A fully automated pipeline for the extraction of pectoralis muscle area from chest computed tomography scans

Daniel Genkin [1], Alex R. Jenkins [2], Nikki van Noord [2], Kalysta Makimoto [3], Sophie Collins [4], Michael K. Stickland [4], Wan C. Tan [5], Jean Bourbeau [6,7], Dennis Jensen [2,6,7,8] and Miranda Kirby [3]

[1]Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, Canada. [2]Clinical Exercise and Respiratory Physiology Laboratory, Department of Kinesiology and Physical Education, McGill University, Montreal, Canada. [3]Department of Physics, Toronto Metropolitan University, Toronto, Canada. [4]Department of Medicine, University of Alberta, Edmonton, Canada. [5]Center for Heart, Lung Innovation, University of British Columbia, Vancouver, Canada. [6]Montreal Chest Institute of the Royal Victoria Hospital, McGill University Health Centre, Montreal, Canada. [7]Respiratory Epidemiology and Clinical Research Unit, Research Institute of McGill University Health Centre, Montreal, Canada. [8]Translational Research in Respiratory Diseases Program, Research Institute of the McGill University Health Centre, Montreal, Canada.

Corresponding author: Miranda Kirby (miranda.kirby@torontomu.ca)

Shareable abstract (@ERSpublications)
A fully automated pipeline for the extraction of pectoralis muscle area measurements from low-dose chest CT was developed that was highly accurate and externally validated https://bit.ly/3R1XgD5

Cite this article as: Genkin D, Jenkins AR, van Noord N, *et al*. A fully automated pipeline for the extraction of pectoralis muscle area from chest computed tomography scans. *ERJ Open Res* 2024; 10: 00485-2023 [DOI: 10.1183/23120541.00485-2023].

Received: 13 July 2023
Accepted: 9 Nov 2023

Abstract

Background Computed tomography (CT)-derived pectoralis muscle area (PMA) measurements are prognostic in people with or at-risk of COPD, but fully automated PMA extraction has yet to be developed. Our objective was to develop and validate a PMA extraction pipeline that can automatically: 1) identify the aortic arch slice; and 2) perform pectoralis segmentation at that slice.

Methods CT images from the Canadian Cohort of Obstructive Lung Disease (CanCOLD) study were used for pipeline development. Aorta atlases were used to automatically identify the slice containing the aortic arch by group-based registration. A deep learning model was trained to segment the PMA. The pipeline was evaluated in comparison to manual segmentation. An external dataset was used to evaluate generalisability. Model performance was assessed using the Dice–Sorensen coefficient (DSC) and PMA error.

Results In total 90 participants were used for training (age 67.0±9.9 years; forced expiratory volume in 1 s (FEV_1) 93±21% predicted; FEV_1/forced vital capacity (FVC) 0.69±0.10; 47 men), and 32 for external testing (age 68.6±7.4 years; FEV_1 65±17% predicted; FEV_1/FVC 0.50±0.09; 16 men). Compared with manual segmentation, the deep learning model achieved a DSC of 0.94±0.02, 0.94±0.01 and 0.90±0.04 on the true aortic arch slice in the train, validation and external test sets, respectively. Automated aortic arch slice detection obtained distance errors of 1.2±1.3 mm and 1.6±1.5 mm on the train and test data, respectively. Fully automated PMA measurements were not different from manual segmentation ($p>0.05$). PMA measurements were different between people with and without COPD ($p=0.01$) and correlated with FEV_1 % predicted ($p<0.05$).

Conclusion A fully automated CT PMA extraction pipeline was developed and validated for use in research and clinical practice.

Introduction

COPD is a lung condition characterised by airflow limitation. Recent studies show extrapulmonary comorbidities, such as involuntary loss of muscle mass (sarcopenia), begin to manifest in at-risk individuals [1] and have been linked with all-cause mortality [2]. In people diagnosed with COPD, low fat-free muscle mass (FFM) measurements are associated with impaired exercise tolerance [3] and increase the risk of mortality, irrespective of pulmonary function [4]. However, FFM measurements are not typically acquired as part of COPD clinical care.

Computed tomography (CT) images are routinely collected in at-risk individuals who smoke as part of lung cancer screening trials [5] and in people with pulmonary abnormalities [6]. The pectoralis muscle area (PMA) is a measure that can be extracted from chest CT, and has been correlated with FFM [7] and associated with prognosis in a variety of populations, such as at-risk for [8] and with COPD [9], and interstitial lung disease (ILD) [10]. The established methodology to extract PMA measurements is: 1) manually identify the slice containing the aortic arch; and 2) perform semi-automated segmentation of the pectoralis at that slice [7–10]. Although the PMA has been manually extracted in large cohorts such as COPD Genetic Epidemiology study (COPDGene) [11] and ECLIPSE (Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points [12]) [7–9, 13], and is a target biomarker for unexplored cohorts, such as CanCOLD (Canadian Cohort of Obstructive Lung Disease [14]), it is impractical to perform segmentation in routine clinical care as the process is time-consuming and introduces intra/interobserver variability. Therefore, a fully automated method to extract the PMA is required for use in routine clinical care.

Convolutional neural networks (CNNs) are deep learning-based tools that are capable of generating predictions using image inputs. Specifically, the U-Net model is a commonly used CNN architecture that specialises in image segmentation and produces highly accurate segmentation results as compared to manual segmentations [15, 16]. Multiple studies have trained such models to automatically extract body composition measurements in abdominal CT with high accuracy [17, 18]. However, few studies have developed an automated method to quantify the PMA from chest scans [19, 20], and no studies have performed both automatic identification of the aortic arch and pectoralis segmentation.

Therefore, our primary objective was to develop a fully automated pipeline using the CanCOLD study of people with or without primarily mild COPD and test its performance in an external group of research study participants with COPD [21]. To accomplish this objective, we aimed to: 1) create an algorithm to automatically identify the aortic arch slice; and 2) train a deep learning model to automatically segment the pectoralis muscle in the identified slice. Our secondary objective was to investigate the discriminative and concurrent validity of the automated PMA measurements by quantifying differences between subgroups and correlations to pulmonary function test variables respectively.

Materials and methods

Dataset

CanCOLD is a multi-centre cohort of ∼1800 participants aged ⩾40 years originally sampled through random dialling across nine sites in Canada [14]. A subset of 90 participants (10 per site) from the baseline visit were randomly selected for the development of the automated pipeline; the participants maintained an equal ratio of males/females, people with/without COPD, and smokers/never-smokers. Informed written consent from the CanCOLD participants and approval by institutional review boards was obtained at each site. All demographics, pulmonary function and chest CT scans were retrospectively collected at baseline. CT analysis of all CanCOLD participants has been previously reported [22–29]; however, the prior studies dealt with development of lung CT measurements, whereas in this manuscript we report development of muscle CT measurements. A group of people with COPD studied at the University of Alberta (UofA) (iNO-NCT03679312) were used to externally test the pipeline [21]. The UofA cohort derives from a single-centre interventional study of people 18–85 years old that focused on examining the effects of inhaled nitric oxide on exercise capacity in COPD patients. Participants were excluded if there was no evidence of left and/or right pectoralis minor and major muscle at the level of the aortic arch.

