which, in turn, will increase the number of images and improve the accuracy of DSC score prediction at the expense increase in computation time. The mRCA method (mean dice score of 82%) for predicting DSC score shows far better performance compared to the RCA method (mean dice score of 52%) on testing dataset of 2,000 images but that was at the expense of increase in computation time. The process of classifying an image as containing LA, segmenting LA from the image and predicting its dice score takes 805 ms with RCA method while it takes 4,385 ms with mRCA method.

The degree of linear correlation between the label and predicted quality-controlled volumes, $R^2$, is 0.968 for the study population of 337 patients. As neighboring slices are similar, for LAV estimation with QC, we have assessed the segmentation...
quality of slices at eight slice intervals for computing \( q_{\text{mean}}, q_{\text{perc}}, \) and \( q_{\text{no}} \). The number of slice intervals to estimate dice scores was chosen to be eight and that has worked for our volume computation problem where the interslice distance ranges from 2.2 to 4.5 mm. In practice, the number of slice intervals could be varied depending on the interslice distance settings of the acquisition machine. Generally, a small value could improve quality assessment at the expense of increase in computation time of LAV estimates.

The scores \( q_{\text{mean}}, q_{\text{perc}}, \) and \( q_{\text{no}} \) were used to develop a QC mechanism by imposing certain conditions on them in order to determine whether to accept or reject an estimated LAV computation. The approach correctly identified 9 out of 10 poor LAV estimations from a total of 337 patients as LAV estimation of poor quality (missing 1 out of 10). Of course, this condition may have worked for our task but, in practice, the constraint of this condition could be made more stringent or relaxed to suit a given problem or dataset. Moreover, the estimation of LAV without QC took \( \sim 23.32 \) s per patient. With QC, the estimation took 74.25 s per patient.

In addition, in relation to the Figures 7A,B, apart from the cases of the very small volumes (four out of nine) that would have been identified by simple thresholding even without the proposed quality control approach, the main reason that explains why our QC mechanism flagged these nine cases are that these patients have anatomical abnormalities that were unfamiliar to the ML models (that is, the dataset used for the training do not contain many of these abnormal cases and, therefore, the model does not generalize to these cases). Such abnormalities include large left atrial appendage, atypical atrial dimensions and geometries, and mispositioning of the pulmonary arteries.

For the classification and segmentation models, we have used the 150 patients (Dataset 1) and, for the volume computation, we have used the whole of the 337 patients. We note here that, in an attempt to make the classification and segmentation models more generalizable, we trained these models further using transfer learning approach (i.e., by training the models further but fixing certain part of the model weights) with the 151–337 datasets but these subsequent models were not superior to the initial models trained on Dataset 1. Furthermore, since model training was not carried out during volume estimation (that is,
Abdulkareem et al. Framework for Atrial Volume Estimation

FIGURE 7 | The regression (A,B) and the Bland-Altman (C,D) plots of the label volumes vs. the predicted volumes using the proposed framework. The data points are shown in blue and green in panel (A); the green data points (9 out of 337) in the figure have been flagged by the QC condition as poor LAV estimates therefore eliminated in panel (B). The red data point in panel (B) indicates one poor volume estimate (out of a total of 10) was missed by the imposed condition. In panels (C,D), the upper and lower dashed horizontal lines represent the confidence interval at 95%.

for example, we are not fitting a regression-like model where the input variables are the images and the output variable is the volume), it was meaningful to compare the computed volumes for the whole dataset of 337 patients using our approach with the results obtained by manual computation from an expert. In other words, our deep learning models are oblivious to the values of the volumes for all the 337 patients, and we are able to show that our approach is able to predict the volumes with high accuracy.

