Radiology

ORIGINAL RESEARCH • GASTROINTESTINAL IMAGING

Automated Abdominal Segmentation of CT Scans for Body Composition Analysis Using Deep Learning

Alexander D. Weston, BS • Panagiotis Korfiatis, PhD • Timothy L. Kline, PhD • Kenneth A. Philbrick, PhD • Petro Kostandy, MD • Tomas Sakinis, MD • Motokazu Sugimoto, MD • Naoki Takahashi, MD • Bradley J. Erickson, MD, PhD

From the Department of Biomedical Engineering and Physiology (A.D.W.) and Department of Radiology (P.K., T.L.K., K.A.P., P.K., T.S., M.S., N.T., B.J.E.), Mayo Clinic, 200 First St SW, Rochester, MN 55905. Received June 21, 2018; revision requested August 15; revision received September 12; accepted October 10. **Address correspondence to** B.J.E. (e-mail: *bje@mayo.edu*).

A.D.W. is a member of the Biomedical Engineering and Physiology graduate program and is supported by Mayo Clinic Graduate School of Biomedical Sciences.

Conflicts of interest are listed at the end of this article.

See also the editorial by Chang in this issue.

Radiology 2019; 00:1–10 • https://doi.org/10.1148/radiol.2018181432 • Content codes: CT GI

Purpose: To develop and evaluate a fully automated algorithm for segmenting the abdomen from CT to quantify body composition.

Materials and Methods: For this retrospective study, a convolutional neural network based on the U-Net architecture was trained to perform abdominal segmentation on a data set of 2430 two-dimensional CT examinations and was tested on 270 CT examinations. It was further tested on a separate data set of 2369 patients with hepatocellular carcinoma (HCC). CT examinations were performed between 1997 and 2015. The mean age of patients was 67 years; for male patients, it was 67 years (range, 29–94 years), and for female patients, it was 66 years (range, 31–97 years). Differences in segmentation performance were assessed by using two-way analysis of variance with Bonferroni correction.

Results: Compared with reference segmentation, the model for this study achieved Dice scores (mean ± standard deviation) of 0.98 ± 0.03, 0.96 ± 0.02, and 0.97 ± 0.01 in the test set, and 0.94 ± 0.05, 0.92 ± 0.04, and 0.98 ± 0.02 in the HCC data set, for the subcutaneous, muscle, and visceral adipose tissue compartments, respectively. Performance met or exceeded that of expert manual segmentation.

Conclusion: Model performance met or exceeded the accuracy of expert manual segmentation of CT examinations for both the test data set and the hepatocellular carcinoma data set. The model generalized well to multiple levels of the abdomen and may be capable of fully automated quantification of body composition metrics in three-dimensional CT examinations.

Online supplemental material is available for this article.

Body composition, defined as the amount and distribution of fat and muscle in the body, is linked to clinical outcomes in a number of conditions, including cancer (1–3), cardiovascular disease (4), and after major surgery (5,6). Despite these associations and the worldwide prevalence of obesity (7), the impact of body composition on these diseases remains poorly understood.

This is partly due to the lack of simple, precise clinical tools. CT and MRI examinations enable accurate and precise body composition measurement (8) and are already part of the clinical workup for many conditions (9) (we focus on CT examinations because they are more commonly performed than MRI, and because the Hounsfield unit range of adipose tissue facilitates its differentiation from other tissues). However, body composition is not routinely calculated as it requires laborious segmentation (tracing) of abdominal compartments.

Several fully automated segmentation methods to assess body composition using CT examinations have been proposed; however, limitations of traditional image-processing techniques and the complexity of abdominal imaging have prevented widespread use. Adipose tissue is primarily identified with threshold-based techniques (10–15), and atlas-based techniques are used to isolate the abdominal muscles (16–19). However, anatomic variability in the abdomen poses a substantial challenge to automated segmentation, and manual correction is almost always required. Because of the time burden for manual correction, a common workaround is to use a single two-dimensional (2D) transverse section to approximate three-dimensional (3D) volume. However, this has been shown to be a poor measure of actual body composition (20,21).

We know of two other publications using a deep learning–based approach for abdominal segmentation for body composition analysis (22,23). Wang et al used thresholding to identify adipose tissue and then a convolutional neural network based on LeNet (24) to separate subcutaneous and visceral fat voxels. Lee et al used VGG-Net (25) to segment the abdominal muscle compartment. Both methods use thresholding to identify adipose tissue, which has many of the same limitations as other automated approaches.

Abbreviations

HCC = hepatocellular carcinoma, STAPLE = simultaneous truth and performance level estimation, 3D = three-dimensional, 2D = two-dimensional

Summary

Fully automated and two-dimensional abdominal segmentation of CT scans performed by using a deep convolutional neural network for assessment of body composition.

Implications for Patient Care

- In a clinical setting, this tool will allow highly accurate body composition information to be calculated automatically from CT examinations to better inform individual care.
- In a research setting, the time required to analyze body composition is the limiting factor in studying its influence on many clinical outcomes; this tool can generate accurate body composition metrics in seconds.

To overcome the limitations of current studies, we propose fully automated abdominal segmentation using a deep convolutional neural network. Because a deep learning approach does not rely on hand-crafted features, it is capable of "learning" information that might not be identified by the image analyst. Furthermore, our model is capable of segmenting subcutaneous and visceral adipose tissue, muscle, abdominal organs, and bone; most fully automated algorithms are demonstrated on adipose tissue and muscle alone. Our hypothesis was that body composition segmentation based on deep learning is as accurate as manual segmentation, even on difficult-to-segment cases.

Materials and Methods

Data Set

The data used in this study were collected retrospectively for CT examinations performed between 1997 and 2015. Data from 2715 examinations were collected from 1429 patients who had been imaged for treatment of pancreatic cancer, renal cell carcinoma, transitional cell carcinoma, or gastrointestinal cancer. Fifteen examinations were excluded due to poor image quality. The study protocol was approved by our local institutional review board and was performed under a Health Insurance Portability and Accountability Act waiver of consent. There was no direct industry support for this study.

To assess the ability of our automated approach to generalize to other data sets, segmentation performance was analyzed on a secondary data set of patients undergoing treatment for hepatocellular carcinoma (HCC). Between 1997 and 2015, 2369 CT examinations from 1083 patients who were being treated for HCC were identified. No examinations from this data set were excluded.

Ground Truth Segmentation

A single transverse section of the abdomen at the level of the L3 vertebra was segmented by one of two expert radiologists (M.S., with 12 years of experience and expertise in general surgery, and N.T., with 18 years of experience and expertise in genitourinary radiology). Segmentation was performed by using a custom semiautomated approach (26). Examinations were segmented into four compartments—subcutaneous adipose tissue, muscle, viscera, and bone—and pixels external to the body. The visceral compartment was further separated into visceral fat-free tissue and visceral adipose tissue by using thresholding. Visceral fat-free tissue is primarily composed of abdominal organs, vessels, and the contents of the digestive tract. Following semiautomated segmentation and thresholding, examinations were visually inspected and manually retraced as needed (see Figure E1 [online]). Figure E2 (online) shows a flowchart of the study design.

