**Automated quality-controlled cardiovascular magnetic resonance pericardial fat quantification using a convolutional neural network in the UK Biobank**

**Brief title: Automated CMR pericardial fat quantification**

Andrew Bard†1,2, Zahra Raisi-Estabragh†1,2, Maddalena Ardissino3, Aaron Mark Lee1, Francesca Pugliese1, 2, Damini Dey4, Sandip Sarkar2, Patricia B. Munroe1, Stefan Neubauer5, Nicholas C. Harvey6, 7, Steffen E. Petersen1, 2

†These authors have contributed equally to this work and share first authorship

1.William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University of London, Charterhouse Square, London, EC1M 6BQ, UK

2. St Bartholomew’s Hospital, Barts Health NHS Trust, West Smithfield, EC1A 7BE, London, UK

3. Imperial College London, Exhibition Road, SW7 2AZ, London, UK

4. Biomedical Imaging Research Institute, Cedars-Sinai Medical Centre, USA

5. National Institute for Health Research Oxford Biomedical Research Centre, Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

6. MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton, SO16 6YD, UK

7. NIHR Southampton Biomedical Research Centre, University of Southampton and University Hospital Southampton NHS Foundation Trust, Southampton, SO16 6YD, UK

**Address for correspondence:** Professor Steffen E. Petersen. William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University of London, London, UK; Email: s.e.petersen@qmul.ac.uk; Telephone: +44-2078826902

**Abstract**

**Background:** Pericardial adipose tissue (PAT) may represent a novel risk marker for cardiovascular disease. However, absence of rapid radiation-free PAT quantification methods has precluded its examination in large cohorts.

**Objectives:** We developed a fully automated quality-controlled tool for cardiovascular magnetic resonance (CMR) PAT quantification in the UK Biobank (UKB).

**Methods:** Image analysis comprised contouring an en-bloc PAT area on 4-chamber cine images. We created a ground truth manual analysis dataset randomly split into training and test sets. We built a neural network for automated segmentation using a Multi-residual U-net architecture with incorporation of permanently-active dropout layers to facilitate quality control of the model’s output using Monte-Carlo sampling. We developed an in-built quality control feature, which presents predicted Dice scores. We evaluated model performance against the test set (n=87), the whole UKB imaging cohort (n=45,519), and an external dataset (n=103). In an independent dataset, we compared automated CMR and cardiac computed tomography (CCT) PAT quantification. Finally, we tested association of CMR PAT with diabetes in the UKB (n=42,928).

**Results:**Agreement between automated and manual segmentations in the test set was almost identical to inter-observer variability (mean Dice score=0.8). The quality control method predicted individual Dice scores with Pearson r=0.75. Model performance remained high in the whole UKB Imaging cohort and in the external dataset, with medium-good quality segmentation in 94.3% (mean Dice score=0.77) and 94.4% (mean Dice score=0.78) respectively. There was high correlation between CMR and CCT PAT measures (Pearson r=0.72, p-value 5.3 10-18). Larger CMR PAT area was associated with significantly greater odds of diabetes independent of age, sex, and body mass index.

**Conclusions:** This novel fully automated method for CMR PAT quantification demonstrates good model performance on independent and external datasets, correlation with reference standard CCT PAT measurement, and expected clinical associations with diabetes.

**Keywords:** cardiovascular magnetic resonance; pericardial fat; automated image analysis; neural network; machine learning

**Introduction**

Pericardial adipose tissue (PAT), which surrounds the surface of the heart and adventitia of the coronary arteries, has been linked to a range of important cardiovascular and metabolic conditions, including atrial fibrillation(1), diabetes(2), and coronary artery disease(3). These relationships appear independent of subcutaneous fat, total body weight, and classical cardiovascular risk factors(4), suggesting distinct biological significance of PAT. Indeed, it has been proposed that adipocytokines and proinflammatory markers secreted by metabolically active PAT may act as mediators for these associations through promotion of a milieu for disease development at both local and systemic levels(5,6). Thus, PAT may provide novel insights into disease processes and has a potential role as a marker of cardiovascular risk.

Technical difficulties in quantification of PAT in an efficient and radiation-free manner have precluded its systematic study in large cohorts. Whilst cardiac computed tomography (CCT) PAT quantification is well-established(3,7–9), exposure of large population cohorts to ionising radiation is not ethically permissible(10). Cardiovascular magnetic resonance (CMR) is the reference imaging modality for assessment of cardiac structure and function and has been used in several large population studies, including the Multi-ethnic Study of Atherosclerosis(11), the Framingham Heart Study(12), and the UK Biobank (UKB)(13). Thus, CMR PAT quantification would have high utility for research, with potential for translation into clinical care; however, existing methods require dedicated acquisitions and, often, arduous manual image analysis(14,15), limiting their applicability to large datasets with standard sequence acquisitions.

We present a novel fully automated method for PAT quantification using standard of care CMR images with in-built quality control (QC) developed and tested in the UKB. We test the correlation of this CMR PAT metric with reference standard CCT PAT quantification in an external dataset and investigate clinical validity through consideration of associations with diabetes in UKB. Reporting in this paper is in accordance with relevant aspects of the Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME) guidance(16).**Materials and Methods**

**Setting and study population**

The UKB incorporates data from over half a million participants recruited between 2006-2010 from across the UK. Individuals aged 40-69 years old were identified through National Health Service (NHS) registers and requested to participate via postal invites. There was detailed baseline characterisation of participant demographics, lifestyle, and medical history(17). The UKB protocol is publicly available(18). The UKB Imaging Study, which includes detailed CMR imaging, aims to scan 100,000 of the original participants (>48,000 completed, January 2021)(19).