Researchers have successfully used DL for the segmentation and estimation of quantifiable measures of the different structures of the heart from CT images (18, 20). This set of algorithms include DL segmentation models which can find certain images challenging to segment owing to the considerable variation of cardiac anatomy among individuals. Importantly, these algorithms are not fully automated to compute quantifiable measures from the output of segmentation tasks. Algorithms that do, such as those given in Chen et al. (21), used ad hoc approaches that do not include mechanisms to detect or handle cases where image segmentation exercise fails which will inevitably affect the results of the computed quantities or the results of further analysis that are based on such quantities. This can be observed, for example, in our estimation of LAV in this study where DL without QC mechanism failed to correctly compute 10 out 337 LAV values. Practically, this could be costly in clinical decision support for diagnosis and treatments.

There are LAV estimation ML methods that have been proposed for other modalities. For example, automated volume estimation methods for the heart ventricles for cardiac magnetic resonance (CMR) images, such as Bayesian method (7), method of linear support vector machines (8), and method of multi-scale deep belief networks and regression forests (9), have been proposed that are capable of estimating the volumes directly when given the set of images containing LA and without the image classification or segmentation processes. Despite the promise of these approaches, they have only been developed with very limited amount of data [e.g., 56 subjects in each of Wang et al. (7) and Zhen et al. (9) and 58 subjects in Afshin et al. (8)], meaning that, the generalizability and evaluation of the methods to large datasets have not been reported.

QC of medical images involves identifying those images with problems that resulted from the image acquisition process, image reconstruction algorithm or even in a step of the image processing pipeline. QC tasks are important to ensure the usability of these images. QC assessment via visual inspection
FIGURE 8 | The top row shows (A) the regression plot and (B) the kernel density estimate plot along the two variables (label volume and predicted volume) marginal distributions. It also shows (C) the regression plot along the two variables’ boxplots and (D) the kernel density estimate and histogram plots of the variables with the dashed vertical lines representing the arithmetic mean of the distributions. The symbols $\rho$ and $p$ represent the p-value and Pearson correlation coefficient, respectively. The plots (A–D) correspond to the label volumes vs. the predicted volumes without quality control (QC) and the plots (E–H) correspond to these variables using the proposed QC framework.

TABLE 11 | Examples of label and predicted LAV with QC scores for 5 selected patients with No. 1–4 being of good quality and No. 5 flagged as being of poor quality.

| No. | Label volume | Predicted volume | $q_{\text{mean}}$ | $q_{\text{perc}}$ | $q_{\text{no}}$ |
|-----|--------------|------------------|-------------------|-----------------|-------------|
| 1   | 146.6        | 146.8            | 0.87              | 96              | 22          |
| 2   | 130.8        | 128.9            | 0.80              | 75              | 6           |
| 3   | 141.7        | 143.1            | 0.83              | 90              | 18          |
| 4   | 113.7        | 114.4            | 0.78              | 89              | 8           |
| 5   | 108.04       | 185.08           | 0.62              | 50              | 5           |

is highly subjective, time consuming, and infeasible in large-scale image acquisition or analysis settings. Methods that have been developed for QC assessment of medical images include the estimation of image noise level (22, 23), random forest supervised learning approaches (24, 25), simple label voting scheme (26), probabilistic schemes (27, 28), and Bayesian neural networks learning (29–31). Many of these methods are inherently specific to the tissues being segmented or to the modalities used for image acquisition.

With advancement in the applications of AI in medical imaging, it is important to have QC methods that can assess the quality of the analyses being carried out by AI techniques. Current studies on this subject have only focused on the image segmentation part of the process. Researchers, such as in Valindria et al. (16), Robinson et al. (17), Biasiolli et al. (32), and Hann et al. (33) reduce the problem to developing classifiers for estimating Dice Similarity Coefficient (DSC), one of the common metrics for image segmentation accuracy, in the absence of manual segmentation as ground truths. The classifiers are often trained using random forests (16, 32) or convolutional neural network (16, 33) on a fixed number of space (k-space) of segmentation features with cases where ground truths are already known. In Robinson et al. (17), QC of image segmentation task of CMR images was proposed. The method used the RCA technique (16, 17) to predict the DSC score. QC is carried out by classifying the segmentation mask of an image as good quality if the accuracy estimate is above a pre-defined threshold value, and as poor quality otherwise.

