Abdominal musculature segmentation and surface prediction from CT using deep learning for sarcopenia assessment

Paul Blanc-Durand, Jean-Baptiste. Schiratti, Kathryn Schutte, Paul Jehanno, Paul Herent, Frédéric Pigneur, Olivier Lucidarme, Y. Benaceur, Alexandre Sadate, Alain Luciani, et al.

Version of Record: https://www.sciencedirect.com/science/article/pii/S2211568420301224
Manuscript_0a27313181fd8101711c3dfe73fcf4f5

Abdominal musculature segmentation and surface prediction from CT using deep learning for sarcopenia assessment

Short title:

Deep learning for sarcopenia assessment

Paul BLANC-DURAND[a,b,c,d*], Jean-Baptiste SCHIRATTI[d], Kathryn SCHUTTE[d], Paul JEHANNO[d], Paul HERENT[d], Frédéric PIGNEUR[e], Olivier LUCIDARME[f,g], Yassine BENACEUR[h], Alexandre SADATE[h], Alain LUCIANI[e], Olivier ERNST[i], Aymeric ROUCHAUD[j], Maud CREUZE[k,l], Axel DALLONGEVILLE[m], Nathan BANASTE[n], Medhi CADI[o], Imad BOUSAID[p], Nathalie LASSAU[l,q], Simon JEGOU[d]

[a]Department of Nuclear Medicine, CHU Henri Mondor, AP-HP, 94010 Créteil, France

[b]INSERM IMRB, Team 8, U-PEC, 94000 Créteil, France

[c]INRIA Epione Team, 06410 Sophia Antipolis-, France

[d]Owkin, 12 rue Martel, 75013 Paris, France

[e] Department of Radiology, CHU Henri Mondor, AP-HP, 94010Créteil, France

[f] Department of Radiology, CHU Pitié Salpétrière-Charles Foix, AP-HP, Paris, 75013, France

[g]Sorbonne Université, CNRS, INSERM, Laboratoire d'Imagerie Biomédical

[h] Department of Radiology, Nimes University Hospital , 34000 Nimes France, F34010

[i] Department of Radiology, Centre Hurriez, CHU de Lille, Université de Lille, 59000 Lille, France

[j] Department of Radiology, Hospices Civils de Lyon, 69000 Lyon, France

[k] Department of Radiology, Hôpitaux Universitaires Paris-Sud, AP-HP, 94270 Le Kremlin Bicêtre, France

BIOMAPS. Université Paris-Saclay, Inserm, CNRS, CEA, Laboratoire d'Imagerie Biomédicale Multimodale Paris-Saclay, 94805, Villejuif, France

[m] Department of Radiology, Hopital Saint Joseph, 75014 Paris, France

[n] Department of Radiology, Centre Léon Bérard, 69000 Lyon, France

[o]Radiologie Paris Ouest, 92200 Neuilly sur Seine, France

[p]Département de la transformation numérique et du système d'information. Gustave Roussy Cancer Campus. Université Paris Saclay, 94805Villejuif, France

[q] Department of Radiology, Gustave Roussy Cancer Campus. Université Paris Saclay, 94805Villejuif, France

*Corresponding author: paul.blancdurand@aphp.fr

Abstract

Purpose: The purpose of this study was to build and train a deep convolutional neural networks (CNN) algorithm to segment muscular body mass (MBM) to predict muscular surface from a two-dimensional axial computed tomography (CT) slice through L3 vertebra.

Materials and methods: An ensemble of 15 deep learning models with a two-dimensional U-net architecture with a 4-level depth and 18 initial filters were trained to segment MBM. The muscular surface values were computed from the predicted masks and corrected with the algorithm's estimated bias. Resulting mask prediction and surface prediction were assessed using Dice similarity coefficient (DSC) and root mean squared error (RMSE) scores respectively using ground truth masks as standards of reference.

Results: A total of 1,025 individual CT slices were used for training and validation and 500 additional axial CT slices were used for testing. The obtained mean DSC and RMSE on the test set were 0.97 and 3.7 cm^2 respectively.

Conclusion: Deep learning methods using convolutional neural networks algorithm enable a robust and automated extraction of CT derived MBM for sarcopenia assessment, which could be implemented in a clinical workflow.

Keywords: Tomography, X-ray computed; Deep learning; Muscular body bass; Sarcopenia; Convolutional neural networks (CNN)

Abbreviations

2D: Two-dimensional

BMI: Body mass index

CNN: Convolutional neural network

CT: Computed tomography

DSC: Dice similarity coefficient

MBM: Muscular body mass

ME: Mean error

MRI: Magnetic resonance imaging

RMSE: Root mean squared error

SD: Standard deviation

TAMA: Total abdominal muscle area

Introduction

Body composition including assessment of muscular body mass (MBM) is of increased importance in oncology and several chronic diseases. In addition, body composition is associated with treatment toxicity that can affect patient survival [1–3]. More specifically, sarcopenia, which is defined by a loss of muscular mass and function, is frequently observed in up to 50% of patients with cancer [4].

Both physical examination and imaging techniques can be used to diagnose sarcopenia. Anthropometry, which generically refers to body measurements during clinical examination includes body mass index (BMI), skin-fold thickness and body circumference. In addition to physical examination imaging techniques such as computed tomography (CT), dual energy x-ray absorptiometry, ultrasonography and magnetic resonance imaging (MRI) are useful to assess body composition [5]. CT allows using three main indices to diagnose sarcopenia, which are psoas index, total abdominal muscle area (TAMA) and Hounsfield unit (HU) average calculation of body composition through estimation of tissue densities expressed in HU. For TAMA, some thresholds have been suggested [6]. Furthermore, calculation of these indices requires a delineation of muscles (also referred as segmentation) usually at the L3 or L4 vertebra level, which is time consuming and possibly affected by inter-observer variability. To make CT derived anthropometry clinically applicable, automatic approaches are mandatory.