Model

Prior to training our model, images were preprocessed with windowing. Values outside the range, −400 to 600 HU, were clipped and the remaining values were scaled between zero and one. The data set was randomly partitioned into training (90%; 2430 of 2700) and test (10%; 270 of 2700) sets. The training set was further partitioned into training (90%; 2187 of 2430) and validation (10%; 243 of 2430) sets. No patient had examinations in both the test and the training and validation data sets.

A convolutional deep neural network referred to as a U-Net (27) was adapted to perform segmentation. The U-Net architecture consists of cascading layers of learnable convolutional filters (Fig 1). Our configuration consisted of five down-sampling and five up-sampling steps, which reduced the 512 × 512 input image to a 16 × 16 × 196 representation and then upsampled it into a 512 × 512 × 6 output. Each step consisted of two consecutive 3 × 3 convolutions (padded), each followed by a hyperbolic-tangent (tanh) activation function, which normalized the values to the range of −1 to +1, followed by max-pooling with a kernel size of 2 × 2 pixels. In the second half of the network, up-sampling operations were performed via 2 × 2 nearest-neighbor interpolation followed by two convolutional layers with tanh activation. A key feature of the U-Net architecture was that the convolutional kernel output from the encoding half of the network was concatenated with each corresponding decoding step, which helped preserve the detail of the original image. The final layer consisted of a convolution with a 1 × 1 × 6 kernel followed by a sigmoid function, which output a score for each of the six tissue classes. The final segmentation was achieved by selecting the tissue class with the highest score for each pixel.

The optimal hyperparameters were experimentally determined. Glorot uniform initialization for the weights of the network was used (28). Weighting was implemented to account for differences in class prevalence. Specifically, the loss contribution of each class was weighted in inverse proportion to its prevalence across the training set. A batch size of 32 sections was used, and the learning rate was kept at a constant 0.0001 with Adam optimization (29). Performance on the training and validation sets was assessed by using categorical cross-entropy. The model was trained until validation loss failed to improve (196 epochs). The network was written in Python 2.7.5 (Python Software Foundation, Beaverton, Ore) using Keras 2.1.5 (open source) with Tensorflow 1.8.0 (open source, Google, Mountain View, Calif)

backend (30). Hyperparameter details are given in Table E1 (online).

Segmentation Performance

Predictions generated by our model were compared against segmentations generated with a custom semiautomated approach with manual correction on the test set (n = 270) (26). Performance was analyzed by using the Dice score, Jaccard score, true-positive fraction (metric for segmentation accuracy; ratio of positive pixels which were correctly classified as positive over total pixels in the compartment), and false-positive fraction (metric for segmentation accuracy; ratio of positive pixels which were falsely classified as negative over total pixels in the compartment) on a pixel-wise basis, as well as the difference in total compartment area calculated from the segmentation. With respect to regions A and B, the Dice score is defined as $2*|A\cap B|/|A+B|$, and the Jaccard score is defined as $|A\cap B|/|A\cup B|$.

Predictions generated by our model on a subset of test images (n = 30) were compared against four-compartment segmentations generated manually by two radiology fellows with expertise in neuroradiology (P.K. and T.S.) (due to time constraints, the visceral compartment was not segmented into visceral adipose and visceral fat-free tissue). Tracers were blinded to the reference segmentation and to the segmentation of the other tracer. These two segmentations were combined with the segmentation taken from the test data set (N.T.) by using simultaneous truth and performance level estimation (STAPLE) (31). Performance of the algorithm compared with these STAPLE-derived expert segmentations was assessed by using the Dice score, Jaccard score, true-positive fraction, false-positive fraction, and difference in area.

Generalizability of Automated Approach

To assess the ability of our automated approach to generalize to other data sets, segmentation performance was analyzed on a secondary data set of patients undergoing treatment for HCC. Predictions were compared with four-compartment reference standard segmentations that had been generated by using a custom semiautomated approach (26) with manual correction (N.T.) (due to time constraints, the visceral compartment was not segmented into visceral adipose and visceral fat-free tissue). Some examinations from the HCC data set were taken at the level of the L4 vertebra in addition to the L3 level; performance on the two levels was also compared.

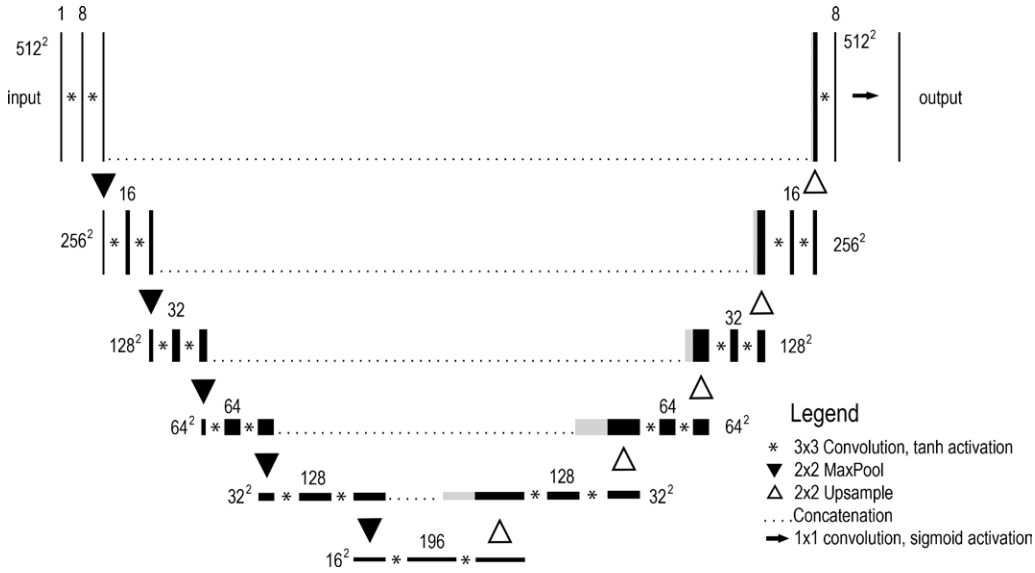

Figure 1: Schematic of U-Net architecture used to generate segmentations.

Table 1: Acquisition Parameters for Training Data Set

Parameter	Value (n = 2707)
CT tube current (mA)*	362 ± 148
Peak tube voltage (kVp)†	
100	171 (6)
120	2384 (88)
130	53 (2)
140	89 (3)
Other	10 (1)
Section thickness (mm)†	
<2	37 (1)
2–3	1105 (41)
3–4	277 (10)
5	1152 (43)
>5	136 (5)

Note.—Data in parentheses are percentages.

* Data are mean ± standard deviation.

† Data are no. of images.

Three-dimensional Segmentation

To determine whether a model trained on a 2D section at the level of the L3 transverse processes could generalize across the entire abdomen, 12 complete examinations of the abdomen from the inferior endplate of the L1 vertebra to the superior endplate of the L5 vertebra were used. Each section in this range was segmented using our model. Segmentation results are reported qualitatively; reference standard 3D segmentations were not available for this data set.