The existing DL algorithms dealing with automated segmentation of CT images do not come with segmentation
QC mechanisms. This is an important setback in practice as it makes the algorithms unsuitable for fully automated analysis tasks in clinical applications. Moreover, in a problem like LAV estimation in which segmentation is part of the process and not itself the end goal of the analysis, the errors at the segmentation level add up cumulatively in the computation of the volumes. Thus, a QC method for DL-based segmentation models that can identify cases that are difficult or inaccurate as presented in this paper will be immensely useful.

Limitation
Our models were trained on data from a single hospital system, limiting their generalizability. Data augmentation may have improved the chance that performance may not deteriorate significantly from datasets from elsewhere. It would therefore be worthwhile assessing the performance of these models using an external dataset. Accordingly, the performance of the models may then be improved, if needed, with training on homogenous multi-center datasets. In addition, the framework we have proposed is not limited by data from a single hospital. We believe our approach is generalizable for building models with multi-center datasets with little or no modification. Importantly, the cases that are flagged as inaccurate LAV estimates in our approach could be interpreted as samples where there is little dataset on which the models have learnt from. The issue can then be addressed by providing training data that covers those areas of interest. This approach could well-prove useful in the future given the current trend in AI research is toward developing data-centric AI algorithms (34).

Future Work
The mRCA could be further optimized to reduce computational time, possibly by reducing the sizes of the transformation matrices and the warped segmentation masks. The quality assessment method is not based on checking the segmentation quality of every slice, it should therefore be worthwhile exploring the optimal interslice interval that will reduce computational time without degrading the quality of LAV prediction. Although our framework focused on LAV estimation with QC for CT images, the framework can be adapted to estimate volumes of other tissues and for images from other modalities.

CONCLUSION
We proposed a novel comprehensive framework that consists of DL models and computational methods for LAV estimation. Although the individual DL model architectures in this work are not new, putting them together in a single framework with a novel QC mechanism as we have done is novel and more clinically useful. The framework provides an efficient and robust strategy for QC assessment of the accuracy for DL-based image segmentation and volume estimation tasks. It therefore efficiently fully automates the process of estimating LAV and is able to flag cases where it fails to accomplish high quality prediction, perhaps, better described as a framework that allows DL models to “grade their own homework.” The framework we have proposed in this paper will be highly appreciated by clinicians for carrying out LAV estimation, particularly for large cohorts and high-throughput applications, and practitioners involved in developing digital solutions around CCT image applications such as CCT image acquisition that will allow better analysis results.

DATA AVAILABILITY STATEMENT
The datasets presented in this article are not readily available because restrictions apply to the availability of these raw data, which were used under license for the current study from Georgetown University Institutional Review Board, and so are not publicly available. Generated anonymized dataset are however available from the authors upon reasonable request and with permission of Georgetown University Institutional Review Board. Requests to access the datasets should be directed to Jose D. Vargas, jose.vargas@nih.gov.

ETHICS STATEMENT
The studies involving human participants were reviewed and approved by Georgetown University Institutional Review Board (STUDY-0400, approved 7/20/2017). Dataset were fully anonymized. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS
MA, MB, JV, and SP conceived the idea and contributed to the analysis. MB, JV, FZ, ATa, ATh, PB, and MS developed the contouring method. MA led on the machine learning methodology and the main mathematical and statistical analysis of CT data and images. MB and JV advised on cardiac CT analysis and validation. MA, AL, and SP advised on data governance and computing infrastructure. MA drafted the first version of the manuscript. SP and JV provided overall supervision. All authors contributed to the content, the writing of the final version, and provided critical feedback.

