

External validation of a deep learning model for automatic segmentation of skeletal muscle and adipose tissue on abdominal CT images

David P.J. van Dijk , MD, PhD^{*,†,1,2,3}, Leroy F. Volmer, MSc^{†,4,5}, Ralph Brecheisen, PhD^{1,2}, Bibi Martens, MD, PhD^{5,6}, Ross D. Dolan, MD, PhD⁷, Adam S. Bryce, MD^{8,9}, David K. Chang, MD, PhD^{8,9}, Donald C. McMillan, PhD⁷, Jan H.M.B. Stoot, MD, PhD³, Malcolm A. West, MD, PhD¹⁰, Sander S. Rensen, PhD^{1,2}, Andre Dekker, PhD^{4,5}, Leonard Wee , PhD^{‡,4,5}, Steven W.M. Olde Damink, MD, PhD^{1,2,11,‡} and the Body Composition Collaborative[§]

¹Department of Surgery, Maastricht University Medical Center, Maastricht, 6200 MD, The Netherlands

²NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, 6229 ER, The Netherlands

³Department of Surgery, Zuyderland Medical Centre, Geleen, 6162 BG, The Netherlands

⁴Department of Radiotherapy (MAASTRO), Maastricht University, Maastricht, Maastricht, 6229 ET, The Netherlands

⁵GROW School for Oncology and Reproduction, Maastricht University, Maastricht, 6229 ER, The Netherlands

⁶Department of Radiology, Maastricht University Medical Center, Maastricht, Maastricht, 6200 MD, The Netherlands

⁷Academic Unit of Surgery, School of Medicine, University of Glasgow, Glasgow Royal Infirmary, Glasgow, Glasgow, G31 2ER, United Kingdom

⁸Wolfson Wohl Cancer Research Centre, School of Cancer Sciences, University of Glasgow, Glasgow, Glasgow, G61 1BD, United Kingdom

⁹West of Scotland Pancreatic Unit, Glasgow Royal Infirmary, Glasgow, Glasgow, G12 0YN, United Kingdom

¹⁰Academic Unit of Cancer Sciences, Faculty of Medicine, University of Southampton, Southampton, Southampton, SO16 6YD, United Kingdom

¹¹Department of General, Visceral and Transplant Surgery, University Hospital Aachen, Aachen, 52074, Germany

*Corresponding author: David P.J. van Dijk, MD, PhD, Department of Surgery, Maastricht University Medical Center, P.O. Box 616, Maastricht, 6200 MD, The Netherlands (d.vandijk@maastrichtuniversity.nl)

†D.P.J. van Dijk and L.F. Volmer contributed equally to this work.

‡L. Wee and S.W.M. Olde Damink contributed equally to this work.

§The collaborators of the Body Composition Collaborative are listed in the Acknowledgments section.

Abstract

Objectives: Body composition assessment using CT images at the L3-level is increasingly applied in cancer research and has been shown to be strongly associated with long-term survival. Robust high-throughput automated segmentation is key to assess large patient cohorts and to support implementation of body composition analysis into routine clinical practice. We trained and externally validated a deep learning neural network (DLNN) to automatically segment L3-CT images.

Methods: Expert-drawn segmentations of visceral and subcutaneous adipose tissue (VAT/SAT) and skeletal muscle (SM) of L3-CT-images of 3187 patients undergoing abdominal surgery were used to train a DLNN. The external validation cohort was comprised of 2535 patients with abdominal cancer. DLNN performance was evaluated with (geometric) dice similarity (DS) and Lin's concordance correlation coefficient.

Results: There was a strong concordance between automatic and manual segmentations with median DS for SM, VAT, and SAT of 0.97 (IQR: 0.95–0.98), 0.98 (IQR: 0.95–0.98), and 0.95 (IQR: 0.92–0.97), respectively. Concordance correlations were excellent: SM 0.964 (0.959–0.968), VAT 0.998 (0.998–0.998), and SAT 0.992 (0.991–0.993). Bland-Altman metrics indicated only small and clinically insignificant systematic offsets; SM radiodensity: 0.23 Hounsfield units (0.5%), SM: 1.26 cm².m⁻² (2.8%), VAT: −1.02 cm².m⁻² (1.7%), and SAT: 3.24 cm².m⁻² (4.6%).

Conclusion: A robustly-performing and independently externally validated DLNN for automated body composition analysis was developed.

Advances in knowledge: This DLNN was successfully trained and externally validated on several large patient cohorts. The trained algorithm could facilitate large-scale population studies and implementation of body composition analysis into clinical practice.

Keywords: body composition; deep learning; convolutional neural networks; image segmentation; CT.

Introduction

Body composition assessment using routine abdominal CT images is increasingly applied in clinical and translational research.¹ By measuring the tissue area at the level of the third

lumbar vertebra (L3) and scaling for subject height, precise assessments of total body mass of skeletal muscle (SM), visceral adipose tissue (VAT), and subcutaneous adipose tissue (SAT) can be made.² Body composition has been found to be

Received: 16 April 2024; Revised: 6 September 2024; Accepted: 11 September 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the British Institute of Radiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

highly independently predictive of survival, especially among cancer patients. In particular, low SM mass (ie, sarcopenia), low adipose tissue mass, and decreased SM radiodensity (ie, myosteatosis) have been shown to be associated with shorter overall survival in various cancer types.³⁻⁵

Body composition exhibits substantial heterogeneity among people due to variations in age, sex, race, build, and lifestyle.⁶ These intrinsic inter-personal differences are unrelated to disease and may therefore obscure disease-related body composition effects, necessitating large population-based data cohorts to adjust for them.

Manual segmentation of body compartments on L3 CT images is time-consuming and requires significant expertise.⁷ Therefore, robust high-throughput automated segmentation is key to body composition assessment in large patient cohorts and ultimately, to support implementation of body composition assessment into routine clinical practice. A deep learning neural network (DLNN) can be an essential part of such an automated workflow.