As muscle and fat have different attenuation values on CT, they can be separated from one another using simple thresholding strategies [7]. However, because thresholding may result in a high level of noise, alternative methods have been developed such as thresholding and morphological operations [8], atlas methods [9] or conditional random fields [10]. More recently, convolutional neural networks (CNN), which originated from the deep learning community have been adopted by the radiological community because of their ability to learn spatial features from medical images. They are particularly effective for lesion segmentation or detection tasks in various medical applications [11, 12,13]. For the specific body composition estimation, CNN have already been employed to automatically identify the axial

slice at the mid-L3 level thus facilitating manual segmentation of MBM [14] or impulse the training of a second CNN for segmentation purposes [15]. Several studies have evaluated different CNN architectures on single CT slices for binary segmentations of adipose tissue or muscles [16,17]. Finally, Weston et al. trained a CNN with a U-net architecture on a large cohort to segment four compartments and used another independent cohort for further validation [18].

The purpose of this study was to build and train a deep CNN algorithm to segment MBM, in order to predict muscular surface from a two-dimensional (2D) axial CT slice at the level of L3 vertebra.

Materials and Methods

Study population

The CT data were provided as part of the "Sarcopenia Challenge" organized during the 2019 edition of the Journées Françaises de Radiologie, which is the annual meeting of the French Society of Radiology (Société Française de Radiologie). The complete dataset was composed of 1025 axial CT slices for training and validation and was released in two times. A first set was initially available, and a second set was released one month later. Five hundred additional axial CT slices were made available for testing. Inference had to be made in one hour. CT examinations came from multiple French institutions with different acquisition parameters and hardware. All CT examinations and ground truth masks were visually inspected and three of them were excluded because the ground truth masks were not correctly co-registered to their respective original CT examinations. Out of the 1,022 axial CT slices were finally included, 40 were randomly assigned to the validation set with stratification on the quality label.

Ground truth generation

The 2D ground truth masks T, which included abdominal belt and psoas muscles at the level of the mid L3 vertebra were manually annotated by eleven expert radiologists (O.L., N.L, Y.B, A.S., A.L., O.E., L.B., M.C.,A.D., N.B. et M.C.), using the public freeware 3DSlicer [19] with a standardized protocol ,which included a manual contouring of muscles followed by a fixed thresholding where pixels with attenuation value < -29 HU and > 150 HU were excluded. Each ground truth segmentation was reviewed by a third-party expert (F.P. ;

with an experience of over 1000 manual segmentations of MBM from CT data for sarcopenia assessment) and was assigned to a quality label ranging from 'A' to 'D' where : “A” corresponded to perfect segmentation ; “B” corresponded to a segmentation with only few pixels in the ground truth mask out of muscles and of which surface was < 0.5 cm^2; “C” for masks with a significant number of pixels > 0.5 cm^2 were misclassified ; “D” when CT were of non-diagnostic quality (noisy images or containing artifacts), if ground truth masks included non-muscular structure or if large muscular portions were missed.

Model

Pre-processing

All available individual CT slices and ground truth masks were resampled to a 512 x 512 tensor of 1 mm^2 pixel size, using a linear interpolation of order 3. Pixel values were clipped between -150 HU and 300 HU and normalized using a min-max normalization from [-150, 300] HU to [0, 1]. Data was stored with 16-bits floating-point precision.

Model architecture

We used an ensemble of 15 similar models. Each model was built on a 2D-Unet architecture with a 4-level depth and 18 initial filters. For all resolutions, a block of layers was designed as follows: three convolutional layers with a filter size of 3 and a rectified linear unit (ReLU) activation layer (that breaks the linearity of the model). The number of filters in the encoding path was doubled in each block. Finally, a decoding path mirrored the encoding path for the upsampling part. Each network counted 2,953,409 weights and had a receptive field of 10.7 × 10.7 cm^2. Therefore, the 15 models only differed by their weights as they have been trained separately. The average of the 15 outputs (one for each model) was used as final prediction. An overview of the network can be seen in Figure 1.

Network training

Training was performed on an Ubuntu workstation 16.04 with a 11-Go graphical processing unit GTX1080Ti for 150 epochs of 200 iterations of batch size 12. Initial learning rate was set to 10^{-3}. An adaptive learning rate scheme with a polynomial degree of order 5 was used. Weights were updated with the Adam optimizer. Main augmentation strategies included rotations (-10, +10) and scaling (0.9, 1.1). The Dice similarity coefficient (DSC) was

used as a loss function and was back-propagated to the network weights with 1 corresponding to a perfect overlap.

Post-processing

The predicted masks were then resampled back to the image original resolution using a linear interpolation of order 3. Only pixels with a probability of being muscle tissue over 0.5 were kept. The final mask was then computed using a morphological dilation with a squared connectivity of 1, and thresholded at [-29; 150] HU to follow the ground truth procedure.

Surface estimation

From the final mask (P), the predicted surface was estimated as the sum of non-zero pixels multiplied by the pixel size. On the validation set, the mean error (ME) was subtracted from each predicted surface of examinations with quality label A, B or C. This ME was defined as follows:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(\widehat{S_i} - S_i)$$

Where $\widehat{S_i}$ (resp. S_i) denotes the predicted (resp. true) muscle surface for the i-th individual.

Results

We analyzed the impact of our pre and post-processing on both DSC and root mean squared error (RMSE) in the validation set. As a baseline without preprocessing, DSC and RMSE were respectively 0.93 ± 0.03 (SD) (range: 0.85 - 0.97) and 4.7 ± 3.1 (SD) (range: 0.0 - 11.7). DSC and RMSE evolution as a function of the number of U-net model are shown in Figure 2. They reached a plateau at the tenth model, with further marginal improvement thereafter. It led to a 1.1% DSC rise and a 4.7% RMSE decrease. The post-processing, which included a dilation and a thresholding, improved DSC of 2.3% and decreased RMSE of 17.7%. Finally, after subtracting ME in the validation set to the predicted surface, RMSE decreased from 3.4 cm^2 ± 3.1 (SD) (range: 0.1 - 13.4) to 2.8 cm^2 ± 2.7 (SD) (range: 0.0 - 10.7). Waterfall plots of those different steps leading to the final performances reported are shown in Figure 3.

Some examples of predictions are shown in Figure 4. DSC in the validation and testing sets were respectively 0.96 ± 0.02 (SD) (range: 0.86 - 0.98) and 0.97 ± 0.02 (SD) (range: 0.89 - 0.99). Distribution of DSC for each quality label is displayed in Figure 5.