**CMR image acquisition**

CMR imaging was with 1.5 Tesla scanners (MAGNETOM Aera, Syngo Platform VD13A, Siemens Healthcare, Erlangen, Germany) using a standardised acquisition protocol, which is detailed elsewhere(13). Cardiac function was assessed using balanced steady state free precession cine sequences with standard long axis cuts and a complete short axis stack. No signal or image filtering was applied, with exception of distortion correction.

**Standard operating procedure for PAT segmentation**

The analysis protocol comprised segmentation of an en bloc PAT area from standard 4-chamber cine images (single 2D slice), a universal component of standard CMR studies and one which demonstrates less variability in cut plane positioning compared to other acquisitions (e.g. short axis slices). For consistency, we measured PAT at phase 1 of the imaging cycle (approximately end-diastole). A single contour was drawn to select areas of high signal intensity adjacent to the epicardial surface of the left and right ventricular myocardium, resulting in output of an area measure in cm2 (Figure 1). Areas of high signal intensity over the liver were not included, as this almost always represents adipose tissue below the diaphragm (Figure 1B).

**Creation of a ground truth manual segmentation dataset**

We selected 500 random UKB participants with record of an imaging centre visit using the random number generator package in R. We excluded participants with missing (n=45) or inadequate quality (n=23) CMR images. Thus, PAT contours were manually drawn for n=432 participants. For the purposes of model training and evaluation, the sample was randomly split into training (n=345) and test (n=87) sets. Image analysis was performed blind to participant details using CVI42® post-processing software (Version 5.11, Circle Cardiovascular Imaging Inc., Calgary, Canada). Contours were drawn by AB and cross-checked by ZRE.

**Neural network architecture**

As the size of PAT does not alter during the cardiac cycle, it may be quantified on static images, without consideration to cardiac motion. Thus, PAT quantification can be framed as a simple foreground-segmentation problem using a single frame per experimental subject, from which the area it occupies can be extracted. The task is to predict whether individual pixels represent a point of interest (i.e., within PAT) or are a part of the background.

Great progress has been made with automated medical image segmentation using fully-convolutional neural networks(20,21), particularly, using encoder-decoder architectures(22). We developed a neural network using a Multi-residual U-net (MultiResUNet) base architecture(23) with the incorporation of a permanently-active dropout layer(24) at the end of each multi-residual block (**Central illustration**). The best trade-off between overall segmentation accuracy and prediction of that accuracy was obtained with a dropout rate r=0.3, which we found to be optimal for model performance. This was selected as the largest r value at which the segmentation quality was not statistically reduced relative to a non-stochastic network (Supplementary figure 1). To incorporate a measure of uncertainty which can be used for QC, we used permanently-active dropout layers to add a stochastic component to their network outputs, meaning that multiple Monte-Carlo (MC) samples can be drawn for any given input(24). This MC sampling from a stochastic neural network generates samples of predicted probability maps , from which thresholding at 0.5 can generate Boolean segmentation maps . For our foreground detection problem, the final segmentation for each voxel () is defined by thresholding the voxelwise mean of :

**Network implementation**

The neural network was implemented and trained using the TensorFlow 2.0 Python API(25), software available from: <https://www.tensorflow.org>. A combination of resampling images to enforce uniform resolution, robust data augmentation and intensity normalisation has previously been shown to increase the generalisability of segmentation networks(26). Inspired by this approach, all image data was first resampled to a uniform resolution (1.82×1.82mm pixel spacing) and cropped/padded to a size of pixels, including the test set. During training, data was augmented with rotation (up to 25°), altered resolution (resizing of up to 20%), random shearing up to 20% and random panning of up to 25% of the image dimension. All data augmentation was performed on-the-fly, meaning that each complete training epoch utilised a different set of images. All images were normalised such that their pixel intensity range was between 0 and 1.

Training proceeded for a maximum of 300 complete epochs on a NVIDIA Tesla M40, using a batch size of 8 images. The loss function used was the binary cross-entropy, and this was optimised using the Adam optimizer(27), with an initial learning rate of 0.01, . The learning rate was decreased by a factor of 0.3 if, for 10 consecutive epochs, the loss was not decreasing. If 20 epochs elapsed with no decrease in the loss function, training was ceased and the weights yielding the lowest loss were restored.

**Metrics for assessment of segmentation agreement, inter-observer variability and model performance**

We evaluated agreement between manual segmentations of different expert observers (AB, ZRE, SEP) and between manual and automated segmentation. When multiple MC samples are drawn from the stochastic neural network, their level of agreement is correlated with the quality of the consensus segmentation(24). We expressed level of segmentation agreement using metrics based on the well-known intersection-over-union (or Jacard Index) and the Dice score. Thus, we used four metrics for agreement between segmentations: the Dice score, Intersection-over-Union (IoU) metrics for overlap, the mean contour distance, and the symmetric Haussdorff distance (compares the closeness of foreground voxels borders). Both of the overlap metrics are bounded between 0 and 1, with 0 representing no overlap and 1 representing perfect overlap. For both of these distance metrics, lower distances represent closer agreement between segmentation results. Furthermore, in line with previous literature pertaining to QC (24, 35), we classify segmentation accuracy as poor if the Dice score <0.6, medium for 0.6-0.8 and good if the Dice score is ≥0.8.

**Correlation of automated CMR PAT quantification with CCT measured PAT**

We tested the correlation of our derived CMR PAT measurement with established CCT PAT measurements. We utilised the Barts Health NHS Trust local sub-study, from the EValuation of INtegrated Cardiac Imaging for the Detection and Characterization of Ischaemic Heart Disease (EVINCI) dataset(28), a clinical trial dataset including n=109 participants with paired CMR and CCT imaging performed within a maximum interval of 37 days. We used the QFAT software (version 2.0, Cedars-Sinai Medical Center) for CCT PAT quantification (7,8), which segments and quantifies epicardial and thoracic fat from non-contrast calcium scoring CCT. We utilised deep learning-based contouring with no manual adjustment. Voxels containing thoracic fat deposits were defined as those with a radiodensity of between -190 and -30 Hounsfield units. The total PAT volume was measured in cm3. For the CMR analysis, we used our automated pipeline: 4-chamber cine images were resampled to a resolution of 1.82 x 1.82mm, padded/cropped to 208 x 208 pixels, normalised to have pixel intensities ranging between 0 and 1, and our stochastic network segmentation and quality control applied with . Finally, the segmented area was extracted as the mean foreground area of the MC samples. Thus, we were able to test the performance of our automated image analysis pipeline on CMR studies within the EVINCII cohort and also to assess the correlation between these measures and CCT PAT quantification.

