scientific reports

Check for updates

OPEN

Technical performance of L3 skeletal muscle area (SMA) measurement on CT for L3 skeletal muscle index (L3SMI) assessment

Hubert Beaumont✉, Emilie Khayat, Alexandre Thinnes & Antoine Iannessi

This study aims at documenting the technical performances of the Skeletal Muscle Index at the L3 vertebra (L3SMI). In this retrospective multicenter study, we used TotalSegmentator freeware to segment muscles in axial CT scans at the L3 vertebra level and measure Skeletal Muscle Area (SMA). A dataset of 100 cancer patients (lung, colorectal, adrenal, pancreatic) was analyzed for accuracy, reproducibility, and linearity, using expert-segmented ground truth. A second dataset of 205 prone/supine colorectal cancer patients assessed repeatability in a test–retest setting. Outliers were reviewed to define exclusion criteria for L3SMI assessment. We computed within Coefficient of Variation (wCV), Bland–Altman analysis for variability and repeatability, and measured absolute/relative bias with confidence intervals from bootstrapped data. Bablok regression assessed measurement linearity. Intra and inter readers variability were, respectively, 5% and 7.6%. Relative bias was − 8.5% (95%CI: − 9.8, − 7.3), absolute bias was − 11.3cm^2 (95%CI: − 13.2, − 9.4) indicating the need for linear correction. The linear relationship between measurement and ground truth was found with a Pearson's coefficient of 0.97. Repeatability gives wCV of 3.2% (95%CI: 2.8; 3.5), with decreasing wCV value of [4.0; 2.7] % for MSA ranges [66.3; 213.9] cm^2. L3SMI evaluations were not applicable to ill positioned patients, truncated Field of View for steatosis patients in non-enhanced imaging. L3SMI and longitudinal changes of L3SMI can be reliably measured. Exclusion criteria have been listed.

Keywords Skeletal muscles, Image segmentation, Third lumbar vertebra (L3), CT images, Sarcopenia

Abbreviations

AP	AbdominoPelvic
BMI	Body mass index
BPR	Body part regression
CI	Confidence interval
CRC	Colorectal cancer
CT	Computed tomography
DXA	Dual-energy X-ray absorptiometry
EWGSOP	European working group on sarcopenia in older people
FoV	Field of view
HU	Hounsfield units
LoA	Limit of agreement
L3	Third vertebra of the lumbar spine
L3SMI	Skeletal muscle index at the third vertebra of the lumbar spine
MRI	Magnetic resonance imaging
NSCLC	Non-small cell lung cancer
PABAK	Prevalence and bias adjusted kappa-ordinal scale
QIB	Quantitative imaging biomarker
QIBA	Quantitative imaging biomarker alliance
RC	Repeatability coefficient
RECIST	Response evaluation criteria in solid tumor
SMA	Skeletal muscle area at L3

Median Technologies, Les Deux Arcs, 1800 Route Des Crêtes, 06560 Valbonne, France. ✉email: hubertbeaumont@hotmail.com

TA	Thoraco-abdominal
TAP	Thoraco-AbdominoPelvic
TCIA	The cancer imaging archive
TS	TotalSegmentator
wCV	Within coefficient of variance
wSD	Within standard deviation

Background

The understanding of skeletal muscle has evolved from its traditional association with mobility to a more comprehensive view, recognizing it as a complex organ with crucial endocrine functions and metabolic roles[1], as well as significant interactions with other organs[2]. Clinicians now place a stronger focus on body composition—the relative distribution of muscle, fat, and bone in the body[3]—as it reveals important insights into muscle health and overall patient condition. This shift has driven the development of various measurement indexes and techniques to quantify muscle-related conditions, most notably sarcopenia, which has substantial clinical implications for patient outcomes.

The European Working Group on Sarcopenia in Older People (EWGSOP) defines sarcopenia as a condition of low muscle mass and function[4], with severe sarcopenia leading to frailty or cachexia, the latter being characterized by muscle wasting often seen in advanced cancer and chronic disease. While sarcopenia typically develops with age due to a natural decline in lean body mass, it can also result from other factors like malnutrition, obesity, inactivity, chronic illness, and certain treatments. Sarcopenia is associated with a range of consequences, including increased risks of osteoporosis[5], endocrine disorders[6], and mental health conditions like depression. In oncology, sarcopenia is particularly relevant as it affects the metabolism of cytotoxic agents, leading to variability in chemotherapy toxicity[7–9] and, critically, is associated with reduced survival rates[10–12]. Imaging plays a pivotal role in assessing sarcopenia in clinical settings.

Non-imaging methods like questionnaires (e.g., SARC-F), anthropometric measures (e.g., BMI), and physical tests (e.g., hand grip strength) are simple but lack accuracy and specificity[13]. Imaging techniques, in contrast, effectively assess both muscle mass and quality[14]. DXA provides a quick, low-radiation whole-body scan but struggles with muscle quality estimation due to factors like hydration, body thickness, and certain pathologies[15]. MRI is radiation-free and allows repeated evaluations but is costly, less available, and lacks standardization[16]. CT, despite radiation exposure, remains preferred for oncological and non-oncological imaging due to its high reproducibility and role in quantitative imaging biomarker (QIB) assessments.

The advancement of QIBs has been driven by medical imaging technologies, providing objective indicators of biological processes, disease, and treatment responses. While various imaging biomarkers exist, rigorous qualification protocols for QIBs have emerged recently, led by initiatives like the Quantitative Imaging Biomarker Alliance (QIBA)[17]. QIB qualification follows three stages: (1) technical performance validation (accuracy, precision, and linearity), (2) clinical qualification (assessing differentiation between health states), and (3) clinical application[18].

The Skeletal Muscle Index at the L3 vertebra (L3SMI), a CT-derived (or MRI-derived) biomarker based on skeletal muscle area (SMA) normalized by height, holds significant promise in sarcopenia evaluation.

$$L3SMI = (Total\ Skeletal\ Muscle\ Area\ at\ L3)\ /Height^2 \quad (1)$$

This study aims to document the technical performance of L3SMI as a reliable biomarker in sarcopenia assessment using an advanced CT imaging software.

Material and method

Data

Two datasets were used:

Dataset 1

Dataset 1 was used for visually evaluating the quality of segmentations, accuracy and linearity. This dataset gather two private and 3 subsets from The Cancer Imaging Archive (TCIA)[19] representing a sample of 100 multicentric scans. Patients' diseases were, Non-Small Cell Lung Cancer (NSCLC), Colorectal Cancer (CRC), adrenocortical carcinoma with Ki-67 expression, advanced pancreatic cancer. IV contrast phase was identified using DICOM metadata associated with each CT series, including acquisition protocol information and series descriptions. When needed, this was verified by reviewing acquisition parameters and image characteristics consistent with the expected enhancement pattern of the corresponding phase.