Pulmonary function measurements

Spirometry was performed according to American Thoracic Society (ATS) guidelines [30–32] for measurement of the postbronchodilator forced expiratory volume in 1 s (FEV_1) and forced vital capacity (FVC) [13]. The residual volume/total lung capacity (RV/TLC) ratio was measured using whole-body plethysmography.

Image acquisition

Chest CT images were acquired at full inspiration using various systems across all CanCOLD sites with the following parameters: 120 kVp, 1.0- or 1.25-mm slice thickness, and 0.52–0.90 mm^2 pixel spacing. The external test set from the UofA was acquired using a similar protocol: 120 kVp, 0.75 mm slice thickness, and 0.55–0.75 mm^2 pixel spacing. A standard or soft tissue kernel was used for image reconstruction. All key acquisition parameters from both training and test images are listed in supplementary table E1.

Ground truth pectoralis muscle segmentation

Two observers (A.R. Jenkins, a postdoctoral student with 1 year of experience, and N. van Noord, a PhD student with 1 year of experience) performed ground truth pectoralis muscle segmentations (45 per observer) using the Chest Imaging Platform extension in 3D Slicer (https://cip.bwh.harvard.edu/index.html) [7]. Briefly, the observers manually identified the axial slice containing the aortic arch. The pectoralis muscle was then segmented five slices below to five slices above the aortic arch using a manual brush tool, creating 11 slices/participant (990 slices total) for deep learning model training. Both observers then repeated segmentations from a random subset of 15 participants to quantify intra- and interobserver variability. In the external dataset, each observer segmented the pectoralis muscle at the aortic arch slice from 16 CT scans (32 slices total) for testing.

Aortic arch slice determination

Aorta segmentation

All proceeding pipeline development was performed in MATLAB R2021a (v. 9.10). First, all CT volumes were resized to a resolution of 1 mm^3 using bicubic interpolation. The lungs were then automatically segmented using thresholding and volumes cropped to localise the top of the heart. In the remaining slices, contrast was rescaled and images were smoothed of high-frequency artefacts using a median filter. Ray casting was used to extract the heart cavity between the lungs. The heart muscle was then automatically segmented using K-means segmentation (n=4 centres) and cleaned of peripheral blood vessels, creating 90 aorta atlases. A visual representation of the aorta segmentation process is shown in supplementary figure E1. The observers also manually delineated the aortic arch from the aorta segmentations to create 90 aortic arch atlases using the “VolumeSegmenter” application, as visualised in supplementary figure E2. A more detailed description is provided in the online supplementary material.

Aortic arch detection

The aortic arch was detected in the aorta segmentations using an image-registration-based approach. A candidate aorta segmentation was compared to all aorta atlases using a cross-correlation of their voxel counts per slice, as shown in supplementary figure E3. The five most correlated atlas aortas were registered to the candidate using an affine algorithm. Resulting warp matrices were used to register the corresponding aortic arch atlases to the candidate. The top-most slice from the consensus volume, created using a majority vote of the warped atlases, was labelled as the top of the aortic arch.

Deep learning model

Image preprocessing

The input images and ground truth segmentations were resized to a matrix dimension of 256×256 for model training using bicubic and nearest neighbour interpolation, respectively. Any pixels below −190 HU and above 150 HU were set to those thresholds, and the result was rescaled to [0,1].

Network architecture

A diagram of pipeline development during training and evaluation is shown in figure 1. The 2D U-Net architecture was selected for the deep learning model [15]. The convolutional path for the model contained four down-sampling and four up-sampling blocks. A 1×1 kernel convolution followed by sigmoid activation was used to create pectoralis muscle probability maps. At the start of each epoch, data augmentation was applied to mimic natural variations in pectoral anatomy. 10-fold cross-validation was implemented during training, with convergence occurring once validation performance failed to improve compared to the previous epoch. A more detailed listing of the hyperparameters is available in supplementary table E2. The model was trained using an NVIDIA RTX A2000 12 GB GPU.

Image postprocessing

A diagram of the complete automated pipeline is shown in figure 2. The resulting probability maps were resized back to their 1 mm^3 matrix size. A threshold of 0.4 was used to binarise the maps into pectoralis segmentations. Any objects <0.5 cm^2 or one-third of the mean area of all components was treated as noise and removed. As a final measure, the resulting segmentations were morphologically closed (structure=square, radius=2 mm). The PMA was defined as the total area of all remaining structures. The code used for the development of the complete pipeline is available upon request.

Statistical analysis

Statistical analyses were performed using MATLAB R2021a (v. 9.10). Group differences were evaluated using the Mann–Whitney U-test or the Kruskal–Wallis test (Tukey correction). Aortic arch detection error was assessed through a difference in millimetres between the slice predicted by the pipeline and the human observers. The Sorensen–Dice Coefficient (DSC) and Jaccard index are two metrics that are typically used

