Body composition assessment by artificial intelligence from routine computed tomography scans in colorectal cancer: Introducing BodySegAI

Dena Helene Alavi1*, Tomas Sakinis2, Hege Berg Henriksen1, Benedicte Beichmann1, Ann-Monica Fløtten3, Rune Blomhoff1,4 & Peter Mæhre Lauritzen3

1Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway; 2Medical Division, Radiology and Nuclear Medicine, Neuroimaging Research Group, Oslo University Hospital, Oslo, Norway; 3Division of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway; 4Department of Clinical Service, Division of Cancer Medicine, Oslo University Hospital, Oslo, Norway

Abstract

Background Body composition is of clinical importance in colorectal cancer patients, but is rarely assessed because of time-consuming manual segmentation. We developed and tested BodySegAI, a deep learning-based software for automated body composition quantification from routinely acquired computed tomography (CT) scans.

Methods A two-dimensional U-Net convolutional network was trained on 2989 abdominal CT slices from L2 to S1 to segment skeletal muscle (SM), visceral adipose tissue (VAT), subcutaneous adipose tissue (SAT), and intermuscular and intramuscular adipose tissue (IMAT). Human ground truth was established by combining segmentations from three human readers. BodySegAI was tested using 154 slices against the human ground truth and compared with a software named AutoMATiCA.

Results Median Dice scores for BodySegAI against human ground truth were 0.969, 0.814, 0.986, and 0.990 for SM, IMAT, VAT, and SAT, respectively. The mean differences per slice for SM were −0.09 cm³, IMAT: −0.17 cm³, VAT: −0.12 cm³, and SAT: 0.67 cm³. Median absolute errors for SM, IMAT, VAT, and SAT were 1.35, 10.54, 0.91, and 1.07%, respectively. When analysing different anatomical levels separately, L3 and S1 demonstrated the overall highest and lowest Dice scores, respectively. On average, BodySegAI segmented 148 times faster than human readers (4.9 vs. 726.5 seconds, \( P < 0.001 \)). Also, BodySegAI presented higher Dice scores for SM, IMAT, SAT, and VAT than AutoMATiCA (slices = 154).

Conclusions BodySegAI rapidly generates excellent segmentation of SM, VAT, and SAT and good segmentation of IMAT in L2 to S1 among colorectal cancer patients and may replace semi-manual segmentation.

Keywords Artificial intelligence; Computed tomography (CT); Body composition; Skeletal muscle (SM); Visceral adipose tissue (VAT); Colorectal cancer (CRC)

Received: 18 March 2022; Accepted: 28 May 2022
*Correspondence to: Dena Helene Alavi, Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway. Email: d.h.alavi@medisin.uio.no

Introduction

Computed tomography (CT) examinations are crucial for the diagnosis and follow-up of patients with colorectal cancer (CRC). In addition, the routine CT scans contain valuable and high-precision information about body composition such as skeletal muscle (SM), visceral adipose tissue (VAT), subcutaneous adipose tissue (SAT), and intermuscular and intramuscular adipose tissue (IMAT). Because low SM and high IMAT are predictors of reduced survival in CRC patients,1,2
body composition data that may be collected from standard CT scans should be used in clinical practice. High amounts of VAT may pose a larger risk for CRC patients than total fat mass, and VAT is also negatively associated with survival in CRC patients. IMAT itself is associated with insulin resistance and loss of muscle strength and function. Although body composition data from clinically acquired CT scans may be used to tailor nutritional interventions and optimize cancer treatment, it is currently not part of standard cancer treatment.

The lack of utilization of CT-based body composition data may be due to the absence of accurate and automated tools. Even though CT is described as one of the golden standards for the measurement of body composition, segmentation has typically been performed manually or semi-manually. Both manual and semi-manual processes are time consuming and requires extensive resources, anatomical knowledge, and software training. This limits the use of CT for body composition purposes in clinical practice and large-scale clinical trials.

The ongoing rapid development of deep learning has revolutionized the field of automated image segmentation, including quantitative CT-based body composition assessment. In particular, convolutional neural networks (CNNs), directly inspired by brain neurons, are well fitted to process large amounts of imaging data. Several studies have shown promising results using deep learning-based methods for quantification of abdominal SM or adipose tissues. As CNNs require large amounts of data, it is common to use multiple available data sets for training purposes. However, no software is trained and tested on CT slices from different anatomical levels acquired from CRC patients. Previous studies typically rely on single slices at L3. This may limit the usefulness of these models as clinically acquired CT scans differ according to the protocol used or the region of interest.