**Association with Diabetes**

Given the established association between diabetes and increased PAT, we tested the clinical validity of our PAT measures through consideration of associations with this condition. We applied our automated CMR PAT analysis tool to the entire UKB Imaging cohort for whom adequate imaging was available (n=42,928). Diabetes was coded based on self-report of the diagnosis, self-reported use of “medication for diabetes”, or serum glycosylated haemoglobin >48mmol/mol. We tested the association of PAT area with diabetes status in multivariable logistic regression models with adjustment for age, sex, and body mass index (BMI). We present the results as odds ratio associated with 10cm2 increase in PAT with corresponding 95% confidence intervals and p-values.

**Results**

**Model training**

We trained a MultiResUNet(23) with a Bayesian modification; such that, multiple MC samples are drawn for each input, in order to perform quality control and derive measures of uncertainty(24).Within this context, there is one hyperparameter that must be optimised after model training - that of the number () of MC samples drawn when segmenting an unseen image. When multiple MC samples are drawn during segmentation, they can be summarised in a number of ways. Firstly, they can be used to produce a single “best-guess” segmentation, via a simple voxel-wise voting procedure. It is expected that drawing more samples from a well-trained network will increase its accuracy, but with diminishing returns. Where the area of “foreground” pixels is particularly of interest (as in this use case, quantifying the area of PAT), we can report the mean and standard deviation of the areas across the samples, which can be used for propagating uncertainty in downstream calculations. was set to 15 for all further work, for the following reasons: Comparisons of segmentation accuracy with a deterministic neural network showed that consistent with prior work(24), there was no sacrifice in segmentation quality by using a stochastic network relative to a deterministic one when N was set to an appropriate level **(Supplementary Figure 1A**)**.** Additionally, increasing N beyond 15 gave very little extra segmentation accuracy **(Supplementary Figure 1A**) or estimated standard deviation of area.

There are a number of different metrics which were proposed as correlates of final segmentation accuracy, however, it was concluded that of these, both the most conceptually convenient and the most easily interpretable are those corresponding to often-used segmentation accuracy metrics – the Dice score and the IoU of the MC samples(24). We tested calculation of both IoU and Dice score globally or mean-pairwise over the MC samples, finding that the best predictor of true segmentation accuracy was the mean pairwise Dice score between the MC samples, assessed on quantitative measures of agreement with the true Dice score of the test set (**Supplementary Figure 2**). Further details, as well as relevant equations, are detailed in the **Supplementary Material**.

**Evaluation of automated CMR PAT segmentation model performance**

The performance of the automated segmentation within the test set relative to manual segmentations was good and very similar to the agreement between human observers (mean Dice score=0.8). This was the case both for raw segmentation metrics (**Table 1**), and under Bland-Altman analysis (**Figure 2E-H**). Arguably, this is the best performance that may be achieved by an automated segmentation algorithm and reflects the inherently challenging nature of the PAT segmentation task. A few cases (n=4, 4.5%) had poor segmentation quality (Dice score <0.6) (**Figure 2A-D)** and very large Hausdorff distances. This underlines the importance of the in-built QC feature, which would flag such cases. We also successfully applied the automated segmentation to the whole UK Biobank Imaging cohort (n=45,519), 94.3% of cases (n=42,928) had predicted Dice score of medium or good quality (mean predicted Dice score 0.77). The automated segmentation also performed well in the external EVINCII dataset, with the majority of studies having medium/good segmentation quality (n=103, 94.4%), with overall mean predicted Dice score of 0.78 (**Figure 3A**). Running on a laptop PC with an Intel®️ Core™️ i7-1165G7 processor, using a Monte Carlo sample size (N) of 15, the model and quality control step took 2.1 seconds, including image pre-processing and final estimation of Dice score.

**Correlation of automated CMR PAT with CCT PAT quantification**

Within the EVINCII dataset, we tested the correlation of CMR PAT measures derived using our automated analysis with PAT derived using the QFAT tool from paired CCT scans. CMR studies with poor segmentation quality (predicted Dice score <) were excluded from the analysis (n=6). For illustration, we present example segmentations with a range of Dice scores in **Figure 3C**. There was a strong, statistically significant correlation between CCT PAT volume and CMR PAT area (Pearson r=0.72, p-value 5.3 10-18, **Figure 3B**).

**Application to the UK Biobank imaging cohort and association with diabetes**

We applied our neural network to CMR scans from 45,519 UK Biobank participants. We excluded cases with a segmentation Dice score of <0.6 (n= 2,591, 5.7%). The remaining 42,928 participants were included in the analysis, of these 2,529 were diabetic. Consistent with existing evidence(2), larger PAT area was associated with greater risk of diabetes in univariate and multivariable models (**Table 2**). In fully adjusted models, every 10cm2 increase in PAT was associated with approximately 7% greater likelihood of diabetes independent of age, sex, and BMI.

**Discussion**

**Summary of findings**

We present a novel method for PAT quantification using standard of care CMR images, fully automated through a convolutional neural network with an in-built QC algorithm. The automated segmentation tool performed well within the test set, the whole UKB Imaging cohort, and an external dataset, producing segmentation agreement close to that of human observers. Our segmentation method demonstrates validity against CCT PAT quantification, with a strong statistically significant correlation between paired CMR and CCT PAT measurements. Furthermore, we demonstrate, within the UK Biobank, expected clinical association of CMR PAT with diabetes independent of age, sex, and BMI. Thus, the proposed CMR PAT method has great potential in facilitating investigation of the clinical significance of PAT in large population cohorts.