After post-processing operation, a surface overestimation was observed in the 40 CT slices used for validation (validation set) that is consistent with the dilation that justified our choice to correct our algorithm bias (Figure 6A). Finally, in the testing set MSE was 13.6 ± 23.6 (SD) (range: 0.0 - 169.5) and RMSE was 3.7 cm^2 ± 2.3 (SD) (range: 0.0 - 13.0). A scatter plot of predicted surfaces and ground truth surface for the testing set is shown in Figure 6B.

Discussion

This study proposed and validated a deep learning algorithm to automatically segment MBM from CT data using an ensemble of U-net architectures. We obtained a DSC of 0.97 ± 0.02 (SD) in the testing set resulting in optimal performances by comparison with prior studies [17, 20]. Using fully connected network Lee et al. reported a DSC of 0.93 for MBM segmentation [17] while using a similar architecture Park et al. reported DSC a 0.96 [20]. More recently, Weston et al. trained and validated a U-net and reached a 0.96 DSC in a large cohort for MBM segmentation but also achieved high performances both for intra-abdominal and subcutaneous fat segmentations [18].

From a methodological point of view, pre-processing was limited as it was only composed of resampling to a 1mm^2 isotropic pixel and a min-max normalization without whitening. We stored our data with 16-bits floating-point precision as Lee et al. demonstrated for the specific body composition segmentation task that neither min-max normalization (referred as windowing) nor a minimum of 256 grey levels (stored on at least 8 bits) impacted DSC in their model [17]. After the post-processing operation, a surface overestimation was observed in the 40 CT slices used as validation set. As a result, we decided to correct the algorithm's bias by subtracting the validation's mean error. We also chose to exclude images with low quality segmentations (denoted by the "D" label quality) for mean error estimation as we believe that ground truth surface may not be reliable enough in that population. We showed that our algorithm was robust to poor label quality. However, the DSC associated with lower quality exams is slightly lower and with higher variance, which can be explained either by an incorrect algorithm prediction or an incorrect ground truth segmentation.

Main limitations of this work include the fact that, even if a large dataset for testing was available, a proper independent cohort would be mandatory to validate the algorithm.

Also because of the anonymization process nor height nor weight were available which did not allow us to compare performances between groups such as sex or stratified on body mass index.

In conclusion, deep learning makes CT derived MBM calculation clinically feasible within the daily radiological workflow. The latter will allow routinely obtaining new biomarker that may provide better diagnosis of sarcopenia and which may benefit to patients within tailored and personalized medicine.

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship. All the authors had fully participated to the study and approved the final draft.

Disclosures

JBS, KS, PJ, PH, SJ, are employees of Owkin. The other authors do not declare any conflict of interest.

Acknowledgements

We would like to thank the Société Française de Radiologie for the opportunity to organize these challenges during the Journées Francophones de Radiologie.

We would like to thank Gustave-Roussy for the resources mobilized and the data hosting.

We would like to thank the Commission Nationale de l'Informatique et des Libertés (CNIL) for their support.

We would like to thank Jean François Raffier and Easys Consulting for the technical support during the challenge.

References

[1] Shachar SS, Williams GR, Muss HB, Nishijima TF. Prognostic value of sarcopenia in adults with solid tumours: a meta-analysis and systematic review. Eur J Cancer 2016;57:58–67.

[2] Hopkins JJ, Sawyer MB. A review of body composition and pharmacokinetics in oncology. Expert Rev Clin Pharmacol 2017;10:947–56.

[3] Madico C, Herpe G, Vesselle G, Boucebci S, Tougeron D, Sylvain C, et al. Intra peritoneal abdominal fat area measured from computed tomography is an independent factor of severe acute pancreatitis. Diagn Interv Imaging 2019;100:421–6.

[4] Hilmi M, Jouinot A, Burns R, Pigneur F, Mounier R, Gondin J, et al. Body composition and sarcopenia: the next-generation of personalized oncology and pharmacology? Pharmacol Ther 2019;196:135-159.

[5] Boutin RD, Yao L, Canter RJ, Lenchik L. Sarcopenia: Current Concepts and Imaging Implications. AJR Am J Roentgenol 2015;205:W255–266.

[6] Prado CM, Lieffers JR, McCargar LJ, Reiman T, Sawyer MB, Martin L, et al. Prevalence and clinical implications of sarcopenic obesity in patients with solid tumours of the respiratory and gastrointestinal tracts: a population-based study. Lancet Oncol 2008;9:629–635.

[7] Mendez J, Keys A. Density and composition of mammalian muscle. Metabolism 1960. https://eurekamag.com/research/024/450/024450136.php (accessed May 17, 2019).

[8] Mensink SD, Spliethoff JW, Belder R, Klaase JM, Bezooijen R, Slump CH. Development of automated quantification of visceral and subcutaneous adipose

tissue volumes from abdominal CT scans. Med Imaging 2011; Computer-aided Diagnosis (Proceedings of SPIE): 79632Q.

[9] Decazes P, Tonnelet D, Vera P, Gardin I. Anthropometer3D: automatic multi-slice segmentation software for the measurement of anthropometric parameters from CT of PET/CT. J Digit Imaging 2019;32:241-250.

[10] Hussein S, Green A, Watane A, Papadakis G, Osman M, Bagci U. Context driven label fusion for segmentation of subcutaneous and visceral fat in CT volumes. ArXiv151204958 Cs 2015.

[11] Couteaux V, Si-Mohamed S, Nempont O, Lefevre T, Popoff A, Pizaine G, et al. Automatic knee meniscus tear detection and orientation classification with mask-RCNN. Diagn Interv Imaging 2019;100:235–242.

[12] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.

[13] Colevray M, Tatard-Leitman VM, Gouttard S, Douek P, Boussel L. Convolutional neural network evaluation of over-scanning in lung computed tomography. Diagn Interv Imaging 2019;100:177-183.

[14] Belharbi S, Chatelain C, Hérault R, Adam S, Thureau S, Chastan M, et al. Spotting L3 slice in CT scans using deep convolutional network and transfer learning. Comput Biol Med 2017;87:95–103.

