의학박사 학위논문

복부 전산화단층촬영 영상에서 인공지능을 이용한 자동 L3 level 탐색모델

A fully automatic deep learning system for L3 slice selection and body composition assessment on abdominal computed tomography

울산대학교대학원

의 학 과

하지연

복부 전산화단층촬영 영상에서 인공지능을 이용한 자동 L3 level 탐색모델

지도교수 김경원

이 논문을 의학박사 학위 논문으로 제출함

2021년 02월

울산대학교 대학원
의 학 과
하지연

하지연의 의학박사학위 논문을 인준함

심사위원 김정곤 (인)

심사위원 김경원 (인)

심사위원 최상현 (인)

심사위원 우동철 (인)

심사위원 이정진 (인)

울 산 대 학 교 대 학 원

2021 년 2 월

국문요약

연구배경

근감소증 연구의 중요성이 대두되며 복부 전산화 단층 촬영영상에서 근육량을 정량적으로 분석해야하는 증례의 수 역시 가파르게 증가하고 있다. 이러한 정량적 측정의 자동화에 있어 가장 첫번째 만나는 장애물은 연구에 적절한 단면영상을 선택하는 단계로 세번째 요추(L3)의 아래쪽 끝단이 연구에 적절한 단면으로 알려져 있다. 척추변이는 보고에 따라 일반인구의 4-30% 정도에서 보이는 것으로 알려져 있으며 이러한 척추의 변이는 정확한 L3 CT slice 선택에 고려되어야 하는 중요한 인자 중 하나이다.

연구목적

본 연구에서는 L3 CT slice 를 선택하여 근육량을 측정하는 완전 자동화 인공지능모델을 개발하고 임상적으로 검증하고자 한다.

연구방법

L3 slice selection 모델은 YOLOv-3 를 기반으로 하는 모델로 L3SEG-net 으로 명명하였

다. 이 과정 후 FNC 을 기반으로 하는 또 다른 인공지능 모델을 연결하여 근육과 내장/피하 지방을 분할하였다. 이 두 과정은 수행자의 개입없이 자동으로 이어져 하나의 모델로 구동될 수 있도록 개발하였다. 총 922 명의 환자에서 1496 개의 복부 전산화단층촬영영상을 developmenta dataset 으로 사용하였으며 이를 8:2 로 분할하여 training 및 tunning set 으로 활용하였다. 검증을 위한 dataset 은 서울아산병원에서 추출한 internal validation dataset (n=496) 과 서로 다른 세 개 병원에서 추출한 external validation dataset (n=586) 으로 구성하였다. 요추에 수술로 인한 하드웨어 있는 환자는 dataset 에서 제외하였다. 인공지능 모델에서 선택한 level 과 ground truth level 과의 차이를 mm 단위로 계산하였고 10mm 이하를 technical success 로 정의하였다. 전반적은 분할의 정확성는 cross sectional area error 로 평가하였다. 척추변이에 따른 인공지능모델의 성능 역시 평가하였다.

연구결과

전체 dataset 에서 인공지능모델의 정확도는 모델이 선택한 level 과 ground truth 사이 거리 3.7±8.4 mm 에서 4.1±8.3 mm 로 통상 5mm thickness 로 촬영하는 전산화단층촬영에서 CT slice 1 장이하의 오차를 보였다. Technical success rates 는 93.1% 에서 92.3% 로

높았다. 척추변이가 있는 군을 따로 떼서 결과를 얻었을 때 distance differencess 는 12.4±15.4 mm 에서 12.1±14.6 mm 로 증가하였으며 technical success rates 는 67.2% 에서 67.9% 로 감소하다. 전체 dataset 에서 근육량 segmentation 의 accuracy 는 anatomic variation 과 관계없이 우수한 결과를 보여 CSA error 값은 1.38–3.10 cm2 로 측정되었다.

결론

L3 selection 과 복벽의 근육량 측정을 측정하는 완전 자동화 인공지능모델이 개발되었고 실제 연구에 사용가능 한 performance 를 보였다.

중심단어: 근감소증, 자동화, 인공지능모델, 복부전산화단층촬영

차 례

표 목차

그림 목차

INTRODUCTION

The association of body morphometry, the quantitative analysis of muscle and fat mass, with various clinical outcomes from various diseases has been well established by prior research. Different muscle and adipose tissue mass compositions are associated with differences in surgical complication rates, chemotherapy-related toxicity, recurrence, and survival across the various cancer types.[2-4] Computed tomography (CT) is part of the standard-of-care management of various malignancies. CT imaging allows for surgical planning, accurate diagnosis, and treatment response evaluation. CT scans can be used secondarily for body composition measurements to evaluate risk among oncology patients. CT body composition measurement is accurate and reliable.[5-7] The muscle area calculated with the help of a single third lumbar vertebral–level CT scan can be used as a representative value of whole-body muscle mass.[8-11]

Although body composition is a potential risk stratification indicator to guide prognostication and treatment for various diseases, such assessments are not readily available. The assessment of body composition through CT scanning involves intricate steps

for selecting the proper level for analysis and for manual segmentation of body composition. The first hurdle in efforts toward automating body composition measurements is achieving accurate and reliable localization of the L3 level to extract appropriate CT slices for segmentation.

Several previous studies have reported the performance of automatic segmentation of body composition using various systems.[12-18] These studies involved manual selection of single CT slices for analysis, and human experts were required for the segmentation process. Two publications describe automatic L3 spotting systems,[19, 20] but the studies that informed these articles either were not clinically validated or were associated with an unacceptable maximum error value between the ground truth and results generated using the automatic spotting system.

The primary objective of this study was to develop a deep learning model (DLM) to automatically select L3 slices on abdominal CT scans and then automatically segment areas of abdominal muscle, visceral fat, and subcutaneous fat. The secondary objective was to validate the DLM's accuracy in terms of selecting L3 slices and segmenting muscle and fat

areas by comparing artificial intelligence–derived results and manually performed ground truths. The tertiary objective was to evaluate clinical effectiveness in terms of conserving time and human resources.

MATERIALS and METHODS

This study was approved by the institutional review boards of Asan Medical Center (AMC), Kyung Hee University Hospital (KHUH), Ajou University Hospital (AUH), and Ulsan University Hospital (UUH). The informed consent requirement was waived by the institutional review board. This article reports on and complies with the methods and terms described in the most recently published guidance on reading literature about machine learning for medical applications.[21]

I. Data acquisition: study subjects

The datasets used for this study were as follows: (1) development dataset used for

developing the DLM, which was further split into the training set and tuning set; (2) validation dataset for independent testing of model performance, including an internal validation set (acquired from AMC) and an external validation set (acquired from KHUH, AUH, and UUH).

The development dataset was composed of 922 patients (560 men and 362 women; mean age, 54.4 ± 14.0 years), with 1496 abdominal CT scans of patients who were referred to our central imaging core lab from various AMC physicians for body morphometric analyses. The development dataset was used in our previous study.[12] The development dataset included patients with various diseases (pancreatic cancer, gastric cancer, chronic kidney disease, sepsis from any cause), and healthy subjects who underwent CT scanning for potential kidney donation were included. The development dataset was further divided into a training set and a tuning set, with a ratio of 8 to 2.

The validation dataset was composed of internal and external validation sets. The internal validation set was composed of 500 CT scans of healthy individuals from AMC. Four subjects who underwent interbody lumbar vertebra fusion surgery were excluded, and a total

of 496 CT scans, captured from March through December 2012, were used for validation (301 men and 195 women; mean age, 3.7 ± 8.7 years [range, 24-88 years]). The external validation dataset included 600 CT scans, captured between September 2011 and March 2019, obtained from three other institutions, and a total of 586 CT scans were included after excluding those of subjects who underwent lumbar interbody fusion surgery (347 men and 239 women; mean age, 58.5 ± 12.3 [range, 18-88 years]). The clinical characteristics of subjects included in the validation dataset are summarized in Table 1. An overview of dataset composition is described in Figure 1.