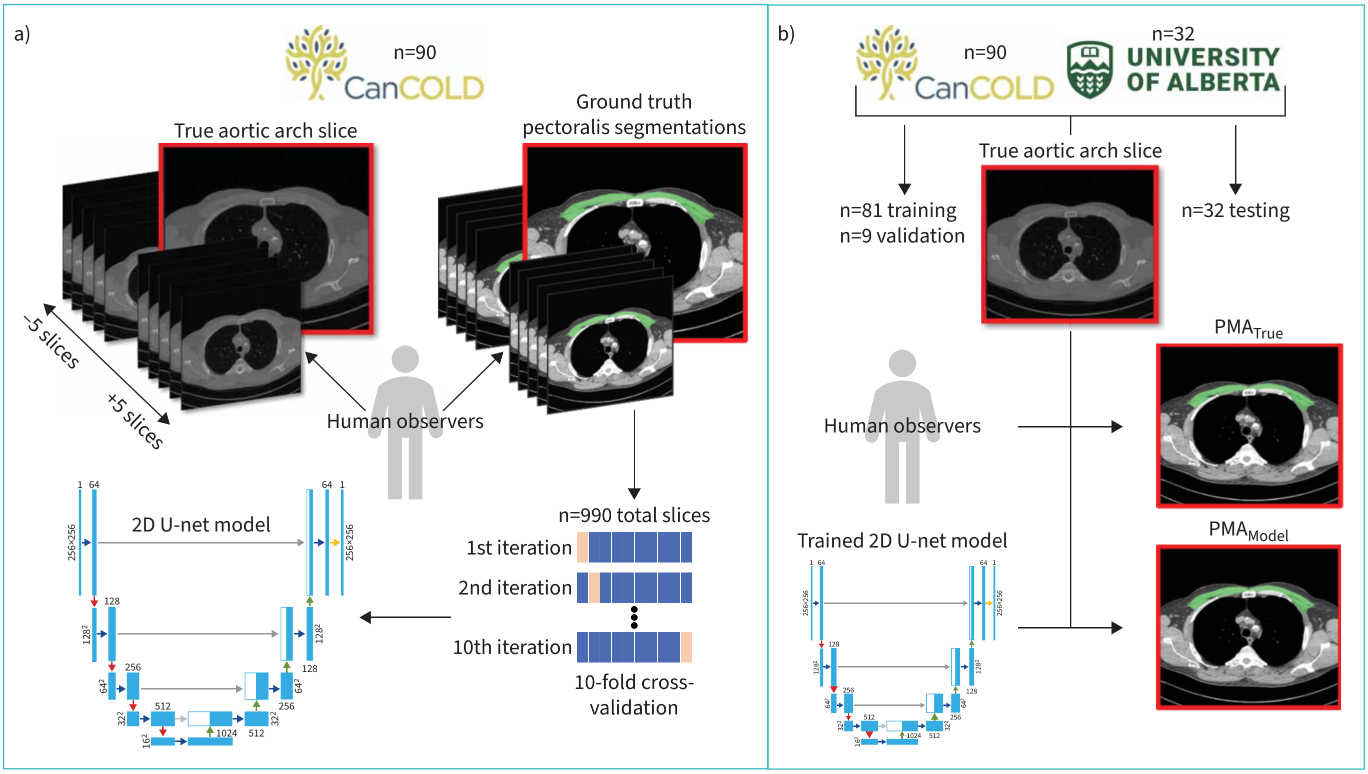

FIGURE 1 Deep learning model training and evaluation data flow diagram. a) Model training: 90 CanCOLD participants had the aortic arch slice labelled by human observers, and ground truth segmentations were performed between the five slices before and after the aortic arch slice (creating a total of 990 slices). 10-fold cross-validation was implemented during model training; b) Model evaluation: model performance was then strictly evaluated on the aortic arch slice labelled by the human observers on the 90 CanCOLD and 32 UofA participants. CanCOLD: Canadian Cohort Obstructive Lung Disease study; UofA: University of Alberta; PMA: pectoralis muscle area.

to evaluate predicted segmentation overlap with manual segmentation as the ground truth (0: no overlap; 1: complete overlap). Therefore, intra- and interobserver variability was evaluated using the DSC, Jaccard index and absolute (cm^2)/relative (%) PMA difference. Deep learning performance at the slice labelled by the observers (PMA$_{Model}$) and predicted by the pipeline (PMA$_{Pipeline}$) were compared to the ground truth (PMA$_{True}$) using the same metrics and a Bland–Altman analysis. Differences in the PMA$_{Pipeline}$ measurements for various subgroups (male *versus* female, body mass index (BMI) $>25\ \text{kg}\cdot\text{m}^{-2}$ *versus* BMI $\leqslant 25\ \text{kg}\cdot\text{m}^{-2}$, COPD *versus* no COPD) in the entire dataset (n=122) were quantified using the Mann–Whitney U-test or an ANCOVA to evaluate discriminant validity. Correlations between PMA$_{True}$ and PMA$_{Pipeline}$ with BMI, FEV$_1$ % predicted, FEV$_1$/FVC and RV/TLC were assessed using the Spearman's rank correlation coefficient to evaluate convergent validity. Statistical significance was evaluated at a type I error of 5%.

Results

Participant characteristics

A single participant was excluded from the UofA test cohort due to the absence of a pectoralis muscle, and no participants in either cohort had visible breast implants at the level of the aortic arch. The demographics and pulmonary function measurements for the training and testing cohorts are reported in table 1. Age, sex, race and BMI were similar between training and test sets ($p>0.05$). However, compared to the CanCOLD training cohort, the UofA test cohort had a greater history of smoke exposure, with evidence of airflow obstruction, and pulmonary gas trapping ($p<0.05$). The demographics and pulmonary function for participants stratified by Global Initiative for Obstructive Lung Disease (GOLD) stage [33] in the combined cohorts are shown in supplementary table E3.

Inter- and intraobserver variability

Supplementary table E4 shows the intra- and interobserver variability for the manual PMA measurements. The mean intra- and interobserver absolute slice difference between the aortic arch labels were 0.2±0.6 mm

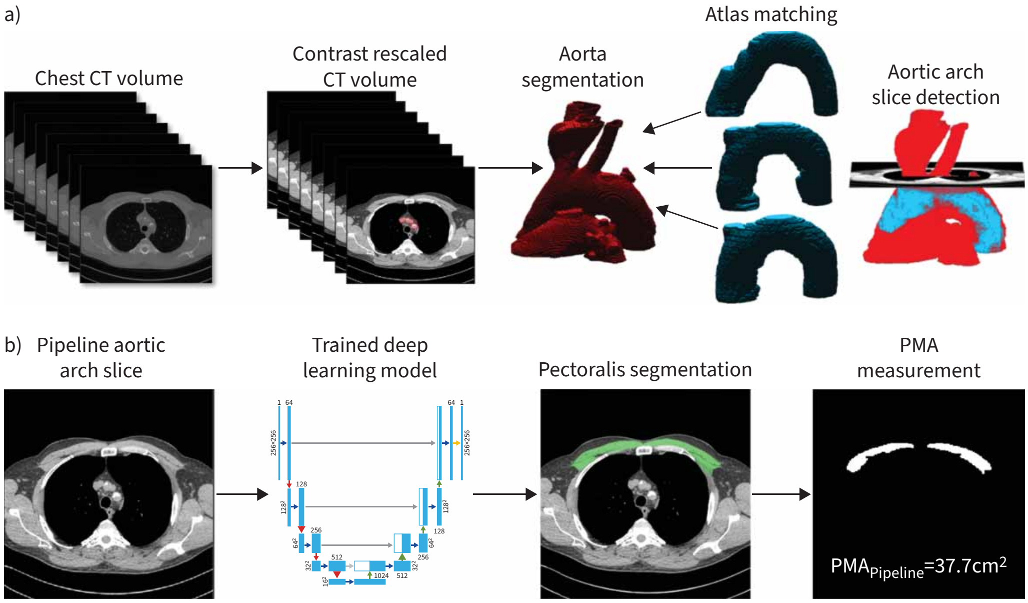

FIGURE 2 The complete fully automated pectoralis segmentation pipeline: a) Aortic arch slice detection: first, the aorta is automatically segmented from the contrast rescaled chest computed tomography (CT). Next, the five most matching atlas arches are registered to the segmented aorta. Majority voting between the five registered atlases is used to quantify the top of the aortic arch. b) Pectoralis muscle area (PMA) extraction: the CT slice identified in the earlier step is then fed into the trained deep learning model to obtain the pectoralis muscle segmentation and subsequent pectoralis muscle area measurement.

TABLE 1 Demographics and pulmonary function for the participants used during model training (CanCOLD) and external testing (UofA)

	CanCOLD	UofA
Participants n	90	32
Demographics		
Age years	67.0±9.9	68.6±7.4
Female sex	43 (48)	16 (50)
Caucasian	86 (96)	31 (97)
BMI kg·m^{-2}	27.7±4.6	27.9±5.7
Smoking pack-years	14.2±18.2	43.2±25.5$^{\#}$
COPD	44 (49)	32 (100)$^{\#}$
GOLD I	26 (29)	10 (31)
GOLD II+	18 (20)	22 (69)$^{\#}$
Pulmonary function		
FEV_1 L	2.5±0.7	1.7±0.5$^{\#}$
FEV_1 % pred	92.6±21.0	64.5±16.5$^{\#}$
FVC L	3.6±0.9	3.4±0.9
FVC % pred	100.9±17.5	99.5±17.9
FEV_1/FVC	0.69±0.10	0.50±0.09$^{\#}$
RV L	2.4±0.7	3.1±1.2$^{\#}$
TLC L	6.1±1.3	6.5±1.4
RV/TLC	0.40±0.10	0.47±0.13$^{\#}$

Data are shown as n (%) or mean±SD. CanCOLD: Canadian Cohort Obstructive Lung Disease study; UofA: University of Alberta; BMI: body mass index; COPD: chronic obstructive pulmonary disease; GOLD: Global Initiative for Obstructive Lung Disease; FEV_1: forced expiratory volume in 1 s; % pred: per cent predicted; FVC: forced vital capacity; RV: residual volume; TLC: total lung capacity. $^{\#}$: significantly different from CanCOLD ($p<0.05$).

and 0.5±0.6 mm, respectively. Intra- and intersegmentation overlap was 0.96±0.02 and 0.95±0.02 as measured by the DSC. The intra- and interabsolute (and relative) PMA error was 0.88±0.57 cm^2 (2.4 ±1.5%) and 1.50±0.97 cm^2 (4.3±2.5%), respectively.