One challenge for developing a robust DLNN is that CT scans can differ in quality due to patient-, scanner-, and contrast media-related parameters.⁸⁻¹⁰ For example, patient positioning can significantly affect image quality (eg, parts of the patient, such as the arms, may be outside the scanning field of view),¹¹ (moving) artefacts can occur, and different scanners are used in daily clinical practice around the world, heavily impacting the radiation dose. This variation might result in poor or differences in performance of an automated segmentation algorithm.¹² A systematic review revealed that 1 in 3 DLNN studies of body composition segmentation has been developed with <100 unique human subjects, and more than half of the reviewed studies used exclusively single-institutional datasets.¹³ Computerized applications trained and validated in closely-related datasets are at high risk of reporting overly-optimistic metrics of performance. Healthcare artificial intelligence (AI) guidelines in general,^{14,15} and, specifically, the “Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis” (TRIPOD) guidelines,¹⁶ recommend that a fully independent external validation—with entirely different clinicians, scanners, and hospitals—be used for a robust estimation of how a computerized system performs in truly realistic clinical settings.

Robust DLNNs need to be trained on datasets that are large enough to incorporate the heterogeneity created by a variety of scanners, image acquisition settings, image reconstruction kernels, imperfect adherence to patient positioning protocols, and sufficiently high heterogeneity of subject clinical presentations. If training datasets are overly restrictive or too small to capture the diversity of clinical work, this introduces a type of selection bias and hence performance will be poor when encountering real-life clinical situations.

In previous work, the DLNN that is the subject of this paper had been independently validated using a large poly-trauma patient cohort extracted from the same university hospital, albeit at a different department (emergency department) and for a clinically distinct setting.¹⁷ This was nonetheless considered a challenging validation attempt due to the large variation in patient positioning (including arms and hands appearing inside the field of view) as well as radiation artefacts (eg, from metal devices attached to the patient). Even with this challenging cohort, the present DLNN model performed very well.

A robust, fully inter-institutional, and large-scale external testing with unseen datasets is needed for developing a quality AI tool for potential clinical use. By training a DLNN with several large trial cohorts and performing external validation on a large independent cohort, we aim to demonstrate the robust performance of our automatic body composition segmentation tool for future use in patients. This paper presents the first fully-independent external validation of the Mosamatic DLNN in a surgical oncology cohort with clinical imaging data from different hospitals, with independent radiology scan protocols, and comparing against reference delineations provided by independent clinicians.

Methods

Patients

A total of 3187 patients requiring abdominal surgery who had undergone a CT scan prior to surgery contributed by 32 distinct centres (located in The Netherlands, Germany, and the United Kingdom) were used for DLNN development (see general patient characteristics in [Table 1](#)). These comprised of de-identified data abstracted from previously ethics board-approved clinical studies; permission for secondary analysis was obtained via the principal investigators of the respective studies. We used L3 CT slices from: 3 colorectal liver metastases trials—2 from multiple sites across the United Kingdom and a single-institution study in The Netherlands; 2 ovarian cancer trials among 5 participating Dutch centres; and 1 pancreatic cancer trial of patients operated either in Aachen, Germany, or in Maastricht, The Netherlands.

An independent external validation set comprised 2535 L3 CT slices at different time intervals taken from 1054 unique subjects diagnosed with either resectable colorectal or pancreatic cancer (see [Table 1](#)).^{18,19} Ethical approval was granted by the West of Scotland Research Ethics Committee, Glasgow.

Image acquisition and reference segmentations

The aforementioned datasets comprised CT scans from a broad range of equipment vendors and image acquisition settings, which enables development of a more robust algorithm. Images were archived in DICOM (digital imaging and communications in medicine) format. [Table S1](#) (see online [supplementary materials](#)) summarizes the diverse imaging settings as recorded in DICOM metadata.

L3 images were obtained at the level of the transverse processes. All human-made segmentations in this study were created with *Slice-o-matic* (Tomovision, Quebec, Canada). Regions of interest (ROIs) were defined using standardized Hounsfield unit (HU) ranges (SM: −29 to +150, VAT: −150 to −50, SAT: −190 to −30). Absolute areas were normalized by physical height squared to derive skeletal muscle index (SMI), visceral adipose tissue index (VATI), and subcutaneous adipose tissue index (SATI). Mean HU in SM at L3 was used as the skeletal muscle radiation attenuation (SMRA). All human reference segmentations were made by 6 clinical researchers extensively trained under supervision of a radiologist to perform body composition analysis in *Slice-o-matic* in concordance with the Alberta Protocol.²⁰ Inter-observer variability for manual segmentations was excellent with 2-way random-effects intra-class coefficient of correlations of 0.96, 1.00, 0.99, and 0.99 for SMI, VATI, SATI, and SMRA, respectively.⁷ Intra-observer variability was low, with a percentage coefficient of variation in

Table 1. General patient characteristics for the DLNN development sets and the external test set.

Study ID	Model development sets					External validation set
	FROGS ^a	New EPOC ^a	Zuyd ^b	MUMC ^c	MUMC/Aachen ^d	UG ^e
Diagnosis	Emergency laparotomy (benign and malignant disease)	Colorectal liver metastases		Ovarian cancer	Pancreatic cancer	Pancreatic cancer + colorectal cancer
Time interval	2017-2019	2007-2012	2013-2017	2002-2015	2015-2019	2008-2019
Sample size	804	153	1587	339	304	1054 (147 pancreatic, 907 colorectal)
No. male (%)	374 (47%)	–	883 (56%)	0 (0%)	161 (53%)	567 (54%)
No. female (%)	430 (53%)	–	704 (44%)	339 (100%)	143 (47%)	487 (46%)
Ages (median)	25-95 (68)	–	32-98 (70)	30-101	10-88 (74)	23-93 (69)
Range BMI in kg.m ⁻² (median)	14-58 (26)	–	15-53 (26)	–	– (25.4)	14-59 (27)

^aBristol, Poole, Bournemouth, Royal Marsden, Surrey, Portsmouth, Velindre, Sheffield, Imperial Charing Cross, Imperial St Mary, Christie, Southend, Yeovil, North Middlesex, Southampton, Guys, Aintree, Winchester, Cambridge, Princess Alexandra, Bedford, Salisbury, University College London, Basingstoke, Pennine (United Kingdom).

^bZuyderland Medical Centre Geleen/Heerlen (The Netherlands).

^cMaastricht University Medical Centre, Radboud University Medical Centre Nijmegen, Bernhoven Medical Centre Uden, St Jansdal Medical Centre Ede (The Netherlands).

^dMaastricht University Medical Centre (Netherlands), RWTH Uniklinik Aachen (Germany).

^eGlasgow Royal Infirmary (United Kingdom).

-No individual values extracted.