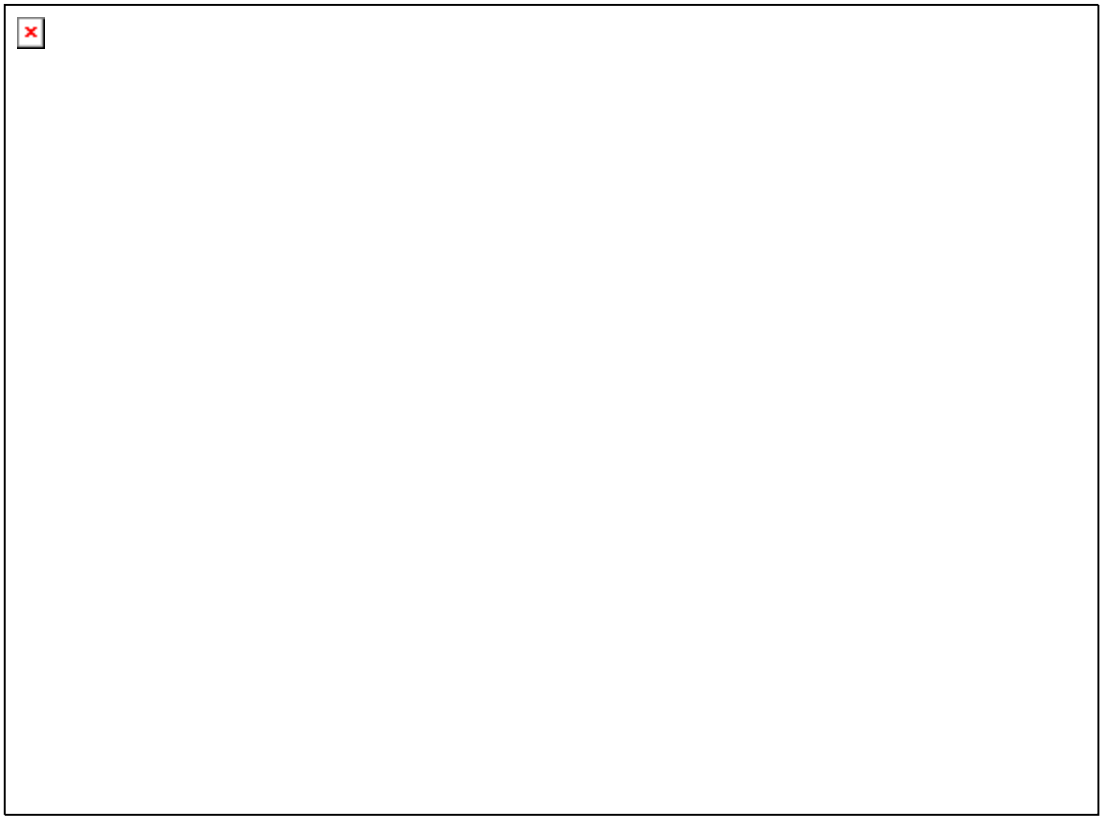

Figure 1. An overview of dataset composition.

Table 1. Subject characteristics of internal and external validation cohorts

Characteristics	Development dataset	Internal validation dataset	External validation dataset
Number of subjects	922	496	586
Age (years)	54.4±14.0	53.7±8.7	58.5±12.3
Female (%, female:male)	39.3% (362:560)	39.3% (195:301)	40.8% (239:347)
Anatomic variation			
Normal anatomy group	807 (87.5%)	438 (88.3%)	505 (86.2%)
Anatomic variants group	115 (12.5%)	58 (11.7%)	81 (13.8%)
Thoracolumbar variant	48 (5.2%)	20 (4.0%)	26 (4.4%)
Lumbosacral variant	43 (4.7%)	29 (5.8%)	43 (7.3%)
Numeric variant	12 (1.3%)	4 (1.4%)	7 (1.2%)
Combined variant	12 (1.3%)	5 (1.7%)	5 (0.9%)
Institution	AMC	AMC	UUH, KHUH, AUH
Underlying disease (n)			
None	87	496	586
Gastric cancer	436	0	0
Sepsis	245	0	0
Pancreatic cancer	154	0	0

Note.—**AMC** = **Asan Medical Center**, **AUH** = **Ajou University Hospital**, **KHUH** = **Kyung Hee University Hospital**, **UUH** = **Ulsan University Hospital**

CT scanners from various manufacturers (Sensation 16, Sensation 64, Somatom Definition, Somatom Definition Flash, and Somatom Definition AS+, Somatom Definition Edge, Somatom Plus 4, Definition, Definition AS, Definition AS+, Volume Zome scanners [Siemens Medical Systems, Erlangen, Germany]; BrightSpeed, LightSpeed 16, LightSpeed plus, LightSpeed VCT, LightSpeed QX/i, Optima CT 660 and Discovery 750 HD scanners [GE Healthcare, Milwaukee, WI, USA]; Aquilion PRIME, Aquilion [TOSHIBA, Tokyo, Japan]; Brilliance 64, iCT256, Ingenuity Core 128, Ingenuity CT [Philips Healthcare, Amsterdam, Netherlands]; and Presto [Hitachi Medical System, Tokyo, Japan]) were used during the recruitment period. Abdominopelvic CT scans, with or without contrast enhancement, which captured from the diaphragmatic dome to the symphysis pubis, were included. Intravenous contrast medium (120-150 mL of 300-370 mgI/mL non-ionic contrast [iopromide, Ultravist 300 or Ultravist 370; Bayer Healthcare, Berlin, Germany]) was administered at a rate of 2–3 mL/sec for enhanced scanning through an automatic power injector using an 18-gauge angiographic catheter. CT scanning was performed after 70 to 90 seconds of contrast medium injection. The images were reconstructed in the axial plane,

ranging in thickness from 2.5 mm to 5 mm. Detailed specifications of the CT scanners used for image acquisition are summarized in Table 2.

Table 2. Summarization of CT acquisition protocols

CT vendors	Model name	Filter type	Convolution kernel	Training set	Internal validation set	External validation set
Siemens	Definition	0	B30f	9		
	Definition AS	0	B30f	5		
	Definition AS +	0	B30f	38		
	Emotion	0	B40s	1		
	Emotion 6	1	B30s	1		
	Emotion 16	1	B41s	1		
	Sensation 16	0	B30f, B31f	272	2	142
	Sensation 64	0	B31f	1		
	SOMATOM Definition	FLAT	B30f	18	21	
	SOMATOM Definition AS	FLAT	B30f, B40f, I30f	49		
	SOMATOM Definition AS+	FLAT	B30f, I30f, I40f	410		1
	SOMATOM Definition Flash	WEDGE, WEDGE_3	B30f, I40f	24	9	9
	SOMATOM Definition Edge	FLAT	B30f, I30f	70		
	SOMATOM PLUS 4	N/A	AB40, AB50	2		
	Volume Zoom	0	B40f	1		
Siemens_sum				**902**	**32**	**152**
GE	BrightSpeed	BODY FILTER	SOFT	1		
	Discovery CT750 HD	BODY FILTER	STANDARD	51	97	
	HiSpeed	N/A	STANDARD	1		
	HiSpeed CT/i	LARGE BOWTIE FILTER	STANDARD	20		
	LightSpeedn Plus	BODY FILTER	STANDARD, SOFT	33		
	LightSpeed QX/i	BODY FILTER	STANDARD	19		
	LightSpeed VCT	BODY FILTER	STANDARD, SOFT	145	337	
	LightSpeed16	BODY FILTER	STANDARD	34	29	
	Optima CT660	BODY FILTER	STANDARD	57	1	
GE_sum				**361**	**464**	
Philips	Brilliance 64	C	C			10
	iCT 256	A,YA	A,YA	12		50
	Ingenuity Core 128	B	B			126
	Ingenuity CT	YA	YA			194
Philips_sum				**12**		**380**
Hitachi	Presto	N/A	4	1		
TOSHIBA	Aquilion PRIME					54
	Aquilion	LARGE,EC	FC 13, FC 08, FC 04, FC 18	10		
Others_sum				**11**		**54**
Sum				**1286**	**496**	**586**

II. Generation of the ground truth

For each CT scan, the axial CT slice number of the third lumbar vertebra inferior endplate was annotated, and the lumbar vertebral anatomic variant was identified by a board-certified radiologist (J.H.) and double-checked by another radiologist (K.W.K.). Disagreement was resolved by reaching consensus through discussion. At first, the morphologic lumbar vertebrae were counted. There are usually five lumbar spines, but some people have four or six lumbar vertebrae, as illustrated in Figure 2. Then, anatomic variants were identified and categorized into four groups as follows: (1) thoracolumbar variant (twelfth rib aplasia/hypoplasia or rudimentary rib attached at the first lumbar vertebra), (2) lumbosacral variant (lumbarization of S1 and sacralization of L5), (3) numeric variant (four or six lumbar vertebrae without transitional vertebra or rib anomaly), and (4) combination of two different variants.[22-24] Morphologically normal ribs were defined as a pair of ribs that were 3.8 cm in length or more and originated from the facet between the pedicle and vertebral body. Unilateral or bilateral short ribs (<3.8cm), presence of ossification centers without ribs, or unfused transverse processes were regarded as rudimentary ribs. Lumbosacral transitional

vertebrae were identified based on the criteria described by Castellvi et al. in 1984 [25]. Unilateral or bilateral dysplastic lumbar vertebral transverse processes (at least 19 mm in the craniocaudal dimension), an enlarged transverse process forming a diarthrodial joint with the sacrum, and osseous fusion of a transverse process to the sacrum were classified as lumbosacral junction transitional vertebrae. Lumbar vertebrae without rudimentary or normal ribs and showing normal transverse processes were regarded as morphologically normal lumbar vertebrae and classified as numeric variants if there were not exactly five of these vertebrae.

A single axial image was extracted from each CT scan at the L3 inferior endplate for analysis. An expert image analyst (S.J.H.) manually generated the ground-truth segmentation map for total skeletal muscle, visceral fat, and subcutaneous fat. The segmentation map was double-checked by a supervising radiologist (K.W.K.). The rectus abdominis, external/internal obliques, transverse abdominis, quadratus lumborum, psoas major/minor, and erector spinae muscles were included in the analysis. The cross-sectional area (CSA, cm2) and the total number of pixels were calculated for the segmentation map of total skeletal muscle, visceral

fat, and subcutaneous fat.