Deep learning segmentation model performance

During training

All deep learning models trained using the data successfully reached the conversion criteria. The time for training the models for 10-fold cross-validation was 19.4±0.4 min per fold, with a total training time of 193.7 min. Convergence during training occurred at 6.1±1.3 epochs. Lastly, the final model mean DSC was 0.94±0.01 for both the training (n=891 slices) and validation (n=99 slices) sets across all 10-folds. The best performing model over the training and validation folds was chosen to proceed with final pipeline creation.

After postprocessing

Figure 3 visually represents the trained model's segmentation capabilities on three representative participants with varying sex and COPD status in all three training, validation and testing sets. The deep learning model performance for PMA$_{\text{Model}}$ measurements was compared to PMA$_{\text{True}}$ measurements on the single slice containing the aortic arch as labelled by the human observers (table 2). The model obtained an average DSC of 0.94±0.02 and 0.94±0.01 on the training and validation data, respectively. Absolute (and relative) PMA error was 1.33±1.42 cm^2 (4.1±4.1%) and 0.99±0.51 cm^2 (2.8±1.2%) on the training and validation data, respectively. The test data obtained slightly lower performance metrics, with a DSC of 0.90±0.04 and absolute (and relative) PMA error of 1.39±1.30 cm^2 (5.5±5.7%). Comparison between

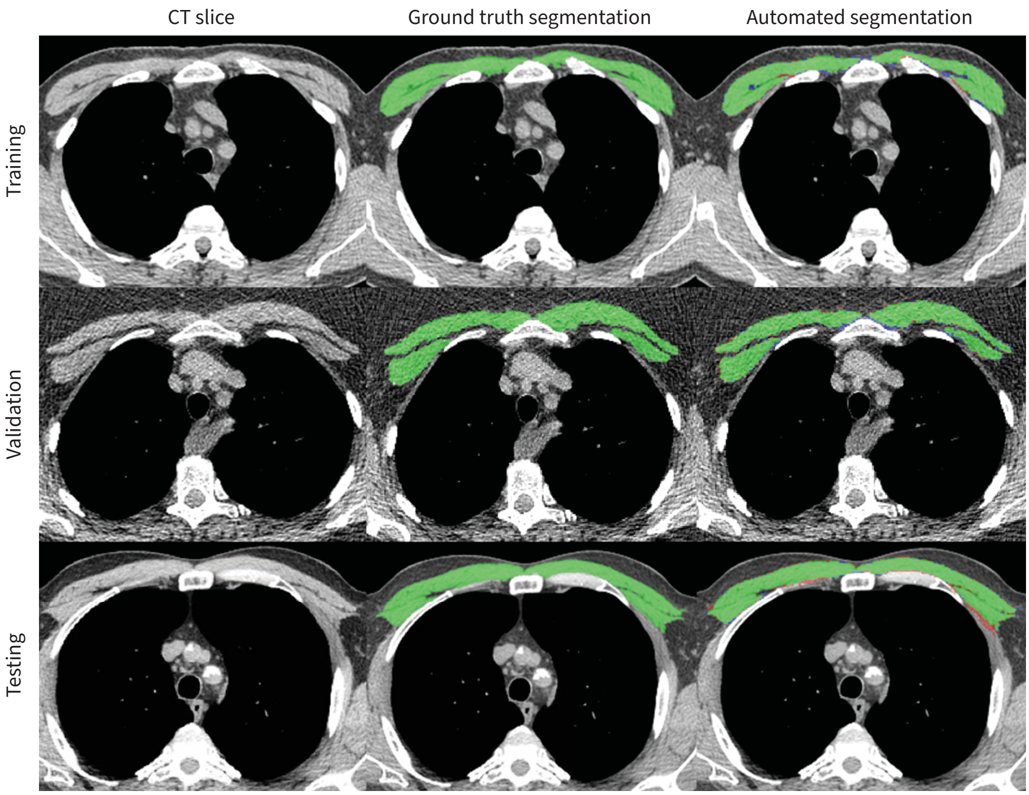

FIGURE 3 Deep learning model performance for an example participant from all three training, validation and testing sets. From left to right: contrast rescaled chest computed tomography (CT) containing the aortic arch as labelled by the human observers. Ground truth pectoralis muscle segmentation. Fully automated segmentation using the developed pipeline (red), ground truth segmentation (blue) and overlap between the two (green). Top row: 66-year-old male with COPD from the training cohort; Sorensen–Dice coefficient (DSC): 0.94, Jaccard: 0.89, pectoralis muscle area (PMA) error: 1.69 cm^2 (4.4%). Middle row: 67-year-old female without COPD from the validation cohort; DSC: 0.94, Jaccard: 0.89, PMA error: 1.16 cm^2 (3.2%). Bottom row: 62-year-old male with COPD from the test cohort; DSC: 0.93, Jaccard: 0.88, PMA error: 2.27 cm^2 (5.7%).

PMA_{Model} and PMA_{True} measurements in all participants (n=122) is shown in figure 4. The PMA_{Model} and PMA_{True} measurements were similar (p=0.94), highly correlated (ρ=0.98, p<0.01) and had negligible bias (b=0.49 cm^2, 95% confidence interval (CI) −3.1–4.1 cm^2). Individual training and test cohort comparisons, as well as model performance stratified by sex and CanCOLD centre ID, are shown in supplementary figure E4 and tables E5 and E6. We note that the pipeline performance was comparable between the sexes and between CanCOLD centres.

Complete pipeline performance

Supplementary table E7 describes the time taken to complete each step of the pipeline for the training and testing cohorts. The average absolute slice error for the aortic arch detection algorithm was 1.2±1.3 mm and 1.6±1.5 mm for training and testing sets, respectively. Performance for $PMA_{Pipeline}$ measurements using the automated detection algorithm in comparison to PMA_{True} measurements is also shown in table 2. The absolute (and relative) PMA error was 1.56±1.36 cm^2 (4.6±4.1%), 1.11±0.80 cm^2 (3.1±2.0%) and 1.50±1.43 cm^2 (6.2±6.8%) on the training, validation and test data, respectively. Comparison between $PMA_{Pipeline}$ and PMA_{True} measurements for all participants (n=122) is shown in figure 4. The $PMA_{Pipeline}$ and PMA_{True} measurements were similar (p=0.93), highly correlated (ρ=0.98, p<0.01), and had negligible bias (b=0.23 cm^2, 95% CI −3.7–4.2 cm^2). Figure E5 summarises comparisons between $PMA_{Pipeline}$ and PMA_{Model} measurements.