[15] Bridge CP, Rosenthal M, Wright B, Kotecha G, Fintelmann F, Troschel F, et al. Fully-automated analysis of body composition from CT in cancer patients using convolutional neural networks. In: Stoyanov D, Taylor Z, Sarikaya D, McLeod J, González Ballester MA, Codella NCF, et al., editors. 20 Context-Aware Oper. Theaters Comput. Assist. Robot. Endosc. Clin. Image-Based Proced. Skin Image Anal., Cham: Springer International Publishing; 2018, p. 204–213.

[16] Wang Y, Qiu Y, Thai T, Moore K, Liu H, Zheng B. A two-step convolutional neural network based computer-aided detection scheme for automatically segmenting adipose tissue volume depicting on CT images. Comput Methods Programs Biomed 2017;144:97–104.

[17] Lee H, Troschel FM, Tajmir S, Fuchs G, Mario J, Fintelmann FJ, et al. Pixel-level deep segmentation: artificial intelligence quantifies muscle on computed tomography for body morphometric analysis. J Digit Imaging 2017;30:487–498.

[18] Weston AD, Korfiatis P, Kline TL, Philbrick KA, Kostandy P, Sakinis T, et al. Automated abdominal segmentation of CT scans for body composition analysis using deep learning. Radiology 2019;290:669–679.

[19] Kikinis R, Pieper SD, Vosburgh KG. 3D Slicer: A platform for subject-Specific image analysis, visualization, and clinical support. In: Jolesz FA, editor. Intraoperative Imaging Image-Guid. Ther., New York, NY: Springer; 2014, p. 277–289.

[20] Park HJ, Shin Y, Park J, Kim H, Lee IS, Seo DW, et al. Development and validation of a deep learning system for segmentation of abdominal muscle and fat on computed tomography. Korean J Radiol 2020;21:88–100.

Figure Legends

Figure 1. Diagram shows the architecture of the model used. It consists of an ensemble of 15 U-net architectures where the mean of predictions is used as final prediction.

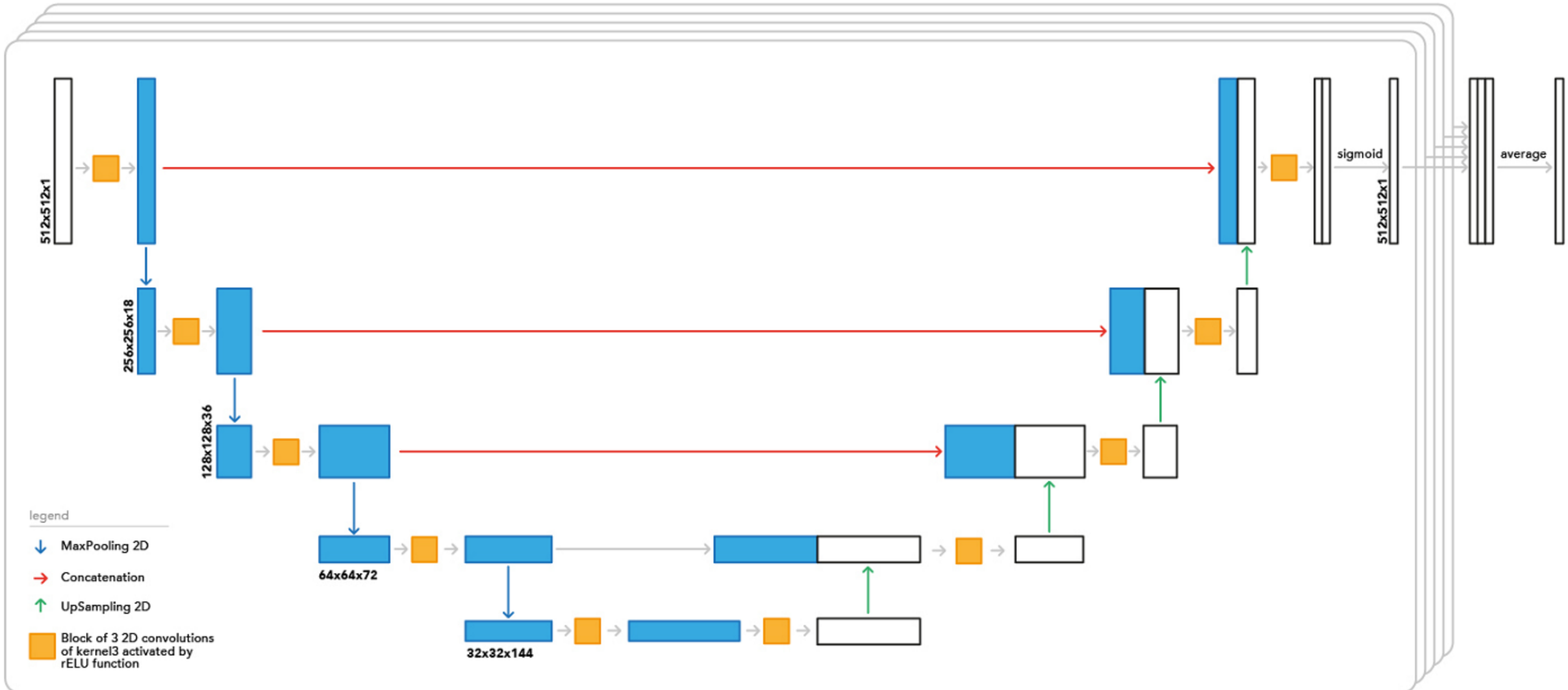
Figure 2. Graphs show evolution of Dice similarity coefficient (DSC) in the validation set as a function of the number of U-net used in the ensemble of models. **A,** Mean DSC in the validation set of the 15 trained models with min-max DSC for each epoch. For clarity, we set the epochs axis to 50 instead of 150. The black dashed line corresponds to the 0.95 DSC. **B,** Evolution of DSC when the number of U-net increased. **C,** Evolution of root mean squared error (RMSE) when the number of U-net increased.