The distribution of images acquisition parameters was mainly General Electric (GE) manufacturer (58.7%), "standard" reconstruction kernel (81.3%), kilovoltage peak of 120 kVp (78.7%), slice thickness of 5 mm (36%) and voxel size 0.8 mm and 1.0 mm both representing 26% (full details provided in supplementary figure S1).

Body coverage was distributed as: TAP (29%); AP (47%); A (13%) and TA (11%).

Visual evaluation relied on a subset of 80 segmentations after having validated the robustness of our evaluation scale on a subset of 60 segmentations.

Reference measurements were acquired after subjective selection of the most representative slice then manual segmentation of Dataset 1 (thresholding: [− 10 a 33]HU) of 100 CT scans. Drawn from distribution of SMA values (See supplementary figure S2), the average value of SMA at L3 was 120.8 cm^2; [Min; Max] = [61.9;

201.7] cm^2. As an example, for patients' height being all 1.7 m, average L3SMI would be of 41.8 cm^2/m^2; [Min; Max] = [21.4; 69.8] cm^2/m^2.

Dataset 2

These data were used for evaluating the repeatability.

We used an original dataset comprising 760 patients with colorectal cancer (CRC). For each patient, two CT series were acquired during the same examination, corresponding to supine and prone positions, as routinely performed in CT colonography. Thus, the two series represent the standard positional acquisitions obtained within a single imaging session, rather than scans from separate studies or different time points.

All examinations were acquired according to a standardized CT colonography protocol, including consistent patient preparation. In accordance with routine clinical practice, the time interval between the supine and prone acquisitions was typically less than 10 min, corresponding primarily to the time required for patient repositioning.

From this original dataset, we excluded scans with truncated Field of View (FoV), tilted positioning, or displaying significant tagged area for colonoscopy. Therefore, repeatability was evaluated using a test–retest data set from 205 patients from five different institutions.

Due to the unstandardized anonymization process, part of DICOM tags were removed at some sites. Kernel were Soft 56.7%; standard 40.3%; Slice thickness were 3 mm (38.9%), 2.5 mm (29.8%); 1.25 mm (23.4%) and manufacturer were Siemens (39.3%), Ge (31.3%) and unknown (29.4%); 89.6% of scan were 120KVp.

Reference data

Reference data (hereafter referred to as the ground truth) were derived from expert manual segmentations. After identification of the L3 vertebral level, a technologist with over 20 years of experience (T1) subjectively selected the most representative slice then manually delineated the skeletal muscle area (SMA). The segmented muscle groups included the psoas, quadratus lumborum, erector spinae, transversus abdominis, internal and external oblique muscles, and the rectus abdominis.

For each patient, a single CT series was selected. The operator was instructed to choose the intravenous contrast phase that best delineates muscle contours, preferentially the portal venous phase when available, as it represents the standard phase in oncologic follow-up.

All reference segmentations performed by T1 were subsequently reviewed by a board-certified radiologist with more than 20 years of experience. The resulting manual segmentations served both to construct the reference dataset and to evaluate intra- and inter-reader variability.

Assessments

In the sections below, we present our evaluations using SMA as a practical surrogate, because patient height was not available for a significant proportion of the data. Since converting SMA to L3SMI is straightforward—by dividing by the square of the patient's height (Eq. (1))—the corresponding L3SMI values and their the SMA measurements.

1. Intra and inter-readers variabilities

Technologist T1 repeated the segmentation after a two-month washout period to assess intra-variability. A second technologist (T2), 12-year experienced, repeated the same segmentations.

2. Visual quality of computer-generated SMA

A subjective segmentation quality scoring scale was designed with five categories: 1 = Perfect, 2 = Good, 3 = Fair, 4 = Poor, and 5 = Fail. The scoring scale was reviewed and endorsed by an experienced radiologist. Inter-reader agreement for this scale was evaluated by two imaging experts following a dedicated training session supported by standardized guidance and representative examples.

The outcomes of the visual assessment included the distribution of quality scores, a focused analysis of segmentations with scores > 3, and an overall evaluation of segmentations rated between 2 and 3.

Prior to analyzing the visual quality of the segmentations, we assessed the reliability of the scoring scale by computing inter-reader agreement after dichotomizing the scores into two categories: "Good" (scores 1–2) and "Not Good" (scores 3–5).

3. Accuracy

We computed the distribution of the signed difference and the relative difference between computed and reference SMA. We computed the average bias with confidence interval and performed an outliers' analysis to document system specifications.

4. Precision of computer-generated SMA

Repeatability was evaluated by repeating SMA measurements on each patient under test–retest conditions[20]. For each patient, the mean SMA value across the repeated measurements was calculated.
Reproducibility was assessed by investigating several sources of variability, including:

a) *Body Coverage*: SMA was computed on scans with varying body coverage and then recomputed on the same scans after cropping to the abdominal region. The paired SMA measurements were subsequently compared.

5. *Contrast phases*: Absolute errors in SMA measurements were computed for patients acquired in each of the four contrast phases. The mean absolute errors across the four groups were then compared.

Paired statistical comparisons were performed to assess the significance of factors potentially contributing to measurement variability. When applicable, bias analyses were conducted within each subgroup.

6. Linearity

We assessed the linearity between measured SMA (SMA_{Meas}) and reference SMA (SMA_{Ref}).

7. Exclusion criteria

We summarize a list of exclusion criteria that encompass visual and numerical outliers reported though trough the validation process.

Measurements procedures

Subjective selection of the axial slice considered most representative of the L3 vertebral level, as well as the manual segmentations used to construct the reference dataset, were performed using LifeX software[21] (v7.7.0). Multiplanar reconstruction (MPR) views were used to identify the L3 vertebra. In addition to manual contouring tools, intensity thresholding was applied to assist in SMA measurements.

Automatic segmentations were performed using a two-step pipeline. First, for optimization purposes, the Body Part Regression (BPR) mode[22] (v1.1) was applied to translate the radiological volume into a machine-interpretable representation, enabling detection of the L3 vertebral level and cropping of the region of interest. Second, the cropped region was processed using the TotalSegmentator freeware[23] (v2.2.0), with the muscle segmentation option enabled. A post-processing step including intensity thresholding in the range [−29; 150] HU was applied to exclude adipose tissue from the segmented muscle structures.

Statistical data analysis

All statistical analyses were performed using the R software environment[24], with a significance level set at 5%.

Inter-rater agreement was evaluated using the Prevalence- and Bias-Adjusted Kappa (PABAK) for ordinal scales[25] from the "epiR" package. PABAK is particularly suitable when the prevalence of categories is unbalanced.

Relative bias was calculated as follows:

$$Relative\ Bias\,(\%) = 100 * \frac{SMA_{Computed} - SMA_{Reference}}{SMA_{Reference}} \tag{2}$$

Outlier detection was performed using Grubbs' test (via the "envoutliers" package), and the normality of distributions was assessed with the Shapiro–Wilk test.

Confidence intervals of the mean were computed by bootstrapping the data with 1,000 replications using the "boot" package.