Group PMA differences: discriminant validity

Using the fully automated PMA measurements generated by the pipeline, a comparison between participants stratified by sex, BMI and COPD status is shown in figure 5. PMA measurements were significantly greater in males than females (p<0.01), people with BMI >25 $kg \cdot m^{-2}$ *versus* BMI ⩽25 $kg \cdot m^{-2}$ (p<0.01), and people without *versus* with COPD (p=0.01, after adjusting for age, sex and BMI). Supplementary figure E6 shows that PMA decreased significantly across the COPD continuum: from non-COPD to GOLD II+ (p<0.05, after adjusting for age, sex and BMI). Finally, PMA was significantly lower in people with a greater amount of airflow limitation and pulmonary gas trapping (p<0.05; supplementary figure E7).

PMA correlations: concurrent validity

Increased PMA was significantly correlated with increased BMI (PMA_{True}: ρ=0.18, p<0.05; $PMA_{Pipeline}$: ρ=0.21; p<0.05), increased FEV_1/FVC (PMA_{True}: ρ=0.37; $PMA_{Pipeline}$: ρ=0.32, p<0.05) and reduced RV/TLC (PMA_{True}: ρ=−0.33, p<0.05; $PMA_{Pipeline}$: ρ=−0.31, p<0.05) (table 3). Correlations between PMA with the rest of the variables and in various subgroups of the data are shown in supplementary tables E8 and E9.

Discussion

We developed a fully automated PMA extraction pipeline from chest CT images. Our deep learning model was able to segment the pectoralis muscle with high accuracy in both the validation and external test sets. Our aortic arch detection algorithm was accurate to <2 slices, and the PMAs generated by the complete pipeline on the predicted aortic arch slice was highly correlated, not significantly different and showed negligible Bland–Altman bias in comparison to the ground truths. Our fully automated PMA measurements were validated discriminately (by reporting significant differences between participants

TABLE 2 Deep learning model segmentation performance after postprocessing in comparison to the ground truth (PMA_{True}) for the single slice containing the aortic arch as labelled by the observers (PMA_{Model}) and as predicted by the aortic arch detection algorithm in the complete automated pipeline ($PMA_{Pipeline}$)

	DSC	Jaccard Index	Absolute PMA error cm^2	Relative PMA error %
PMA_{Model}				
Training (n=81)	0.94±0.02	0.88±0.04	1.33±1.42	4.1±4.1
Validation (n=9)	0.94±0.01	0.88±0.02	0.99±0.51	2.8±1.2
Testing (n=32)	0.90±0.04	0.82±0.06	1.39±1.30	5.5±5.7
$PMA_{Pipeline}$				
Training (n=81)	N/A	N/A	1.56±1.36	4.6±4.1
Validation (n=9)	N/A	N/A	1.11±0.80	3.1±2.0
Testing (n=32)	N/A	N/A	1.50±1.43	6.2±6.8

Data are shown as mean±SD. DSC: Sorensen–Dice coefficient; PMA: pectoralis muscle area.

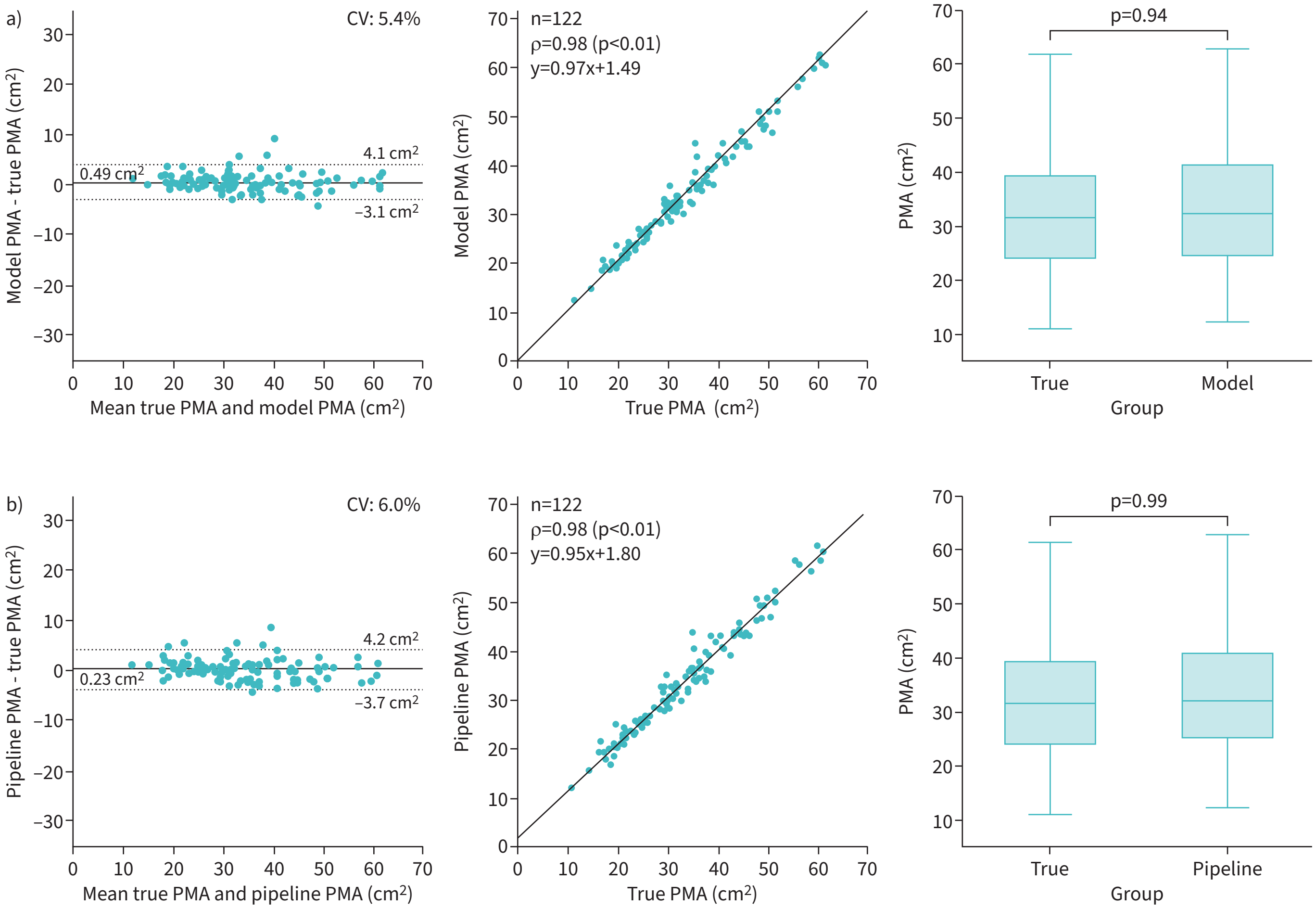

FIGURE 4 Bland–Altman plots, correlations and measurement differences between the pectoralis muscle areas (PMA) of all participants (n=122). a) PMA predicted by the deep learning model ($\text{PMA}_{\text{Model}}$) *versus* PMA segmented by the observers (PMA_{True}) on the human labelled aortic arch slice. b) PMA predicted by the complete automated pipeline on the detected aortic arch slice ($\text{PMA}_{\text{Pipeline}}$) *versus* PMA segmented by the observers (PMA_{True}) on the human labelled aortic arch slice. The Bland–Altman plots contain the coefficient of variation (CV) (top-right), and mean difference and 95% confidence intervals (1.96*SD) overlaid. The correlation plots contain the Spearman's ρ (along with the p-value) and linear regression for the two sets of data at the top left corner. The box-and-whisker plots show the median and 95% confidence interval for each measurement, along with the p-value from the measurement differences.

stratified by sex, BMI and COPD status) and concurrently (by reporting significant correlations with lung function indices).