**Comparison with existing work**

Limited studies have attempted to quantify and study the clinical associations of CMR PAT. In a study from 1992, Ross et al.(29) proposed a method for quantification of abdominal and subcutaneous fat on spin echo T1-weighted magnetic resonance imaging (MRI) sequences. They propose a signal intensity threshold for defining adipose tissue pixels, the area of adipose tissue regions was then calculated by summing adipose tissue pixels and multiplying by the pixel surface area, this area was then multiplied by slice thickness to derive the volume of the adipose tissue. Unsupervised approaches for quantification of abdominal fat using this method have been developed using small datasets (30). More recently, these principles have been repurposed to derive CMR measures of thoracic fat using spin echo T1-weighted CMR acquisitions with zero slice gap and to investigate clinical associations in small cohorts (31–33). Whilst this approach has had some utility, there are two fundamental limitations. Firstly, as the thresholding is based on pixel intensity levels, this is subject to variation based on technical (magnet strength, acquisition sequences, vendor) factors; as such, the threshold would need to be re-established depending on technical parameters. Secondly, because the methodology aims to derive a volumetric measure of adipose tissue, dedicated acquisitions with zero slice gap have been obtained. As standard protocols do not have zero slice gap (usually 5-8mm slice gap), this approach, as it stands, is not suitable for application to standard of care CMR images.

Ding et al.(14) propose another approach for CMR PAT quantification; they present a limited study demonstrating feasibility of a fully automated pericardial fat quantification method from water/fat-resolved whole-heart non-contrast coronary magnetic resonance angiography. The very small sample size (n=10) in this limited feasibility study precludes any meaningful assessment of model performance and the clinical validity of the proposed measurement is not known. Furthermore, as fat/water sequences are not routinely acquired as part of standard CMR studies, this methodology is unlikely to have wide application.

In a similar approach to our work, Rado et al.(15) quantify epicardial and pericardial fat areas from 4-chamber cine images. They use a manual analysis protocol taking measurements in end-diastole and end-systole and making distinction between epicardial and pericardial fat area. They use their manual analysis measures (n=374) to investigation associations with impaired glucose metabolism and left ventricular function. In developing our SOP, we also experimented with distinguishing between epicardial and pericardial fat areas. However, on inspection of a large number of studies, it became apparent that reliable distinction of these two areas was not possible for a substantial number of cases. Hence, we opted for a simpler approach of using a single en bloc contour. The strong correlation of our measure with CCT PAT quantification and observed associations with diabetes suggest that quantification according to our SOP does not detract from the potential utility of the measurement. Furthermore, the simplicity of our method enabled development of a fully automated analysis tool, which is essential for study of CMR PAT in large datasets.

**Technical implications**

In terms of the technical details of our neural network, the Multi-Residual U-net architecture(23) was vital, yielding far better results than “vanilla” U-nets(21) (data not shown). Meanwhile, a quality control method has been demonstrated using an extension of a stochastic network, which approximates Bayesian MC sampling(24). Consistent with prior work, we find that measures of similarity between MC samples are correlated with segmentation quality; intuitively this corresponds to how “sure” the network is of the output. However, in contrast, we found that the mean pairwise Dice score yielded best prediction, in contrast to the global intersection-over-union used in previous work, and that an additional linear correction was required.

A potential consideration is whether better segmentation accuracy could be obtained via the removal of the stochastic component, thereby providing a single prediction. This would be undesirable for a number of reasons. Firstly it is important to have some estimate of segmentation quality, which can only be provided if the actual segmentation is derived from our stochastic process. Secondly, a comparison with a non-deterministic MultiresUNet is provided in Supplementary Figure 1, and the accuracy is comparable with our stochastic model. However, note that dropout was not used within training of this network for the following reason: The MultiResUNet architecture makes extensive use of batch normalisation. Because of a phenomenon known as variance shift, the combination of batch normalisation and dropout often produces reduced accuracy once the dropout is “turned off” (34) for a full explanation of this phenomenon. However, this problem does not apply to our results, as the dropout is kept permanently active.

**Strengths and limitations**

Using a modest manually-annotated dataset, we have achieved good segmentation accuracy, with a mean Dice score of 0.80. However, the performance of machine learning tools may be reduced when applied to external datasets (decline in generalisability), to minimise this effect we made use of robust data augmentation procedures during training (26). We are reassured by the good performance of our tool on the whole UKB Imaging cohort and on the external EVINCII CMR dataset. We use a very simplified SOP taking PAT area measurement from a single 2D slice, clearly this approach does not accurately quantify the volume of mediastinal fat. However, we demonstrate correlation of our measurement with established volumetric CCT PAT measure and replicate known clinical associations with diabetes. This suggests that our CMR PAT measure is valid as a marker to study associations with the PAT exposure. Indeed, we would argue, that complicated acquisitions to quantify thoracic fat have hampered previous attempts to make wide practical use of this measure. A potential limitation of the method is that the model cannot distinguish between specific pathologies – e.g., fat or fluid. In the UK Biobank, and similar cohorts, we do not expect this to be a significant source of error, as there are very few participants with pericardial effusions. However, with broader application of the tool to clinical cohorts, such considerations may be more relevant. Further studies in large cohorts are now needed to establish the clinical utility of this CMR PAT measure in different settings and patient cohorts, the proposed automated tool will facilitate such studies in large (and small) cohorts. As we use standard of care images, the CMR PAT measurement can be retrospectively applied to any existing dataset and, furthermore, if clinical value of this metric is established, it could be readily integrated into clinical practice.