Repeatability was evaluated using the Within-Case Coefficient of Variation (wCV) from the "agRee" package, which also allowed derivation of the minimal detectable change (MDC)—the value below which the absolute difference between two repeated measurements is expected to lie with 95% probability. Since wCV is unaffected by linear transformations, it provides a robust measure for SMA measurements and ensures validity for L3SMI.

Repeatability was further illustrated using sample Bland–Altman plots (via the "blandr" package).

Reproducibility was assessed through paired nonparametric Wilcoxon tests.

Dunn's test was used for multiple pairwise comparisons.

Linearity of SMA—and consequently L3SMI—estimates was assessed using Passing-Bablok regression (via the "mcr" package), with the regression coefficient and R^2 value reported.

Given a linear relationship (b_1, b_0) between the reference SMA_{Ref} and the measured SMA_{Meas}:

$$SMA_{Ref} = \beta_1 * SMA_{Meas} + \beta_0 \tag{3}$$

Dividing both sides of the equation by the square of the patient's height (h^2) yields:

$$L3SMI_{Ref} = \beta_1 * L3SMI_{Meas} + \beta_0 / h^2 \tag{4}$$

Finally, we provided point estimates of:

- Confidence Interval attached to a SMA estimation is given by:

$$SMA \pm 1.96 \times wCV \times SMA \tag{5}$$

wCV computed within the adequate SMA range.

- A SMA value has significantly changed if the percentage change is greater than:

$$RC = 2.77 \times wCV \times 100 \tag{6}$$

wCV computed within the adequate SMA range.

- The change in time of SMA measurements is given by:

$$(SMA_2 - SMA_1) \pm 1.96 * \sqrt{[SMA_1 * wCV_1]^2 + [SMA_2 * wCV_2]^2} \tag{7}$$

wCV_1 and wCV_2 are drawn from respective ranges of SMA1 and SMA2 values.

This retrospective study was conducted in accordance with the Declaration of Helsinki. All imaging data were fully anonymized prior to analysis. According to local regulations, the study was exempt from informed consent and institutional review board (IRB) approval.

Results

Intra/inter readers agreement

Intra-reader variability was wCV = 2.6% (95%CI: 1.7; 3.4), therefore, considering wCV free from scaling effect, reference L3SMI values measured by Technologist T1 can be approximated as:

$$L3SMI_{Manual-Intra} \pm 0.05 \times L3SMI_{Manual-Intra} \tag{8}$$

Inter-reader variability was wCV = 3.9% (95%CI: 2.6; 5.3).

Bland Altman display for intra and inter reader variability are presented in Fig. 1.

Visual scoring

The paired scoring of the 60 segmentations by the two readers showed a substantial agreement of 0.67 (95%CI; 0.43; 0.83).

Overall, 90% of the segmentations were deemed satisfactory (Perfect: 12.5%; Good: 77.5%). Full detail in supplementary figure S3.

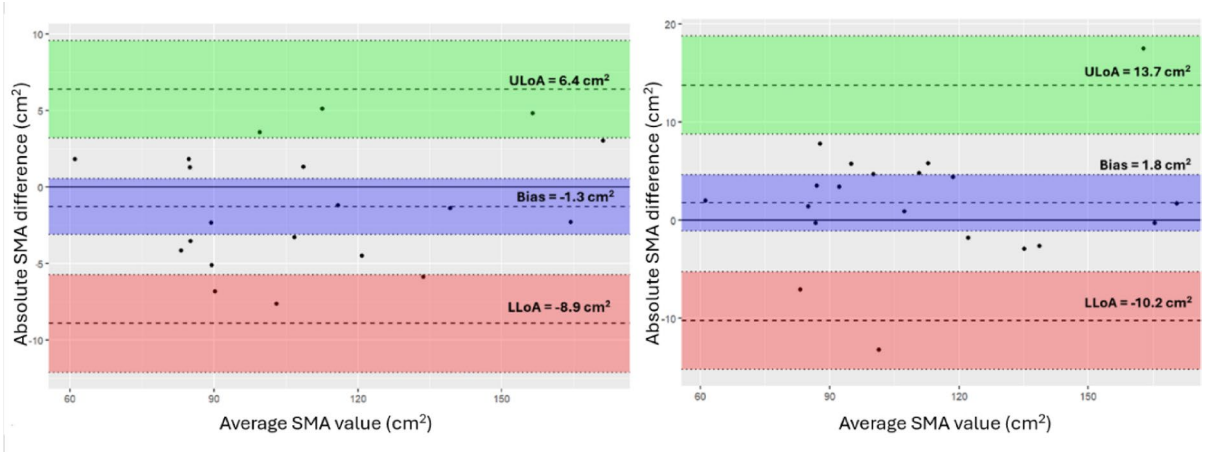

Fig. 1. Reader agreement of reference SMA values. Bland Altman displays of repeated (N = 20) reference SMA estimations. Left: A seasoned technologist (T1) repeated the segmentations, the minimal detection change was 7.8% (95%CI: 6.0; 11.4) and RC = 7.15 cm^2. Right: Two seasoned technologists (T1 & T2) reproduced the segmentations, the minimal detection change was 10.9% (95%CI: 7.9; 15.4) and RC = 12.2cm^2.

In the following, we excluded 5 scans for clinical or conformity reasons: patients' arms were set along the thorax, suspected severe liver disease or enlarged liver. These observations will contribute to documenting exclusion criteria.

Accuracy

Out of the original 100 patient scans, 3 were excluded due to anatomical limitations: an enlarged liver extending to L3 (with similar density to muscle) and arms positioned alongside the body. Additionally, two scans were identified as outliers for which segmentation results diverged for unknown reasons. The final analysis included 95 patients.

The relative bias was not normally distributed (Shapiro–Wilk, $p = 0.007$), with a mean of − 8.5% (95% CI: − 9.8, − 7.3). As shown in Fig. 2, the distribution of relative differences indicates a quasi-systematic under-segmentation of the SMA masks. Visual inspection revealed that under-segmentation was most pronounced in the rectus abdominis, internal oblique, and transversus abdominis muscles.

Bland Altman analysis gives an average bias of − 11.3cm^2 (95%CI: − 13.2, − 9.4) with upper and lower limits of agreement of 7.0 cm^2 (95%CI: 3.8, 10.3) and -29.7cm^2 (95%CI: − 33.0, − 26.4) and a regression equation of the bias as: "y(differences) = − 0.2 x(means) + 12" indicating a significant proportional bias.

Precision

Repeatability The mean SMA values of the 205 repeated measurements ranged from 66.3 cm^2 to 213.9 cm^2, with a median of 107.3 cm^2 (Supplementary Figure S4). The raw differences in SMA values were normally distributed (Shapiro–Wilk, $p = 0.66$).

The within-case coefficient of variation (wCV) was 3.2% (95% CI: 2.8–3.5%), and the smallest detectable change, also referred to as the Repeatability Coefficient (RC), was 8.8% (95% CI: 7.9–9.7%). The mean bias did not differ from zero, indicating no systematic difference between prone and supine measurements.