Abbreviations: DLNN = deep learning neural network.

measurement error of 0.65% for tissue areas and 0.60 for tissue radiation attenuation.⁵

Previously published analyses on the external validation dataset had been made with ImageJ (National Institutes of Health, v1.47, <http://rsbweb.nih.gov/ij/>), but this method was shown to overestimate adipose tissue areas relative to other software.²¹ Every validation subject in this study was therefore independently re-annotated in Slice-o-matic by the original data owners. To ensure consistency for direct comparison, we re-computed areas and mean HU for all subjects with independent Python code, and confirmed equivalent values with each version of Slice-o-matic used to 2 decimal places or better.

Deep learning neural network

A DLNN for multi-label segmentation of SM, VAT, and SAT was built from a canonical 2D U-Net,²² with minor changes in the input layer to match the dimensions of a CT slice (512 × 512). An essential development for this work was to chain 2 independently-trained U-Net networks; the first U-Net was developed to segment the whole abdomen, whilst ignoring hands, arms, CT mattress, and extraneous medical devices that sometimes appeared in the CT field of view. The second U-Net was specialized for segmenting SM, VAT, and SAT within the abdominal outline detected by the first U-Net (see online [Figure S1](#) and its accompanying text).

Pixel intensities were clipped to the range [−500, +500] HU for the abdomen segmentation network. The reference abdominal region was generated by computing the outermost continuous contour of the human expert's SAT region before morphologically filling in every pixel inside. The range of intensities was further clipped to [−200, +200] HU to train the multi-label segmentation of muscle and fat. In each network, clipped intensities were scaled between [0,1] via standard min-max normalization. Pre-processed CT images were stored and handled in DICOM format. Human expert segmentations were extracted from Slice-o-matic in its

proprietary TAG format and converted to Python (NumPy) array objects before training the deep learning model.

Hands, arms, and other extraneous objects were rare within the training set; thus, we synthetically over-sampled images with extraneous objects outside the abdomen until they comprised 50% of each training batch while developing the abdomen U-Net. To train the muscle and fat multi-level segmentation network, all available 3187 subjects were randomly shuffled and split into 80% for training and 20% for validation. Given the relatively large sample size, a (non-overlapping) 80-20 split is superior to alternative methods like K-fold cross-validation where each validation block ultimately ends up being “seen” by the training algorithm, potentially introducing bias due to data leakage. More details of DLNN construction have been provided in [online supplementary materials](#).

CT slices and human-drawn (reference) annotations for the external validation were not revealed until the final DLNN model had been selected and all its model weights permanently fixed. Pre-processing of the test set followed the same steps as aforementioned. The full DLNN code (stripped of all trained models and patient data) is made open access (see Data availability). The trained algorithm can run easily on a conventional office laptop with standard specifications. An example of the DLNN output is shown in [Figure 1](#).

Automatic L3-selection

For use on large cohorts and for ease of future clinical implementation, automatic vertebra localization is necessary. We have integrated a state-of-the-art externally validated and open-source tool known as TotalSegmentator (<https://github.com/wasserth/TotalSegmentator>).^{23,24} In keeping with the “narrow AI” paradigm, we have chained together highly specialized AI tools for each task. TotalSegmentator was first used for automated segmentation of all visible vertebrae in a volumetric CT study. The resulting labelled masks were used to locate all the slices intersecting L3, and then we selected

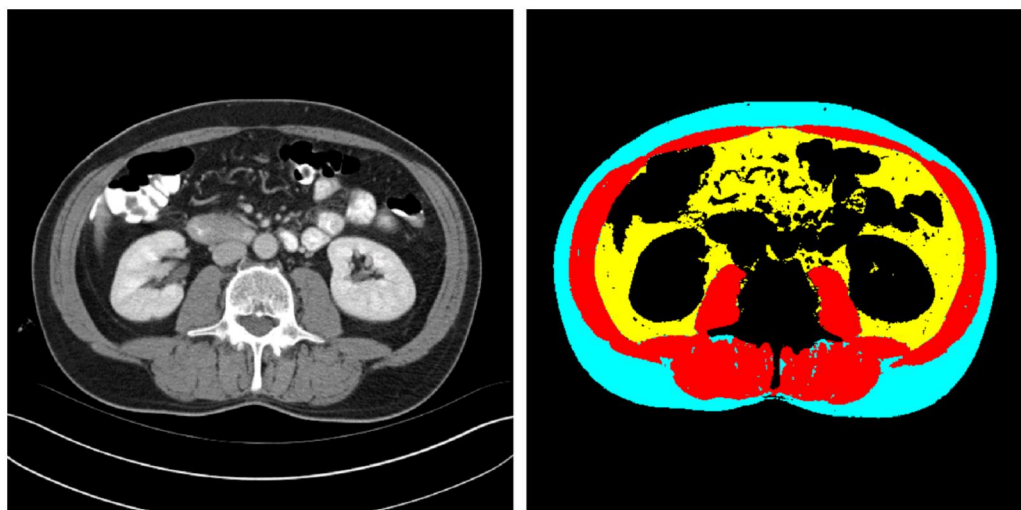


Figure 1. Segmentation of skeletal muscle (red), visceral adipose tissue (yellow), and subcutaneous adipose tissue (blue) on a single CT slice at the level of the third lumbar vertebra.

the CT slices closest to the centre of the segmented object (see Figure S2).

Analysis

Geometric agreement was evaluated by using 2D dice similarity (DS) comparing the DLNN segmentations of SM, SAT, and VAT against the corresponding annotation made by human experts. DS computes the area of the intersection between human and DLNN segmentations as a fraction of half the summated area (human-drawn area plus DLNN-drawn area). Perfect geometric agreement implies $DS = 1$, and if the intersection area is zero then $DS = 0$. Agreement of SMI, VATI, SATI, and SMRA between manual and automatic annotations was quantitatively evaluated in the test set using Lin's concordance correlation coefficient (CCC) and Bland-Altman's Limits of Agreement (LoA) (with and without repeated measurements). By using the human-drawn annotations in the test set as reference and then applying the risk classification supplied by Martin et al,³ we computed the diagnostic performance (sensitivity, specificity, balanced accuracy, and agreement kappa) of the DLNN results. Statistical analyses were performed in R (version 4.2.0).

Results

Model training

Total loss and DS curves in the training dataset show DLNN model convergence within about 40 000 steps (see Figure S3). There was rapid improvement within the first 10 000 steps but DS was largely stable thereafter. Total (Dice+L3) loss continued to decrease gradually but we stopped model training after 38 000 steps, since there was very little to gain with further training. The DLNN weights after the last training step were thus fixed as the "final model" for subsequent testing. The established segmentation tool was named MosaMatic.