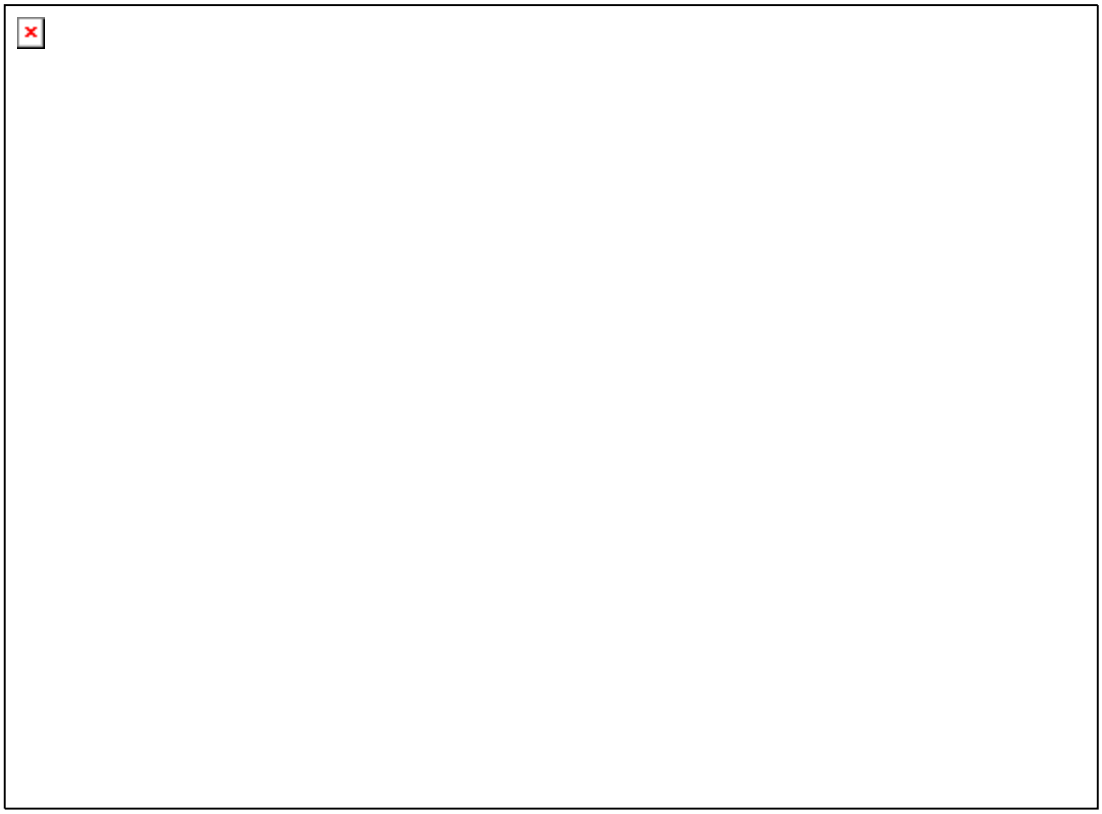

Figure 2. Anatomic lumbar spine variants. Examples of normal, thoracolumbar, lumbosacral, numeric, and combined variations are presented.

III. Deep learning model development

The DLM was composed of two parts, as follows: (1) a YOLOv3-based component for automatically selecting the L3 slice on abdominal CT scans [26] and (2) a fully convolutional network (FCN)-based component for automatically segmenting areas of muscle, visceral fat, and subcutaneous fat. Of these, the FCN-based component is described elsewhere [12].

Several pre-processing steps were used to generate input data for training our YOLOv3-based L3 selection model. Using Otsu thresholding, a region of a patient's image was extracted from the background. Then, hole filling and noise removal were performed on the region by seeded region growing and morphological filtering. The histogram distribution of the region was normalized to generate consistent grayscale information, irrespective of the scanner type and protocol. Then, maximum intensity projection (MIP) coronal images were generated to efficiently evaluate the lumbar spines.

To increase the effective dataset size and improve overfitting and accuracy, data augmentation was performed to generate 15,226 MIP images from 922 CT scans. Of these,

12,180 MIP images were used as a training set, and 3,046 images were used as a tuning set. The data augmentation was performed using random combinations of affine transformations, elastic distortion, edge extraction, blurring, and cropping.

i. YOLOv3-based L3 slice selection algorithm

A YOLOv3-based model was adopted because such a model can detect objects and extract features more efficiently than conventional convolution neural networks, accomplished via object detection and classification [26]. With our YOLOv3-based model, we aimed to predict bounding boxes using anchor boxes which were configured via dimension clustering. As illustrated in Figure 3, multiple bounding boxes were generated in the MIP images based on the following prerequisites: (1) the L4 vertebra was located at the iliac crest level, (2) the L3 vertebra was located superiorly to the L4 vertebra, (3) the morphologies of the lumbar vertebrae were the same. The YOLOv3-based model used an objectness score for each bounding box obtained from logistic regression to predict the width and height of the box as well as its location relative to grid cell. The sum of the squared error loss was used to train the model for minimizing differences between the ground-truth object and the bounding box.

Any error between the bounding box over the ground-truth object was incurred for both classification and detection loss.

Our model extracted features of the bounding boxes using the network architecture illustrated in Figure 3. Our network architecture used successive 3×3 and 1×1 convolution layers and a set of residual blocks with shortcut connections. A total of 53 convolutional layers were formed like Darknet-53. YOLOv3 predicted boxes at three different scales to support detection on varying scales. Using a similar concept to feature pyramid networks,[27] our model extracted features from multi-scales of the bounding boxes and took the feature maps. Then, we merged them with upsampled features using concatenation. For training, we used 416×416 MIP images with no hard-negative mining. We also used multi-scale training, data augmentation, and batch normalization.

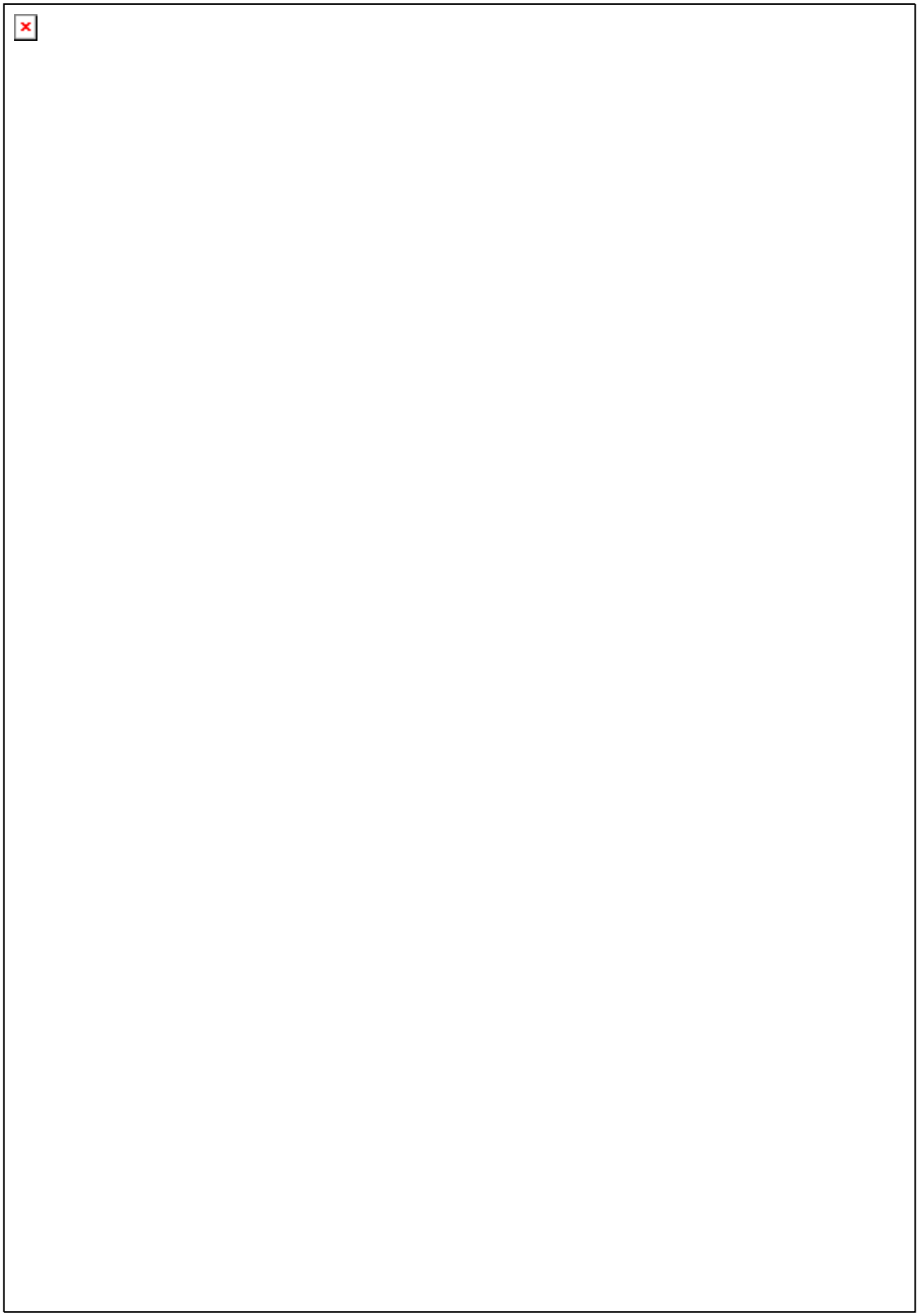

Figure 3. Example of multiple bounding boxes for training of the YOLOv3-based model and architecture of our YOLOv3-based network.

ii. FCN-based segmentation algorithm

Our FCN-based model for automatically segmenting areas of muscle, visceral fat, and subcutaneous fat is described elsewhere [12]. We adopted the same model for automatic segmentation of muscle and fat areas. In this study, we added post-processing based on Hounsfield units (HU) to divide the segmented muscle areas into skeletal muscle areas (SMAs, −29 to 150 HU) and inter-/intra-muscular adipose tissue (IMAT, −190 to -30 HU). The network architecture of our FCN-based model was illustrated in Figure 4.