The corresponding Bland–Altman plot is presented in Fig. 3.

We further sampled wCV according to the quartiles of the mean of the repeated measurements.

Reproducibility

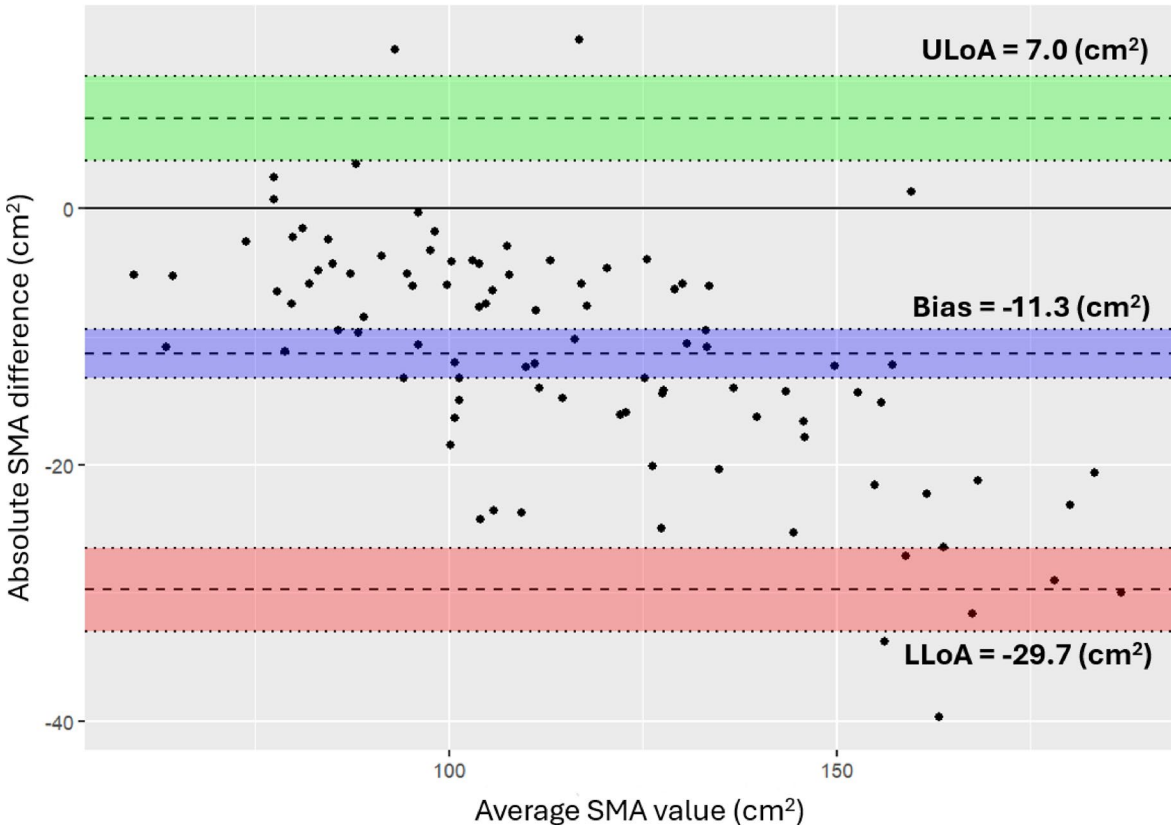

Fig. 2. Bland–Altman analysis of bias between manual and software-based SMA measurements. Bias value (dashed line; 95%CI purple area) is significantly different from zero (− 11.3 cm^2). Higher boundary (dashed line; 95%CI green area), Lower boundary (dashed line; 95%CI red area). We found a LoA of + /- 18.35 cm^2.

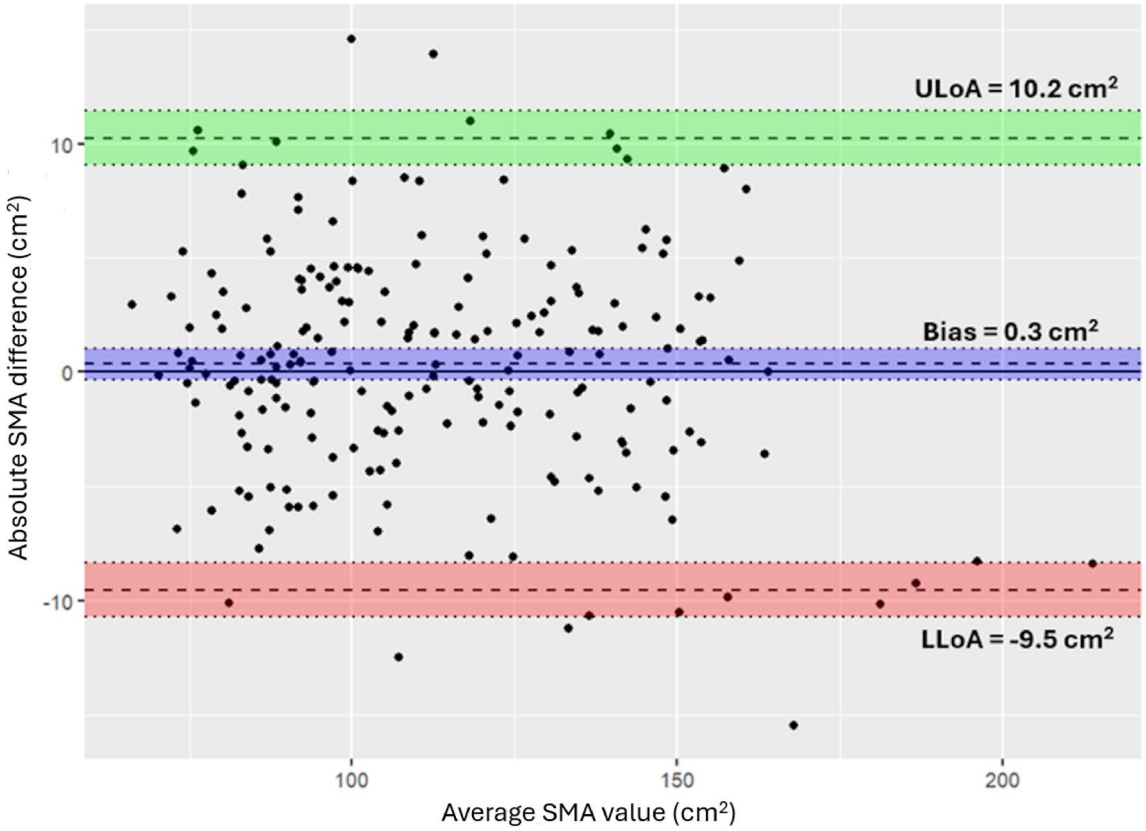

Fig. 3. Bland Altman analysis of repeated SMA measurements. Bias value (dashed line; 95%CI purple area) is not significantly different from zero. Higher boundary (dashed line; 95%CI green area), Lower boundary (dashed line; 95%CI red area). We found a LoA of + / − 9.8 cm^2 with Min/max = [− 15.5; 14.6] cm^2.