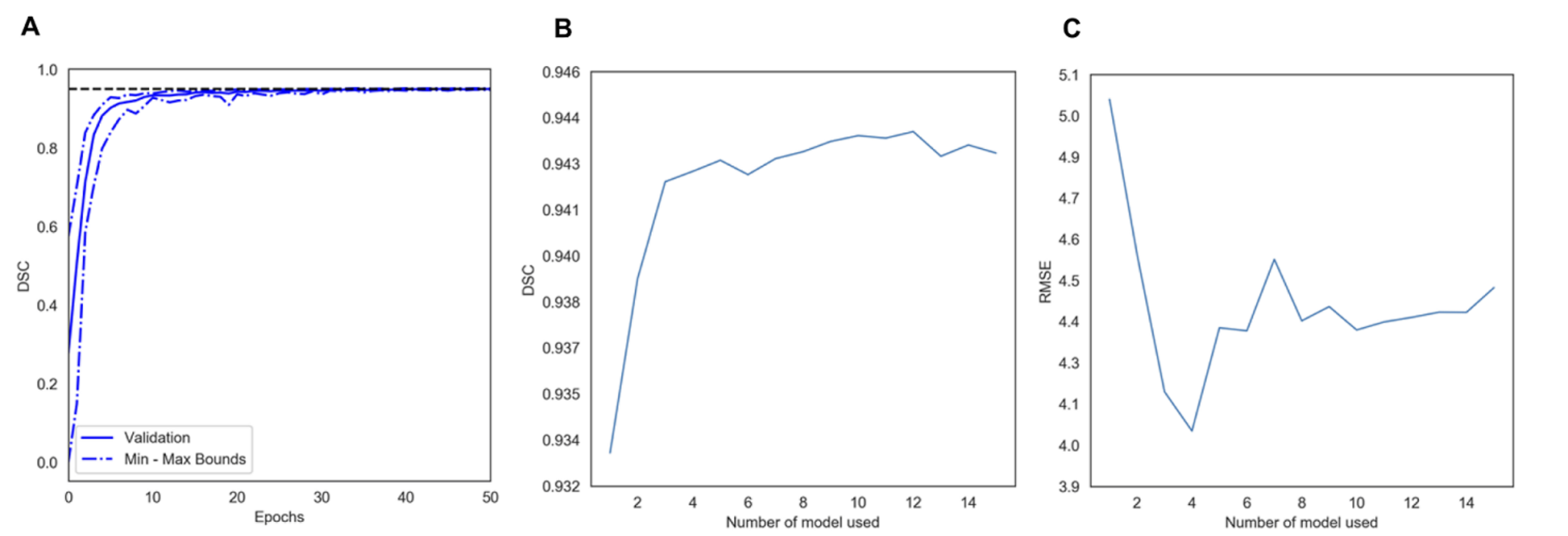
Figure 3. Waterfall plots show impact on Dice similarity coefficient(DSC) (**A),** and root mean squared error (RMSE) **(B),** when multiple models are used, with post-processing (including a dilation and a fixed thresholding), and after correction of models bias for RMSE.

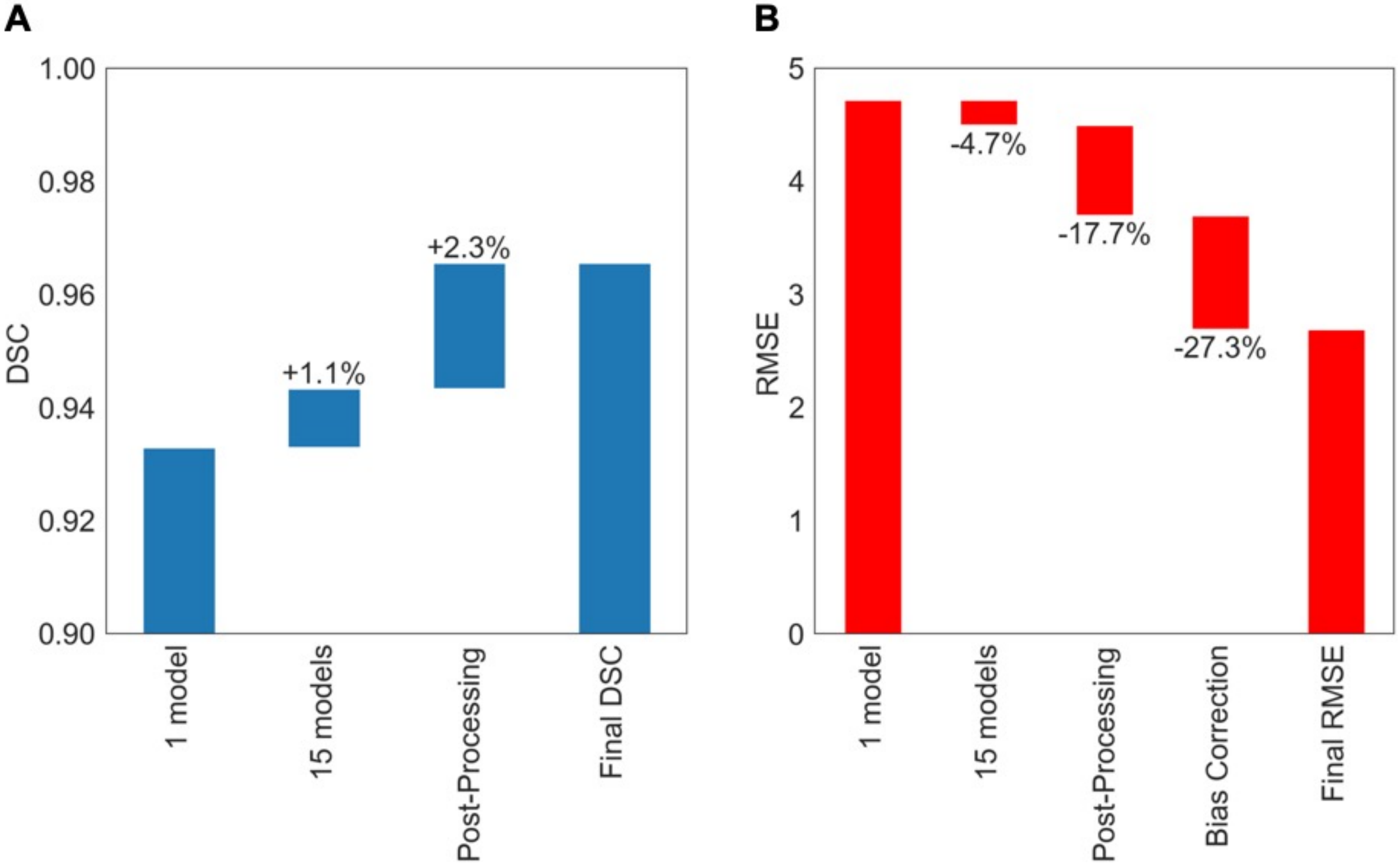
Figure 4. Left part of figure shows examples of three axial CT slices in the validation set overlaid (*in red*) with their ground truth masks. Right part of figure shows CT overlaid with in green the true-positive pixels, in blue the false-positive pixels and in red the false-negative pixels. Dice similarity coefficients are 0.971 for **A**, 0.962 for **B,** and 0.962 for **C**.