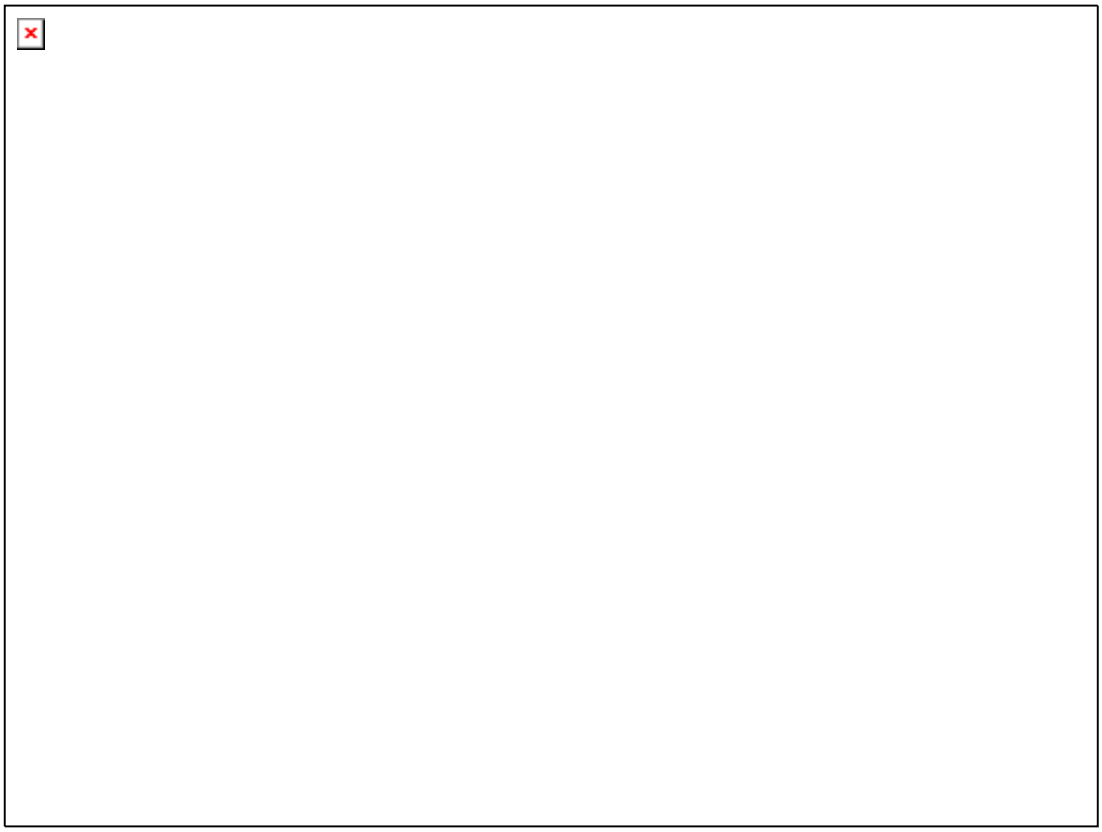

Figure 4. The network architecture of a fully convolutional network–based segmentation model. Post-processing based on Hounsfield units (HU) was added to separate the intramuscular adipose tissue from skeletal muscle area. Red and blue areas indicate skeletal muscle area (−29 to 150 HU) and area in yellow indicates intramuscular adipose tissue (−190 to −30 HU). Brown area indicates subcutaneous fat area and area in purple represents visceral fat area.

IV. Validation of deep learning model

i. Accuracy of automatic L3 slice selection

The accuracy of the DLM selection of the third lumbar vertebra inferior endplate was evaluated using internal and external validation cohorts. Two board-certificated abdominal radiologists evaluated the accuracy of the DLM-derived results.

The differences in CT slice numbers between the ground truth and the DLM-derived results were calculated and multiplied by slice thickness to generate the actual difference in millimeters. Mean differences were calculated for each internal and external validation group. Technical success was calculated as the percentage of cases that showed differences of less than 10 mm between the ground truth and the DLM-derived results for each validation group.

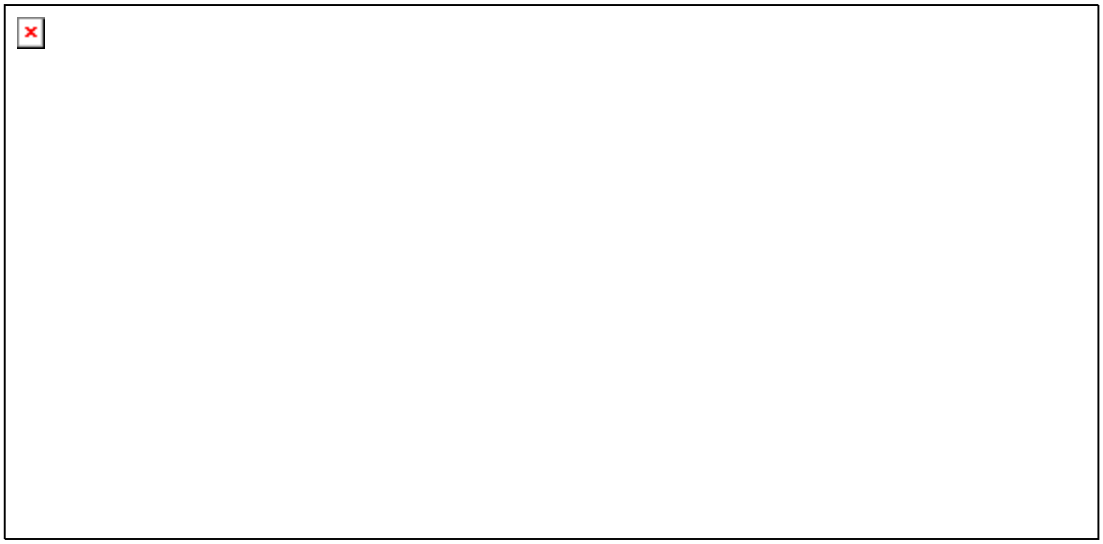

Figure 5. Examples of technical success and technical failure.

(A) Technical success when L3 CT slice numbers are identical between the ground truth (GT) and the deep learning model (DLM)–derived results.

(B) Technical success when the distance difference between the GT and the DLM-derived results is less than 10 mm.

(C) Technical failure when the distance difference between the GT and the DLM-derived results is greater than 10 mm.

ii. Segmentation accuracy of the DLM

The axial CT images were extracted at the level of the inferior endplate of the third lumbar vertebra selected by manually and using the DLM. Areas of each tissue type—skeletal muscle, subcutaneous fat, and visceral fat—were segmented manually and using the DLM system.

DSC and CSA error values were used to evaluate technical performance. The DSC is an index of spatial overlap ranging from 0 to 1. Completely overlapping areas yield a DSC value of 1, whereas the absence of overlap yields a DSC value 0. DSCs were calculated using the following formula:

$$DSC = \frac{2 \times |\text{ground truth} \cap \text{FCN}|}{|\text{ground truth}| + |\text{FCN}|}$$

$$DSC = \frac{2 \times TP_P}{2 \times TP_P + FP_V + FN_V}$$

TPp denotes the number of pixels correctly included in both the ground truth and the DLM-derived result. FPv denotes the number of pixels included in the DLM-derived result but not in the ground truth. FNv represents the number of pixels included in the ground truth but not

the DLM-derived results.

DSCs were used for the subgroups that yielded identical third lumbar vertebral inferior endplates from both the ground truth and the DLM-derived results. SMA, subcutaneous fat area (Sfat), and visceral fat area (Vfat) at the L3 inferior endplate level were manually measured by an expert image analyst (ground truth) and using the automatic DLM model. The areas were compared using DSCs.

CSA in cm2 and the total pixel number of the segmentation map (total skeletal muscle, subcutaneous fat, and visceral fat) were analyzed. CSA error was used for the subgroups that yielded different L3 inferior vertebra levels between the ground truth and the DLM. CSA error is a standardized measure of percentage differences in measured areas. The ground-truth segmentation maps for SMA, Sfat, and Vfat were generated manually by an expert image analyst at the ground-truth L3 inferior endplate level, and these were compared with the DLM-measured segmentation map at the DLM-derived L3 endplate level.

$$\text{CSA error (\%)} = \frac{|\text{ground truth}_{CSA} - \text{FCN}_{CSA}|}{\text{ground truth}_{CSA}} \times 100$$

DSC values were compared in the concordant group with identical CT slice numbers from both the ground truth and the DLM-derived results. CSA error and Bland–Altman analyses were used to assess agreement in the discordant group with acceptable differences between the ground truth and the DLM-derived results, producing technical success and technical failure groups. Examples from the concordant group, discordant group, and a maximal difference case of technical failure are presented in Figure 5.