Segmentation speed

The DLNN was able to segment a single CT-image in around 2 s and the whole external validation cohort ($n = 2535$) in around 90 min.

Concordance between manual and DLNN segmentations

The overall distribution of DS for SM, VAT, and SAT in the quarantined validation dataset is summarized in the box-whisker plot shown in Figure 2A. The median DS for SM was 0.97 (IQR: 0.95-0.98), with a tail of outliers down to a minimum DS of 0.45. The distributions of DS for VAT (median: 0.98, IQR: 0.95-0.98) and SAT (median: 0.95, IQR: 0.92-0.97) were highly skewed, with extreme outliers landing near zero (these were patients with very small amounts of total adipose tissue). The DS is known to be overly sensitive for small volumes, and this can also be seen in our results—Figure 2B-D.

Lin's CCC evaluation of SMRA, SMI, VATI, and SATI comparing expert segmentations (as reference) and DLNN results (as test) was excellent, as shown in Figure 3A-D. Numerical measures of the CCC, bias correction factor for slope of agreement, and finally the Bland-Altman intervals of agreement without repeated scans are provided in Table 2. The CCC ranges from 0.964 (for SMI) up to 0.998 (for VATI). The errors in the agreement slope, as indicated by deviation from the dotted line in Figure 3, were all close to unity, indicating no major deviations from the ideal, which is supported by bias correction multipliers being better than 0.991 (ie, no correction implies 1.00). Based on our large cohort, median *in vivo* values (which are in reality age- and sex-dependent) of SMRA, SMI, VATI, and SATI roughly fall in the vicinity of 50 HU, 45, 60, and 70 $\text{cm}^2 \cdot \text{m}^{-2}$. The Bland-Altman metrics (with percentages in parentheses) indicate only small systematic offsets of 0.23 HU (1.0%), 1.26 $\text{cm}^2 \cdot \text{m}^{-2}$ (2.9%), 1.02 $\text{cm}^2 \cdot \text{m}^{-2}$ (2.5%), and 3.24 $\text{cm}^2 \cdot \text{m}^{-2}$ (4.9%) for SMRA, SMI, VATI, and SATI, respectively. The upper and lower limits of the Bland-Altman tests indicate SATI had the widest random variation component (-6.7 to $13 \text{ cm}^2 \cdot \text{m}^{-2}$). Most importantly, for risk stratification by muscle fat content, the random noise component of SMRA was estimated at about 2-3 HU in magnitude, and correspondingly for SMI about 3-5 $\text{cm}^2 \cdot \text{m}^{-2}$ in magnitude.

Consistent concordance for consecutive measurements

In 449 subjects, we obtained a consecutive CT image at varying time intervals ranging from within a month up to

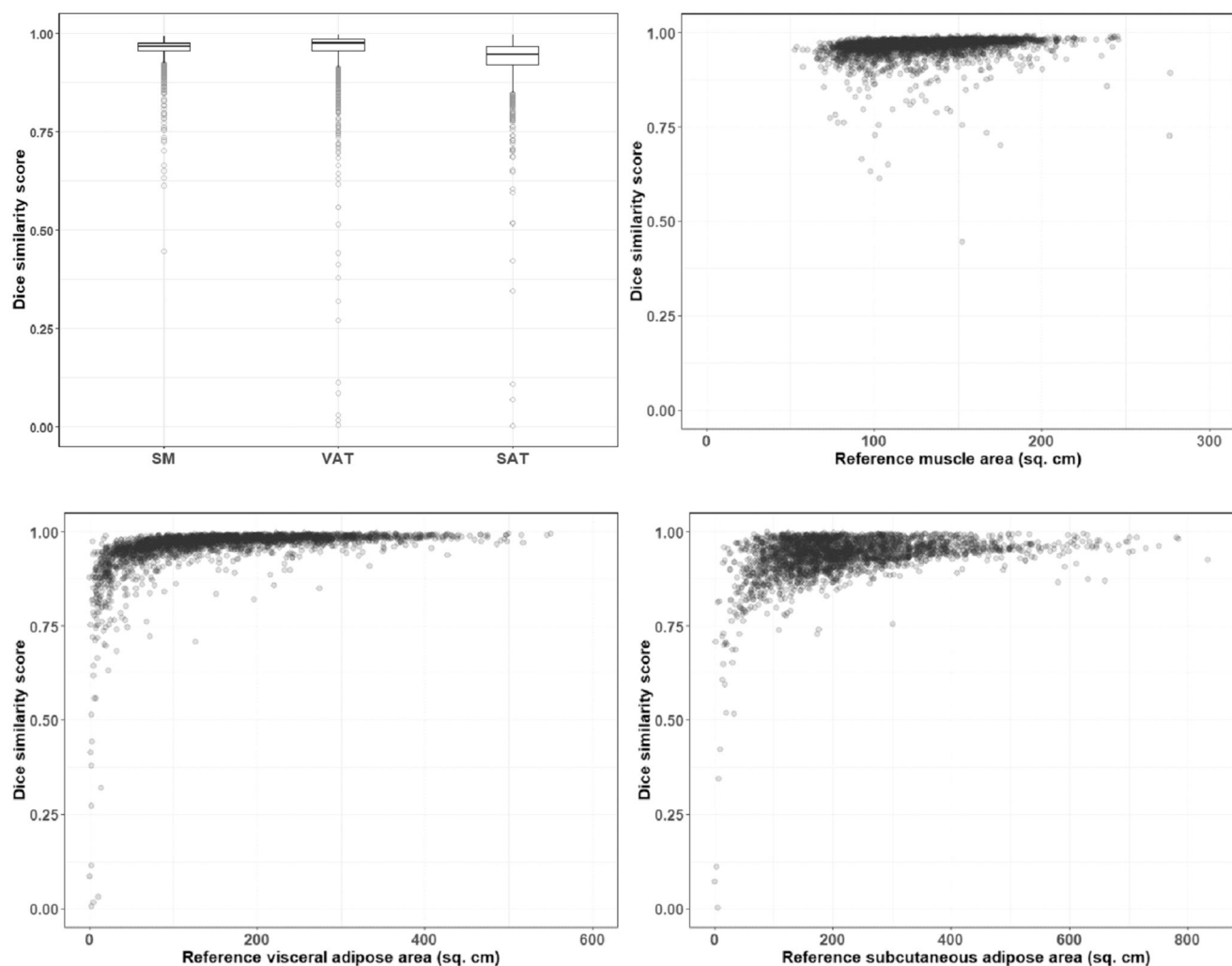


Figure 2. Distribution of geometric dice similarity (DS) on L3 slice for skeletal muscle (SM), subcutaneous fat (SAT), and visceral fat (VAT). (A) Box-whisker plot showing the median DS as the solid horizontal line and the interquartile range as the upper and lower limits of the box. The vertical line ends indicate 1%-tile and 99%-tile, and outliers outside this range are shown as individual dots. (B)-(D) show the distribution of DS as a function of SM area, VAT area, and SAT area, respectively.