1) *Field of View*: We selected 30 scans from Dataset 1 and cropped them from the liver tip (ensuring the entire liver was included) down to the sacrum (S1). Anatomical coverage of the cropped scans was as follows: TAP: 13/30, AP: 16/30, TA: 1/30.

No significant differences were observed in SMA values between the original and cropped scans (paired comparison, $p = 0.06$).

1) *Contrast agent*: Comparing the average SMA values across different contrast phases (Fig. 4: No IV = 14, Arterial = 50, Portal = 14, Late = 2), we found no significant differences (Kruskal–Wallis test, $p = 0.18$), nor in multiple pairwise comparisons using Dunn's test with Bonferroni correction.

Linearity Considering that the overall ratio of intra-reader wCV to repeatability wCV is 1.2 (Fig. 5), the figure illustrates the linear relationship between SMA measurements (SMA_{Meas}) and the reference SMA (SMA_{Ref}).

As shown in Eq. (4), because L3SMI is proportional to SMA divided by the square of the patient's height, the regression from SMA to L3SMI can be directly derived.

Exclusion criteria Exclusion criteria are summarized in Fig. 6 below.

Discussion

Visual inspection indicated that 90% of segmentations were acceptable, with no failures. Manual segmentation exhibited intra- and inter-reader variability of 5% and 7.6%, respectively. Automated segmentation achieved a global wCV of 3.2%, decreasing from 4.0 to 2.7% with increasing SMA. Computed SMA showed a linear relationship with the reference standard. No significant reproducibility issues were observed for body region ($p = 0.06$) or contrast phase ($q > 0.05$). These results also informed the definition of exclusion criteria for L3SMI assessment using the "TotalSegmentator" software.

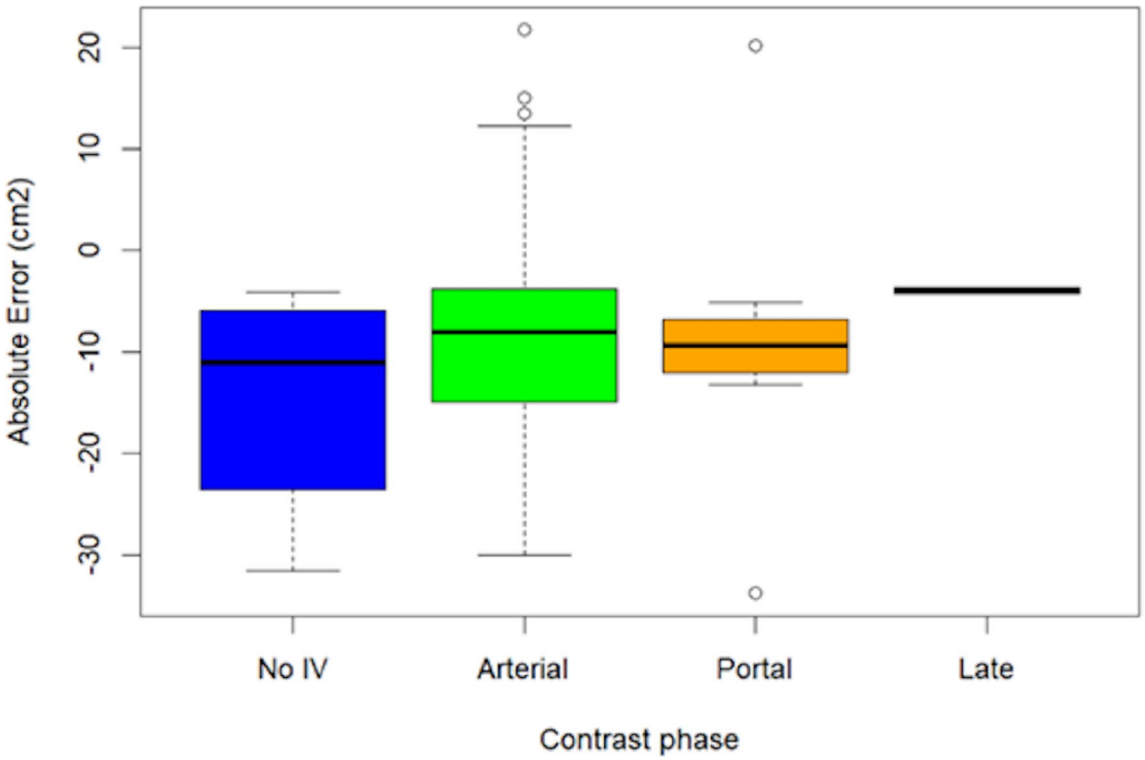

Fig. 4. Distribution of absolute error across contrast phases. We considered 80 absolute errors of the SMA estimations (Pre: 14; Arterial: 50; Portal: 14 and late: 2) and found no significant differences between the average values.