The deep learning model generated pectoralis segmentation with high accuracy in comparison to the ground truth segmentation. The performance of the model was also similar to the interobserver variability of the segmentation performed by the human observers (Observers: 0.95±0.02; Model: 0.94±0.02). When combined with the significant time-savings provided by the automated segmentation in comparison to the human observers, it provides substantial evidence for the use of this pipeline to analyse large cohorts, such as CanCOLD. In addition, our model accuracy was either similar (magnetic resonance imaging-based; DSC: 0.94±0.01) [20] or slightly higher (CT-based; DSC: 0.93±0.04) [19] than other state-of-the art pectoralis segmentation networks from the literature. Unlike the previously published work, we also quantified model performance on an external never-before-seen dataset of research study participants with COPD, which is important for testing generalisability. Furthermore, we believe the slight decrease in average absolute PMA error (−1.4%) on the test cohort may be explained by the variable CT scan settings relative to the training data. This is moreover evidenced by the increased systematic bias in the Bland–Altman plots for the test cohort $\text{PMA}_{\text{Model}}$ measurements as compared to the training cohort (1.1 cm^2 *versus* 0.28 cm^2). Nonetheless, we believe the high model-ground truth DSC and low absolute PMA error on all datasets confirm the utility and generalisability of our deep learning model.

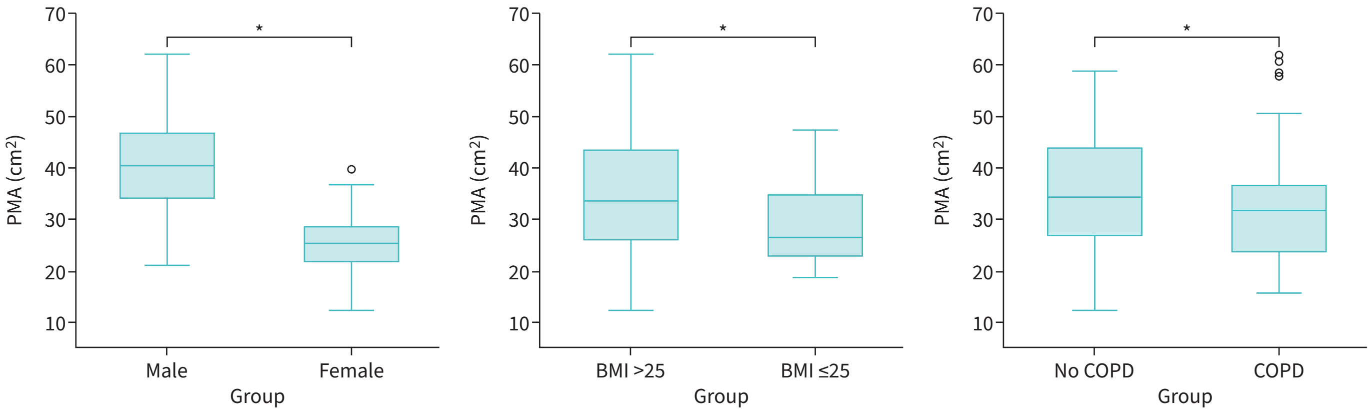

FIGURE 5 Box-and-whisker plots (median, 95% confidence intervals) of fully automated pectoralis muscle area (PMA) measurements for various subgroups in the combined CanCOLD and UofA datasets (n=122): male (n=59), 41.1±9.4 cm^2 *versus* female (n=63), 25.5±5.6 cm^2, p<0.01; BMI >25 kg·m^{-2} (n=85), 35.3±11.7 cm^2 *versus* BMI ⩽25 kg·m^{-2} (n=37), 29.3±8.0 cm^2, p<0.01; no COPD (n=46), 35.5±10.9 cm^2 *versus* COPD (n=76), 32.3±11.0 cm^2, p=0.02. CanCOLD: Canadian Cohort Obstructive Lung Disease study; UofA: University of Alberta; BMI: body mass index. COPD group differences were adjusted for age, sex and BMI. *: indicates a significant difference (p<0.05).

Our study was the first to create and report on a fully automated algorithm and complete pipeline for identifying the chest CT image slice containing the aortic arch. Similar slice localisation techniques for automated body composition measurements in abdominal CT have reported average errors of 4–5 mm using deep learning-based methods [17, 18]. In our work, the average error was <2 mm, indicating that our simpler atlas-matching approach was comparable to the state-of-the-art and sufficiently accurate to identify the aortic arch slice. Furthermore, we found that the average PMA error increased slightly when implementing the complete pipeline on the train (+0.5%) and test (+0.7%) sets. However, this did not necessarily reflect a worsening in measurement accuracy; there may have been natural minor variations in pectoral anatomy between the predicted and true aortic arch slice. In addition, $\mathrm{PMA_{Pipeline}}$ measurements were highly comparable to $\mathrm{PMA_{True}}$ as indicated by the Bland–Altman analysis and nearly linear positive correlation. These findings indicate that our aortic arch detection algorithm did not significantly affect the final model performance.

Finally, the results showed that PMA measurements generated by the complete pipeline were able to detect significant differences between various subgroups. In agreement with Diaz *et al*. [34] and McDonald *et al*. [7], we showed that PMA was significantly lower in: females *versus* males; and people with compared to without COPD. We also showed that PMA was significantly greater in people with BMI >25 kg·m^{-2} *versus* BMI ⩽25 kg·m^{-2}. Both $\mathrm{PMA_{True}}$ and $\mathrm{PMA_{Pipeline}}$ measurements showed similar and significant correlations with $\mathrm{FEV_1}$ % predicted, $\mathrm{FEV_1/FVC}$ and RV/TLC for all participants, indicating that pipeline automation captures the same trends in the data as manual segmentation performed by human observers.

TABLE 3 Spearman correlations between BMI and pulmonary function indices with the true pectoralis muscle area (PMA) measured by the human observers ($\mathrm{PMA_{True}}$) and the predicted PMA measured by the automated pipeline ($\mathrm{PMA_{Pipeline}}$)

Parameter	$\mathrm{PMA_{True}}$		$\mathrm{PMA_{Pipeline}}$	
	ρ	p-value	ρ	p-value
Demographics				
BMI	0.18	0.04	0.21	0.02
Pulmonary function				
$\mathrm{FEV_1}$ % pred	0.22	0.01	0.18	0.04
$\mathrm{FEV_1/FVC}$	0.37	<0.01	0.32	<0.01
RV/TLC	−0.33	<0.01	−0.31	<0.01

BMI: body mass index; $\mathrm{FEV_1}$: forced expiratory volume in 1 s; % pred: per cent predicted; FVC: forced vital capacity; RV: residual volume; TLC: total lung capacity.