Statistical Analysis

Differences in area calculated with the various segmentation methods were assessed by using two-way analysis of variance with Bonferroni correction for multiple comparisons. A P value less than .004 after Bonferroni correction was indicative

Table 2: Patient Characteristics for Training Data Set

Characteristic	Male Patients (n = 878)	Female Patients (n = 551)	All Patients (n = 1429)
Mean age (y)	66.9 ± 10.5	65.9 ± 11.7	66.5 ± 11.0
Age range (y)	29–94	31–97	
Mean weight (kg)*	88.4 ± 16.8	72.0 ± 16.9	82.7 ± 18.6
Mean BMI (kg/m^2)*	28.0 ± 5.5	27.0 ± 6.8	27.6 ± 6.0
BMI classification†	472	255	727
Underweight	7 (1)	9 (4)	16 (2)
Healthy	123 (26)	88 (35)	211 (29)
Overweight	194 (41)	93 (36)	287 (39)
Obese	148 (31)	65 (25)	213 (29)
Diagnosis†	520	287	807
Pancreatic cancer	212 (41)	173 (60)	385 (48)
Bladder cancer	175 (34)	84 (29)	205 (25)
Healthy	133 (26)	30 (10)	217 (27)

Note.—Unless otherwise indicated, data are mean ± standard deviation. BMI = body mass index.

* BMI was available for a subset of patients (n = 727).

† Data are number of patients and data in parentheses are percentages.

Table 3: Algorithm Performance Compared with Semiautomated Segmentation on the Test Data Set, Manual Segmentation on the Test Data Set, and Semiautomated Segmentation on the HCC Data Set

Compartment	Dice Score	Jaccard Score	True-Positive Fraction	False-Positive Fraction	Area Difference (%)
	Prediction vs Semiautomated Approach with Manual Correction (n = 270)				
Subcutaneous adipose tissue	0.98 ± 0.03	0.96 ± 0.04	0.98 ± 0.03	0.02 ± 0.03	1.1 ± 2.0
Muscle	0.96 ± 0.02	0.92 ± 0.04	0.96 ± 0.03	0.04 ± 0.03	2.1 ± 2.5
Visceral fat-free tissue	0.97 ± 0.01	0.94 ± 0.02	0.97 ± 0.02	0.03 ± 0.02	1.4 ± 1.8
Visceral adipose tissue	0.94 ± 0.12	0.90 ± 0.13	0.94 ± 0.12	0.05 ± 0.07	3.7 ± 12.2
Bone	0.98 ± 0.02	0.97 ± 0.03	0.99 ± 0.01	0.03 ± 0.04	2.2 ± 4.3
	Prediction vs STAPLE-derived Manual Expert Segmentation (n = 30)				
Subcutaneous adipose tissue	0.98 ± 0.01	0.96 ± 0.02	0.98 ± 0.01	0.02 ± 0.01	1.2 ± 1.2
Muscle	0.95 ± 0.02	0.91 ± 0.03	0.95 ± 0.02	0.05 ± 0.02	2.1 ± 1.8
Viscera	0.99 ± 0.00	0.98 ± 0.01	0.99 ± 0.01	0.01 ± 0.01	0.7 ± 0.4
Bone	0.95 ± 0.02	0.91 ± 0.04	0.99 ± 0.01	0.10 ± 0.05	9.4 ± 4.9*
	Prediction vs Semiautomated Approach with Manual Correction, HCC Cases (n = 2369)				
Subcutaneous adipose tissue	0.94 ± 0.05	0.89 ± 0.07	0.97 ± 0.05	0.09 ± 0.09	6.6 ± 8.1*
Muscle	0.92 ± 0.04	0.86 ± 0.07	0.92 ± 0.04	0.08 ± 0.07	4.3 ± 4.8*
Viscera	0.98 ± 0.02	0.96 ± 0.03	0.98 ± 0.02	0.02 ± 0.03	1.7 ± 2.9*
Bone	0.97 ± 0.04	0.95 ± 0.05	0.99 ± 0.03	0.05 ± 0.08	4.2 ± 7.3*
	Manual Interobserver Variability (n = 30)				
Subcutaneous adipose tissue	0.95 ± 0.02	0.91 ± 0.03	0.94 ± 0.02	0.03 ± 0.02	3.1 ± 2.5*
Muscle	0.93 ± 0.02	0.87 ± 0.03	0.93 ± 0.02	0.07 ± 0.02	3.8 ± 3.2*
Viscera	0.99 ± 0.01	0.97 ± 0.01	0.98 ± 0.01	0.01 ± 0.01	0.9 ± 1.0
Bone	0.95 ± 0.02	0.91 ± 0.03	0.95 ± 0.03	0.04 ± 0.03	4.3 ± 3.6

Note.—Data are mean ± standard deviation. HCC = hepatocellular carcinoma, STAPLE = simultaneous truth and performance level estimation.

* Statistically significant.

of statistical significance. The Pearson correlation coefficient was used to assess the relationship between compartment area and body mass index. Bland-Altman plots were used to compare the prediction versus 30 STAPLE-derived reference standards, and also the two manual segmentations to evaluate interobserver variation. Statistical analysis was performed by using LibreOffice Calc 5.0.6.2 (open source, The Document Foundation, Berlin, Germany).

Results

A total of 2707 CT images (one axial section from each examination) were collected from 1429 patients treated for pancreatic can-