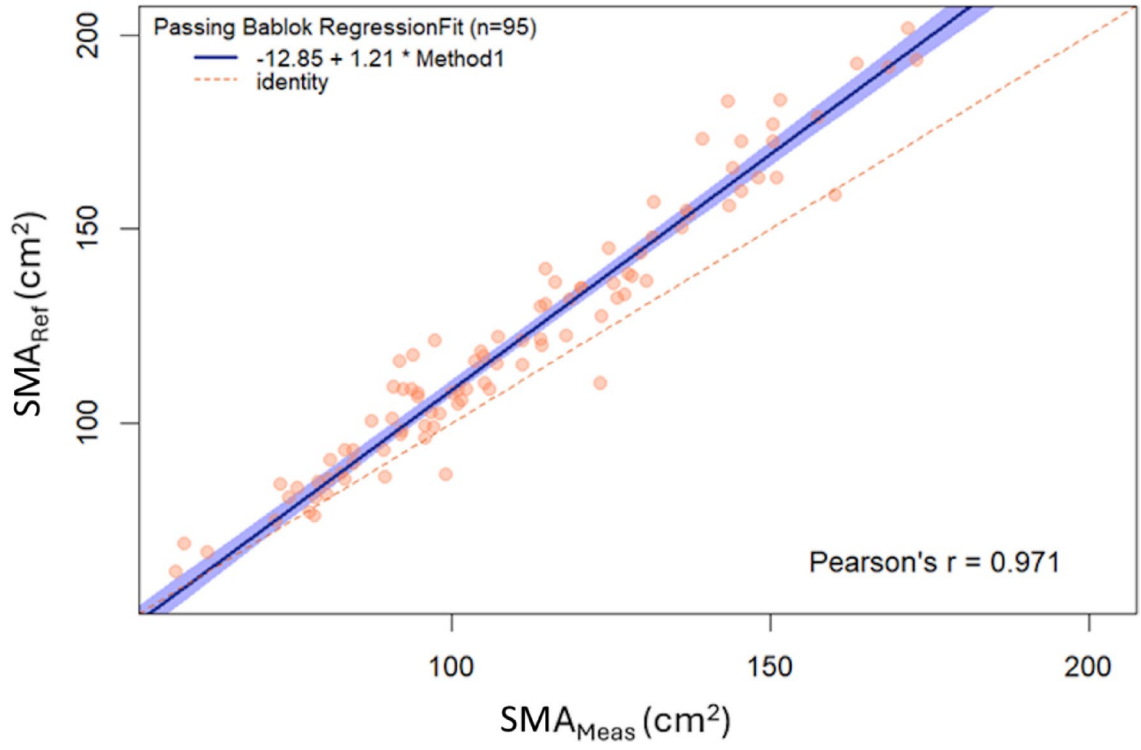

Fig. 5. Regression between SMA_{Meas} and SMA_{Ref}. We overlay the data with the regression line (blue) that featured a Pearsons' correlation coefficient of 0.97, $b_0 = -12.8$ (95%CI: – 19.6, -6.4) intercept and a $b_1 = 1.21$ (95CI: 1.16, 1.27) slope and the line of equivalence (dashed orange line).

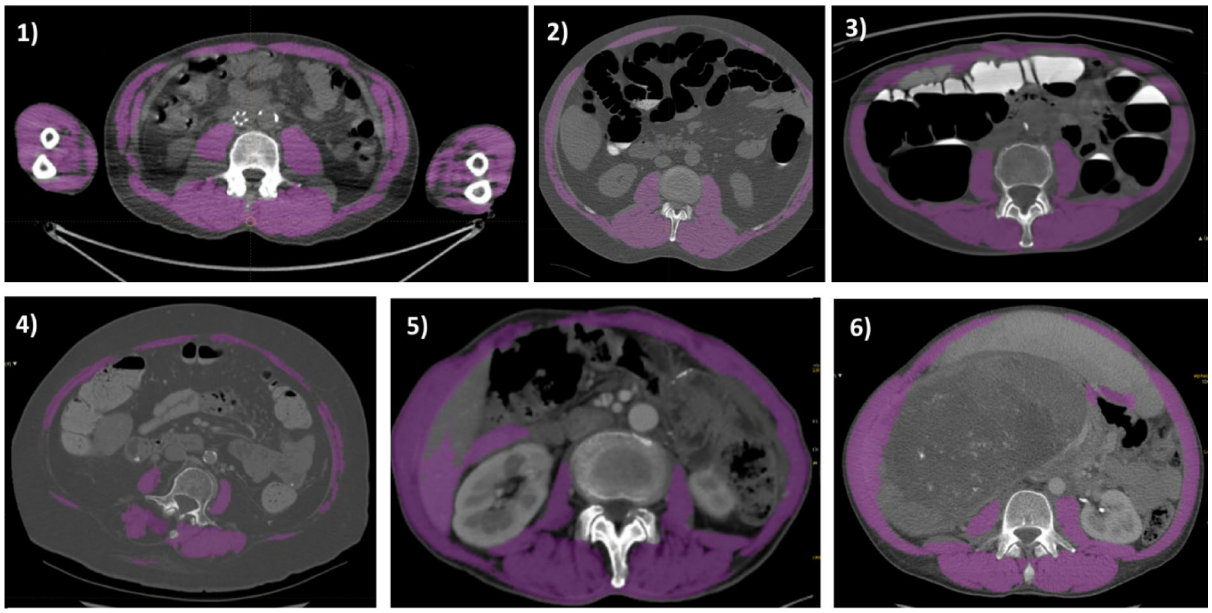

Fig. 6. Exclusion criteria. Visual inspection confirmed four main exclusion criteria: (**1**) Arms along the body, (**2**) Truncated FoV, (**3**) Significant tagged area, (**4**) Tilted body, (**5**) Enlarged liver/suspected steatosis, 6) Severe liver disease. Muscles are displayed in pink.

Comparison to literature

Intra/inter variability

We observed an intra-reader wCV of 2.6% and an inter-reader wCV of 3.9%. The corresponding inter-reader variability of 7.6% was slightly higher than the 5.1% reported by Borrelli et al.[26], in contrast, Park et al.[27] reported an inter-reader wCV of 8.2%. These differences may be partially explained by variations in the proportion of artifacts within the field of view and the distribution of SMA values: in Park et al., 28% of scans contained artifacts and the mean SMA was 137 cm^2, whereas in our study only 5% of scans were affected and the mean SMA was 120 cm^2.

Bias

Nowake et al.[28], applied a similar pipeline for L3 detection and muscle segmentation, reporting SMA differences of 3.7 ± 4.1 cm^2 ($2.7\% \pm 4.4\%$) and 5.4 ± 6.4 cm^2 ($3.5\% \pm 4.0\%$) across two centers, without assessing linearity. In comparison, our study observed less variable bias but a larger mean bias, which we corrected using a linear transformation.

Gomez-Perez et al.[29] performed a Bland–Altman analysis comparing neural network-derived SMA measurements[30] to 420 manual references, reporting limits of agreement (LoA) of [−29.9; 37.7] cm^2 and a proportional bias with a mean of 3.85 cm^2. Our analysis found narrower LoA ([−29.7; 7.0] cm^2) but also a proportional bias, which required linearity estimation. On average, our bias had a larger absolute value (−11.3 cm^2).

In repeatability analysis, Borrelli et al. reported a relative SMA difference of 3.9% (Min/Max: 0.0–15.9) between two time points at least three days apart, approximating a test–retest scenario.

A reproducibility meta-analysis by Lortie et al.[31] reported that contrast in CT images caused only a minimal SMA change of 2.58%, while variations in kilovoltage (80–140 kVp) had little effect on segmentation. Similarly, Fuchs et al.[32] observed negligible influences (~1%) from contrast, tube voltage, and slice thickness. In our study, we found no detectable effect of IV contrast on SMA measurements, with most scans acquired at 120 kVp and 5 mm slice thickness. Our primary aim was to validate the BPR software and assess the impact of body coverage. The results indicate no compromise in L3 vertebra detection, consistent with current state-of-the-art performance standards 33, and confirm the robustness of our segmentation pipeline.

Our visual inspection identified conditions where L3SMI evaluations may be suboptimal. As reported by others, limitations arise from patient positioning issues, such as truncated FoV, partial waist exclusion[33], or lateral decubitus, which can squeeze muscles and arms into the field of view[34]. We also observed suspected steatosis on non-enhanced scans. Normal liver attenuation is ~64 HU, while fatty liver is 39.3 ± 6.4 HU[35], overlapping with muscle attenuation (30.3 ± 8.8 HU)[36]. This overlap may lead to SMA overestimation in non-enhanced scans, highlighting a potential limitation in segmentation accuracy under these conditions.

Given that the prevalence hepatic steatosis into the obese population is reported as 65% of subjects with grade I-II obesity and in 85% of patients with grade III[37], the use of non-enhanced scans would be discouraged in the obese population.