Therefore, the main aim of this study was to develop and test the performance of an automatic software named BodySegAI for the quantification of abdominal SM, IMAT, VAT, and SAT in CT scans from CRC patients. To overcome the limitation of using a single slice at L3, we aimed to train the model using multiple CT slices, from both the abdominal and the pelvic area (L2 to S1). Additionally, we compared the performance of BodySegAI against a similar software named AutoMATICA. Furthermore, we also aimed to investigate the time effectiveness of BodySegAI compared with semi-manual segmentation by human readers.

Methods

Network set-up and training

We used a two-dimensional U-Net adapted in a similar fashion as by Weston et al. with six down-sampling and up-sampling steps. Each down-sampling step had two convolutions (3 × 3), a hyperbolic-tangent (tanh) activation function, followed by max pooling (2 × 2). Each up-sampling step had interpolation (2 × 2) and two convolutions (3 × 3) with tanh activation. In each step, the convolutions output from the down-sampling side was concatenated with the corresponding up-sampling step. The last layer was a convolution (1 × 1 × 3) followed by a sigmoid function. The first and last steps had 16 kernels, and the bottleneck layer of the U-net had 16 × 2^16 = 1024 kernels, resulting in an 8 × 8 × 1024 representation. The model had 31 109 747 trainable weights. The highest score for each pixel above a threshold of 0.5 determined the final segmentation. The input image size was one channel of size 512 × 512, and the output was three channels of the same size. VAT and SAT were classified in separate output channels. The third output channel represented a combined class of IMAT and SM, which was then split into separate IMAT and SM masks by thresholding according to the Alberta protocol (described in the ‘image segmentation’ paragraph). All pixels that had corresponding raw image value of −30 HU or less were determined to be IMAT. Any segmented pixels that were outside the predefined threshold values for the four segmentation labels were discarded. Hyper-parameters and pre-processing and post-processing steps were selected experimentally using 300 validation CT slices from the training data set. The model was later trained using all training and validation data. Input images were clipped to −400–600 HU and normalized by subtraction of the mean and division by the standard deviation of the training data set (values −536 and 493, respectively). We used the DICE loss function, a batch size of 16, and Adam optimization with a descending learning rate of 0.0001 down to 0.000001. The model was trained for 250 epochs. During training, there were simple linear augmentations used with translation by up to 50 pixels in any direction, up to 5% rotation, 10% scale, and 40% right/left flipping. The programming environment was Python v. 3.6.8, Keras v. 2.2.0 with TensorFlow v. 1.12.0, all open source. The model was trained using NVIDIA RTX 2070 GPU.

Training data set

A flow chart of the CT data used for model training, validation and testing is shown in Figure 1. BodySegAI was trained on a total of 2989 CT slices from 286 subjects recruited from three independent studies (Table 1). The bulk part of the training data set contained pre-operative and post-operative scans of CRC patients in the randomized controlled trial CRC-NORDIET. From 153 CRC patients, we acquired mid-vertebral single CT slices from L2 to S1, in addition to volumes (20% of the distance from the iliac crest towards the lowest point of the mandible, corresponding to the abdominal volume defined for dual-energy X-ray absorptiometry.
(DXA) by the International Society for Clinical Densiometry \textsuperscript{29}). From two previous studies, we also acquired single slices from 94 healthy\textsuperscript{30} and 39 diabetic subjects.\textsuperscript{31} A portion of the training data set was used to validate the model (number of slices = 300) and was later reinserted into the training data set for training purposes (Figure 1). All CT examinations were performed between 1991 and 2020. Thus, BodySegAI was trained on multi-centre CT slices from several various scanners and protocols with different contrast phases, slice thicknesses, tube currents, and voltages.

**Test data set and human ground truth**

An independent test data set consisting of 154 post-operative CT slices from 32 CRC patients from the CRC-NORDIET study\textsuperscript{28} was created to test the performance of BodySegAI (Table 2). The median age and BMI in the test data set were 66 years and 27.6 kg/m\(^2\). Fifty-three per cent of the population were men (Table 2). The slice thickness was 3 mm for all the test data. The test data set was designated for testing exclusively and had never been introduced to BodySegAI before testing. Performance of BodySegAI was tested against human ground truth and compared with AutoMATICA: a CNN-based open-source software for automatic segmentation at level L3.

| Training data set n = 286, number of slices = 2989 | Colorectal cancer subjects (Henriksen et al.) | Diabetic subjects (Wium et al.) | Healthy subjects (Skarn et al.) |
|---|---|---|---|
| Subjects, n | 153 | 39 | 94 |
| Slices | 2,654 | 176 | 159 |
| Age (y) | 64 (60, 70) | 42 (37, 43) | 40.4 (36.2, 43.4) |
| Males n (%) | 84 (48) | 19 (49) | 94 (100) |
| Weight (kg) | 75.0 (69, 89) | 97.3 (86.2, 108.7) | 85.9 (78.7, 94.9) |
| BMI (kg/m\(^2\)) | 25.9 (22.6, 29.0) | 35.6 (29.8, 38.4) | 25.8 (23.6, 27.8) |
| CT peak tube voltage (kVp) n (%) | | | |
| 90 | 36 (1) | | |
| 100 | 368 (14) | | |
| 110 | 73 (3) | | |
| 120 | 2,177 (82) | 176 (100) | 159 (100) |
| Slice thickness (mm) | | | |
| 3 | | 4 | 5 |
| Use of intravenous contrast agent | | | |
| Both | No | No |