Several limitations of our work must be acknowledged. Our model was trained on a relatively small number of participants with and without COPD (n=90), which could potentially limit the generalisability for future use. However, the countermeasures undertaken by the pipeline (data augmentation, hold-out validation sets and testing on a never-before-seen external dataset) helped offset this supposed limitation. Nonetheless, future work should look to expand the pipeline using more participants with varying anatomy, including those with breast implants. We acknowledge that the pipeline's predictions may deviate from the manual observer's segmentations. However, the fully automated pipeline would reduce any measurement variability, which may allow for more subtle changes over time to be measured. The intra-/interobserver variability for a manual PMA segmentation undertaken by the same (0.9±0.6 cm^2 or 2.4±1.5%) or different observer (1.5±1.0 cm^2 or 4.3±2.5%), as reported in this study, would introduce measurement error, thereby reducing the potential to measure longitudinal change. For example, with manual segmentation, the reproducibility coefficient was found to be 3.5 cm^2, therefore, any longitudinal change detected that is <3.5 cm^2, as measured by different observers, could be attributed to measurement error [35]. Additionally, Mason *et al*. [13] showed that PMA longitudinal change over a 5-year span was −1.8 cm^2 (−6.0%) in women and −2.8 cm^2 (−6.0%) in men. This longitudinal change is within the measurement error for manual segmentation of the PMA as reported in our study. Therefore, although our pipeline slightly deviates from the ground truth, a fully automated approach reduces measurement variation error and therefore allows for smaller changes to be measured over time. We also acknowledge that other deep learning segmentation models such as YOLO [36] and DeepLab [37] should be explored. Future work should investigate the performance of each for this application. We acknowledge that there are further improvements to a fully automated PMA measurement pipeline that can potentially increase speed and accuracy. For example, developing an automated method to segment the aorta directly would remove potential errors introduced during atlas registration [38, 39]. A 3D deep learning network could bypass slice detection altogether, allowing for the entire segmentation of the pectoralis muscle volume. However, the current standard in the field is 2D segmentation at the aortic arch slice, and therefore such work would be an avenue of future research.

Limitations notwithstanding, our work is an important step forward in automated PMA extraction from chest CT scans. For example, this pipeline can be used in research settings to dichotomise people into low *versus* high skeletal muscle mass, to assess the efficacy of interventions or drug trials, and significantly decrease the manual burden when analysing large cohorts. Clinically, this pipeline can be used to assess longitudinal changes in skeletal muscle mass over time, evaluate the efficacy of pulmonary rehabilitation and assess the effectiveness of therapeutic intervention on skeletal muscle mass (*e.g.*, nutrition, exercise training, anabolic steroids). Future research should develop age, sex and BMI-specific normal reference values and prediction equations for PMA so that researchers and/or clinicians can better assess the (ab)normality of individual measurements. Other potential future directions include automated pectoralis sub-muscle segmentation, and measurements of different muscle groups, such as the paraspinals. Finally, future work should explore the application of this tool in individuals with other respiratory-related diseases where measurements of skeletal muscle have been associated with outcomes, such as ILD [10] and lung cancer [40].

Conclusions

We developed a fully automated pipeline for aortic arch localisation and PMA segmentation from routinely collected chest CT scans. Our pipeline was accurate and highly generalisable to an external never-before-seen set of CT images from a group of clinical research study participants with COPD. Our findings motivate the use of this fully automated pipeline for pectoralis segmentation in both clinical and research settings.

Provenance: Submitted article, peer reviewed.

Acknowledgements: The authors would like to thank Benjamin Smith and Adamo Donovan (McGill University) for their technical input for the PMA segmentations.

Conflicts of interest: D. Genkin reports no conflicts of interest. A.R. Jenkins acknowledges salary support from Le Fonds de Recherche du Québec Santé - Postdoctoral Training fellowship. N. van Noord reports no conflicts of interest. K. Makimoto reports no conflicts of interest. S. Collins reports no conflicts of interest. M.K. Stickland acknowledges salary support from the Canadian Institutes of Health Research, Natural Sciences and Engineering Council of Canada, and University Hospital Foundation. W.C. Tan reports no conflicts of interest. J. Bourbeau acknowledges salary support from the Canadian Institute of Health Research, Réseau en santé respiratoire du FRQS, McGill University Health Centre Foundation, AstraZeneca Canada Ltd, Boehringer Ingelheim Canada Ltd, GlaxoSmithKline Canada Ltd, Grifols, Novartis, Sanofi, Trudell Canada Ltd, Pfizer Canada Ltd and COVIS Pharma Canada Ltd. D. Jensen reports no conflicts of interest. M. Kirby is a paid consultant at VIDA Diagnostics.

Support statement: D. Genkin acknowledges salary support from the Natural Sciences and Engineering Research Council (NSERC). M. Kirby acknowledges support from the NSERC Discovery Grant, the Early Researchers Award Program and the Canada Research Chair Program (Tier II). A.R. Jenkins acknowledges salary support from Le Fonds de Recherche du Québec Santé - Postdoctoral Training fellowship. D. Jensen holds a Canada Research Chair in Clinical Exercise and Respiratory Physiology (Tier 2) from the Canadian Institutes of Health Research. This study was supported by a Team Project Grant from the Québec Respiratory Health Research Network to D. Jensen as Principal Investigator, and A. Jenkins, M. Kirby and J. Bourbeau as co-investigators. Funding information for this article has been deposited with the Crossref Funder Registry.

Ethics statement: Institutional review board approval was obtained at each site.

References

1 Degens H, Gayan-Ramirez G, van Hees HW. Smoking-induced skeletal muscle dysfunction: from evidence to mechanisms. *Am J Respir Crit Care Med* 2015; 191: 620–625.

2 Zhang X, Wang C, Dou Q, *et al.* Sarcopenia as a predictor of all-cause mortality among older nursing home residents: a systematic review and meta-analysis. *BMJ Open* 2018; 8: e021252.

3 Gaynor-Sodeifi K, Lewthwaite H, Jenkins AR, *et al.* The association between fat-free mass and exercise test outcomes in people with chronic obstructive pulmonary disease: a systematic review. *COPD* 2022; 19: 182–205.

4 Schols AM, Broekhuizen R, Weling-Scheepers CA, *et al.* Body composition and mortality in chronic obstructive pulmonary disease. *Am J Clin Nutr* 2005; 82: 53–59.

5 de Koning HJ, van der Aalst CM, de Jong PA, *et al.* Reduced lung-cancer mortality with volume CT screening in a randomized trial. *N Engl J Med* 2020; 382: 503–513.

6 Coxson HO, Leipsic J, Parraga G, *et al.* Using pulmonary imaging to move chronic obstructive pulmonary disease beyond FEV1. *Am J Respir Crit Care Med* 2014; 190: 135–144.

7 McDonald ML, Diaz AA, Ross JC, *et al.* Quantitative computed tomography measures of pectoralis muscle area and disease severity in chronic obstructive pulmonary disease. A cross-sectional study. *Ann Am Thorac Soc* 2014; 11: 326–334.

8 Diaz AA, Martinez CH, Harmouche R, *et al.* Pectoralis muscle area and mortality in smokers without airflow obstruction. *Respir Res* 2018; 19: 1–8.