Artefacts can impair L3SMI evaluations as we observed some cases of fecal tagging "cutting" part of the muscles. Other groups illustrated the impact of CT "streak"[34] or cupping artifacts making impossible to threshold muscular tissue thus invalidating quantification[38].

Hou et al. compared their proprietary software to TS[39] using 882 SAROS dataset patients[40], focusing on segmented axial volumes rather than SMA. They found both methods performed similarly, with TS showing high linearity against GT ($\beta_1 = 0.80$, $R^2 = 0.967$). Our study yielded comparable results ($\beta_1 = 0.83$, Pearson $r = 0.971$). Hou et al. also confirmed TS's tendency for sub-segmentation but analyzed all muscle regions rather than focusing on L3. In contrast, our validation examines a complete pipeline, including BPR software, TS, and a post-processing step, ensuring a more targeted assessment of segmentation performance.

Our study has some limitations

First, repeatability was evaluated using prone–supine image pairs rather than conventional same-position test–retest acquisitions. Changes in posture introduce systematic, non-random anatomical deformation (e.g., gravity-dependent soft-tissue displacement and organ reconfiguration), which differ fundamentally from the stochastic variability typically captured in standard test–retest experiments. As a result, the observed variability reflects a combination of measurement-related uncertainty (segmentation effects) and true posture-induced geometric changes. These sources of variation are not strictly exchangeable with conventional repeatability metrics obtained under identical positioning.

Consequently, our repeatability estimates should be interpreted as a conservative upper bound on measurement variability under controlled same-position conditions, rather than as a pure test–retest metric. While this design may better reflect positional variability encountered in practice, it may also inflate variability estimates relative to standard repeatability frameworks.

Second, dataset 1 had varied acquisition parameters, while dataset 2 was more homogeneous, limiting consistency in accuracy and precision assessments.

Third, dataset 1 included NSCLC, adrenal, pancreatic, and CRC patients, whereas dataset 2 had only CRC patients. Disease-specific factors (e.g., fecal tagging for CRC, steatosis in obesity) may affect inclusion/exclusion criteria, so further validation with diverse disease datasets is needed to ensure generalizability.

Fourth, our work relied on references derived from a subjective expert consensus, involving both the selection of the axial slice considered most representative of the L3 vertebral level and the subsequent manual segmentation of skeletal muscles on that slice. We fully acknowledge the limitations associated with the subjectivity inherent to this approach and the resulting uncertainty regarding the reliability of the reference. However, to the best of our knowledge, no alternative methodology providing a more robust or convincing reference standard has been proposed to date.

Fifth, although third-party software tools were used, all validation, annotation, and analysis were conducted internally without independent external verification. This may introduce observer- or institution-specific bias and limits the generalizability of the results. Independent validation and multi-center studies will be needed to confirm the robustness and reproducibility of the findings.

Body composition measurements are proved to be useful in[8] and outside of oncology. For those applications diverse cut off values for sarcopenia were estimated as for the selection of patient in end-stage liver disease (SMI < 50 cm2/m2 for men and < 39 cm2/m2 for women)[41] or the prognostication of recurrent ovarian cancer (35.6 cm2/m2)[10] or those of CRC patients male ≤ 38.89cm2/m2, female ≤ 33.28cm2/m2)[42]. Other thresholds aimed at selecting.

advanced lung cancer[43] or chronic kidney disease patients[44].

Furthermore, a consensus has been established to adjust these thresholds based on key factors such as gender, age, and ethnicity. Consequently, precise and accurate assessment of L3SMI thresholds is essential to ensure optimal sensitivity and specificity in sarcopenia detection for its various clinical applications.

To go further, some improvements are foreseen: 1) As we identified some exclusion criteria, the implementation of an automatic Quality Control system[28] will alleviate human review of segmentation mask, at least rising warnings; 2) Improving the SMA segmentation in enabling post processing involving, for instance, convex hull[45] processing that can cope with the observed under segmentation.

Finally, our study focused on L3SMI, a biomarker whose clinical relevance has been extensively validated. Nevertheless, several studies have also demonstrated the added value of assessing skeletal muscle density[46]. The technical validation of muscle density as a QIB remains to be performed and represents an important direction for future work, potentially in combination with L3SMI to provide a more comprehensive characterization of muscle status.

Conclusion

We provide data aiming at accurately and precisely evaluating the L3SMI quantitative biomarker. The confidence associated with the measurement allows to involve this biomarker in a plurality of clinical settings where classification thresholds have been designed. Irrespective of the clinical settings, our work will allow to detect any significant change of the L3SMI value. Finally, users can refer to our list of exclusion/inclusion for proper use of L3SMI.

Data availability

The datasets used and analysed during the current study available from the corresponding author on reasonable request. All data generated or analysed during this study are included in this published article and its supplementary information files.