Values are median (25 and 75th percentiles) unless otherwise specified. CT, computed tomography.
AutoMATiCA was trained on L3 slices from renal and liver donors, patients that were critically ill or had liver cirrhosis, pancreatic cancer, or clear cell renal carcinoma. The human ground truth was created as a reference to the test data set. We used ‘Simultaneous truth and performance level estimation’ (STAPLE)—an expert-maximization algorithm that provides a probabilistic estimation of the true segmentation. The STAPLE-based human ground truth was established by combining the segmentations performed by three experienced human readers A-MF (radiographer), PML (radiologist), and DHA (registered dietitian) hereafter referred to as Reader 1, 2, and 3, respectively. Each of the three human readers segmented 154 CT slices included in the test data set in sliceOmatic.

**Image segmentation**

SM, IMAT, VAT, and SAT was segmented using sliceOmatic Version 5.0 Revision 7 (TomoVision, Montreal, Canada). All segmentation was performed in accordance with the Alberta Protocol as defined and used in body composition segmentation studies performed at the Alberta hospital. Segmentation of the training data set was performed by four radiographers, two registered dieticians, one oncologist, and one radiologist. For the segmentation tool using thresholding the following values were used: SM: –29 to 150 Hounsfield Units (HU), IMAT: –190 to –30 HU, VAT: –150 to –50 HU, and SAT: –190 to –30. In case of beam hardening artefacts caused by metal implants or colonography contrast media, the thresholding was turned off to enable correct segmentation. We defined the lowest vertebra above the sacrum as L5.

**Statistical methods**

Statistical analyses were performed in SPSS (IMB SPSS Statistics 27). The performance of BodySegAI was tested against the human ground truth. Dice score coefficients, Hausdorff distances, sensitivity, and specificity were calculated. The Dice score calculates the overlap between two volumes, defined as

\[
\text{Dice} = \frac{2 \times \text{True positive}}{2 \times \text{True positive} + \text{False positive} + \text{False negative} \text{pixels}}
\]

Hausdorff distance is the maximum of distances in mm between points in two point sets, calculated by the following formula: Hausdorff distance\((A,B) = \max(h(A,B), h(B,A))\). Main results were given in median and 25 and 75th percentiles as data were non-normally distributed, unless otherwise specified. The median absolute error (%) was calculated as the median of the absolute difference in percent between BodySegAI and human ground truth, divided by human ground truth. Bland–Altman plots were used to quantify the mean differences per slice between segmentation methods. The y-axis is defined as the difference between methods (BodySegAI – human ground truth), and the x-axis represents the mean values (BodySegAI + human ground truth)/2. Upper and lower limits of agreement were defined as mean difference ± 1.96*standard deviation (SD) (corresponds to 95% confidence intervals). BodySegAI was additionally compared against AutoMATiCA. The performance was investigated both for single slices alone and all the slices combined. Time effectiveness was presented as the average time (in seconds) with SD for segmentation of CT scans selected by chance by Readers 1, 2, and 3 against BodySegAI. The difference was tested using the independent samples t-test. A computer with an Intel® Core™ i5-8500 processor, Intel UHD Graphics 630,16GB of RAM, and Windows 10 operative system was used.

**Results**

**BodySegAI compared with human ground truth**

The median Dice score for BodySegAI compared with the human ground truth was 0.969 for SM, 0.814 for IMAT, 0.986 for VAT, and 0.990 for SAT. Dice scores, Hausdorff distances, sensitivity, and specificity of BodySegAI compared with human ground truth using the test data set are shown in Table 3.

| Segment | Dice Score | Mean Difference | 25th Percentile | 75th Percentile |
|---------|------------|-----------------|----------------|----------------|
| SM      | 0.969      | 1.35            | 0.50           | 2.53           |
| IMAT    | 0.814      | 10.54           | 5.61           | 19.79          |
| VAT     | 0.986      | 0.91            | 0.41           | 1.91           |
| SAT     | 0.990      | 0.91            | 0.41           | 1.91           |

The median volumes per slice for SM, IMAT, VAT, and SAT were 40.67 (25 and 75th percentiles: 32.98, 50.13), 3.47 (2.24, 5.80), 33.72 (19.35, 53.08), and 60.65 cm³ (45.13, 82.33) for SM, IMAT, VAT, and SAT by BodySegAI, respectively. The median absolute errors (%) between BodySegAI and human ground truth for SM, IMAT, VAT, and SAT were 1.35 (0.50, 2.53), 10.54 (5.61, 19.79), 0.91 (0.41, 1.91), and 1.07% (0.47, 1.81), respectively.