V. Subgroup analysis according to anatomic variation

Factors influencing the performance of the DLM at spotting the L3 level were explored by subgroup analysis. Each validation cohort was divided according to spinal anatomic variation and age groups. The L3 selection performance of the DLM was compared between these groups. The thoracolumbar junction variant group included scans of patients with a rudimentary rib attached to the first lumbar vertebra and those with a hypoplastic/aplastic twelfth rib. The lumbosacral junction variant group included scans of patients with sacralization of L5 and those with lumbarization of S1. Subjects with four or six morphologically normal lumbar vertebrae were defined as numeric variant group, and those

with two or more anatomic variants were classified as the combined variation group. The whole validation dataset was divided according to age group, and L3 selection performance was compared between the groups..

VI. Statistical analysis

Differences between the ground truth and the DLM-derived results were compared using the Mann–Whitney and Kruskal–Wallis tests. Logistic regression was conducted to evaluate factors potentially influencing technical success, including anatomic variation and demographic factors. Bland–Altman analysis was done to assess for agreement of body composition between the ground truth and the DLM–derived results. The limits of agreement used in the Bland–Altman plot was defined as the mean difference ± the 95% confidential interval. SPSS Statistics for Macintosh , version 21 (IBM Corp., Armonk, NY, USA), MedCalc 12.7.0 (MedCalc Software, Mariakerke, Belgium), and R version 3.6.3 (R Foundation for Statistical Computing, Vienna, Austria) were used for statistical analysis. A p-value < 0.05 was regarded as statistically significant..

RESULTS

I. Accuracy of automatic L3 slice selection

The YOLOv3-based DLM developed for spotting L3 was successful in both the internal and external validation datasets. The overall validation dataset results are summarized in Figure 6.

The mean differences [median, interquartile range, min-max] between the ground truth and the DLM-derived results were 3.9 ±8.3 [0, 0-5, 0-40], 3.7±8.4 [0, 0-5, 0-40], and 4.1±8.3 [0, 0-4, 0-40], in mm, for the overall, internal, and external validation cohorts, respectively. Subjects with normal spine anatomy yielded smaller differences between the ground truth and the DLM-derived results than those with anatomic variants (2.6±6.0 vs. 12.2±14.8, 2.5±6.1 vs.12.4±15.4, and 2.8±5.9 vs. 12.1±14.6 mm for overall, internal, and external validation datasets; $p < 0.01$). The technical success rates were 92.7%, 93.1%, and 92.3% for the overall, internal, and external validation datasets, respectively. The normal anatomy group yielded higher technical success rates than the anatomic variant group (96.5% vs. 67.4%, 96.6%vs. 67.2% and 96.2% vs. 67.9%).

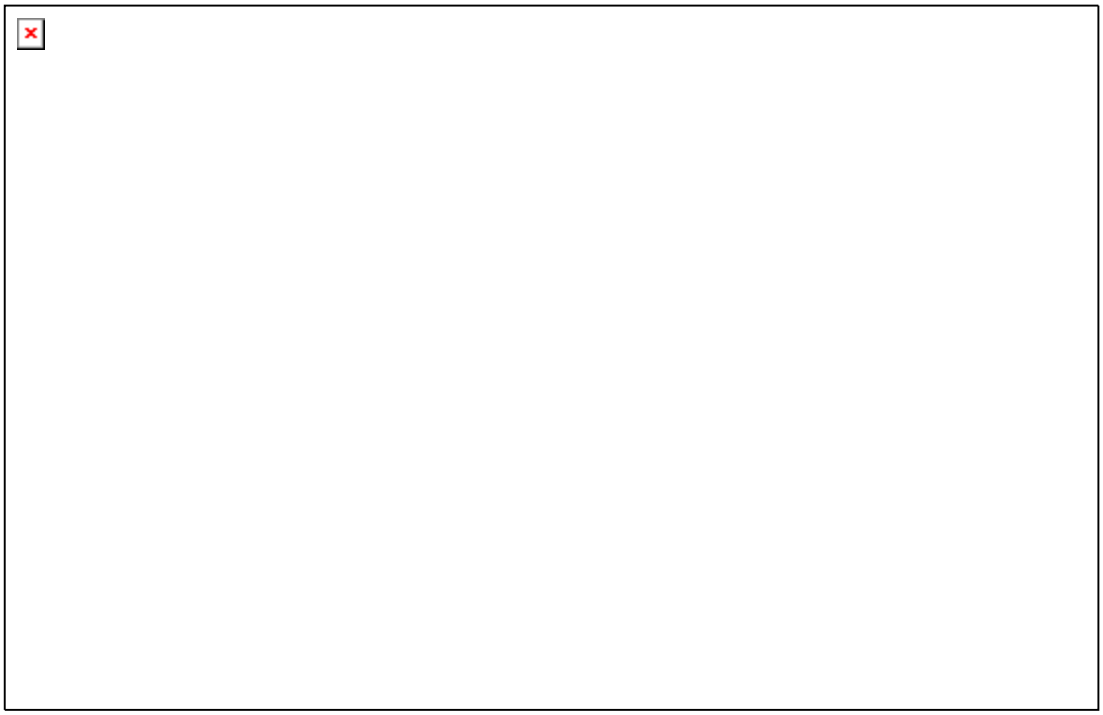

Figure 6. Box plots of distance difference between ground truth and deep learning model (DLM) derived results in (A) internal validation dataset and (B) external validation cohorts. The mean differences between the ground truth and the DLM-derived results were 3.7 mm ± 8.4 and 4.1 mm ± 8.3 for the internal, and external validation cohorts, respectively.

II. Segmentation accuracy of DLM-derived abdominal muscle and fat areas

The area segmentation concordance analysis is summarized in Table 3. The mean DSC values for SMA, Sfat, and Vfat were high in both the internal and external validation datasets. The DSC values of body composition area in concordant subgroups were high in both the internal and external validation datasets. The DSC values of SMA, Sfat, and Vfat for the internal validation group were 0.98, 0.98, and 0.98, respectively. The values for the external validation group were 0.96, 0.97, and 0.97, respectively. The mean CSA error values of the concordant subgroup were low in both the internal and external validation groups. The CSA error values for SMA, Sfat, and Vfat were 1.05%,1.89%, and 2.10%, respectively, in the internal validation dataset; the values in the external validation dataset were 2.71%, 2.75%, and 2.26%, respectively. The CSA error values of the discordant subgroup were low but slightly higher than those of the concordant subgroup. The CSA error values of SMA, Sfat, and Vfat were 1.62%, 3.29%, and 5.02%, respectively, in the internal validation dataset. The values for the external dataset were 3.05%, 4.18%, and 4.71%, respectively.

The Bland–Altman plots of both the internal and external validation datasets reflected good agreement in terms of area segmentation between the ground-truth L3 level and the DLM-derived level, as depicted in Figure 7. The mean differences (±limits of agreement) of SMA, Sfat, and Vfat in the internal validation dataset were 0.33 ± 4.28, −1.62±7.62, and −1.69±6.66 cm2, respectively. The mean differences for the external validation dataset were 2.61±7.61, −2.74±8.65, and −0.25±7.53, respectively. The Bland–Alman plot outlier in the internal validation set was for an Sfat area, and the case achieved technical success, with a difference of 5 mm between the ground truth and the DLM-derived level. However, thick gluteal subcutaneous fat layers were included in the ground-truth level and were not included in the DLM-derived level. Figure 8 is the exact Sfat map segmented at both levels.

The CSA error values of the subgroup that did not achieve technical success were 5.03%, 19.2%, and 16.5% for SMA, Sfat, and Vfat, respectively, in the whole validation dataset. The values were 3.7%, 20.4%, and 18.4% for the internal validation group and 5.0%, 18.3%, and 15.1% for the external validation group, respectively.

The Bland–Altman analysis was done to evaluate for agreement in terms of body

composition area for the technical failure group (Figure 9). The mean difference and limits of agreement for the internal validation group was 3.27±52.8 cm2. The value for the external validation group was 1.82±38.8 cm2.

Table 3. Cross-sectional area segmentation using the ground truth–derived and DLM–derived levels

Parameter	Internal validation dataset			External validation dataset		
	SMA	Sfat	Vfat	SMA	Sfat	Vfat
All subjects (n=1082)						
CSA from GT (cm^2)	140.88±34.53	140.90±56.71	114.53±65.05	132.76±31.25	133.15±62.16	110.59±64.29
CSA from DLM (cm^2)	140.53±34.20	141.98±56.60	115.93±65.40	130.07±31.07	135.54±62.64	110.72±65.19
p value*	0.874	0.764	0.736	0.139	0.492	0.973
CSA error (%)	1.38±1.46	3.51±5.41	4.00±6.35	3.10±2.85	4.54±6.34	4.26±6.47
Subjects with technical success (n= 1004)						
CSA from GT (cm^2)	141.20±34.46	138.85±55.86	112.42±64.73	132.75 ±31.15	133.99 ±62.82	110.88 ±64.18
CSA from DLM (cm^2)	140.87±34.06	140.47 ± 55.72	114.11±64.95	130.14 ±31.00	136.73 ±63.15	111.13 ±65.06
p value*	0.883	0.659	0.692	0.167	0.474	0.950
CSA error (%)	1.22 ±1.08	2.31±2.21	2.97 ±3.21	2.86 ±2.57	3.39±2.78	3.36 ±4.68
Subjects with technical failure (n=78)						
CSA from GT (cm^2)	136.33±35.18	169.78±60.54	144.23± 62.27	132.97±32.3	123.03±52.30	107.10±65.61
CSA from DLM (cm^2)	135.77±35.80	163.24±64.41	141.53± 66.37	129.21±31.93	122.66±54.42	105.78±66.50
p value*	0.949	0.672	0.865	0.579	0.974	0.924
CSA error (%)	3.68±3.19	20.42±8.14	18.37±15.68	6.01±4.18	18.28±15.16	15.06±12.56
p value§	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

Note.—Data are presented as mean ± standard deviation

* The p-value is calculated from Student t-test comparing the GT CSA and the CSA determined using the DLM.