12 months. Whereas the scope of this study was not to objectively quantify longitudinal precision, we can already derive some preliminary insight into stability with repeated imaging over time using these data. The concordance plots for SMRA, SMI, VATI, and SATI for *consecutive scans* are equivalent to Figure 3 (see Figure S4). There was no evidence of divergence from the high concordance observed in the agreement on primary CTs. According to CCC metrics and Bland-Altman limits with repeated measures, there are no notable changes between agreement of body composition indices between primary (top half of Table 2) and repeat scans (bottom half of Table 2).

Accuracy

We tested the clinical significance of using the DLNN segmentations with respect to a change in stratification for low SMI and low SMRA using the widely used thresholds defined by Martin et al.³ Overall accuracy of stratification was 0.93 for low SMI (sensitivity: 0.99, specificity: 0.87) and 0.98 for low SMRA (sensitivity: 0.98, specificity: 0.98). The discretized agreement (Cohen's inter-rater kappa) was 0.85 for low SMI and 0.96 for low SMRA, which is generally considered

as being excellent. For completeness, a 2×2 confusion matrix for low SMI and low SMRA is included in the [online supplemental materials](#) as Figure S5.

In addition, we tested the accuracy of L3 mid-vertebrae localization from TotalSegmentator using a small independent test cohort of 30 subjects completely unrelated to the present study. In this brief quality assurance test, we correctly identified the CT-slice at L3 in 30 out of 30 cases (100%) based on the vertebrae segmentation produced by TotalSegmentator.

Discussion

In this study, we present our TRIPOD Level 3 fully independently and externally validated deep learning model for automated segmentation of CT-based L3 slices. Due to its robust performance in both internal and external validation cohorts, this study shows that DLNN-generated segmentation can reliably replace manual segmentation when performing body composition assessment. This opens up new possibilities both in clinical and scientific settings, such as cost- and time-effective clinical implementation and large cohort/population studies.

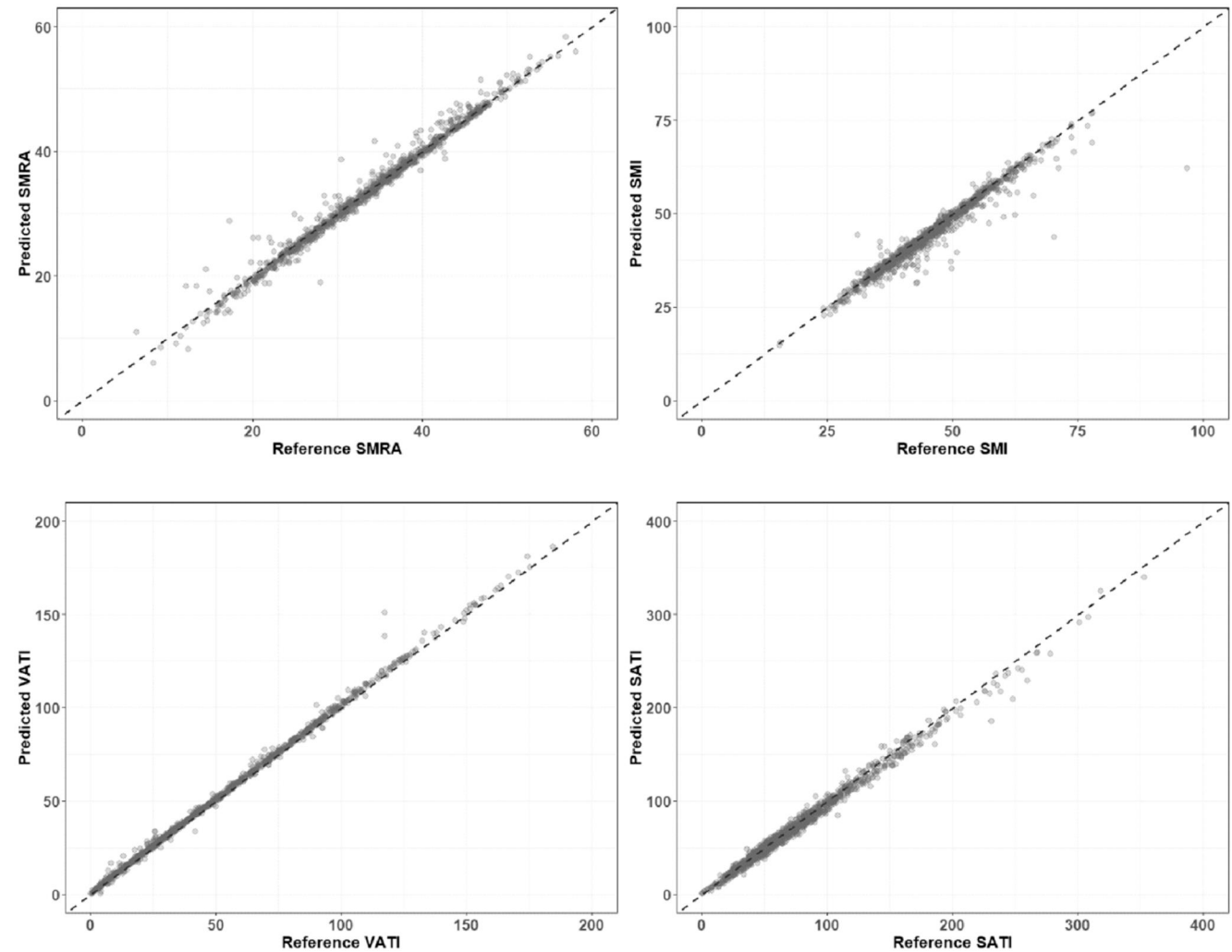


Figure 3. Lin’s concordance correlation (CCC) plots. (A) Skeletal muscle attenuation (SMRA), (B) skeletal muscle index (SMI), (C) visceral fat index (VATI), and (D) subcutaneous fat index (SATI). The units of SMRA are Hounsfield unit (HU). The units of SMI, VATI, and SATI are all cm².m⁻². Reference values were defined as those extracted from human-drawn segmentations. Predicted values were extracted from DLNN-made segmentations. DLNN = deep learning neural network.