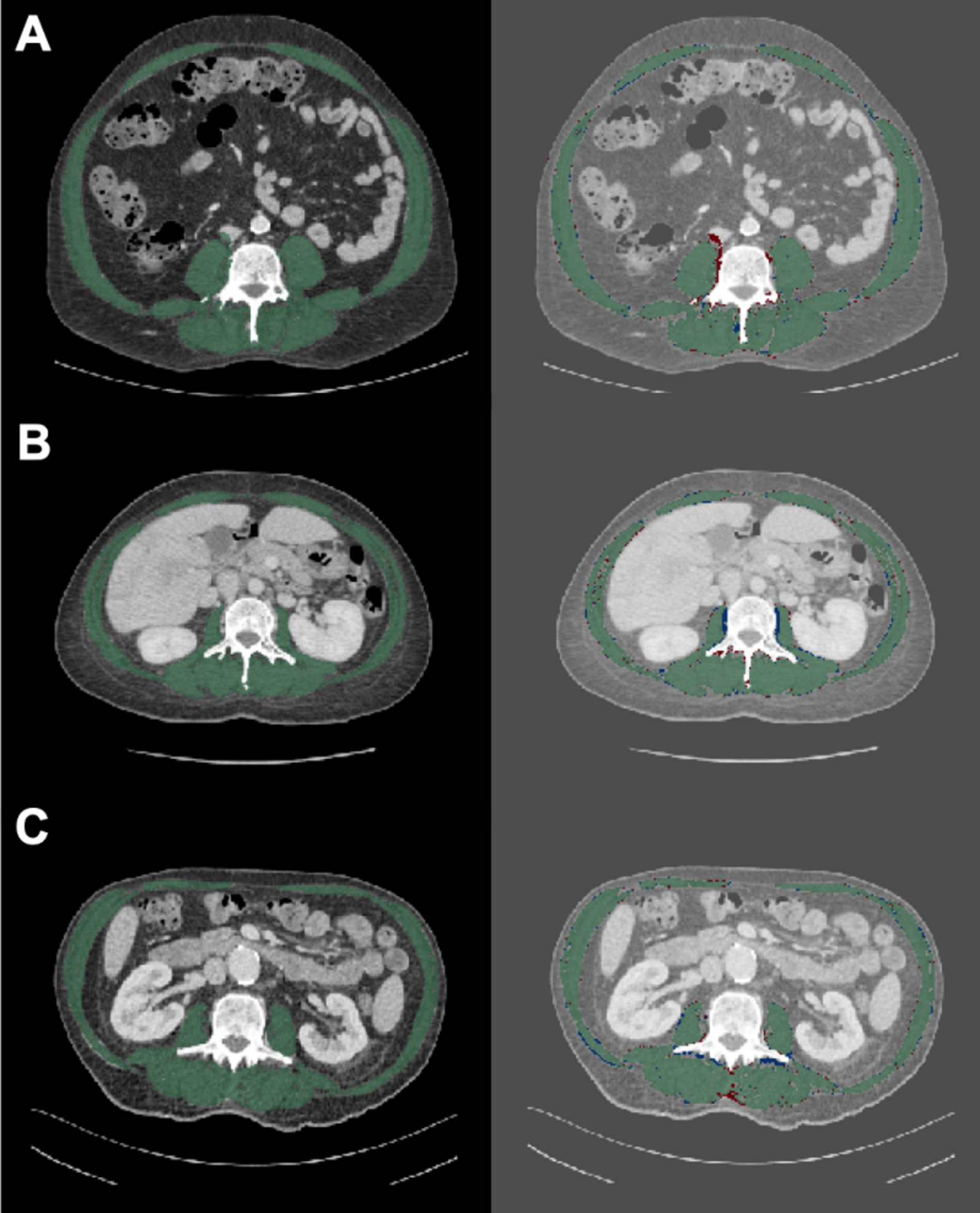
Figure 5. Violin plots show the distributions estimated by kernel density estimate of Dice similarity coefficient (DSC) among the different quality labels of the ground truth masks ranging from better to worse from A to D (x-axis) for the testing set. Medians (white dots), interquartile ranges (thick black line) and 1.5 × interquartile ranges (thin black line) are over-imposed with the kernel density estimate.