**Conclusion**

We present a novel fully automated quality-controlled method for CMR PAT quantification using standard of care 4-chamber cine images. Throughout the study, we demonstrate that our QC method functions as intended, and we demonstrate that the segmentation performance of this method is equivalent to inter-observer variability, and that the area extracted by our method is strongly correlated with measurements taken using reference standard CCT quantification. Finally, we demonstrate that our CMR PAT quantification method can recapitulate known clinical associations with diabetes. Overall, we present a novel tool that is now ready to be used for new research.

**Ethics**

This study was covered by the ethics approval for UKB studies from the NHS National Research Ethics Service on 17th June 2011 (Ref 11/NW/0382) and extended on 10th May 2016 (Ref 16/NW/0274). Use of paired CT and CMR data from the EVINCI study was covered by a Data Protection Impact Assessment by the Data Protection Officer of Barts Health NHS Trust. For the original EVINCII study, local ethical approval was provided (REC Number: 10/H0721/79) and all subjects gave written informed consent.

**Acknowledgments**

This study was conducted using the UK Biobank resource under access application 2964. We would like to thank all the participants, staff involved with planning, collection and analysis, including core lab analysis of the CMR imaging data.

**Funding statement**

AB, SS, and S.E.P. acknowledge support from the CAP-AI programme, London’s first AI enabling programme focused on stimulating growth in the capital’s AI Sector. CAP-AI is led by Capital Enterprise in partnership with Barts Health NHS Trust and Digital Catapult and is funded by the European Regional Development Fund and Barts Charity. This work was supported by Health Data Research 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. ZRE was supported by British Heart Foundation Clinical Research Training Fellowship No. FS/17/81/33318. AML and SEP acknowledge support from the National Institute for Health Research (NIHR Biomedical Research Centre at Barts and from the “SmartHeart” EPSRC programme grant (www.nihr.ac.uk; EP/P001009/1). SEP and PBM acknowledge support from the National Institute for Health Research (NIHR) Cardiovascular Biomedical Research Centre at Barts. SEP has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825903 (euCanSHare project). DD acknowledges support NIH grant no. 1R01HL133616 in development of the QFAT tool. SN and SKP are supported by the Oxford NIHR Biomedical Research Centre and the Oxford British Heart Foundation Centre of Research Excellence. DD is funded by NIH research grant NIH/NHLBI 1R01HL133616. NCH acknowledges support from the UK Medical Research Council (MRC #405050259, MRC LEU), NIHR Southampton Biomedical Research Centre, University of Southampton and University Hospital Southampton. This project was enabled through access to the MRC eMedLab Medical Bioinformatics infrastructure, supported by the Medical Research Council (www.mrc.ac.uk; MR/L016311/1).

**References**

1. Wong CX, Ganesan AN, Selvanayagam JB. Epicardial fat and atrial fibrillation: Current evidence, potential mechanisms, clinical implications, and future directions. *Eur Heart J* (2017) **38**:1294–1302. doi:10.1093/eurheartj/ehw045

2. Li Y, Liu B, Li Y, Jing X, Deng S, Yan Y, et al. Epicardial fat tissue in patients with diabetes mellitus: A systematic review and meta-analysis. *Cardiovasc Diabetol* (2019) doi:10.1186/s12933-019-0807-3

3. Greif M, Becker A, Von Ziegler F, Lebherz C, Lehrke M, Broedl UC, et al. Pericardial Adipose Tissue Determined by Dual Source CT Is a Risk Factor for Coronary Atherosclerosis. *Arterioscler Thromb Vasc Biol* (2009) **29**:781–786. doi:10.1161/ATVBAHA.108.180653

4. Mahabadi AA, Berg MH, Lehmann N, Kälsch H, Bauer M, Kara K, et al. Association of epicardial fat with cardiovascular risk factors and incident myocardial infarction in the general population: The Heinz Nixdorf recall study. *J Am Coll Cardiol* (2013) **61**:1388–1395. doi:10.1016/j.jacc.2012.11.062

5. Iacobellis G, Bianco AC. Epicardial adipose tissue: Emerging physiological, pathophysiological and clinical features. *Trends Endocrinol Metab* (2011) **22**:450–457. doi:10.1016/j.tem.2011.07.003

6. Cheng KH, Chu CS, Lee KT, Lin TH, Hsieh CC, Chiu CC, et al. Adipocytokines and proinflammatory mediators from abdominal and epicardial adipose tissue in patients with coronary artery disease. *Int J Obes* (2008) **32**:268–274. doi:10.1038/sj.ijo.0803726

7. Commandeur F, Goeller M, Razipour A, Cadet S, Hell MM, Kwiecinski J, et al. Fully Automated CT Quantification of Epicardial Adipose Tissue by Deep Learning: A Multicenter Study. *Radiol Artif Intell* (2019) **1**:e190045. doi:10.1148/ryai.2019190045

8. Eisenberg E, McElhinney PA, Commandeur F, Chen X, Cadet S, Goeller M, et al. Deep Learning-Based Quantification of Epicardial Adipose Tissue Volume and Attenuation Predicts Major Adverse Cardiovascular Events in Asymptomatic Subjects. *Circ Cardiovasc Imaging* (2020) **13**: doi:10.1161/CIRCIMAGING.119.009829

9. Spearman JV, Meinel FG, Schoepf UJ, Apfaltrer P, Silverman JR, Krazinski AW, et al. Automated quantification of epicardial adipose tissue using CT angiography: Evaluation of a prototype software. *Eur Radiol* (2014) **24**:519–526. doi:10.1007/s00330-013-3052-2

10. Petersen SE, Matthews PM, Bamberg F, Bluemke DA, Francis JM, Friedrich MG, et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank - rationale, challenges and approaches. *J Cardiovasc Magn Reson* (2013) **15**:46. doi:10.1186/1532-429X-15-46

11. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, et al. Multi-Ethnic Study of Atherosclerosis: Objectives and design. *Am J Epidemiol* (2002) **156**:871–881. doi:10.1093/aje/kwf113