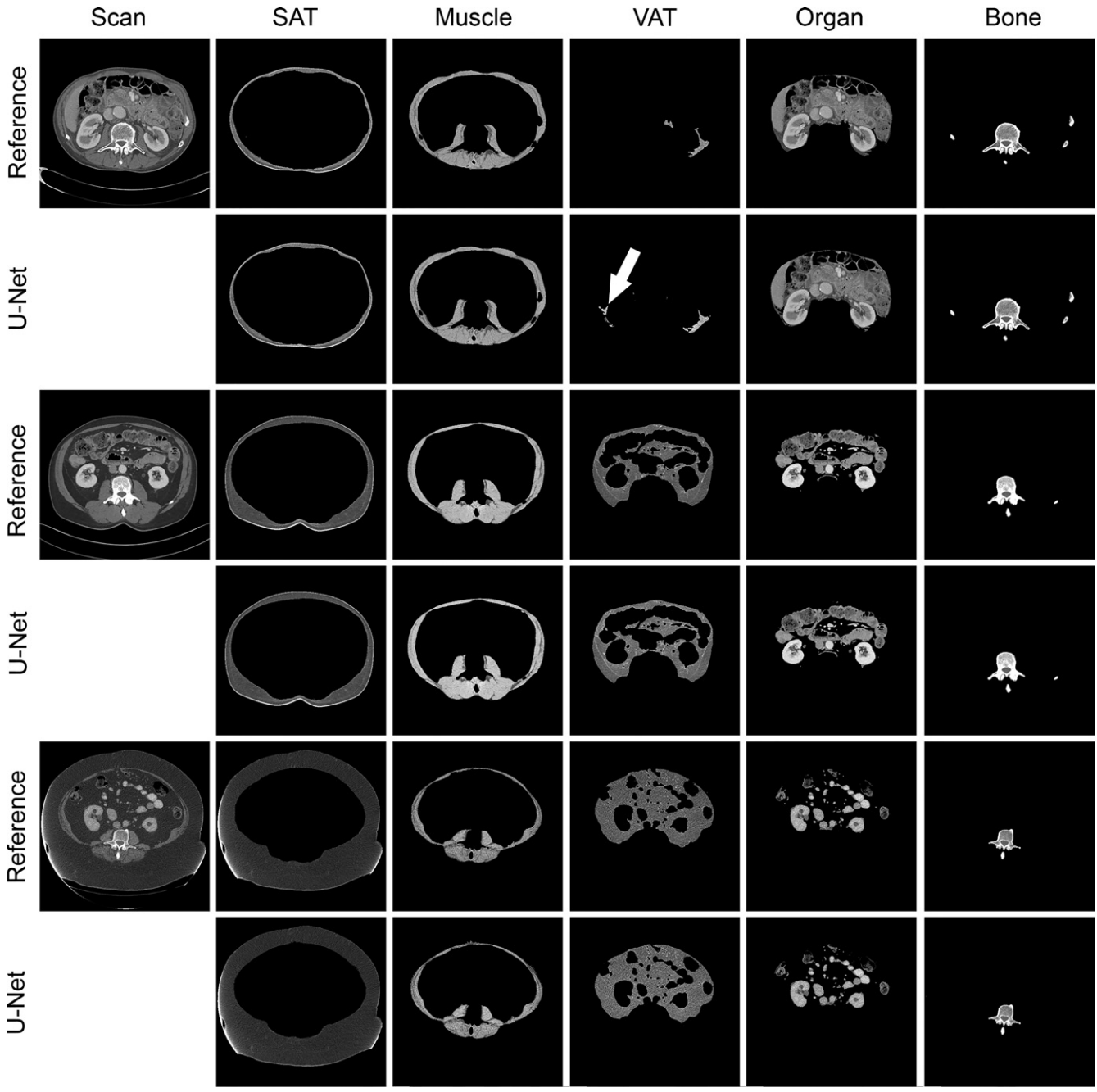

Figure 2: Representative images illustrate deep learning–based body composition segmentation. Original image (*Scan*) is followed by images of subcutaneous adipose tissue (*SAT*), muscle, visceral adipose tissue (*VAT*), visceral organ, and bone compartments. For each scan, segmentation with semiautomated method is shown in the first row (*Reference*); U-Net predictions are shown in the second row. Arrow indicates region of disagreement.

cer, renal cell carcinoma, transitional cell carcinoma, and various gastrointestinal cancers at our institution between 1997 and 2015. The mean age for patients was 67 years; for men, it was 67 years (range, 29–94 years) and for women, it was 66 years (range, 31–97 years). Acquisition parameters taken from the Digital Imaging and Communications in Medicine headers are shown in Table 1. Characteristics of patients are shown in Table 2.

Segmentation Performance

Accurate segmentation of body composition was achieved by using our automated approach (Table 3). Dice and Jaccard scores were highest in the subcutaneous adipose tissue compartment and bone compartment and lowest in the visceral adipose tissue compartment. Variability in the visceral adipose tissue compartment was somewhat greater than in other compartments.

Figure 2 shows representative segmentations produced by the automated approach on the test data set. The patient with healthy weight had less visceral fat than other tissues; even small absolute errors in classification substantially reduced the Dice score (0.66). The difference in compartment area calculated by using both methods was approximately the same (7.7 vs 8.3

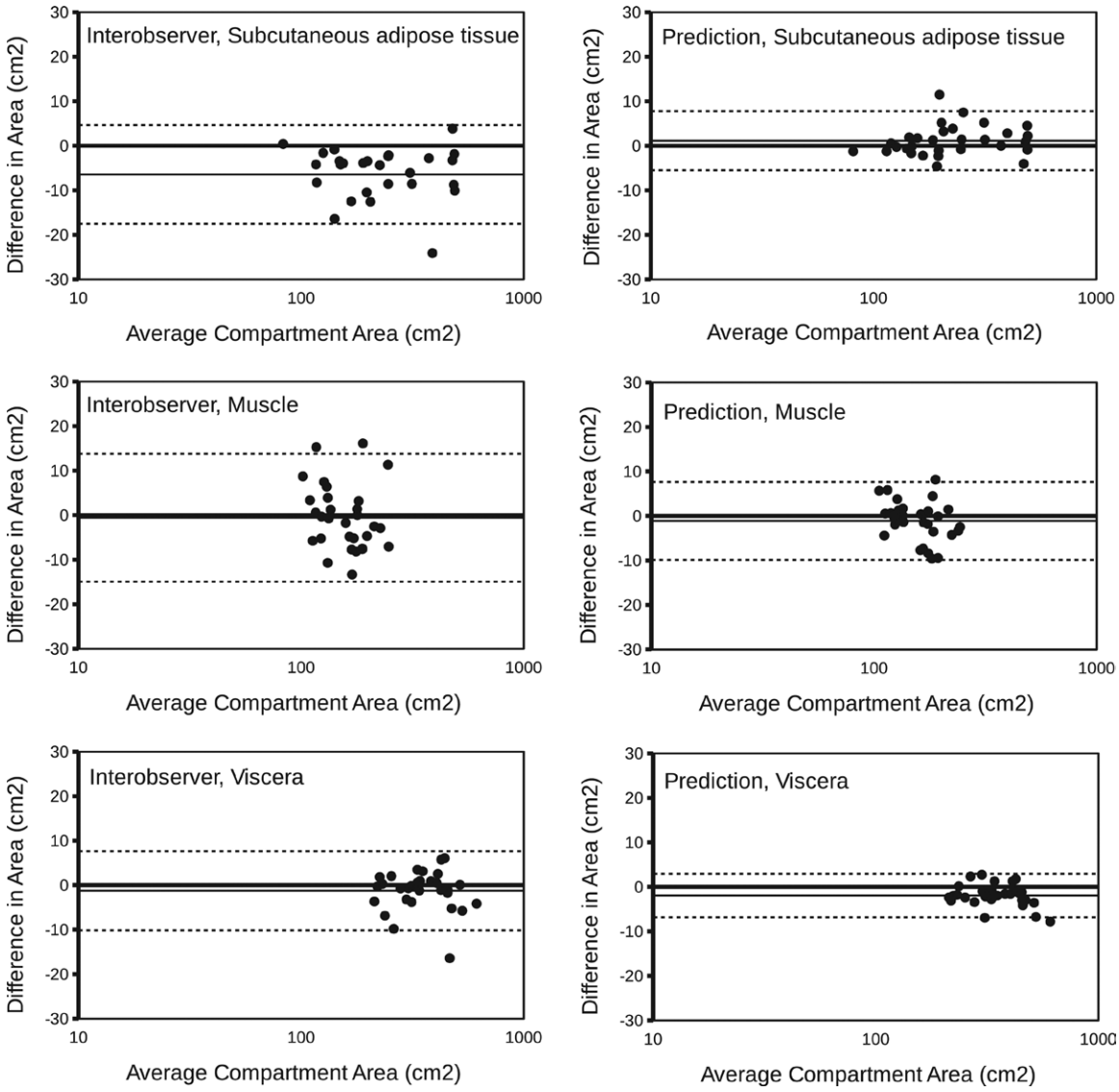

Figure 3: Bland-Altman plots for difference in area for subcutaneous adipose tissue, muscle, and visceral adipose tissue compartments. The difference between two expert manual segmentations (*Interobserver*) is shown on the left, and the difference between our algorithm and expert segmentation derived from simultaneous truth and performance level estimation (*Prediction*) is shown on the right. The mean is indicated by the solid line, and 95% confidence intervals are indicated by the dashed lines.