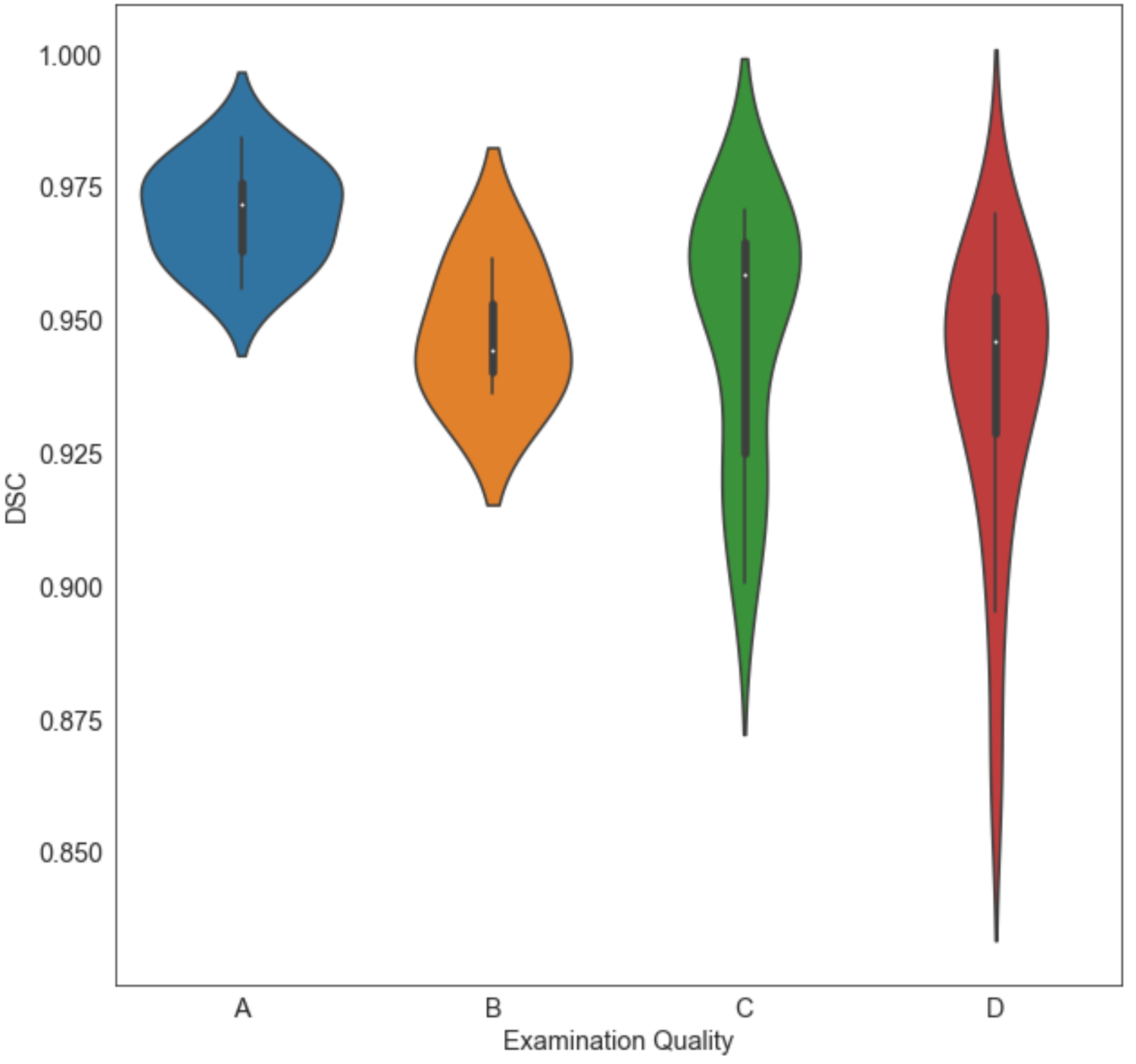
Figure 6. Graphs show scatter plot of predicted area and ground truth area before (*blue dots)* and after (*green stars*) algorithm bias correction using linear regression in the validation set **A** and a scatter plot of predicted area and ground truth area in the testing set **B.** Of note, on A, most of the blue dots are at the right part of the identity line, indicating surface overestimation.