Bland–Altman plots comparing the segmentation volume between BodySegAI and human ground truth per slice are shown in Figure 2. The mean differences between BodySegAI and human ground truth were –0.09 (limits of agreement: –2.37–2.19) for SM, –0.17 (–1.69–1.36) for IMAT, –0.12 (–1.34–1.09) for VAT, and 0.67 cm³ (–1.03–2.36) for SAT. All Bland–Altman plots showed an even distribution of the differences above and below the mean difference.

The performance of BodySegAI against human ground truth for different anatomical slice locations is shown in Table 4. Although L2, L3, and L4 demonstrated similar Dice scores, L3 presented the overall highest median Dice, and S1 demonstrated the lowest.

The Dice score of each human reader vs. the STAPLE based human ground truth for L2 to S1 is presented in Table 5. Figure 3 shows sample cases of the performance and typical errors in segmentation by BodySegAI.

DOI: 10.1002/crt2.53
Figure 2  Bland–Altman plots showing differences between BodySegAI and human ground truth for (A) skeletal muscle, (B) intermuscular and intramuscular adipose tissue, (C) visceral adipose tissue, and (D) subcutaneous adipose tissue in slices L2 to S1. The solid line represents the mean difference and dotted line the upper and lower limits of agreement.

Table 3  The segmentation performance of BodySegAI against human ground truth and AutoMATICA against human ground truth for multiple slices in L2 to S1 using the test data set, number of slices = 154

|                  | Dice score | Hausdorff distance mm | Sensitivity | Specificity |
|------------------|------------|------------------------|-------------|-------------|
| BodySegAI vs. human ground truth, number of slices = 154 |            |                        |             |             |
| SM               | 0.969 (0.957,0.979) | 14.177 (10.050,21.006) | 0.969 (0.956,0.982) | 0.998 (0.996,0.998) |
| IMAT             | 0.814 (0.740,0.880) | 56.974 (38.432,84.305) | 0.814 (0.729,0.886) | 0.999 (0.998,0.999) |
| VAT              | 0.986 (0.977,0.991) | 25.040 (17.951,30.761) | 0.986 (0.976,0.992) | 0.999 (0.999,1.000) |
| SAT              | 0.990 (0.985,0.994) | 5.657 (4.123,9.849)    | 0.996 (0.993,0.998) | 0.998 (0.997,0.999) |
| AutoMATICA vs. human ground truth, number of slices = 154 |            |                        |             |             |
| SM               | 0.933 (0.812,0.961) | 32.795 (19.177,56.813) | 0.980 (0.769,0.996) | 0.992 (0.989,0.999) |
| IMAT             | 0.772 (0.655,0.843) | 57.511 (43.635,80.120) | 0.810 (0.696,0.876) | 0.998 (0.997,0.999) |
| VAT              | 0.976 (0.962,0.987) | 38.942 (27.152,54.481) | 0.992 (0.980,0.996) | 0.998 (0.997,0.999) |
| SAT              | 0.985 (0.974,0.991) | 7.211 (4.123,15.549)   | 0.997 (0.994,0.998) | 0.996 (0.993,0.998) |

Values are median (25 and 75th percentiles).

IMAT, intermuscular and intramuscular adipose tissue; SAT, subcutaneous adipose tissue; SM, skeletal muscle; VAT, visceral adipose tissue.

Table 4  The segmentation performance of BodySegAI against human ground truth for different anatomical slice locations

| Anatomical slice level | Dice score | Dice score | Dice score | Dice score | Dice score | Dice score |
|------------------------|------------|------------|------------|------------|------------|------------|
|                        | L2         | L3         | L4         | L5         | S1         |
| Number of slices       | 32         | 31         | 31         | 30         | 30         |
| SM                     | 0.972 (0.964,0.980) | 0.977 (0.968,0.983) | 0.975 (0.967,0.980) | 0.952 (0.937,0.967) | 0.964 (0.956,0.973) |
| IMAT                   | 0.876 (0.842,0.906) | 0.857 (0.792,0.893) | 0.833 (0.753,0.884) | 0.791 (0.746,0.837) | 0.714 (0.646,0.773) |
| VAT                    | 0.985 (0.978,0.990) | 0.987 (0.983,0.992) | 0.987 (0.977,0.991) | 0.986 (0.975,0.992) | 0.984 (0.972,0.989) |
| SAT                    | 0.990 (0.984,0.994) | 0.992 (0.988,0.994) | 0.992 (0.987,0.994) | 0.991 (0.986–0.995) | 0.985 (0.981,0.992) |

Values are median (25 and 75th percentiles).