12. Salton CJ, Chuang ML, O’Donnell CJ, Kupka MJ, Larson MG, Kissinger KV, et al. Gender differences and normal left ventricular anatomy in an adult population free of hypertension: A cardiovascular magnetic resonance study of the Framingham Heart Study Offspring cohort. *J Am Coll Cardiol* (2002) **39**:1055–1060. doi:10.1016/S0735-1097(02)01712-6

13. Petersen SE, Matthews PM, Francis JM, Robson MD, Zemrak F, Boubertakh R, et al. UK Biobank’s cardiovascular magnetic resonance protocol. *J Cardiovasc Magn Reson* (2015) **18**:8. doi:10.1186/s12968-016-0227-4

14. Ding X, Pang J, Ren Z, Diaz-Zamudio M, Jiang C, Fan Z, et al. Automated pericardial fat quantification from coronary magnetic resonance angiography: feasibility study. *J Med Imaging* (2016) **3**:014002. doi:10.1117/1.jmi.3.1.014002

15. Rado SD, Lorbeer R, Gatidis S, Machann J, Storz C, Nikolaou K, et al. MRI-based assessment and characterization of epicardial and paracardial fat depots in the context of impaired glucose metabolism and subclinical left-ventricular alterations. *Br J Radiol* (2019) **92**: doi:10.1259/bjr.20180562

16. 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. *JACC Cardiovasc Imaging* (2020) **13**:2017–2035. doi:10.1016/j.jcmg.2020.07.015

17. Raisi-Estabragh Z, Petersen SE. Cardiovascular research highlights from the UK Biobank: opportunities and challenges. *Cardiovasc Res* (2020) **116**:e12–e15. doi:10.1093/cvr/cvz294

18. UK Biobank: Protocol for a large-scale prospective epidemiological resource. (2007) Available at: https://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf [Accessed December 13, 2019]

19. Raisi-Estabragh Z, Harvey NC, Neubauer S, Petersen SE. Cardiovascular magnetic resonance imaging in the UK Biobank: a major international health research resource. *Eur Hear J - Cardiovasc Imaging* (2020)jea297. doi:10.1093/ehjci/jeaa297

20. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson* (2018) **20**:65. doi:10.1186/s12968-018-0471-x

21. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* (2015) **9351**:234–241. doi:10.1007/978-3-319-24574-4\_28

22. Drozdzal M, Vorontsov E, Chartrand G, Kadoury S, Pal C. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. (2016). Available at: https://link.springer.com/bookseries/558 [Accessed October 31, 2020]

23. Ibtehaz N, Rahman MS. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks* (2020) **121**:74–87. doi:10.1016/j.neunet.2019.08.025

24. Roy AG, Conjeti S, Navab N, Wachinger C. Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *Neuroimage* (2019) **195**:11–22. doi:10.1016/j.neuroimage.2019.03.042

25. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. (2016) Available at: http://arxiv.org/abs/1603.04467

26. Chen C, Bai W, Davies RH, Bhuva AN, Manisty CH, Augusto JB, et al. Improving the Generalizability of Convolutional Neural Network-Based Segmentation on CMR Images. *Front Cardiovasc Med* (2020) **7**:1–17. doi:10.3389/fcvm.2020.00105

27. Kingma DP, Ba JL. Adam: A method for stochastic optimization. *3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc* (2015)1–15.

28. Neglia D, Rovai D, Caselli C, Pietila M, Teresinska A, Aguadé-Bruix S, Pizzi MN, Todiere G, Gimelli A, Schroeder S, et al. Detection of Significant Coronary Artery Disease by Noninvasive Anatomical and Functional Imaging. *Circ Cardiovasc Imaging* (2015) **8**: doi:10.1161/CIRCIMAGING.114.002179

29. Ross R, Leger L, Morris D, De Guise J, Guardo R. Quantification of adipose tissue by MRI: Relationship with anthropometric variables. *J Appl Physiol* (1992) **72**:787–795. doi:10.1152/jappl.1992.72.2.787

30. Positano V, Gastaldelli A, Sironi AM, Santarelli MF, Lombardi M, Landini L. An accurate and robust method for unsupervised assessment of abdominal fat by MRI. *J Magn Reson Imaging* (2004) **20**:684–689. doi:10.1002/jmri.20167

31. Sicari R, Sironi AM, Petz R, Frassi F, Chubuchny V, De Marchi D, et al. Pericardial rather than epicardial fat is a cardiometabolic risk marker: An MRI vs echo study. *J Am Soc Echocardiogr* (2011) **24**:1156–1162. doi:10.1016/j.echo.2011.06.013

32. Sironi AM, Gastaldelli A, Mari A, Ciociaro D, Positano V, Buzzigoli E,et al. Visceral fat in hypertension: Influence on insulin resistance and β-cell function. *Hypertension* (2004) **44**:127–133. doi:10.1161/01.HYP.0000137982.10191.0a

33. Sironi AM, Petz R, De Marchi D, Buzzigoli E, Ciociaro D, Positano V, et al. Impact of increased visceral and cardiac fat on cardiometabolic risk and disease. *Diabet Med* (2012) **29**:622–627. doi:10.1111/j.1464-5491.2011.03503.x

34. X. Li, S. Chen, X. Hu and J. Yang, "Understanding the Disharmony Between Dropout and Batch Normalization by Variance Shift," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2677-2685, doi: 10.1109/CVPR.2019.00279.

35. Valindria VV, Lavdas I, Bai W, et al. Reverse Classification Accuracy: Predicting Segmentation Performance in the Absence of Ground Truth. IEEE Transactions on Medical Imaging. 2017 Aug;36(8):1597-1606. DOI: 10.1109/tmi.2017.2665165.