§ The p-value is calculated from Student t-test comparing CSA errors between subjects with technical success and subjects with technical failure.

CSA = cross-sectional area, DLM = deep learning model, GT = ground truth, Sfat = subcutaneous fat area, SMA = skeletal muscle area, Vfat = visceral fat area

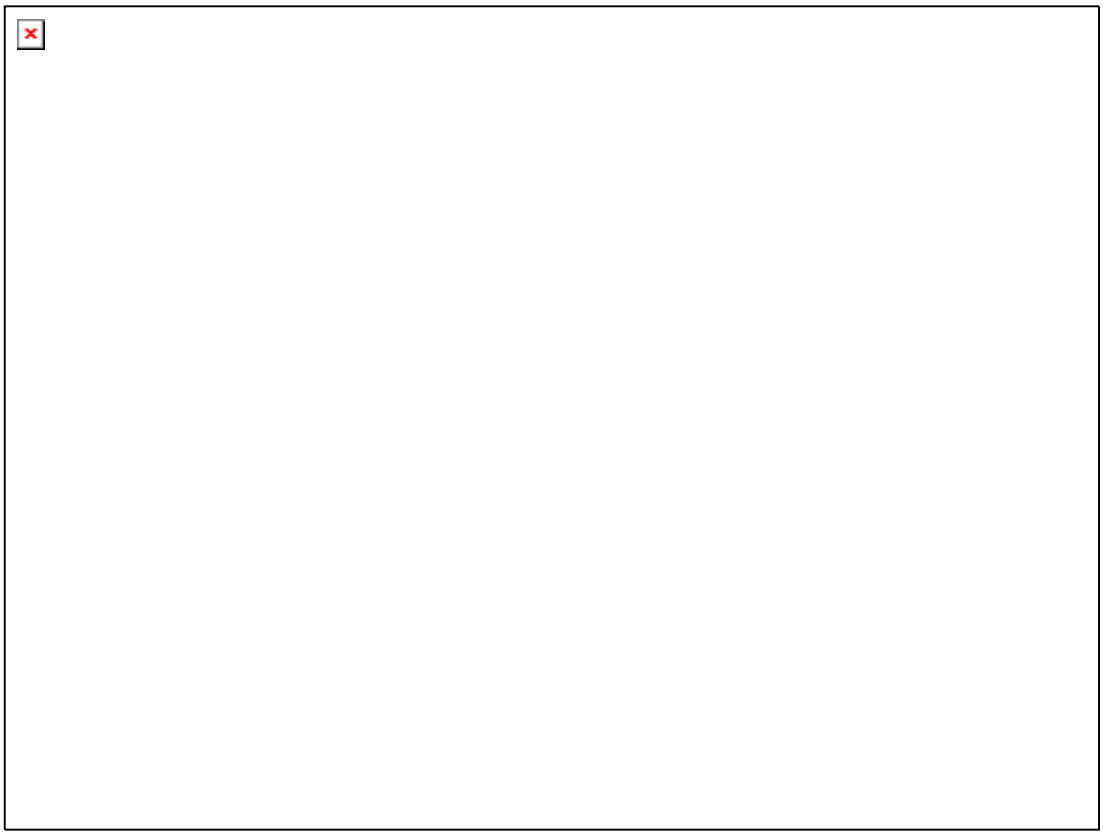

Figure 7. Bland Altman plots to evaluate agreement of SMA between the GT and DLM

(A) In subjects with technical success in the internal validation cohort

(B) In subjects with technical failure in the internal validation cohort

(C) In subjects with technical success in the external validation cohort

(D) In subjects with technical failure in the external validation cohort

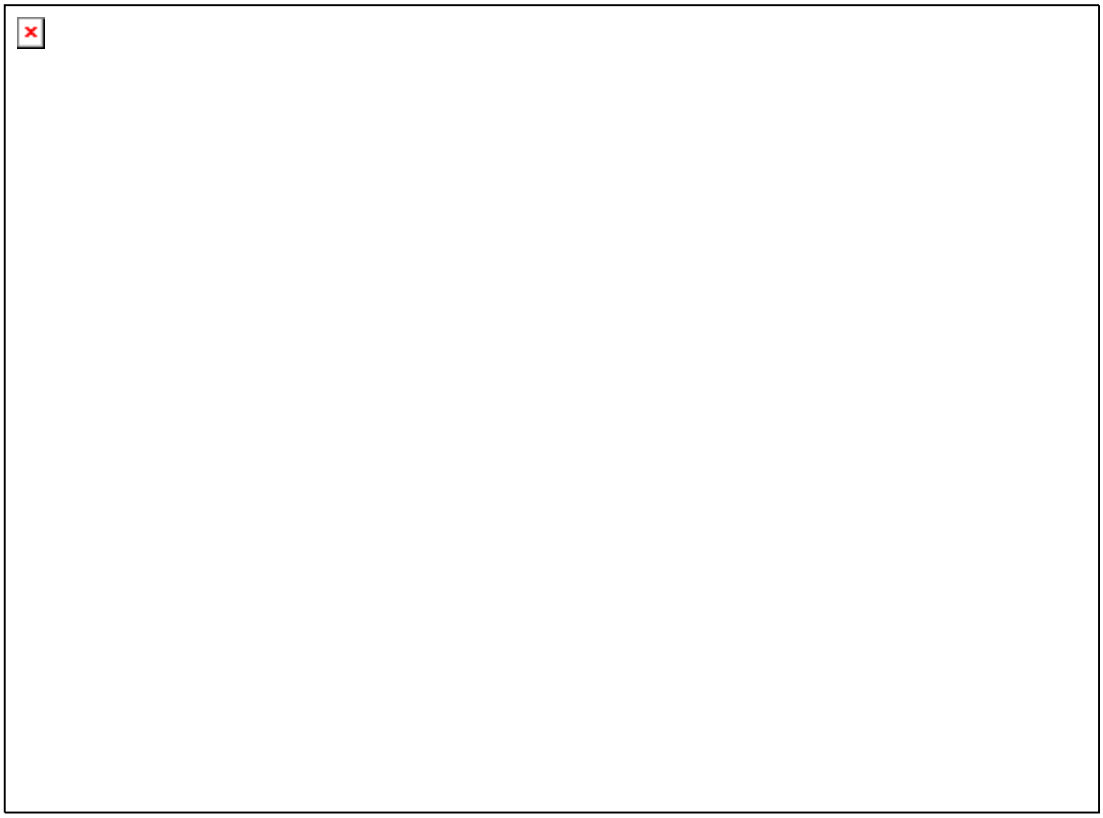

Figure 8. Bland Altman plots to evaluate agreement of Sfat between the GT and DLM

(A) Sfat in subjects with technical success in the internal validation cohort

(B) Sfat in subjects with technical failure in the internal validation cohort

(C) Sfat in subjects with technical success in the external validation cohort

(D) Sfat in subjects with technical failure in the external validation cohort

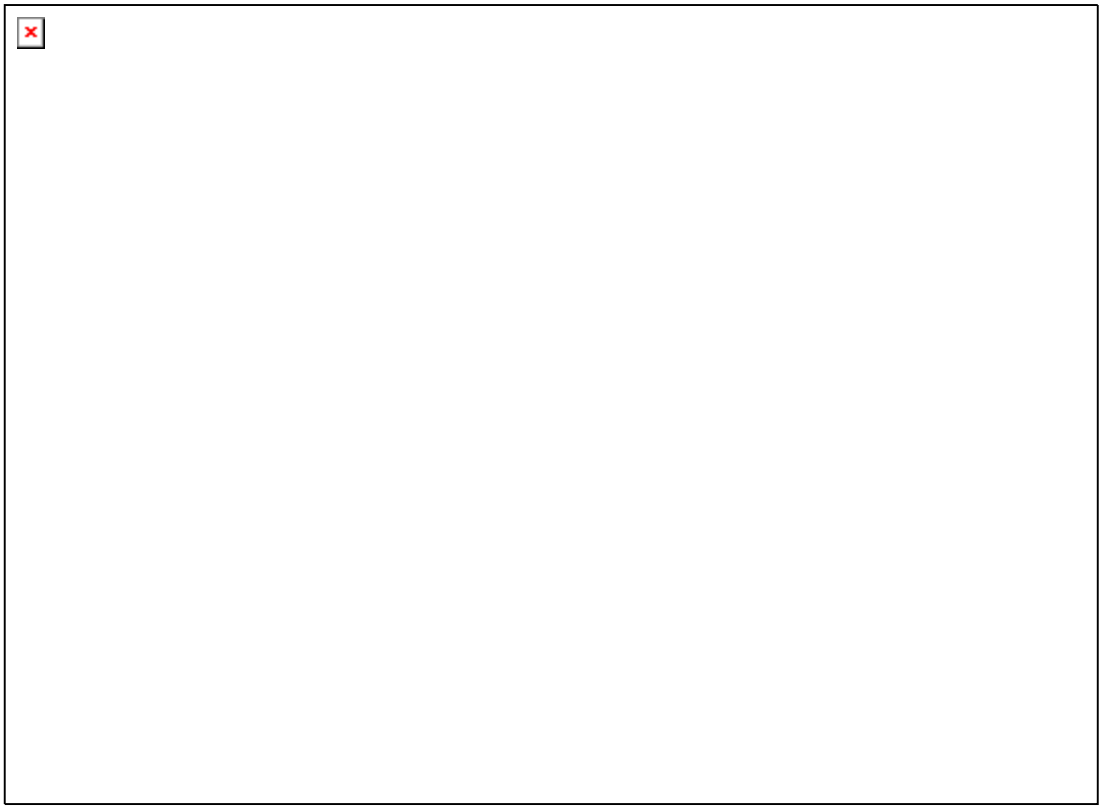

Figure 9. Bland Altman plots to evaluate agreement of Vfat between the GT and DLM