IMAT, intermuscular and intramuscular adipose tissue; SAT, subcutaneous adipose tissue; SM, skeletal muscle; VAT, visceral adipose tissue.
Table 5  Dice score for each human reader vs. human ground truth for L2 to S1

| Number of slices | Reader 1 vs. human ground truth | Reader 2 vs. human ground truth | Reader 3 vs. human ground truth |
|------------------|---------------------------------|---------------------------------|---------------------------------|
| SM               | 0.989 (0.985,0.993)              | 0.989 (0.984,0.992)             | 0.985 (0.977,0.989)             |
| IMAT             | 0.933 (0.893,0.954)              | 0.930 (0.9,0.949)               | 0.889 (0.842,0.925)             |
| VAT              | 0.993 (0.988,0.996)              | 0.994 (0.990,0.997)             | 0.990 (0.985,0.993)             |
| SAT              | 0.995 (0.993,0.997)              | 0.996 (0.991,0.997)             | 0.994 (0.990,0.996)             |

Values are median (25 and 75th percentiles).

IMAT, intermuscular and intramuscular adipose tissue; SAT, subcutaneous adipose tissue; SM, skeletal muscle; VAT, visceral adipose tissue.

Figure 3  Sample cases demonstrating the segmentation performance by BodySegAI, by human ground truth and the unsegmented CT slice. Dice score, the difference between BodySegAI and human ground truth and the mean total body compartments is shown. The CT slices with the green arrows show (A) particularly good performance by BodySegAI, (B) a mistake in the segmentation of musculus transversus abdominis by BodySegAI, (C) error in the segmentation of patient with ostomy by BodySegAI, and (D) error in the segmentation of VAT by BodySegAI. Skeletal muscle is segmented in red, intermuscular and intramuscular adipose tissue in green, visceral adipose tissue in yellow, and subcutaneous adipose tissue in blue. Mean difference represents BodySegAI minus human ground truth, and total is the segmented volume by human ground truth.
Body composition from CT scans using artificial intelligence

**BodySegAI compared with AutoMATiCA**

When testing our test data set against human ground truth, AutoMATiCA showed lower Dice scores than BodySegAI for all body compartments in L2 to S1 (Table 3). AutoMATiCA also showed higher Hausdorff distances compared with BodySegAI in SM, VAT, and SAT. Median Dice scores by AutoMATiCA for the L3 slice only was 0.962 for SM, 0.843 for IMAT, 0.986 for VAT, and 0.986 for SAT (slices tested = 31).

**Time effectiveness of BodySegAI**

The average time used for automatic segmentation of one slice by BodySegAI was 4.9 seconds (SD: 0.7, number of slices = 20). Semi-manual segmentation by Reader 1, 2, and 3 required on average 723.5 s (SD: 232.3, number of slices = 99) and BodySegAI segmented on average 148 times faster ($P < 0.001$) than human readers.

**Discussion**

The deep learning-based segmentation tool BodySegAI showed excellent segmentation performance for SM, VAT, and SAT (Dice scores: >0.969) and good results for IMAT (Dice score: 0.814) in multiple abdominal CT slices in CRC patients. These results are in line with other deep learning models tested on cancer patients, only using a single slice at L3 or L4.14,16,20,23,24 In the present study, BodySegAI demonstrated higher Dice scores than AutoMATiCA when using our test data set. However, in the original article by Paris et al., AutoMATiCA outperforms BodySegAI’s Dice scores in the present study for SM and IMAT, but not for VAT and SAT.17 This suggests that the CT test data used in our study had different characteristics from the data used to train AutoMATiCA. Lack of generalization is a well-known challenge in deep learning. This may be improved by training the model on heterogeneous data spanning the full range of image characteristics in clinical use.11 BodySegAI was trained on data from several institutions using different CT scanners and protocols, introducing variation itself. Additionally, the CT slices in the training data set originated from three independent groups of subjects with heterogeneous characteristics (i.e., CRC, healthy, and diabetic). Such heterogeneity in imaging data and study population may facilitate a more robust model and better generalizability.