Table 2. Concordance correlation, bias correction factor, and Bland-Altman agreement without repeated measures (*n* = 1054).

Bland-Altman estimates of agreement for primary scan only (<i>n</i> = 1054)			
	Concordance correlation (95% confidence interval)	Bias correction factor	Bland-Altman agreement (95% lower-upper limits)
SMRA	0.991 (0.990-0.992)	0.999	0.23 (−2.06 to 2.52) HU
SMI	0.964 (0.959-0.968)	0.991	1.26 (−3.11 to 5.63) cm ² .m ⁻²
VATI	0.998 (0.998-0.998)	0.999	−1.02 (−4.55 to 2.50) cm ² .m ⁻²
SATI	0.992 (0.991-0.993)	0.997	3.24 (−6.69 to 13.2) cm ² .m ⁻²
Bland-Altman estimates of agreement for repeated scans only (<i>n</i> = 449)			
	Concordance correlation (95% confidence interval)	Bias correction factor	Bland-Altman agreement (95% lower-upper limits)
SMRA	0.991 (0.990-0.992)	0.999	0.18 (−2.08 to 2.45) HU
SMI	0.973 (0.969-0.976)	0.997	0.75 (−3.56 to 5.06) cm ² .m ⁻²
VATI	0.998 (0.998-0.998)	0.999	−1.07 (−4.55 to 2.41) cm ² .m ⁻²
SATI	0.992 (0.991-0.993)	0.998	2.55 (−8.36 to 13.4) cm ² .m ⁻²

Abbreviations: HU = Hounsfield unit; SATI = subcutaneous adipose tissue index; SMI = skeletal muscle index; SMRA = skeletal muscle radiation attenuation; VATI = visceral adipose tissue index.

Clinically, and subject to a future clinical implementation study to follow this work, our automated L3 body composition segmentation tool is intended to be easily implemented in standard practice for all routine CT-scans, which clinicians can then use for prognostic risk assessment and treatment decision making. Changes in body composition over time can be detected during oncologic follow-up, which might provide early indications of treatment effect or disease progression/recurrence. Going from a prognostic tool to a predictive tool—in which the tool is used for treatment decisions—still remains a large step to take as large international datasets are needed to provide clinical reference values. Nevertheless, a recent publication on the increasing incorporation of body composition analysis as confounder or endpoint in clinical trials indicates the need for a fast and easy-to-use CT body composition assessment tool.¹

Body composition is highly variable among sex, age, race, and cancer types.^{4,5,25-27} For this reason, developed clinical cut-offs vary greatly among different patient cohorts, and prognostic models of outcome (eg, survival) are likely to fail during external validation.^{4,28} In addition, body composition can be dependent on other clinical parameters and may have stronger prognostic effects when combined with parameters such as systemic inflammation and weight loss.^{18,29,30} We have previously demonstrated that such combinations or “host phenotypes” are more predictive of overall survival than tumour-based prognostic scores in patients with colorectal liver metastases.²⁹ Larger cohorts are needed for each cancer type, as these could support the use of body composition analysis in the standard diagnostic work-up, and potentially aid in clinical treatment decision-making. Automated body composition analysis is the only way of acquiring sufficient data for adequate Z-scoring and accounting for the aforementioned patient characteristics. While cut-offs are necessary for clinical use, we advocate the development of a clinical risk calculator, as the prognostic effect of body composition variables are incremental⁵ and should therefore not be arbitrarily forced into dichotomic cut-offs. In the end, integrating body composition data with established prognostic factors such as tumour stage may improve prediction of a patient’s prognosis. A combined tumour and host-focused approach would provide a basis for clinical trials aimed at exploring whether body composition-based prognostic information can be used as a basis for treatment decision making (eg, palliative intent instead of curative intent, or indication for/selection of (neo)adjuvant therapy).

Scientifically, our L3 segmentation tool enables assessment of large (incl. historical) cohorts that would be unfeasible to segment manually. In addition, as the AI has learned from multiple observers, it has not learned an expert’s specific signature, ensuring a more stable output. However, the true value of automated segmentation is that it facilitates the inclusion of body composition as a study parameter in randomized controlled trials (RCTs), as the time and effort of analysis is reduced from a couple of months to a few minutes. Segmentation speed is obviously highly dependent on computer hardware, but on a standard entry-level set-up, the automatic segmentation speed was around 2 s as compared to 440 s for a manual segmentation on the same device.⁷ Fast segmentation enables stratification and selection of patients with different body compositions, creating either homogeneous or heterogeneous cohorts as required. Including body composition is particularly important in oncology as it is

related to chemotherapy effectiveness and toxicity.³¹ Ideally, chemotherapy dosing should be based on lean mass to prevent dose-limiting toxicities for which DLNN would be a logical application in the future.

Some other automated segmentation tools have been developed. The largest cohort ($n = 12\,128$) was used for development of the AI tool published by Magudia et al.²⁵ Their tool performed well with similar dice scores to our algorithm. Their training cohort only included 604 pancreatic cancer patients while the large ($n = 12\,128$) hospital dataset was used to derive reference curves. However, the large hospital dataset only included patients without cancer and cardiovascular disease, making it less applicable to a clinical population of subjects with cancer who frequently display body composition alterations. In addition, analysis of CT-scans of cancer patients can be more challenging due to anatomic abnormalities and suboptimal patient positioning. As patients with cancer were excluded, the tool by Magudia et al could perform worse in cancer cohorts. Our analyses did not exclude patients with anatomical variations or unconventional patient positioning, which prevents overfitting the model to a specific patient group and will likely result in a more robust segmentation tool. Dabiri et al³² published an automated segmentation tool that was trained on 2 cohorts of patients with cancer ($n = 2529$). Their segmentation tool performed similarly well compared with our segmentation tool. However, in contrast to our study, they did not perform external validation, making it uncertain how their AI performs in other cohorts.