(A) Vfat in subjects with technical success in the internal validation cohort

(B) Vfat in subjects with technical failure in the internal validation cohort

(C) Vfat in subjects with technical success in the external validation cohort

(D) Vfat in subjects with technical failure in the external validation cohort

III. Subgroup analysis according to anatomic variation

Anatomic variant type was the only factor significantly influencing the technical success of the system ($p = 0.003$). The technical success rate of the whole validation set (n=1082) was 92.7%. The value for the normal anatomy subgroup (n=943) was 96.5%; the values were 82.6%, 63.9%, 54.5%, and 40% for thoracolumbar (n=46), lumbosacral (n=72), numeric variant(n=11), and combined variation (n=10) subgroups, respectively. The mean difference, in the presence of any abnormal anatomic variant, between the ground truth and the DLM-derived was statistically significant (2.6 vs. 12.2 mm, $p < 0.01$). The mean differences according to the specific abnormal variant types were 7.4, 13.4, 16.5, and 21.4 mm for the thoracolumbar, lumbosacral, numeric and combined variant groups, respectively. Distance according to specific variation group was showed in Figure 10. CSA errors of body composition area were significantly different between the subgroups. The mean CSA error values were 2.21 and 3.04 cm2 for SMA, 3.49 and 7.95 cm2 for Sfat, and 3.59 and 8.87 cm2 for Vfat in terms of normal and abnormal anatomy, respectively ($p < 0.01$). Bland–Altman analysis was performed to evaluate CSA concordance between the subgroups. The mean

differences (±limits of agreement) of SMA, Sfat, and Vfat for each subgroup were 1.67±7.19, −2.23±13.61, −0.86±10.08 for the normal anatomy group and 1.33±10.28, 0.71±37.08, 0.33±29.84 for the abnormal anatomy group, respectively. The results of subgroup analysis were summarized in Table 4.

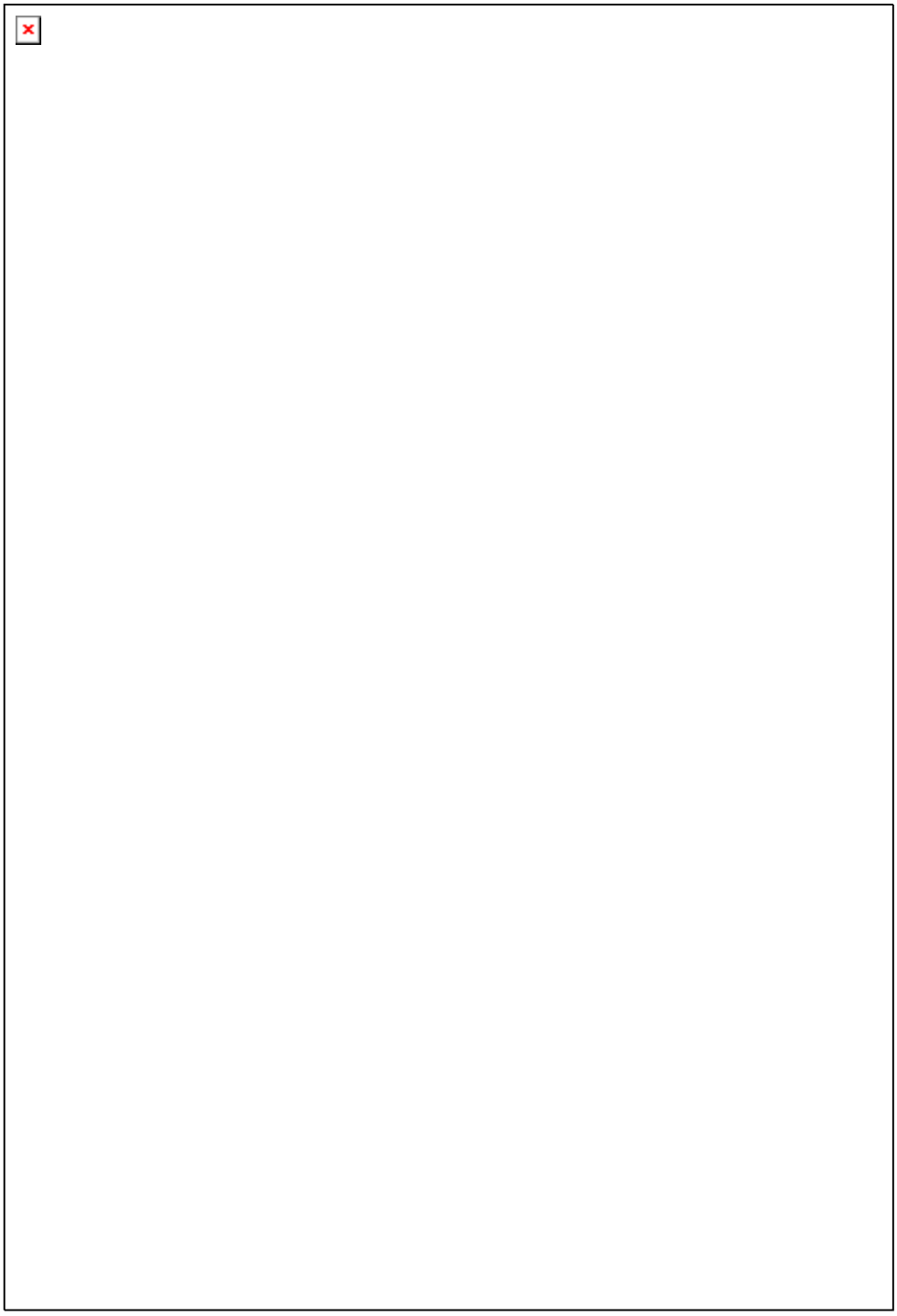

Figure 10. Box plot of distance difference in (A) internal, and (B) external validation dataset according to specific anatomic variation.

Table 4. Subgroup analysis according to spine anatomy

Subgroup	Distance difference (mm)	Technical success (%)	CSA error (%)			Bland-Altman (mean±limits of agreement)		
			SMA	Sfat	Vfat	SMA	Sfat	Vfat
Normal anatomy (n=943)	2.6±6.0	96.5	2.22±2.46	3.46±4.78	3.57±5.58	1.68±7.22	-2.29±13.13	-0.84±10.03
Thoracolumbar variation (n=46)	7.4±11.9	82.6	2.73±2.24	5.83±8.79	5.87±7.04	2.23±7.90	2.41±34.33	-2.69±24.10
Lumbosacral variation (n=72)	13.4±15.2	63.9	3.04±2.49	8.72±10.63	7.94±9.23	1.40±10.56	2.17±35.93	0.86±24.84
Numeric variation (n=11)	16.5±16.1	54.5	2.37±2.11	10.87±7.62	10.36±10.19	-0.22±7.10	-3.93±46.02	-1.82±26.79
Combined variation (n=10)	21.4±17.0	40	4.06±2.92	11.86±12.66	14.95±17.03	-2.53±14.78	-7.15±58.06	10.82±67.40

Note.—CSA = cross-sectional area, Sfat = subcutaneous fat area, SMA = skeletal muscle area, Vfat = visceral fat area.

DISCUSSION

Our YOLOv3-based model allowed accurate automatic localization of the L3 inferior endplate. Technical success was achieved for the majority of cases in both the internal and external validation groups. Technical success rates were higher than 90% for all datasets from all participating institutions. Almost all cases (95%) showed differences of less than 30 mm between the human expert and the DLM-derived results, equivalent to differences of less than one vertebral body height. CSA differences of each body composition area between the human expert and the DLM-derived results were small (less than 10 cm2) in the whole validation group, even in the subset that did not achieve technical success, representing a comparable performance of the L3 spotting system. DSC values were higher than 0.96 in the concordant group, and CSA errors were lower than 5% in both the concordant and discordant groups that achieved technical success. The CSA error rate for SMA in the technical failure group was 5% for the whole validation group; the error rates were 3.7% and 6.0% for the internal and external validation groups, respectively.

Anatomic spinal variation was the only factor that significantly influenced DLM

performance. CSA errors for each body composition segmented area were lower in the normal anatomy group than in the anatomic variant group. However, the CSA error rates for SMA were less than 5%, regardless of the presence or absence of abnormal anatomy. Among the abnormal variant subtypes, the thoracolumbar junction variant subgroup, including T12 rib hypoplasia/aplasia and L1 rudimentary rib, yielded similar performance to the normal anatomy group, whereas the lumbosacral junction variant subgroup and other numeric variant subgroup yielded lower technical success rates. The lower technical success of the lumbosacral junction variant subgroup may be attributable to a specific process component wherein the algorithm assumes the L4 level as the iliac crest.