cm^2). Table E2 (online) shows sensitivity analysis for these results. Figure E3 (online) shows segmentation performance on three examinations taken from the same individual.

Performance of our automated approach on STAPLE-derived expert segmentations was comparable to performance on the test data set. Compartment area calculated on the test data set was not different than the STAPLE-derived expert segmentations for the subcutaneous adipose tissue (P = .91), muscle (P = .87), and visceral compartments (P = .87). There was a difference in compartment area calculated for the bone compartment, with better performance on the test data set compared with the STAPLE-derived expert segmentation ($P < .001$).

The automated approach performed at a level comparable to manual expert segmentation. Our model (evaluated on the test data set) exceeded interobserver variability on the subcutaneous adipose ($P < .001$) and muscle compartments (P = .001). No difference was observed on the visceral (P = .55) and bone (P = .01) compartments after Bonferroni correction.

The automated method performed as well as expert manual segmentation regardless of patient characteristics, as indicated by the Bland-Altman plots (Fig 3). Despite the variation in patient weight, the absolute error between our prediction and the reference standard was constant across all patients for both our model and manual segmentation. A suggestion of bias (P = .01, not significant after Bonferroni correction) was observed between our expert manual segmentations on the subcutaneous compartment; this was not observed between the automated approach and the STAPLE-derived expert segmentation.

Generalizability

The automated approach generalized well to the secondary data set of HCC cases, which were not represented in the training and test data sets. Our model performed slightly worse on the

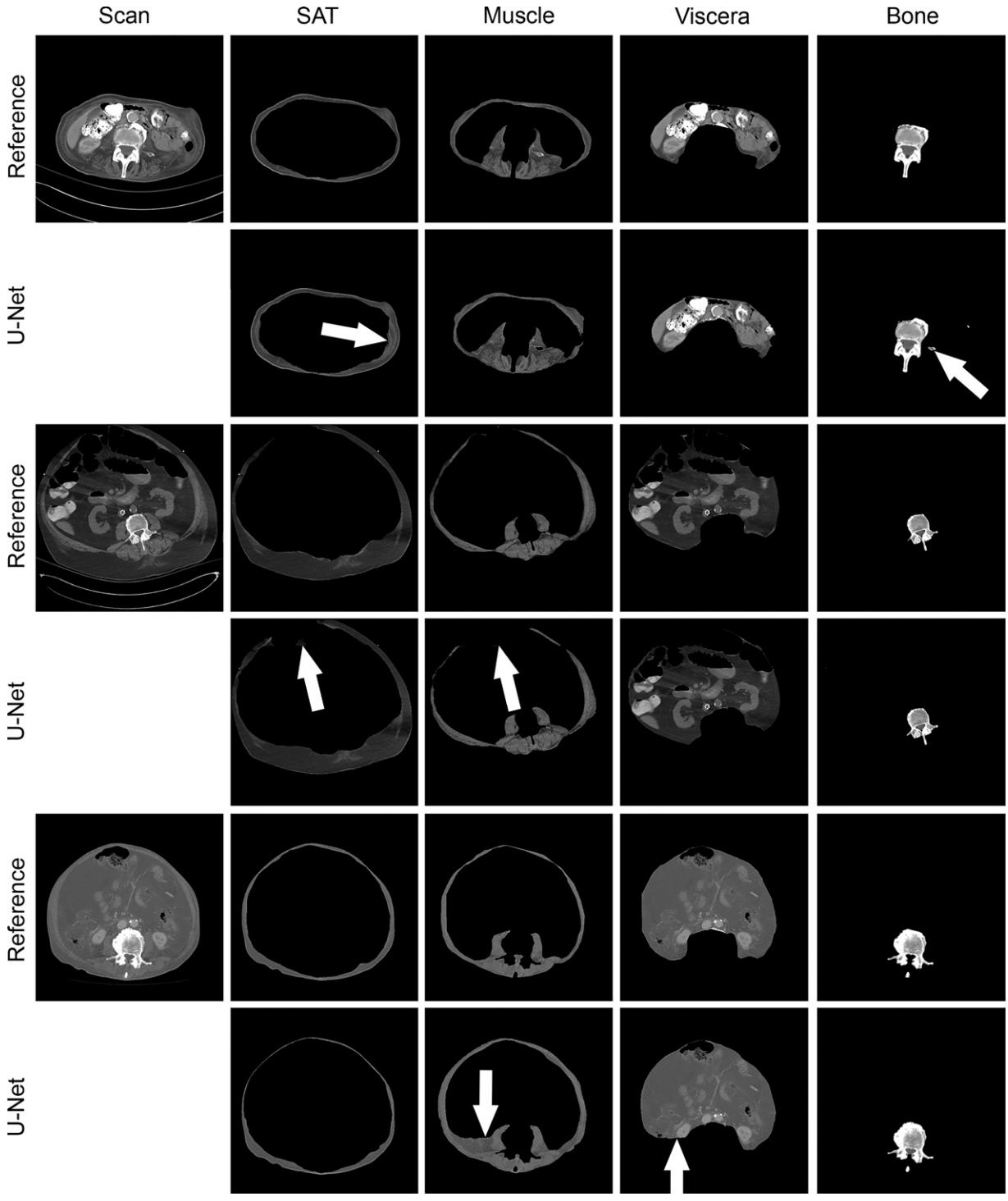

Figure 4: Representative images of body composition segmentation based on deep learning for three difficult-to-segment hepatocellular carcinoma (HCC) scans not taken from the train/test data set. For each HCC scan, reference standard segmentation with a semiautomated approach with expert correction is shown in the first row (Reference) and U-Net predictions, in the second row (U-Net). Arrows indicate regions of disagreement. Rows 1 and 2: For the first scan, the model falsely segmented the subcutaneous adipose tissue (SAT) compartment as belonging to muscle. Additionally, the edge of the transverse process is visible and its classification is included in the prediction, but not in the reference standard segmentation. Dice scores for SAT, muscle, viscera, and bone compartments are 0.84, 0.80, 0.92, and 0.96, respectively. Rows 3 and 4: Images in a patient with thin abdominal wall and ventral hernia, which was incorrectly segmented. Dice scores for SAT, muscle, viscera, and bone compartments are 0.95, 0.86, 0.95, and 0.98, respectively. Rows 5 and 6: Images in a patient with ascites, which distorted anatomy, and subcutaneous edema, which increased the density of adipose tissue. Misclassification of the right dorsal abdominal wall was also observed. Dice scores for SAT, muscle, viscera, and bone compartments are 0.80, 0.75, 0.96, and 0.96, respectively.