FUNDING
MA and SP acknowledge support from the CAP-AI programme (led by Capital Enterprise in partnership with Barts Health NHS Trust and Digital Catapult and funded by the European Regional Development Fund and Barts Charity) and Health Data Research UK (HDR UK—an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities; www.hdruk.ac.uk). SP acknowledges support from the National Institute for Health Research (NIHR) Biomedical Research Center at Barts, from the SmartHeart EPSRC programme grant (www.nihr.ac.uk; EP/P001009/1) and the London Medical Imaging and AI Center for Value-Based Healthcare. SP has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 825903 (euCanSHare project).
REFERENCES

1. Njoku A, Kannabhiran M, Arora R, Reddy P, Gopinathannair R, Lakshkreddy D, et al. Left atrial volume predicts atrial fibrillation recurrence after radiofrequency ablation: a meta-analysis. Europace. (2018) 20:33–42. doi: 10.1093/europace/euy13

2. Hof I, Chilukuri K, Arbab-Zadeh A, Scherr D, Dalal D, Nazarian S, et al. Does left atrial volume and pulmonary venous anatomy predict the outcome of catheter ablation of atrial fibrillation? J Cardiovasc Electrophysiol. (2009) 20:1005–10. doi: 10.1111/j.1540-8167.2009.01504.x

3. Abecasis J, Dourado R, Ferreira A, Saraiva C, Cavaco D, Santos KR, et al. Regional assessment of cardiac biventricular volumes with an adapted bayesian formulation. IEEE Trans Biomed Eng. (2014) 61:1251–60. doi: 10.1109/TBME.2014.229943

4. Abdulkareem M, Petersen SE. The promise of AI in detection, diagnosis and epidemiology for combating COVID-19: beyond the hype. Front Artif Intell. (2021) 4:53. doi: 10.3389/frai.2021.652660

5. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. J Cardiovasc Electrophysiol. (2015) 16:3381–90. doi: 10.1111/jce.12768

6. Robinson R, Valindria VV, Liu C-M, Chang S-L, Chang PY-C, Chen W-S, Pan Y-M, et al. Learning-based quality control for cardiac MR images. IEEE Trans Med Imaging. (2018) 38:1127–38. doi: 10.1109/TMI.2018.287509

7. Lakkireddy D, et al. Left atrial volume predicts atrial fibrillation recurrence after radiofrequency ablation: a meta-analysis. J Cardiovasc Electrophysiol. (2015) 16:536–48. doi: 10.1111/jce.12768

8. Afshin M, Ben AI, Punithakumar K, Law M, Islam A, Goela A, et al. Automated cardiac volume assessment and cardiac long-and-short-axis imaging plane prediction from ECG-gated CT volumes enabled by deep learning. Eur Hear J Digit Heal. (2021) 2:311–22. doi: 10.1093/ehjdh/ztab033

9. Johnson KW, Shameer K, Glicksberg BS, Readhead B, Sengupta PP, Björkegren JLM, et al. Enabling precision cardiology through multiscale biology and systems medicine. JACC Basic to Transl Sci. (2017) 2:311–27. doi: 10.1016/j.jacbts.2016.11.010

10. Doerr P, Arbel T, Carneiro G, Syeda-Mahmood T, Tavares JMR, Moradi M, et al. Deep learning in medical image analysis and multimodal learning for clinical decision support. In: Third International Workshop, DLMIA 2017 and 7th International Workshop, ML-CDS 2017 Held in Conjunction with MICCAI. Quebec City, QC: Springer (2017).

11. Chen Z, Rigolli M, Vigneault DM, Kligerman S, Hahn L, Narezkina A, et al. Automated cardiac volume assessment and cardiac long-and-short axis imaging plane prediction from ECG-gated CT volumes enabled by deep learning. Eur Hear J Digit Heal. (2021) 2:311–22. doi: 10.1093/ehjdh/ztab033

12. Dormer JD, Ma L, Halicek M, Reilly CM, Schreibmann E, Fei B. Heart chamber segmentation from CT using convolutional neural networks. In: Medical Imaging 2018: Biomedical Applications in Molecular, Structural, Functional Imaging. 1057825, Houston, TX (2018).

13. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. Cham (2015). p. 234–41.