**Table 1. Standard segmentation performance metrics** **for pairwise comparisons of manually-contoured PAT by 3 observers (O1-3), and comparing automated segmentation with manual for the test set. All values are mean (standard deviation)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | O1 vs O2 (n=50) | O1 vs O3 (n=50) | O2 vs O3 (n=50) | Automated vs manual (O1) (n=87) |
| Intersection-over-Union | 0.689 (0.133) | 0.636 (0.153) | 0.678 (0.123) | 0.677 (0.116) |
| Dice | 0.808 (0.102) | 0.766 (0.127) | 0.801 (0.096) | 0.800 (0.090) |
| Mean contour distance (mm) | 2.78 (2.44) | 3.83 (3.48) | 3.79 (3.44) | 2.79 (2.35) |
| Hausdorff distance (mm) | 30.1 (23.8) | 37.0 (28.6) | 39.9 (28.8) | 29.9 (22.9) |

**Table 2. Logistic regression for prediction of diabetes in the UK Biobank dataset. The dataset includes 2,529 diabetics, and 40,399 non-diabetics. Odds ratios for PAT area are indicated for an increase of 10cm2.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Variable | Odds ratio | 95% CI | p-value |
| Model 1 | PAT Area (cm2) | 1.44 | 1.40, 1.47 | 8.43 10-164 |
| Model 2 | PAT Area (cm2) | 1.36 | 1.32, 1.40 | 1.26 10-92 |
| Male sex | 1.32 | 1.20, 1.44 | 8.97 10-9 |
| Age (years) | 1.03 | 1.03, 1.04 | 4.66 10-30 |
| Model 3 | PAT Area (cm2) | 1.07 | 1.03, 1.10 | 1.41 10-4 |
| Male sex | 1.80 | 1.63, 1.98 | 2.43 10-32 |
| Age (years) | 1.05 | 1.04, 1.05 | 7.91 10-54 |
| BMI | 1.18 | 1.17, 1.19 | 4.27 10-305 |

**Table 2 footnote:** BMI: body mass index; CI: confidence interval; PAT: pericardial adipose tissue.

**Figure 1. Two examples of PAT contoured in end-diastole on 4-chamber bSSFP cine-CMR, performed using CVI42® software according to the established SOP**

A picture containing text, indoor

Description automatically generated

**A**

**A**

**B**

**B**

**Figure 1 footnote:** bSSFP: balanced steady state free precession; CMR: cardiovascular magnetic resonance; PAT: pericardial adipose tissue; SOP: standard operating procedure.Images reproduced with permission of UK Biobank©.

**Central Illustration. Summary of Model architecture used in the present study**

Timeline

Description automatically generated

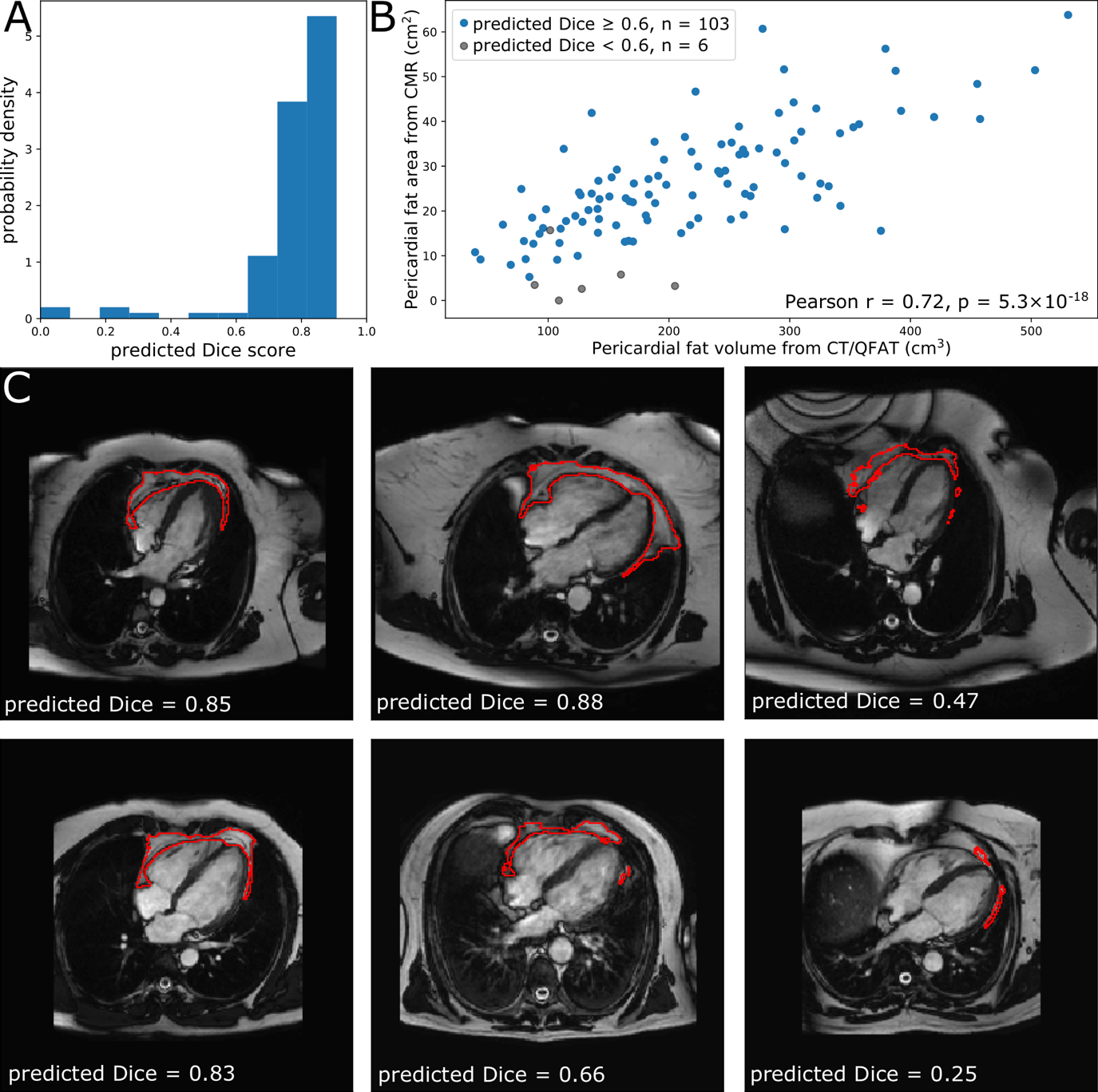
**Central illustration footnote:** The MultiRes blocks (**A**) form the encoder and decoder arms of the network. The number of filters used throughout the different components of the block is parameterised by U. The encoder and decoder arms are joined by Res paths (**B**), which are parameterised by F and L. They are formed of L repeating units, and their convolutional components each have F filters. The complete network is shown in (**C**). In (**C**), the colours indicate the placement of MultiRes blocks (**A**) and Res paths (**B**), while the hyperparameters used in each instance of the blocks are indicated as overlaid text. Because of the permanently-active dropout components, each prediction the network makes is equivalent to a Monte Carlo (MC) sample. **D** shows three such samples drawn based on the same input image. Note the disagreement at the edges of the segmented regions, particularly clear as shown on the overlay (far right). Images reproduced with permission of UK Biobank©