Table 4: Mean Dice Score, Jaccard Score, True-Positive Fraction, and Fase-Positive Fraction Values Comparing Segmentation with U-Net versus Semiautomated Approach with Manual Correction at the L3 and L4 Levels

	Dice Score		Jaccard Score		True-Positive Fraction		False-Positive Fraction	
Compartment	L3	L4	L3	L4	L3	L4	L3	L4
Subcutaneous adipose tissue	0.93 ± 0.06	0.94 ± 0.08	0.87 ± 0.10	0.89 ± 0.10	0.93 ± 0.08	0.94 ± 0.10	0.08 ± 0.08	0.06 ± 0.05
Muscle	0.88 ± 0.07	0.88 ± 0.06	0.79 ± 0.10	0.80 ± 0.09	0.89 ± 0.08	0.89 ± 0.07	0.13 ± 0.09	0.13 ± 0.09
Viscera	0.97 ± 0.02	0.96 ± 0.03	0.94 ± 0.04	0.93 ± 0.05	0.97 ± 0.02	0.96 ± 0.03	0.03 ± 0.04	0.04 ± 0.06
Bone	0.95 ± 0.05	0.95 ± 0.07	0.90 ± 0.08	0.91 ± 0.10	0.99 ± 0.04	0.99 ± 0.02	0.10 ± 0.11	0.11 ± 0.22

Note.—Data are mean ± standard deviation. No difference in performance between the L3 and L4 vertebrae was observed by Student t test using the Bonferroni correction.

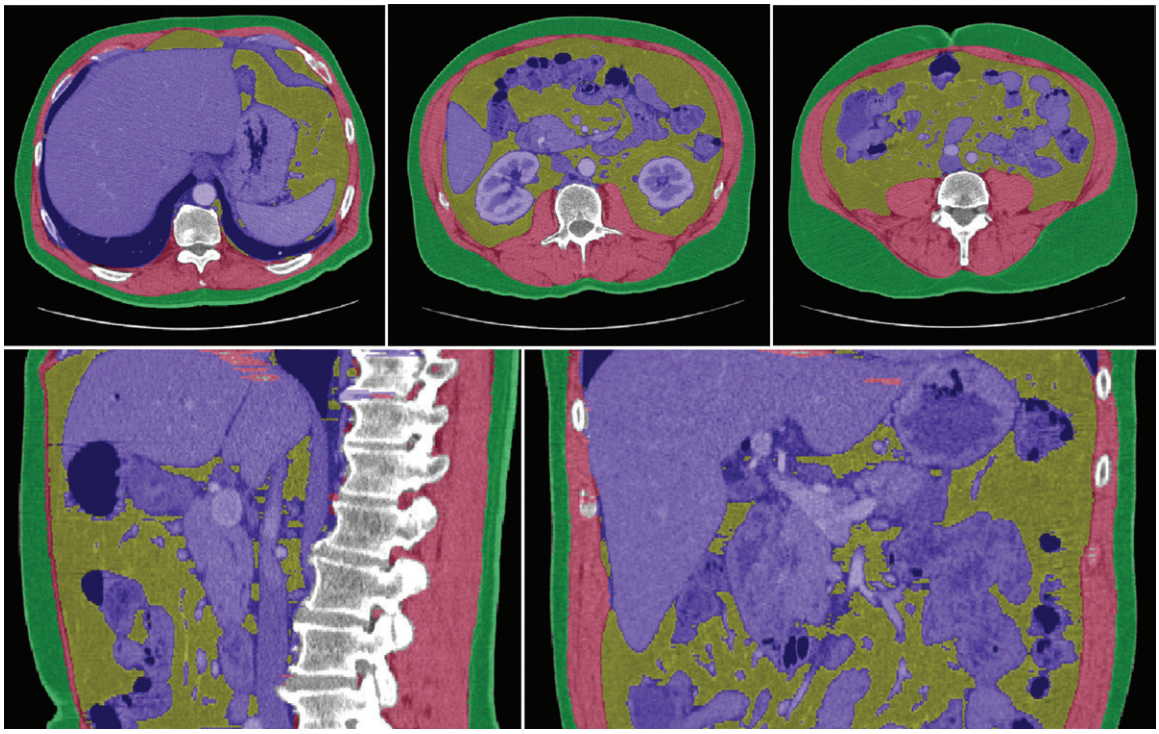

Figure 5: Example of three-dimensional (3D) segmentation using our algorithm. The 3D scan was subsampled into a series of two-dimensional images, and fully automated segmentation based on deep learning was performed on the series of images. Several views of the 3D volume are shown to demonstrate the model's ability to generalize across multiple sections.

subcutaneous adipose tissue ($P < .001$), muscle ($P < .001$), visceral ($P < .001$), and bone compartments ($P < .001$), compared with the 270 test cases. This was most pronounced for the muscle compartment (Dice score, 0.92 compared with 0.96 on the test data set). Compartment area calculated on the HCC data set was also slightly worse for the subcutaneous adipose tissue ($P < .001$), muscle ($P < .001$), visceral ($P < .001$), and bone compartments ($P < .001$). Nevertheless, the average difference in compartment area between the prediction and the reference standard was small, 7% for the subcutaneous adipose compartment and less for other compartments.

Figure 4 shows the representative segmentations produced by our model on the HCC data set. Images were chosen to highlight instances where the model failed.

Our approach was capable of generalizing to other section levels within the abdomen (Table 4). Despite the fact that our algorithm was trained solely on transverse sections at the level of the L3 transverse processes, no significant difference in model performance was measured for the subcutaneous adipose tissue ($P = .32$), muscle ($P = .65$), and visceral adipose tissue ($P = .06$) compartments at the level of the L4 transverse processes.

Three-dimensional Segmentation

Figure 5 shows a representative 3D segmentation performed using our model. Our algorithm correctly classified portions of the sacrum and pelvis in the lower abdomen, and the diaphragm and lungs in the upper abdomen, despite the fact that these structures were not present in the training data set (which consisted of images at the L3 level alone). Figure E4 (online) shows a 3D rendering of body composition segmentation; Figure E5 (online) shows examples of compartment cross-sectional area from the L5 to the L1 vertebra (approximately

15 cm) for an underweight, healthy weight, overweight, and obese individual for the compartments of interest. Figure E6 (online) demonstrates the relationship between body mass index and compartment area.