Received: 16 October 2025; Accepted: 24 April 2026
Published online: 12 May 2026

References

1. Lee, K. et al. Recent issues on body composition imaging for sarcopenia evaluation. *Korean J. Radiol.* **20**, 205–217 (2019).
2. Davydov, D. M., Boev, A. & Gorbunov, S. Making the choice between bioelectrical impedance measures for body hydration status assessment. *Sci. Rep.* **11**, 1–13 (2021).
3. Heymsfield, S. B. Advances in body composition: A 100-year journey. *Int. J. Obes.* https://doi.org/10.1038/s41366-024-01511-9 (2024).
4. Tagliafico, A. S., Bignotti, B., Torri, L. & Rossi, F. Sarcopenia: How to measure, when and why. *Radiol. Med.* **127**, 228–237 (2022).
5. Gielen, E., Dupont, J., Dejaeger, M. & Laurent, M. R. Sarcopenia, osteoporosis and frailty. *Metabolism* **145**, 155638 (2023).
6. Beaudart, C., Rizzoli, R., Bruyère, O., Reginster, J. Y. & Biver, E. Sarcopenia: Burden and challenges for public health. *Arch. Public Health.* **72**, 1–8 (2014).
7. Williams, G. R. et al. The impact of skeletal muscle on the pharmacokinetics and toxicity of 5-fluorouracil in colorectal cancer. *Cancer Chemotherapy Pharmacol.* **81**(2), 413–417 (2018).
8. Hilmi, M. et al. Body composition and sarcopenia: The next-generation of personalized oncology and pharmacology?. *Pharmacol. Ther.* https://doi.org/10.1016/j.pharmthera.2018.12.003 (2019).
9. Schmulenson, E. et al. Influence of the skeletal muscle index on pharmacokinetics and toxicity of fluorouracil. *Cancer Med.* **12**, 2580–2589 (2023).
10. Aichi, M. et al. Sarcopenia shortens overall survival of patients with platinum-resistant recurrent ovarian cancer: inverse probability of treatment-weighting analysis. *Int. J. Gynecol. Cancer* https://doi.org/10.1136/ijgc-2024-005323 (2025).
11. Jogiat, U. M. et al. Sarcopenia determined by skeletal muscle index predicts overall survival, disease-free survival, and postoperative complications in resectable esophageal cancer. *Ann. Surg.* **276**, e311–e318 (2022).
12. Yang, Y. R. et al. The impact of sarcopenia on overall survival in patients with pan-RAS wild-type colorectal liver metastasis receiving hepatectomy. *Sci. Rep.* **13**, 1–7 (2023).
13. Ackermans, L. L. G. C. et al. Screening, diagnosis and monitoring of sarcopenia: When to use which tool?. *Clin. Nutr. ESPEN.* **48**, 36–44 (2022).
14. Chianca, V. et al. Sarcopenia: imaging assessment and clinical application. *Abdom. Radiol.* **47**, 3205–3216 (2022).
15. Messina, C. et al. Prevalence and type of errors in dual-energy x-ray absorptiometry. *Eur. Radiol.* **25**, 1504–1511 (2015).
16. Huber, F. A., Del Grande, F., Rizzo, S., Guglielmi, G. & Guggenberger, R. MRI in the assessment of adipose tissues and muscle composition: How to use it. *Quant. Imaging Med. Surg.* **10**, 1636–1649 (2020).
17. Buckler, A. J., Bresolin, L., Dunnick, N. R. & Sullivan, D. C. A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. *Radiology* **258**, 906–914 (2011).
18. Obuchowski, N. A. et al. Statistical considerations for planning clinical trials with quantitative imaging biomarkers. *J. Natl. Cancer Inst.* **111**, 19–26 (2019).
19. Clark, K. et al. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).
20. Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R. & Andreou, P. The case for using the repeatability coefficient when calculating test-retest reliability. *PLoS ONE* **8**, 1–7 (2013).
21. Nioche, C. et al. Lifex: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Res.* **78**, 4786–4789 (2018).
22. Yan, K., Lu, L. & Summers, R. M. Unsupervised body part regression via spatially self-ordering convolutional neural networks. *Proc. - Int. Symp. Biomed. Imaging*. 1022–1025 (2018).
23. Wasserthal, J. et al. TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiol. Artif. Intell.* **5**, 272–284 (2023).
24. Team, R. D. C. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. https://doi.org/10.1007/978-3-540-74686-7 (2011).
25. Byrt, T., Bishop, J. & Carlin, J. B. Bias, prevalence and kappa. *J. Clin. Epidemiol.* **46**, 423–429 (1993).
26. Borrelli, P. et al. Artificial intelligence-aided CT segmentation for body composition analysis: A validation study. *Eur. Radiol. Exp.* https://doi.org/10.1186/s41747-021-00210-8 (2021).
27. Park, J. et al. Reliable and robust method for abdominal muscle mass quantification using CT/MRI: An explorative study in healthy subjects. *PLoS ONE* **14**, 1–14 (2019).
28. Nowak, S. et al. End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT. *Eur. Radiol.* **32**, 3142–3151 (2022).
29. Gomez-Perez, S. L. et al. Concordance of computed tomography regional body composition analysis using a fully automated open-source neural network versus a reference semi-automated program with manual correction. *Sensors* **22**, 1–15 (2022).
30. Paris, M. T. et al. Automated body composition analysis of clinically acquired computed tomography scans using neural networks. *Clin. Nutr.* **39**, 3049–3055 (2020).
31. Lortie, J. et al. The effect of computed tomography parameters on sarcopenia and myosteatosis assessment: A scoping review. *J. Cachexia Sarcopenia Muscle* **13**, 2807–2819 (2022).
32. Fuchs, G. et al. Quantifying the effect of slice thickness, intravenous contrast and tube current on muscle segmentation: Implications for body composition analysis. *Eur. Radiol.* **28**, 2455–2463 (2018).
33. Zaffina, C. et al. Body composition assessment: Comparison of quantitative values between magnetic resonance imaging and computed tomography. *Quant. Imaging Med. Surg.* **12**, 1450–1466 (2022).
34. Ackermans, L. L. G. C. et al. Clinical evaluation of automated segmentation for body composition analysis on abdominal L3 CT slices in polytrauma patients. *Injury* **53**, S30–S41 (2022).
35. Alshoabi, S. A. et al. Grading of fatty liver based on computed tomography Hounsfield unit values versus ultrasonography grading. *Gastroenterol. Insights* **15**, 588–598 (2024).
36. Van den Broeck, J. et al. The correlation of muscle quantity and quality between all vertebra levels and level L3, measured with CT: An exploratory study. *Front. Nutr.* **10**, 1–8 (2023).
37. Divella, R., Mazzocca, A., Daniele, A., Sabbà, C. & Paradiso, A. Obesity, nonalcoholic fatty liver disease and adipocytokines network in promotion of cancer. *Int. J. Biol. Sci.* **15**, 610–616 (2019).
38. 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.* **31**, 1795–1804 (2021).
39. Hou, B., Mathai, T. S., Liu, J., Parnell, C. & Summers, R. M. Enhanced muscle and fat segmentation for CT-based body composition analysis: A comparative study. *Int. J. Comput. Assist. Radiol. Surg.* **19**, 1589–1596 (2024).
40. Koitka, S. et al. SAROS: A dataset for whole-body region and organ segmentation in CT imaging. *Sci. Data* **11**, 1–10 (2024).
41. Carey, E. J. et al. A multicenter study to define sarcopenia in patients with end-stage liver disease. *Liver Transpl.* **23**, 625–633 (2017).
42. Wang, S. et al. The value of L3 skeletal muscle index in evaluating preoperative nutritional risk and long-term prognosis in colorectal cancer patients. *Sci. Rep.* **10**, 1–11 (2020).

43. Liu, X. et al. The correlation between skeletal muscle index of the L3 vertebral body and malnutrition in patients with advanced lung cancer. *BMC Cancer* **21**, 1–9 (2021).
44. Sun, K. D. & Hwan, P. J. Effect of low muscle mass on total mortality related to metabolic disease in patients with chronic kidney disease in US adults. *Endocr. Abstr.* https://doi.org/10.1530/endoabs.99.ep661 (2024).
45. Toussaint, G. T. A historical note on convex hull finding algorithms. *Pattern Recognit. Lett.* **3**, 21–28 (1985).
46. Lee, K. et al. Skeletal muscle density as a predictor of prognosis and physical reserve in patients with cancer of unknown primary. *J. Clin. Med.* **14**, 2947 (2025).

Author contributions
H.B wrote the main manuscript text E.K. performed measurements E.K. and H.B analysed the data A.T. curated data and run software H.B. and A.I. managed the project and edited the final version of the manuscript.

Funding
The authors state that this work has not received any funding.

Declarations

Ethical approval
Institutional Review Board approval was obtained from Median Technologies.

Competing interests
H.B., E.K., A.T. and A.I. are employees at Median Technologies.

Informed consent
In accordance with the decision of the Median Technologies Ethics Committee, the requirement for informed consent was waived. Written informed consent was not required, as the study involved retrospective analysis of previously collected data.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-51010-2.

Correspondence and requests for materials should be addressed to H.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.