14. Roy AG, Conjeti S, Navab N, Wachinger C. Inherent brain segmentation quality control from fully convnet monte carlo sampling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Shenzhen (2019). p. 119–27.

15. Kriegstein AK, Wozny J, Alber M, et al. Automated identification of the aorta in cine CMR can significantly accelerate processing of the UK Biobank population data. PLoS One. (2019) 14:e0212272. doi: 10.1371/journal.pone.0212272

16. Hann E, Papescu IA, Zhang Q, Gonzales RA, Baruçu A, Neubauer S, et al. Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac MRI T1 mapping. Med Image Anal. (2021) 71:100229. doi: 10.1016/j.media.2021.100229

17. Robinson R, Valindria VV, Liu C-M, Chang S-L, Chang PY-C, Chen W-S, Pan Y-M, et al. Automated extraction of left atrial volumes from two-dimensional computer tomography images using a deep learning technique. Int J Cardiol. (2020) 316:272–8. doi: 10.1016/j.ijcard.2020.03.075

18. Zhen X, Wang Z, Islam A, Bhaduri M, Chan I, Li S. Direct estimation of atrial volume and pulmonary venous anatomy in isolation in catheter ablation of atrial fibrillation. Europace. (2009) 11:1289–94. doi: 10.1093/europace/eup198

19. JohnsonKW, Shameer K, Glicksberg BS, Readhead B, Sengupta PP, Björkegren JLM, et al. Enabling precision cardiology through multiscale biology and systems medicine. JACC Basic Transl Sci. (2017) 2:311–27. doi: 10.1016/j.jacbts.2016.11.010

20. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng. (2017) 19:221–48. doi: 10.1146/annurev-bioeng-071516-044442

21. Abdulkareem M, Petersen SE. The promise of AI in detection, diagnosis and epidemiology for combating COVID-19: beyond the hype. Front Artif Intell. (2021) 4:53. doi: 10.3389/frai.2021.652660

22. Wang Z, Ben SM, Gu R, Islam A, Goela A, Li S. Direct estimation of cardiac bi-ventricular volumes with regression forests. Med Image Comput Comput Assist Interv. (2014) 17:586–93. doi: 10.1007/978-3-319-04070-6_73

23. Sengupta PP, Shrestha S, Berthon B, Messas E, Donal E, Tison GH, et al. Proposed requirements for cardiovascular imaging-related machine learning evaluation (PRIME): a checklist: reviewed by the American College of Cardiology Healthcare Innovation Council. Cardiovasc Imaging. (2020) 13:2017–35. doi: 10.1016/j.crcim.2020.07.015

24. Simonny K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:14091556 [arXiv Preprint] (2014).

25. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition. (2015). p. 770–8. Available online at: http://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

26. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham (2015). p. 234–41.

27. Ioffe S, Szegedy C. Batch normalization: accelerating deep net work training. In: International Conference on Computer Vision. San Diego, CA (2015). p. 448–56.

28. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. (2014) 15:1929–58.

29. Valindria VV, Lavdas I, Bai W, Kamnitas K, Aboagye EO, Rockall AG, et al. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE Trans Med Imaging. (2017) 36:1597–606. doi: 10.1109/TMI.2017.2665165

30. Valindria VV, Lavdas I, Bai W, Kamnitas K, Aboagye EO, Rockall AG, et al. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE Trans Med Imaging. (2017) 36:1597–606. doi: 10.1109/TMI.2017.2665165

31. Robinson R, Valindria VV, Liu C-M, Chang S-L, Chang PY-C, Chen W-S, Pan Y-M, 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. doi: 10.1186/s12968-019-0523-x.
34. Motamedi M, Sakharnykh N, Kaldewey T. A data-centric approach for training deep neural networks with less data. arXiv211003613 [arXiv Preprint] (2021).

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Abdulkareem, Brahier, Zou, Taylor, Thomaides, Bergquist, Srichai, Lee, Vargas and Petersen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.