Discussion

Our algorithm accurately segmented four compartments used for body composition using routine CT of the abdomen. Visceral adiposity had a larger standard deviation than other compartments (standard deviation in the Dice score of 0.12, compared with 0.03 and 0.02 in the subcutaneous and muscle compartments, respectively). Visceral adiposity is particularly challenging to measure in patients with healthy weight where low adiposity results in a poor Dice score even when the absolute error is low. The error in visceral adipose tissue area was less than that for other compartments, which is an indication that our algorithm performs well, and Bland-Altman plots demonstrate that our model was at least as accurate as manual segmentation regardless of patient weight. A particular challenge for this compartment is the partial volume effect of gas with fatty bowel contents, which can cause simple thresholding to fail (21). Muscle can also be difficult to delineate because of its thin width and similar texture and intensity to visceral organs, especially when muscle and organs are adjacent, or in the case of ascites. Nevertheless, the error in calculated compartment area was approximately 2% compared with that in both the test data set (3 cm^2/159 cm^2) and the STAPLE-derived expert segmentation (4 cm^2/161 cm^2). Finally, model performance on the test data set equaled interobserver agreement on the visceral and bone compartments and exceeded interobserver agreement on the subcutaneous adipose tissue and muscle compartments.

Our algorithm was accurate on HCC cases (a pathology cohort not represented in our initial training and test data sets). Our algorithm performed only slightly worse on this data set (area difference of 6.6% vs 1.2% on the subcutaneous adipose tissue compartment, 4.3% vs 2.1% on the muscle compartment, and 1.7% vs 0.7% on the visceral compartment). Although it is not uncommon for deep learning–based methods to perform slightly worse on a secondary data set, several features of the HCC data set, including atrophy of the abdominal musculature, distorted anatomy due to ventral herniation, and ascites and edema in the visceral and subcutaneous compartments, may explain the decreased performance. Ascites can occur in advanced cirrhosis and is particularly prevalent in the HCC data set—the difficulty of delineating the muscle wall in patients with edema in the subcutaneous compartment contributes to the low true-positive fraction of the muscle compartment observed on this data set. These examples of algorithm failure are useful for understanding what features are considered important by the deep learning model. In particular, these results suggest that texture and location, in addition to intensity, are important for adipose tissue classification.

Our model generalized well to transverse sections other than the L3 level on which it was trained. There was no significant difference between performance at the level of the L4 vertebra versus performance at L3. Additionally, in the 3D segmentation, organs not present in the training data, such as lungs, bladder, and pelvis, were correctly segmented. This strong performance on anatomy not represented in the training set indicates our model is learning organ features, which generalize to examinations beyond the training data. Our results suggest that an accurate 3D segmentation is possible by using a simple 2D model, which could vastly reduce both the complexity and the amount of training data required to develop a segmentation tool.

Three-dimensional body composition analysis from CT examinations using our algorithm represents an improvement over current clinical tools. Common anthropometric techniques such as body mass index, skin-fold analysis, bioelectric impedance analysis, and functional metrics to measure sarcopenia, such as gait speed and grip strength, are incapable of discerning adipose tissue distribution in the body (32–34). The current reference standard, dual-energy x-ray absorptiometry, is a 2D projection of the body, which poorly differentiates between subcutaneous and visceral adipose fat and cannot measure muscle quality or organ adiposity (35). CT or MRI examinations alone are capable of measuring these biomarkers. Additionally, for many clinical conditions, CT imaging is already included in the clinical workflow. The main barrier for adoption of CT body composition analysis into clinical practice is the lack of accurate fully automated segmentation techniques. Several fully automated techniques have been proposed in the last decade, including threshold-based approaches (10–15) and atlas-based approaches (16–19); however, the heterogeneous appearance of the abdomen, such as the thin muscle wall, or the similar intensity of gastrointestinal contents to adipose tissue makes this a challenging task.

Because of the time-consuming nature of manual segmentation, it is common to quantify body composition from a single section, usually at the L3 level. However, both the adipose tissue area and the muscle area vary dramatically at different levels of the abdomen, sometimes as much as twofold for muscle and threefold for visceral adipose tissue. Because the contents of the gastrointestinal tract are constantly shifting, there is no guarantee that two repeated L3 sections will capture the same anatomy. For these reasons, 3D analysis of body composition is more accurate than 2D approximation at the L3 level.

Segmentation of adipose and muscle volumes may be valuable for creating new biomarkers of health and disease. Certain body composition markers are known to predict outcomes in major surgery (5,6). It is likely that other metabolic conditions may be better diagnosed or predicted using measures developed here. Research of this nature requires study of large populations, something that is really only feasible with automation.

Limitations of our study include the fact that the model was validated on CT alone at the L3 vertebra. Although the model generalized well to other levels of the abdomen, it is likely to be most accurate at the midabdomen. Qualitative results are reported on the 3D data set in the absence of 3D reference standards. Our future work will include further development of the 3D model, including quantitative comparisons to reference segmentations. Finally, although our study contained a large and diverse data set, it is limited somewhat by characteristics of the patient population, which was skewed toward older, overweight patients often with substantial pathology.

We present a fully automated algorithm for segmenting abdominal CT images to quantify body composition based on a

deep learning approach. Our approach is unique in that it segments subcutaneous adipose tissue, muscle, bone, visceral adipose tissue, and visceral organs simultaneously. Our algorithm meets or exceeds the accuracy of expert manual segmentation. It is generalizable to data sets not represented in our training data, an essential requirement for developing a clinically useful tool. Despite the fact that our model is trained solely on 2D sections at the L3 level, we observed our model at other levels of the abdomen. This suggests that our algorithm is a useful tool for performing fully automated 2D segmentation despite being trained on 2D data. Using our tool to perform body composition analysis on several patients, we observed high variability in anatomy at a single section, suggesting that single-section analysis is of limited utility and that 3D analysis is a more accurate method.

Acknowledgments: The authors would like to thank Zeynettin Akkus, PhD, for intellectual contributions to this project, Bill Ryan, BS, for assistance with data curation, and Scott Squires, BS, for technical support.

Author contributions: Guarantors of integrity of entire study, A.D.W., N.T., B.J.E.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, A.D.W., P.K., T.L.K., N.T.; clinical studies, A.D.W., M.S., N.T.; experimental studies, A.D.W., P.K., T.L.K., K.A.P., P.K., T.S., M.S., B.J.E.; statistical analysis, A.D.W., T.L.K., K.A.P., N.T., B.J.E.; and manuscript editing, A.D.W., P.K., T.L.K., K.A.P., P.K., N.T., B.J.E.

Disclosures of Conflicts of Interest: **A.W.** disclosed no relevant relationships. **P.K.** disclosed no relevant relationships. **T.L.K.** disclosed no relevant relationships. **K.A.P.** disclosed no relevant relationships. **P.K.** disclosed no relevant relationships. **T.S.** disclosed no relevant relationships. **M.S.** disclosed no relevant relationships. **B.J.E.** disclosed no relevant relationships. **N.T.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed a patent in CT segmentation (software) for System and Method for Quantification of Muscle and Fat Using Abdominal CT Imaging. The gold standard of the current study was created using the software, which is not commercially available or licensed, and neither N.T. nor Mayo Clinic have received any royalty. N.T. disclosed that the deep learning software described in this manuscript is superior to his patented software, and thus his software is obsolete.