9 McDonald MN, Diaz AA, Rutten E, *et al.* Chest computed tomography-derived low fat-free mass index and mortality in COPD. *Eur Respir J* 2017; 50: 1701134.

10 Molgat-Seon Y, Guler SA, Peters CM, *et al.* Pectoralis muscle area and its association with indices of disease severity in interstitial lung disease. *Respir Med* 2021; 186: 106539.

11 Regan EA, Hokanson JE, Murphy JR, *et al.* Genetic epidemiology of COPD (COPDGene) study design. *COPD* 2011; 7: 32–43.

12 Vestbo J, Anderson W, Coxson HO, *et al.*; ECLIPSE Investigators. Evaluation of COPD longitudinally to identify predictive surrogate end-points (ECLIPSE). *Eur Respir J* 2008; 31: 869–873.

13 Mason SE, Moreta-Martinez R, Labaki WW, *et al.* Longitudinal association between muscle loss and mortality in ever smokers. *Chest* 2022; 161: 960–970.

14 Bourbeau J, Tan WC, Benedetti A, *et al.* Canadian Cohort Obstructive Lung Disease (CanCOLD): fulfilling the need for longitudinal observational studies in COPD. *COPD* 2014; 11: 125–132.

15 Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)* 2015: 9351: 234–241.

16 Liu X, Song L, Liu S, *et al.* A review of deep-learning-based medical image segmentation methods. *Sustainability* 2021; 13: 1224.

17 Ha J, Park T, Kim H-K, *et al.* Development of a fully automatic deep learning system for L3 selection and body composition assessment on computed tomography. *Sci Rep* 2021; 11: 21656.

18 Belharbi S, Chatelain C, Hérault R, *et al.* Spotting L3 slice in CT scans using deep convolutional network and transfer learning. *Comput Biol Med* 2017; 87: 95–103.

19 Indra ND, Syed AN, Alejandro PC, *et al.* CT-based segmentation of pectoral muscle using deep learning and association of computed metrics with aging and sex. *Proceedings of the SPIE* 2022; 120360R.

20 Godoy IRB, Silva RP, Rodrigues TC, *et al.* Automatic MRI segmentation of pectoralis major muscle using deep learning. *Sci Rep* 2022; 12: 5300.

21 Stickland MK, Collins S, Fuhr DP. The Effect of Inhaled Nitric Oxide on Dyspnea and Exercise Tolerance in COPD (iNO). Identification no. NCT03679312. University of Alberta. 2018. Date last updated: 13 November 2023. Date last accessed: 12 December 2023. https://clinicaltrials.gov/ct2/show/NCT03679312

22 Smith BM, Kirby M, Hoffman EA, *et al.*; MESA Lung, CanCOLD, and SPIROMICS Investigators. Association of dysanapsis with chronic obstructive pulmonary disease among older adults. *JAMA* 2020; 323: 2268–2280.

23 Phillips DB, Elbehairy AF, James MD, *et al.*; CanCOLD Collaborative Research Group and the Canadian Respiratory Research Network. Impaired ventilatory efficiency, dyspnea, and exercise intolerance in chronic obstructive pulmonary disease: results from the CanCOLD study. *Am J Respir Crit Care Med* 2022; 205: 1391–1402.

24 Tan WC, Sin DD, Bourbeau J, *et al.* CanCOLD Collaborative Research Group. Characteristics of COPD in never-smokers and ever-smokers in the general population: results from the CanCOLD study. *Thorax* 2015; 70: 822–829.

25 Abozid H, Kirby M, Nasir N, *et al.*; CanCOLD Collaborative Research Group and the Canadian Respiratory Research Network. CT airway remodelling and chronic cough. *BMJ Open Respir Res* 2023; 10: e001462.

26 Kirby M, Yin Y, Tschirren J, *et al.*; CanCOLD Collaborative Research Group and the Canadian Respiratory Research Network. A novel method of estimating small airway disease using inspiratory-to-expiratory computed tomography. *Respiration* 2017; 94: 336–345.

27 Kirby M, Smith BM, Tanabe N, *et al.* Computed tomography total airway count predicts progression to COPD in at-risk smokers. *ERJ Open Res* 2021; 7: 00307-2021.

28 Kirby M, Tanabe N, Tan WC, *et al.*; CanCOLD Collaborative Research Group; Canadian Respiratory Research Network; CanCOLD Collaborative Research Group, the Canadian Respiratory Research Network. Total airway count on computed tomography and the risk of chronic obstructive pulmonary disease progression. Findings from a population-based study. *Am J Respir Crit Care Med* 2018; 197: 56–65.

29 Makimoto K, Au R, Moslemi A, *et al.* Comparison of feature selection methods and machine learning classifiers for predicting chronic obstructive pulmonary disease using texture-based CT lung radiomic features. *Acad Radiol* 2023; 30: 900–910.

30 Macintyre N, Crapo RO, Viegi G, *et al.* Standardisation of the single-breath determination of carbon monoxide uptake in the lung. *Eur Respir J* 2005; 26: 720–735.

31 Wanger J, Clausen JL, Coates A, *et al.* Standardisation of the measurement of lung volumes. *Eur Respir J* 2005; 26: 511–522.

32 Miller MR, Hankinson J, Brusasco V, *et al.*; ATS/ERS Task Force. Standardisation of spirometry. *Eur Respir J* 2005; 26: 319–338.

33 Agustí A, Celli BR, Criner GJ, *et al.* Global Initiative for Chronic Obstructive Lung Disease 2023 report: GOLD executive summary. *Am J Respir Crit Care Med* 2023; 207: 819–837.

34 Diaz AA, Zhou L, Young TP, *et al.* Chest CT measures of muscle and adipose tissue in COPD: gender-based differences in content and in relationships with blood biomarkers. *Acad Radiol* 2014; 21: 1255–1261.

35 Reinstein DZ, Archer TJ, Silverman RH, *et al.* Accuracy, repeatability and reproducibility of Artemis very high-frequency digital ultrasound arc-scan lateral dimension measurements. *J Cataract Refract Surg* 2006; 32: 1799–1802.

36 Maryam H, Denis F, Razmig K. Data augmentation for multi-organ detection in medical images. 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA), November 2020, Paris, France, pp. 1–6.

37 Tang W, Zou D, Yang S, *et al.* DSL: automatic liver segmentation with Faster R-CNN and DeepLab. *In:* Kůrková V, Manolopoulos Y, Hammer B, *et al.*, eds. Artificial Neural Networks and Machine Learning. ICANN 2018. Lecture Notes in Computer Science, vol 11140. Cham, Springer, 2018.

38 Sedghi GZ, Bons LR, Giordano M, *et al.* Automated 3D segmentation and diameter measurement of the thoracic aorta on non-contrast enhanced CT. *Eur Radiol* 2019; 29: 4613–4623.

39 Xie Y, Padgett J, Biancardi AM, *et al.* Automated aorta segmentation in low-dose chest CT images. *Int J Comput Assist Radiol Surg* 2014; 9: 211–219.

40 Kinsey CM, San José Estépar R, van der Velden J, *et al.* Lower pectoralis muscle area is associated with a worse overall survival in non-small cell lung cancer. *Cancer Epidemiol Biomarkers Prev* 2017; 26: 38–43.