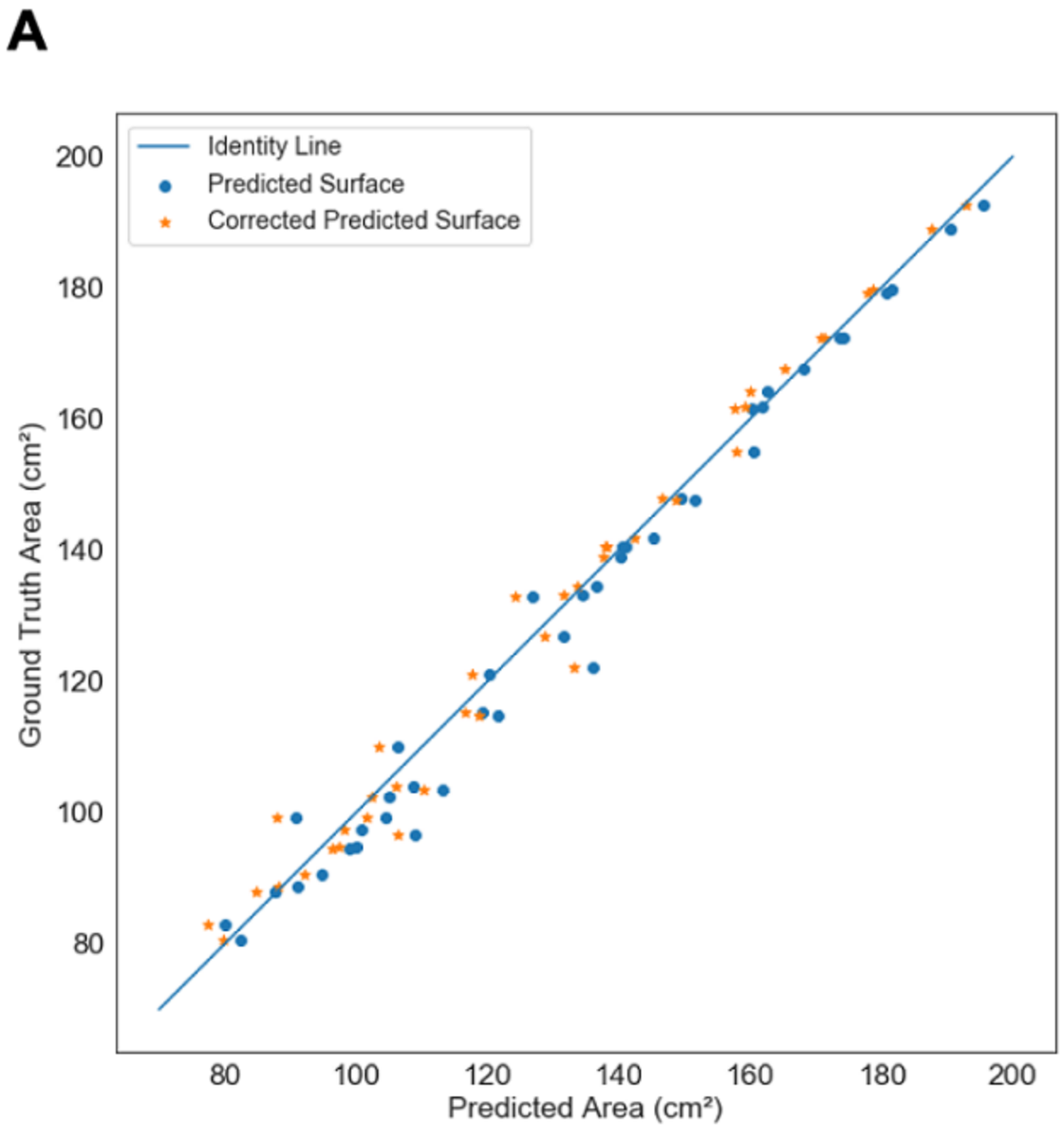

512x512x1
256x256x18
128x128x36
64x64x72
32x32x144
sigmoid
512x512x1
average
legend
MaxPooling 2D
Concatenation
UpSampling 2D
Block of 3 2D convolutions of kernel3 activated by rELU function

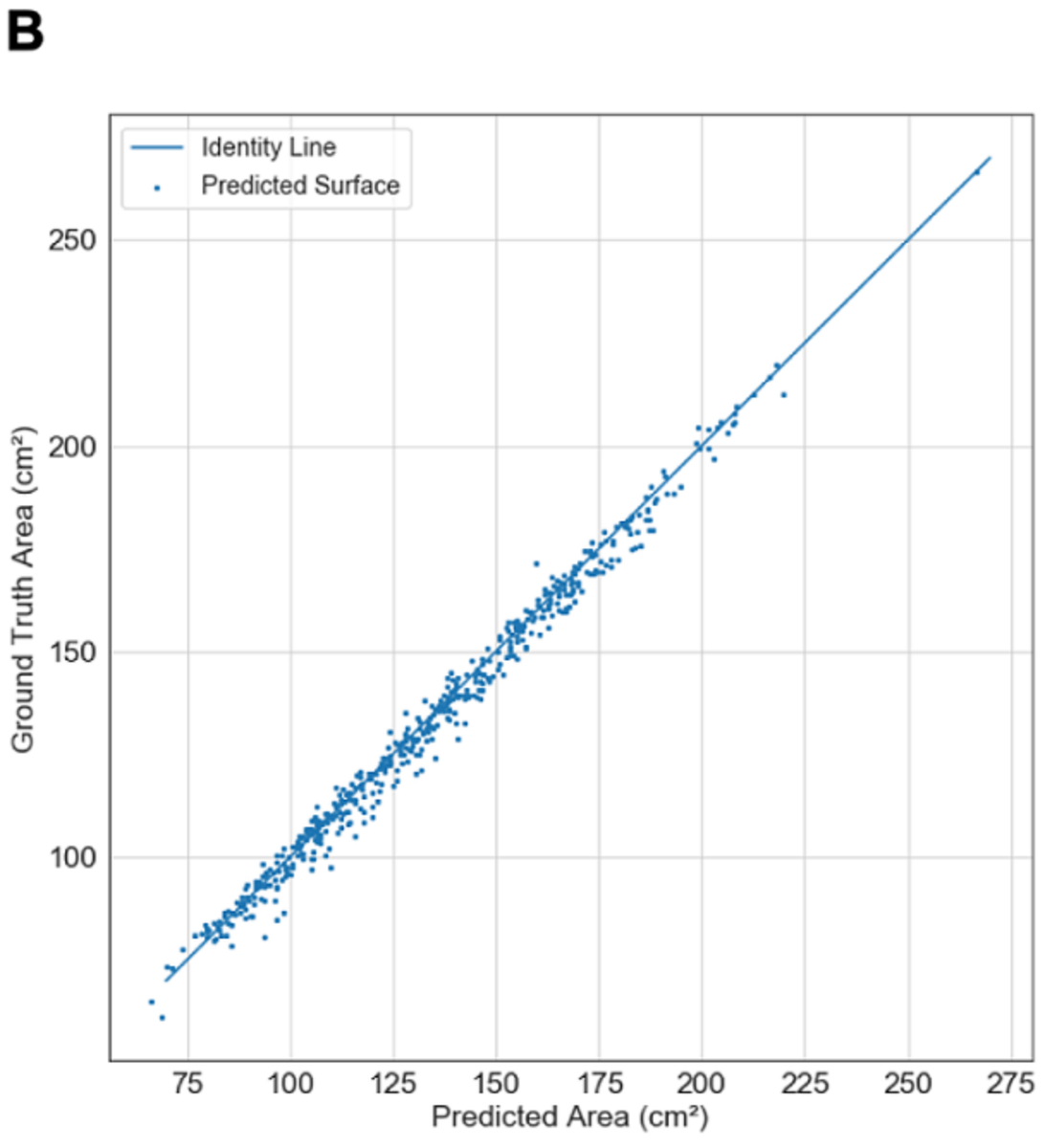

A
B
C
DSC
Epochs
Validation
Min - Max Bounds
Number of model used
RMSE


A
1.00
0.98
0.96
0.94
0.92
0.90
DSC
+1.1%
+2.3%
1 model
15 models
Post-Processing
Final DSC
B
5
4
3
2
1
0
RMSE
-4.7%
-17.7%
-27.3%
1 model
15 models
Post-Processing
Bias Correction
Final RMSE


A
B
C


1.000
0.975
0.950
0.925
0.900
0.875
0.850
DSC
A
B
C
D
Examination Quality


A
Identity Line
Predicted Surface
Corrected Predicted Surface
Ground Truth Area (cm²)
Predicted Area (cm²)
80
100
120
140
160
180
200


B
Identity Line
Predicted Surface
Ground Truth Area (cm²)
Predicted Area (cm²)
100
150
200
250
75
100
125
150
175
200
225
250
275