References

1. Gonzalez MC, Pastore CA, Orlandi SP, Heymsfield SB. Obesity paradox in cancer: new insights provided by body composition. Am J Clin Nutr 2014;99(5):999–1005.
2. Aust S, Knogler T, Pils D, et al. Skeletal muscle depletion and markers for cancer cachexia are strong prognostic factors in epithelial ovarian cancer. PLoS One 2015;10(10):e0140403.
3. Malietzis G, Currie AC, Athanasiou T, et al. Influence of body composition profile on outcomes following colorectal cancer surgery. Br J Surg 2016;103(5):572–580.
4. Jean N, Somers VK, Sochor O, Medina-Inojosa J, Llano EM, Lopez-Jimenez F. Normal-weight obesity: implications for cardiovascular health. Curr Atheroscler Rep 2014;16(12):464.
5. Sheetz KH, Waits SA, Terjimanian MN, et al. Cost of major surgery in the sarcopenic patient. J Am Coll Surg 2013;217(5):813–818.
6. Sugimoto M, Farnell MB, Nagorney DM, et al. Decreased skeletal muscle volume is a predictive factor for poorer survival in patients undergoing surgical resection for pancreatic ductal adenocarcinoma. J Gastrointest Surg 2018;22(5):831–839.
7. Abelson P, Kennedy D. The obesity epidemic. Science 2004;304(5676):1413–1413.
8. Seabolt LA, Welch EB, Silver HJ. Imaging methods for analyzing body composition in human obesity and cardiometabolic disease. Ann N Y Acad Sci 2015;1353(1):41–59.
9. Mourtzakis M, Prado CMM, 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. Appl Physiol Nutr Metab 2008;33(5):997–1006.
10. 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. In: Medical Imaging 2011: Computer-Aided Diagnosis. International Society for Optics and Photonics, 2011; 79632Q. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/7963/79632Q/Development-of-automated-quantification-of-visceral-and-subcutaneous-adipose-tissue/10.1117/12.878017.short. Accessed April 13, 2018.
11. Kim YJ, Lee SH, Kim TY, Park JY, Choi SH, Kim KG. Body fat assessment method using CT images with separation mask algorithm. J Digit Imaging 2013;26(2):155–162.
12. Parikh AM, Coletta AM, Yu ZH, et al. Development and validation of a rapid and robust method to determine visceral adipose tissue volume using computed tomography images. PLoS One 2017;12(8):e0183515.
13. Pednekar A, Bandekar AN, Kakadiaris IA, Naghavi M. Automatic segmentation of abdominal fat from CT data. In: Application of Computer Vision, 2005 WACV/MOTIONS'05 Vol 1 Seventh IEEE Workshops on. IEEE, 2005; 308–315. http://ieeexplore.ieee.org/abstract/document/4129496/. Accessed April 13, 2018.
14. 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. arXiv [cs.CV]. 2015. http://arxiv.org/abs/1512.04958. Accessed April 13, 2018.
15. Kullberg J, Hedström A, Brandberg J, et al. Automated analysis of liver fat, muscle and adipose tissue distribution from CT suitable for large-scale studies. Sci Rep 2017;7(1):10425.
16. Agarwal C, Dallal AH, Arbabshirani MR, Patel A, Moore G. Unsupervised quantification of abdominal fat from CT images using Greedy Snakes. In: Medical Imaging 2017: Image Processing. International Society for Optics and Photonics, 2017; 101332T. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10133/101332T/Unsupervised-quantification-of-abdominal-fat-from-CT-images-using-Greedy/10.1117/12.2254139.short. Accessed April 13, 2018.
17. Popuri K, Cobzas D, Esfandiari N, Baracos V, Jägersand M. Body composition assessment in axial CT images using FEM-based automatic segmentation of skeletal muscle. IEEE Trans Med Imaging 2016;35(2):512–520.
18. Zhang W, Liu J, Yao J, Summers RM. Segmenting the thoracic, abdominal and pelvic musculature on CT scans combining atlas-based model and active contour model. In: Medical Imaging 2013: Computer-Aided Diagnosis. International Society for Optics and Photonics, 2013; 867008. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/8670/867008/Segmenting-the-thoracic-abdominal-and-pelvic-musculature-on-CT-scans/10.1117/12.2007970.short. Accessed April 13, 2018.
19. Chung H, Cobzas D, Lieffers J, Birdsell L, Baracos V. Automated segmentation of muscle and adipose tissue on CT images for human body composition analysis. https://webdocs.cs.ualberta.ca/~dana/Papers/09SPIE_muscleFat.pdf. Accessed April 13, 2018.
20. Shen W, Punyanitya M, Wang Z, et al. Visceral adipose tissue: relations between single-slice areas and total volume. Am J Clin Nutr 2004;80(2):271–278.
21. Potretzke AM, Schmitz KH, Jensen MD. Preventing overestimation of pixels in computed tomography assessment of visceral fat. Obes Res 2004;12(10):1698–1701.
22. 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.
23. Lee H, Troschel FM, Tajmir S, et al. Pixel-level deep segmentation: artificial intelligence quantifies muscle on computed tomography for body morphometric analysis. J Digit Imaging 2017;30(4):487–498.
24. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86(11):2278–2324.
25. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv [cs.CV]. 2014. http://arxiv.org/abs/1409.1556. Accessed May 4, 2018.
26. Takahashi N, Sugimoto M, Psutka SP, Chen B, Moynagh MR, Carter RE. Validation study of a new semi-automated software program for CT body composition analysis. Abdom Radiol (NY) 2017;42(9):2369–2375.
27. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, eds. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham, Switzerland: Springer, 2015; 234–241.
28. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010; 249–256. http://proceedings.mlr.press/v9/glorot10a.html. Accessed November 7, 2017.

29. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv [cs.LG]. 2014. http://arxiv.org/abs/1412.6980. Accessed November 7, 2017.
30. Abadi M, Barham P, Chen J, et al. TensorFlow: A System for Large-Scale Machine Learning. In: OSDI. usenix.org, 2016; 265–283. https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf. Accessed May 4, 2018.
31. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004;23(7):903–921.
32. Prado CM, Gonzalez MC, Heymsfield SB. Body composition phenotypes and obesity paradox. Curr Opin Clin Nutr Metab Care 2015;18(6):535–551.
33. Kutáč P, Gajda V. Evaluation of accuracy of the body composition measurements by the BIA method. Hum Mov Sci 2011;12(1):41–45.
34. Lowry DW, Tomiyama AJ. Air displacement plethysmography versus dual-energy x-ray absorptiometry in underweight, normal-weight, and overweight/obese individuals. PLoS One 2015;10(1):e0115086.
35. Bachrach LK. Dual energy X-ray absorptiometry (DEXA) measurements of bone density and body composition: promise and pitfalls. J Pediatr Endocrinol Metab 2000;13(Suppl 2):983–988.