**Figure 2. Model Performance. A-D** Histograms of standard segmentation performance metrics on the test set (n=87) **E-H** Bland-Altman plots of PAT area between manual measurement between measurements by different human observers, and a human observer (O1) and automated measurement. The x-axis denotes the average of two measurements and the y-axis denotes the difference between them. The dark line is the mean difference, and the dashed lines show standard deviations from the mean. **E-G** show the inter-observer variability evaluated by 3 observers (O1-3) on a randomly selected subset of the manually contoured training set (n=50 subjects). **H** shows the agreement between automated and manual measurements in the manually contoured test set (n=87 subjects). **I** Example segmentations from the test set, with annotations showing Dice score and the predicted Dice score. Images reproduced with permission of UK Biobank©

A picture containing calendar

Description automatically generated

**Figure 3. Comparison of quantified PAT from CT and CMR. A** The predicted Dice scores of the segmented data. **B** Correlation between PAT volume quantified via QFAT software and PAT area quantified using our method. Subjects with a predicted Dice < 0.6 were excluded from Pearson analysis. **C** Some example CMR images, their automatically segmented PAT, and the predicted segmentation quality, are also shown for reference



**Supplementary Material**

**Selection of Monte Carlo sample size:**

It is expected that drawing more samples from a well-trained network will increase its accuracy, but with diminishing returns. Where the area of “foreground” pixels is particularly of interest (as in this use case, quantifying the area of PAT), we can report the mean and standard deviation of the areas across the samples. The standard deviation is useful in that it can be propagated forward and used in downstream analyses. Again, we would expect increasing to increase this standard deviation for any individual input image, but with diminishing returns.

Thus, we examined the effect of changing on the mean segmentation accuracy and the mean estimated standard deviation of segmented areas (**Figure S1)**. Where comparisons with a deterministic neural network are shown (**Figure S1A**), this network was trained separately, using the same split data for training and evaluation - consistent with prior work(24), there was no sacrifice in segmentation quality by using a stochastic method. There is a clear trade-off between computational resources and accuracy - when is increased (requiring more computational time), there is an increase in the segmentation accuracy of the consensus, the uncertainty in the derived areas, and the correlations between accuracy proxies and the true accuracy. However, there are diminishing returns on all of these values. On this basis, was set to 15 for all further work.

**Selection of metric used in quality control:**

When multiple MC samples are drawn from the stochastic neural network, their level of agreement is correlated with the quality of the consensus segmentation (Roy et al. 2019). However, their level of agreement can be measured in many different ways. Here, we use 4 metrics, based on the well-known intersection-over-union (or Jacard Index) and the Dice score. However, there are a number of ways to formulate these for a summary statistic over MC samples (denoted ).

The IoU can be calculated globally (as in (Roy et al. 2019)):

Alternatively, it can be calculated as the mean of pairwise combinations across the MC samples:

Likewise, the Dice score can be calculated globally over all MC samples:

Or, it can be calculated as the mean of pairwise combinations (Roy et al. 2019):

The performance of these metrics can be quantified in a number of ways (after a linear correction is applied):

* The Pearson correlation coefficient between the prediction and the true Dice score.
* The Mean Absolute Error (MAE) between the prediction and the true Dice score.
* How well the predicted categories of segmentation quality match up with those provided by the true Dice score, with descriptors of poor, medium and good corresponding to Dice scores of <0.6, 0.6-0.8 and >0.8 respectively.

These indicators of success can be seen in **Figure S2**, which also shows the linear corrections applied to each metric. Based on these results, the best predictor of segmentation accuracy is the mean pairwise Dice score . This is contrary to previous result(24), which finds the global IoU to be superior to , and additionally did not find a requirement for linear corrections. For the remainder of the paper we use for MC sampling, and the mean pairwise Dice score with linear correction for predicting the Dice score of each individual segmentation.

**Supplementary Figure 1. Effects of on model performance**

Chart, scatter chart

Description automatically generated

***Supplementary Figure 1:*** *Each graph here shows meanstandard deviation of 15 runs for each Monte Carlo , over the test set ( participants).* ***A.*** *Mean Dice score. For comparison, the optimal Dice score obtained with a deterministic MultiResUNet is also shown (dotted line).* ***B.*** *Mean esstimated standard deviation of PAT area.*

**Supplementary Figure 2. Effectiveness of different methods for predicting segmentation quality** for the test set (), as well as statistics describing their performance (Pearson , the mean absolute error of predictions after linear corrections, and the category accuracy).

Chart, scatter chart

Description automatically generated