For volumetric CTs as input, an important consideration is how to select the slice intersecting the middle of the L3 vertebra, and more generally in case the user arbitrarily wishes to select some other vertebra. In keeping with the “narrow AI” paradigm, we have elected to implement a modular software design such that highly specialized DLNNs are joined up sequentially in a workflow to accomplish a meaningful task. Currently, we integrated the state-of-the-art and validated TotalSegmentator tool to automatically localize spinal vertebrae. If a superior vertebrae segmentation tool should emerge in future, we could relatively easily adapt our workflow to incorporate the new tool, compared to “all-in-one” monolithic software design.

The DLNN was developed using a training set consisting of images with a variety of scan parameters (eg, scan manufacturers, tube voltage, contrast protocol), resulting in a more robust algorithm. While the DLNN showed excellent performance, even with challenging CT-scans, it has its limitations. In particular, analysis of CT-scans of patients with anatomical abnormalities (eg, large abdominal hernia, colostomy, profound oedema, and scoliosis) or of patients with abnormal/non-standard positioning in the CT-scanner can lead to (partially) incorrect segmentations. Such challenging CT-images should then be manually corrected and stored prospectively. In due time, this cohort of “challenging CT-images” can be used to retrain and improve the DLNN. In addition, while we successfully tested our algorithm on repeated scans within the same patient, changes in body composition due to cancer progression could have occurred resulting in significant changes in body composition over time. Another limitation of the algorithm is that it has only been trained and validated in a Western European setting. Future studies should validate the algorithm in other parts of the world, as factors such as patient race and scanner

manufacturer could influence the segmentations. Finally, while having included a wide range of acquisition types, certain combinations could be underrepresented in the training set, resulting in potential segmentation errors. Different deep learning segmentation algorithms will have different limitations depending on the cohort. A comparative study using both healthy individuals and different patient groups could provide insight into how these different algorithms perform and if 1 algorithm is preferred over the other in specific cohorts.

The key step forward will be implementing automated segmentation into clinical practice and making it easily accessible for new research initiatives. Our tool was created in such a way that it can be easily integrated in clinical imaging software or work independent alongside existing imaging infrastructure. To ensure easy access for research purposes, both the untrained AI and the automatic DLNN-trained segmentation tool (Mosamatic) will be freely available for scientific use by other research groups. This enables rapid implementation and much-needed data collection to develop clinical prediction tools.

Conclusion

In this study, we developed a reliable deep-learning model that was independently and externally validated for automated analysis of body composition of patients with cancer. To simplify future use and potential integration of the DLNN-based automated segmentation workflow, we have incorporated the steps into a web browser-based graphical user interface which included the open-source TotalSegmentator tool for automatic vertebral localization. The algorithm could be implemented in various clinical infrastructures and used by other research groups to assess large cancer patient cohorts.

Acknowledgements

This manuscript has been pre-printed: van Dijk, D.P.J., Volmer, L.F., Brecheisen, R., Dolan, R.D., Bryce, A.S., *et al.* External validation of a deep learning model for automatic segmentation of skeletal muscle and adipose tissue on L3 abdominal CT images. medRxiv, Preprint posted online 23 April 2023, <https://doi.org/10.1101/2023.04.23.23288981> (2023).

Collaborators

The Body Composition Collaborative: Thais T.T. Tweed^{1,2}, Stan Tummers², Gregory van der Kroft¹, Marjolein A.P. Ligthart¹, Merel R. Aberle¹, Lubbers Tim¹, Bart C. Bongers³, Jorne Ubachs⁴, Roy F.P.M. Kruitwagen⁴, Siân Pugh⁵, John N. Primrose⁶, John A. Bridgewater⁷, Philip H. Pucher⁸, Nathan J. Curtis⁹, Stephan B. Dreyer^{10,11}, Michael Kazmierski¹²

¹Department of Surgery, Maastricht University Medical Center, 6200 MD Maastricht, The Netherlands

²Department of Surgery, Zuyderland Medical Centre, Sittard-Geleen, The Netherlands

³Department of Nutrition and Movement Sciences, School of Nutrition and Translational Research in Metabolism (NUTRIM), Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands

⁴Department of Obstetrics and Gynaecology, Maastricht University Medical Center, Maastricht, The Netherlands

⁵Department of Medical Oncology, University of Southampton, Southampton, United Kingdom

⁶Department of Surgery, University of Southampton, Southampton, United Kingdom

⁷UCL Cancer Institute, University College London, London, United Kingdom

⁸Department of General Surgery, Portsmouth Hospitals University NHS Trust, Portsmouth, United Kingdom

⁹Academic Unit of Cancer Sciences, Faculty of Medicine, University of Southampton, Southampton, United Kingdom

¹⁰Wolfson Wohl Cancer Research Centre, School of Cancer Sciences, University of Glasgow, Glasgow, United Kingdom

¹¹West of Scotland Pancreatic Unit, Glasgow Royal Infirmary, Glasgow, United Kingdom

¹²Department of Radiotherapy (MAASTRO), School of Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands

Author contributions

Conceptualization: David P.J. van Dijk, Sander S. Rensen, Andre Dekker, Leonard Wee, Steven W.M. Olde Damink; Methodology: David P.J. van Dijk, Leroy F. Volmer, Leonard Wee; Software: Leroy F. Volmer, Ralph Brecheisen; Validation: David P.J. van Dijk, Ross D. Dolan, Adam S. Bryce, David K. Chang, Donald C. McMillan, Leonard Wee; Formal analysis: Leroy F. Volmer, Leonard Wee; Investigation: David P.J. van Dijk, Ralph Brecheisen, Ross D. Dolan, Adam S. Bryce, David K. Chang, Donald C. McMillan, Jan H.M.B. Stoot, Malcolm A. West; Writing (original draft): David P.J. van Dijk, Leroy F. Volmer, Leonard Wee; Writing (review and editing): all authors; Visualization: Leroy F. Volmer, Ralph Brecheisen, Leonard Wee; Supervision: Sander S. Rensen, Andre Dekker, Leonard Wee, Steven W.M. Olde Damink.

Supplementary material

Supplementary material is available at BJR online.

Funding

The New EPOC study was supported by Cancer Research UK.

Conflicts of interest

None declared.

Data availability

This work concerns only secondary re-use of clinical study data of patients, which were obtained in de-identified form with permission from the original principal investigators. Each study had previously been reviewed by a competent ethics body. Data may be obtained from the aforementioned principal investigators upon reasonable request. Source code for data preparation of CT slices and human reference annotations, along with the DLNN model architecture, are publicly available here under a Creative Commons 4.0 CC-BY-NC License: https://github.com/MaastrichtU-CDS/BodyCompL3_DLNN_Open_Code.