**Multi-slice modelling**

A single slice at L3 is frequently used in research settings because of its high association with whole-body measurements36,37 and abdominal volume.38 We have previously shown that although a single slice at L3 is highly associated with abdominal VAT volume, the use of multiple slices (L2, L3, and L4) increased the explained variance against VAT volume.38 However, BodySegAI was trained and tested on multiple slices at different anatomical levels (L2 to S1). This may have increased the model’s robustness to anatomical changes after abdominal surgery and enables segmentation of data from other anatomical levels in cases where the L3 slice is unavailable or contains artefacts. A few models have targeted a multi-slice approach, and BodySegAI demonstrates similar or higher Dice scores than these.13,18,22,25,39

**Limitations**

Although our model is trained on multi-centre data, it is tested on a sample of post-operative CT scans from CRC patients. BodySegAI should be tested on other patient groups and settings to assess its external validity. By visual inspection of the segmented CT slices, we observed that BodySegAI under-performs in CT slices that have abnormalities, such as oedema, ostomy, or artefacts. The training set consisted mostly of lumbar slices, which may explain the lower performance at S1 level. This is important to keep in mind if the model is to be utilized at a sacral level. BodySegAI was trained on single slice thicknesses (3–5 mm), but it was only tested on CT slices with 3 mm thickness. Thus, the effect of slice thickness on segmentation performance is unknown.

Due to the comprehensive task of semi-manual segmentation for establishing the human ground truth and lack of standardized regions of interest, BodySegAI was only tested on single CT slices from predefined anatomical regions (L2 to S1), not entire abdominal volumes. However, the model can readily be used for volumetric data, which is proposed to further increase the accuracy compared with the usage of single slices only.38 Although we consider BodySegAI as an automatic software, the axial CT slices still have to be manually extracted from the CT examination and uploaded into BodySegAI in order for the body composition analysis to be conducted.

**Strengths**

To our knowledge, this is the first software developed to automatically segment SM, IMAT, SAT, and VAT in multiple CT
slices from L2 to S1 in CRC cancer patients staged TNM I–III. Due to the time-consuming task of semi-manual segmentation by humans, most previous studies have only used one or two readers to establish the ground truth or reference method.\textsuperscript{17} Although the main variation in segmentation results is due to subject variation and not inter-rater segmentation variability,\textsuperscript{40} a single readers subjective performance may influence the reference standard or ground truth. By using STAPLE, we reduced this vulnerability by providing a ground truth representing the statistically optimized combination of segmentations from several human readers.\textsuperscript{32} Our results also showed that each of the three readers contributed almost equally to the final STAPLE-based human ground truth (Table 5).

**Clinical implications**

There is increasing evidence of the importance of body composition on health outcomes in cancer patients; nevertheless, there is a lack of tools for clinical implementation. Other methods for the assessment of body composition include DXA and bioelectrical impedance analysis. However, these methods are often unavailable at the hospitals and less precise and require additional resources to perform compared with using the existing CT scans.\textsuperscript{10} In clinical practice, weight loss in kilograms or BMI is therefore often used to identify conditions such as malnutrition, sarcopenia, or cachexia, despite not being able to distinguish between fat and muscles.

In comparison with the existing and routinely used methods to estimate body composition or weight, we consider the general uncertainty of BodySegAI to be low. This can be demonstrated by the narrow limits of agreement against human ground truth in the Bland–Altman plots. Median absolute errors (%), which were below 2% for total SM, VAT, and SAT and 10.5% for IMAT also support the high accuracy of BodySegAI. Consequently, we consider the performance of BodySegAI to be acceptable for research purposes and to exceed the methods for body composition evaluation in clinical practice today.

A number of patients perform CT scans as routine care, and we suggest integrating BodySegAI into the clinical workflow. Body composition data could be used to track changes and risk stratify patients, enabling the diagnosis and treatment of malnutrition, sarcopenia, and cachexia. It may also facilitate large-scale studies to increasingly utilize CT-based body composition data and then generate larger sets of normative data to use as reference curves in the clinic.

**Conclusion**

BodySegAI generates excellent and detailed segmentations of SM, VAT, and SAT and good segmentations of IMAT, 148 times faster than human readers. BodySegAI may replace the semi-manual segmentation presently used in research after validation on local data. We suggest that BodySegAI may be integrated into a future clinical workflow for patients that perform CT as standard care.

**Acknowledgements**

We would like to thank Manuela Zucknick and the statistical advising service by the Oslo Centre for Biostatistics and Epidemiology of the University of Oslo and Oslo University Hospital for statistical advise. A special thanks to Heidi Beate Eggesebø and her team for fantastic encouragement and the great collaboration that enabled this comprehensive project. We would also thank Atle Børnerud for the technical support and correction of the final manuscript. The data used in this study was approved by Regional Committee for Medical and Health Research Ethics and conformed to the Helsinki Declaration. Informed written consents were obtained from all patients.\textsuperscript{28,30,31} All authors comply with the Ethical guidelines for authorship and publishing in the Journal of Cachexia, Sarcopenia and Muscle Clinical Reports.\textsuperscript{43}

**Funding**

Funding was received from the Throne Holst Nutrition Research Foundation, Norwegian Cancer Society, and the South Eastern Norway Regional Health Authority.