A recently published study by park et al.[28] indicated that the tissue compositions of the L2 to L4 levels were not significantly different from one another and that the body composition between the L2 and L4 levels could represent the whole body composition area. The maximum difference in our results was 40 mm, which is equivalent to the distance from L2 to L4. In cases of technical failure, the body composition of the DLM-derived level could represent the maximal range of error.

Several researchers have reported adequate performance of automatic L3 level spotting models. Belharbi et al. [20] compared the performance of various convolutional neural networks (CNNs) at spotting the L3 level, including homemade and pre-trained CNNs, with a dataset of 642 CTs. The mean difference was 1.8 to 10.5 CT slices with thicknesses of 2-5 mm, equivalent to 3.6 to 50.5 mm in total distance. The dataset was from one institution, and they did not conduct clinical validation. Additionally, the study was limited to the task of L3 spotting and did not evaluate the accuracy and reliability of segmented body composition at the extracted level. Our YOLOv3-based L3 spotting system had a minimum value similar to that reported by Belharbi el al.[19], with a mean distance of 3.9 mm in the whole validation dataset.

Bridge et al.[19] reported the L3-spotting and automatic segmentation performances of deep learning algorithms, specifically ResNet, DenseNet, and U-net. They treated slice selection and segmentation as one process. The dataset was composed of a training cohort (n=595) and a testing cohort (n=534). The mean localization error was 9.4 mm, and the mean DSCs for muscle, Sfat, and Vfat were 0.97, 0.98, and 0.95, respectively. The DSC values of our

study were similarly high, but the mean difference of our YOLOv3-based model was smaller than that reported by Bridge and colleagues.

Our study had some limitations. First, the study was retrospective, and the subject recruitment process was not consecutive and may have been fraught with selection bias. Second, healthy subjects were only included for the internal and external validation cohorts. The performance of the developed DLM may require validation with large samples of patients with various diseases. Furthermore, our YOLOv3-based model had a moderate technical success rate (67.4%) with the abnormal anatomic variant group. Further training processes using subjects with anatomic variation may improve the performance of the system.

CONCLUSION

In conclusion, our YOLOv3-based L3 spotting model performed well at localizing the L3 inferior endplate. Additionally, segmented body composition identification (especially the muscular area) at the DLM-derived level was acceptable. Therefore, this automated L3

spotting model could be used in various clinical and research applications for body morphometry analysis.

ABSTRACT

Background

Sarcopenia research has been expanding rapidly. The numbers of patients requiring abdominal muscle measurements in each new study have also rapidly increased, requiring human resources and time. Among the hurdles to automating such measurements is L3 vertebral level selection. We aimed to develop and validate a fully automatic system for selecting the L3 level and compared body composition errors between the ground truth and deep learning model (DLM)–derived results.

Methods

A YOLOv3-based DLM automatically spotting L3 CT slice was developed via supervised learning from a training dataset (922 computed tomography [CT] scans). A radiologist provided L3 slice levels as the ground truth(GT). The internal(n=500) and external

validation(n=600) datasets ended up with 496 and 586 scans after excluding 4 and 14 scans of patients who underwent lumbar surgery. The difference between the GT and DLM-selection was calculated. Technical success was evaluated based on a 10 mm cut-off value in difference. Dice similarity coefficient (DSC) and cross-sectional area (CSA) errors were evaluated for segmented body compositions. Bland–Altman analysis was conducted to assess segmented area agreement. Subgroup analysis was performed according to vertebral anatomic variation.

Results

The mean differences between the ground truth and DLM selections were 3.6±8.3 mm, 2.5±6.1 mm, and 12.1±15.1 mm in the internal validation set (n=496), subgroup with normal anatomy (n=438), and subgroup with abnormal anatomy (n=58), respectively. The technical success rates were 93.1% (463/496), 96.6% (422/438), and 67.2% (39/58), respectively. In the external validation set (n=586), with the normal (n=503) and abnormal (n=83) anatomic variant groups, the mean differences were 4.1±8.3 mm, 2.8±5.9 mm, and 12.1±14.6 mm, respectively. The technical success rates were 92.3% (542/586), 96.2 % (486/503), and

67.9% (55/83), respectively. DSC values of segmented body composition were over 0.96 in subgroup with identical CT slice between GT and DLM-derived results. The CSA error rate was less than 5% in technical success group.

Conclusions

The YOLOv3-based DLM system performed well in the automatic L3 level selection, which enables fully automated CT measurement of abdominal muscle area.

REFERENCES

1. von Haehling S, Morley JE, Coats AJS, Anker SD. Ethical guidelines for publishing in the Journal of Cachexia, Sarcopenia and Muscle: update 2019. J Cachexia Sarcopenia Muscle 2019;10:1143-5.

2. Prado CM, Cushen SJ, Orsso CE, Ryan AM. Sarcopenia and cachexia in the era of obesity: clinical and nutritional impact. Proc Nutr Soc 2016;75:188-98.

3. Prado CM, Birdsell LA, Baracos VE. The emerging role of computerized tomography in assessing cancer cachexia. Curr Opin Support Palliat Care 2009;3:269-75.

4. 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.

5. Mitsiopoulos N, Baumgartner RN, Heymsfield SB, Lyons W, Gallagher D, Ross R. Cadaver validation of skeletal muscle measurement by magnetic resonance imaging and computerized tomography. Journal of applied physiology (Bethesda, Md : 1985) 1998;85:115-22.

6. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyere O, Cederholm T, et al. Sarcopenia: revised European consensus on definition and diagnosis. Age Ageing 2019;48:601.

7. Beaudart C, McCloskey E, Bruyere O, Cesari M, Rolland Y, Rizzoli R, et al. Sarcopenia in daily practice: assessment and management. BMC Geriatr 2016;16:170.

8. Shen W, Punyanitya M, Wang Z, Gallagher D, St-Onge MP, Albu J, et al. Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image. J Appl Physiol (1985) 2004;97:2333-8.

9. 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;9:629-35.

10. Muller MJ, Geisler C, Pourhassan M, Gluer CC, Bosy-Westphal A. Assessment and definition of lean body mass deficiency in the elderly. Eur J Clin Nutr 2014;68:1220-7.

11. Mourtzakis M, Prado CM, Lieffers JR, Reiman T, McCargar LJ, Baracos VE. A

practical and precise approach to quantification of body composition in cancer patients using computed tomography images acquired during routine care. Applied physiology, nutrition, and metabolism = Physiologie appliquee, nutrition et metabolisme 2008;33:997-1006.

12. 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.

13. Cespedes Feliciano EM, Popuri K, Cobzas D, Baracos VE, Beg MF, Khan AD, et al. Evaluation of automated computed tomography segmentation to assess body composition and mortality associations in cancer patients. J Cachexia Sarcopenia Muscle 2020; doi:10.1002/jcsm.12573.

14. Decazes P, Rouquette A, Chetrit A, Vera P, Gardin I. Automatic Measurement of the Total Visceral Adipose Tissue From Computed Tomography Images by Using a Multi-Atlas Segmentation Method. J Comput Assist Tomogr 2018;42:139-45.

15. Lee SJ, Liu J, Yao J, Kanarek A, Summers RM, Pickhardt PJ. Fully automated segmentation and quantification of visceral and subcutaneous fat at abdominal CT:

application to a longitudinal adult screening cohort. The British journal of radiology 2018;91:20170968.

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. Kamiya N, Zhou X, Chen H, Muramatsu C, Hara T, Yokoyama R, et al. Automated segmentation of psoas major muscle in X-ray CT images by use of a shape model: preliminary study. Radiological physics and technology 2012;5:5-14.

18. 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. Journal of digital imaging 2017;30:487-98.

19. Belharbi S, Chatelain C, Herault 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.

20. Bridge C, Rosenthal M, Wright B, Kotecha G, Fintelmann F, Troschel F. Fully

Automated Analysis of Body Composition from CT in Cancer Patients Using Convolutional Neural Networks. 2018.

21. Liu Y, Chen PC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. JAMA 2019;322:1806-16.

22. Khalsa AS, Mundis GM, Jr., Yagi M, Fessler RG, Bess S, Hosogane N, et al. Variability in Assessing Spinopelvic Parameters With Lumbosacral Transitional Vertebrae: Inter- and Intraobserver Reliability Among Spine Surgeons. Spine (Phila Pa 1976) 2018;43:813-6.

23. Konin GP, Walz DM. Lumbosacral transitional vertebrae: classification, imaging findings, and clinical relevance. AJNR Am J Neuroradiol 2010;31:1778-86.

24. Park SK, Park JG, Kim BS, Huh JD, Kang H. Thoracolumbar junction: morphologic characteristics, various variants and significance. Br J Radiol 2016;89:20150784.

25. Castellvi AE, Goldstein LA, Chan DPK. Lumbosacral Transitional Vertebrae and Their Relationship with Lumbar Extradural Defects. Spine 1984;9:493-5.

26. Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv 2018. arXiv

preprint arXiv:180402767 2019.

27. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017.

28. Park J, Gil JR, Shin Y, Won SE, Huh J, You MW, et al. Reliable and robust method for abdominal muscle mass quantification using CT/MRI: An explorative study in healthy subjects. PLoS One 2019;14:e0222042.