Mosamatic (the trained DLNN automatic segmentation tool) is available here under a Creative Commons 4.0 CC-BY-NC License: <https://github.com/rbrecheisen/MosamaticDesktop>.

References

1. Brown LR, Sousa MS, Yule MS, et al.; Cancer Cachexia Endpoints Working Group. Body weight and composition endpoints in cancer cachexia clinical trials: systematic review 4 of the cachexia endpoints series. *J Cachexia Sarcopenia Muscle*. 2024;15(3):816-852.
2. 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.
3. Martin L, Birdsell L, Macdonald N, et al. Cancer cachexia in the age of obesity: skeletal muscle depletion is a powerful prognostic factor, independent of body mass index. *J Clin Oncol*. 2013;31(12):1539-1547.
4. van Dijk DP, Bakens MJ, Coolsen MMM, et al. Low skeletal muscle radiation attenuation and visceral adiposity are associated with overall survival and surgical site infections in patients with pancreatic cancer. *J Cachexia Sarcopenia Muscle*. 2017;8(2):317-326.
5. van Dijk DPJ, Zhao J, Kemter K, et al. Ectopic fat in liver and skeletal muscle is associated with shorter overall survival in patients with colorectal liver metastases. *J Cachexia Sarcopenia Muscle*. 2021;12(4):983-992.
6. Heymsfield SB, Gonzalez MC, Lu J, Jia G, Zheng J. Skeletal muscle mass and quality: evolution of modern measurement concepts in the context of sarcopenia. *Proc Nutr Soc*. 2015;74(4):355-366.
7. Ackermans L, Volmer L, Timmermans Q, et al. Clinical evaluation of automated segmentation for body composition analysis on abdominal L3 CT slices in polytrauma patients. *Injury*. 2022;53(Suppl 3):S30-S41.
8. Bae KT. Intravenous contrast medium administration and scan timing at CT: considerations and approaches. *Radiology*. 2010;256(1):32-61.
9. Lell MM, Wildberger JE, Alkadhi H, Damilakis J, Kachelriess M. Evolution in computed tomography: the battle for speed and dose. *Invest Radiol*. 2015;50(9):629-644.
10. Martens B, Hendriks BMF, Muhl C, Wildberger JE. Tailoring contrast media protocols to varying tube voltages in vascular and parenchymal CT imaging: the 10-to-10 rule. *Invest Radiol*. 2020;55(10):673-676.
11. Manava P, Galster M, Ammon J, Singer J, Lell MM, Rieger V. Optimized camera-based patient positioning in CT: impact on radiation exposure. *Invest Radiol*. 2023;58(2):126-130.
12. Ha J, Park T, Kim HK, et al. Development of a fully automatic deep learning system for L3 selection and body composition assessment on computed tomography. *Sci Rep*. 2021;11(1):21656.
13. Bedrikovetski S, Seow W, Kroon HM, Traeger L, Moore JW, Sammour T. Artificial intelligence for body composition and sarcopenia evaluation on computed tomography: a systematic review and meta-analysis. *Eur J Radiol*. 2022;149:110218.
14. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320-1324.
15. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *JAMIA*. 2020;27(12):2011-2015.
16. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.
17. Ackermans L, Volmer L, Wee L, et al. Deep learning automated segmentation for muscle and adipose tissue from abdominal computed tomography in polytrauma patients. *Sensors (Basel, Switzerland)*. 2021;21(6):2083.
18. Dolan RD, Almasaudi AS, Dieu LB, Horgan PG, McSorley ST, McMillan DC. The relationship between computed tomography-derived body composition, systemic inflammatory response, and survival in patients undergoing surgery for colorectal cancer. *J Cachexia Sarcopenia Muscle*. 2018;10(1):111-122.
19. Dolan RD, Abbass T, Sim WMJ, et al. Longitudinal changes in CT body composition in patients undergoing surgery for colorectal cancer and associations with peri-operative clinicopathological characteristics. *Front Nutr*. 2021;8:678410.
20. sliceOmatic. Alberta protocol. Accessed February 11, 2017. https://tomovision.com/Sarcopenia_Help/index.htm
21. Dolan RD, Tien Y-T, Horgan PG, Edwards CA, McMillan DC. The relationship between computed tomography-derived body composition and survival in colorectal cancer: the effect of image software. *JCSM Rapid Commun*. 2020;3(2):81-90.
22. Ronneberger O, Fischer P, Brox T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Springer International Publishing; 2015:234-241.
23. Wasserthal J, Meyer M, Breit H, Cyriac J, Yang S, Segeroth M. TotalSegmentator: robust segmentation of 104 anatomical structures in CT images. arXiv, <https://arxiv.org/abs/2208.05868>, 2022, preprint: not peer reviewed.
24. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211.
25. Magudia K, Bridge CP, Bay CP, 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. *Radiology*. 2021;298(2):319-329.
26. Tweed TTT, van der Veen A, Tummers S, et al.; LOGICA Study Group. Body composition is a predictor for postoperative complications after gastrectomy for gastric cancer: a prospective side study of the LOGICA trial. *J Gastrointest Surg*. 2022;26(7):1373-1387.
27. Rutten IJG, Van Dijk DPJ, Kruitwagen RFP, Beets-Tan RGH, Olde Damink SWM, Van Gorp T. Changes in skeletal muscle mass during neoadjuvant chemotherapy are related to survival in ovarian cancer. *J Cachexia Sarcopenia Muscle*. 2016;7(4):458-466.
28. Petermann-Rocha F, Balntzi V, Gray SR, et al. Global prevalence of sarcopenia and severe sarcopenia: a systematic review and meta-analysis. *J Cachexia Sarcopenia Muscle*. 2022;13(1):86-99.
29. Van Dijk DP, Krill M, Farshidfar F, et al. Host phenotype is associated with reduced survival independent of tumor biology in patients with colorectal liver metastases. *J Cachexia Sarcopenia Muscle*. 2018;10(1):123-130.
30. Martin L, Senesse P, Gioulbasanis I, et al. Diagnostic criteria for the classification of cancer-associated weight loss. *J Clin Oncol*. 2015;33(1):90-99.
31. Hopkins JJ, Sawyer MB. A review of body composition and pharmacokinetics in oncology. *Expert Rev Clin Pharmacol*. 2017;10(9):947-956.
32. Dabiri S, Popuri K, Ma C, et al. Deep learning method for localization and segmentation of abdominal CT. *Comput Med Imaging Graph*. 2020;85:101776.