**Conflict of interest**

RB is a shareholder of AS Vitas. TS is a shareholder of Artificial Intelligence AS. DHA, HBH, BB, A-MF, and PML declare no conflict of interests.
References

1. Hopkins JJ, Reif RL, Bigam DL, Baracos VE, Eurfich DT, Sawyer MB. The impact of muscle and adipose tissue on long-term survival in patients with stage I to III colorectal cancer. Dis Colon Rectum. 2019;62:549–560.

2. Sueda T, Takahasi H, Nishimura J, Hata T, Matsuda C, Mizushima T, et al. Impact of low muscularity and myosteatosis on long-term outcome after curative colorectal cancer surgery: A propensity score-matched analysis. Dis Colon Rectum. 2018;61:364–374.

3. Keum N, Lee DH, Kim R, Greenwood DC, Malietzis G, Aziz O, Bagnall NM, Johns N, Kostandy P, Sakinis T, et al. Automated convolutional networks for biomedical image segmentation. Cham: Springer International Publishing; 2015. p 234–241.

4. Xiao J, Mazurak VC, Olobatuyi TA, Caan BJ, Prado CM. Visceral adiposity and cancer survival: A review of imaging studies. Eur J Cancer Care. 2018;27:e12611. https://doi.org/10.1111/ecc.12611.

5. Addison O, Marcus RL, Lastayo PC, Ryan AS. Intermuscular fat: A review of the consequences and causes. Int J Endocrinol. 2014;2014:309570. https://doi.org/10.1155/2014/309570.

6. Waters DL. Intermuscular adipose tissue: A brief review of etiology, association with physical function and weight loss in older adults. Ann Geriatr Med Res. 2019;23:3–8.

7. Thibault R, Pichard C. The evaluation of body composition: A useful tool for clinical practice. Ann Nutr Metab. 2012;60:6–16.

8. Kvaerner A, Harnaes H, Alavi DH, Bærebring L, Henriksen HB, Guren MG, et al. Should calculation of chemotherapy dosage for bowel cancer be based on body composition? Tidsskr Nor Laegeforen. 2020;140. https://doi.org/10.4045/tidsskr.19.0769.

9. Malletzis G, Aziz O, Bagnall NM, Johns N, Fearon KC, Jenkins JT. The role of body composition evaluation by computerized tomography in determining colorectal cancer treatment outcomes: A systematic review. Eur J Surg Oncol. 2015;41:186–196.

10. Yip C, Dinkel C, Mahajan A, Siddique M, Cook GJ, Goh V. Imaging body composition in cancer patients: Visceral obesity, sarcopenia and sarcopenic obesity may impact on clinical outcome. Insights Imaging. 2015;6:489–497.

11. Wang B, Torriani M. Artificial intelligence in the evaluation of body composition. Semin Musculoskelet Radiol. 2020;24:30–37.

12. Chartrand G, Cheng PM, Vorontsova E, Drozdwal M, Turcotte S, Pal CJ, et al. Deep learning: A primer for radiologists. Radiographics. 2017;37:2113–2131.

13. Edwards K, Chhabra A, Dormer J, Jones P, Boutin RD, Lenchik L, et al. Abdominal muscle segmentation from CT using a convolutional neural network. Proc SPIE Int Soc Opt Eng. 2020;11317. https://doi.org/10.1117/12.2549406.

14. Weston AD, Korfias P, Kline TL, Philbrick KA, Kostandy P, Sakinis T, et al. Automated abdominal segmentation of CT scans for body composition analysis using deep learning. Radiology. 2018;290:669–679.

15. Yan 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.

16. Lee H, Troschel FM, Tajmir S, Fuchs G, Mario J, Fintelmann FJ, et al. Pixel-level deep segmentation: Artificial intelligence quantifies muscle on computed tomography for body morphometric analysis. J Digit Imaging. 2017;30:497–498.

17. Paris MT, Tandon P, Heyland DK, Furberg H, Premji T, Low G, et al. Automated body composition analysis of clinically acquired computed tomography scans using neural networks. Clin Nutr. 2020;39:3049–3054.

18. Koitka S, Kroll L, Malamutmann E, Oezcelik A, Nensa F. Fully automated body composition analysis in routine CT imaging using 3D semantic segmentation convolutional neural networks. Eur Radiol. 2020;31:1795–1804.

19. Nowak S, Faron A, Luukenks JA, Geiilfier HL, Praktinjko M, Block W, et al. Fully automated segmentation of connective tissue compartments for CT-based body composition analysis: A deep learning approach. Invest Radiol. 2020;55:357–366.

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

21. Magudia K, Bridge CP, Bay CP, Babic A, Fintelmann FJ, Troschel FM, et al. Population-scale CT-based body composition analysis of a large outpatient population using deep learning to derive age-, sex-, and race-specific reference curves. Radiol. 2020;298:319–329.

22. Burns JE, Yao J, Chalhoub D, Chen JJ, Summers RM. A machine learning algorithm to estimate sarcopenia on abdominal CT. Acad Radiol. 2020;27:311–320.

23. Amarasinghe KC, Lopes J, Beraldo J, Kiss N, Bucknell N, Everitt S, et al. A deep learning model to automate skeletal muscle area measurement on computed tomography images. Front Oncol. 2021;11:580806. https://doi.org/10.3389/fonc.2021.580806.

24. Ackermans M, Volmer L, Wee L, Brecheneis R, Sánchez-González P, Seiffert AP, et al. Deep learning automated segmentation for muscle and adipose tissue from abdominal computed tomography in polytrauma patients. Sensors (Basel). 2021;21:2083. https://doi.org/10.3390/s21022083.

25. Borrelli P, Kabotche R, Enquist O, Ulén J, Trägårdh E, Kjölhede H, et al. Artificial intelligence-aided CT segmentation for body composition analysis: A validation study. Eur Radiol Exp. 2021;5:11. https://doi.org/10.1186/s41747-021-00210-8.

26. Hsu TH, Schawkat K, Berkowitz SJ, Wei JL, Makoyeva A, Legare K, et al. Artificial intelligence to assess body composition on routine abdominal CT scans and predict mortality in pancreatic cancer. A recipe for your local application. Eur J Radiol. 2021;142:109834. https://doi.org/10.1016/j.ejrad.2021.109834.

27. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Cham: Springer International Publishing; 2015. p 234–241.

28. Henriksen HB, Raeder H, Bohn SK, Paar I, Kvaerner AS, Billinga SJ, et al. The Norwegian dietary guidelines and colorectal cancer survival (CRC-NORDIET) study: A food-based multicentre randomized controlled trial. Br J Cancer. 2017;17:83. https://doi.org/10.1038/s41388-017-0372-0.

29. Hangartner TN, Warner S, Braillon P, Jankowski L, Shepherd J. The official positions of the British Thoracic Society for body composition analysis using deep learning to derive age-, sex-, and race-specific reference curves. Clin Radiol. 2017;72:530–536.

30. Skarn SN, Eggesbo HB, Flaa A, Kjeldsen SE, Rostrup M, Brunborg C, et al. Predictors of abdominal adipose tissue compartments: 18-year follow-up of young men with and without family history of diabetes. Eur J Intern Med. 2016;29:26–31.

31. Wium C, Eggesbo HB, Ueland T, Michelsen AE, Torjesen PA, Aukrust P, et al. Adipose tissue distribution in relation to insulin sensitivity and inflammation in Pakistani and Norwegian subjects with type 2 diabetes. Scand J Clin Lab Invest. 2014;74:700–707.

32. 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;23:903–921.

33. Prado CM, Baracos VE, McCargar LJ, Mauritakis M, Mulder KE, Reiman T, et al. Body composition as an independent determinant of 5-fluorouracil-based chemotherapy toxicity. Clin Cancer Res. 2007;13:3264–3268.

34. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15:29. https://doi.org/10.1186/s12895-015-0068-3.

35. Giavarina D Understanding Bland Altman analysis. Biochem Med. 2015;25:141–151.

36. Mauritakis 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. Appl Physiol Nutr Metab. 2008;33:997–1006.
37. Faron A, Luetkens JA, Schmeel FC, Kuetting DLR, Thomas D, Sprinkart AM. Quantification of fat and skeletal muscle tissue at abdominal computed tomography: Associations between single-slice measurements and total compartment volumes. *Abdom Radiol*. 2019;44:1907–1916.

38. Alavi DH, Henriksen HB, Lauritzen PM, Kværner AS, Sakinis T, Langleite TM, et al. Quantification of adipose tissues by dual-energy X-ray absorptiometry and computed tomography in colorectal cancer patients. *Clin Nutr ESPEN*. 2021;43:360–368.

39. Weston AD, Korfias P, Philbrick KA, Conte GM, Kostandy P, Sakinis T, et al. Complete abdomen and pelvis segmentation using U-net variant architecture. *Med Phys*. 2020;47:5609–5618.

40. Kjønigsen LJ, Harneshaug M, Fløtten A-M, Karterud LK, Petterson K, Skjolde G, et al. Reproducibility of semiautomated body composition segmentation of abdominal computed tomography: A multiobserver study. *Eur Radiol Exp*. 2019;3:42. https://doi.org/10.1186/s41747-019-0122-5

41. 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;10:1143–1145.