Validation of a Body-Composition Segmentation Software (Soma) on a Diverse Cohort of Publicly Available CT Scans

Protocol Version	1.0
Date	April 10, 2026
Sponsor	Nucleo Research, Inc.
Principal Investigator	Luca Pegolotti, CTO

Nucleo Research, Inc.

Contents

1. Purpose **3**
Intended use . . . 3

2. Device Description **3**
Segmentation model . . . 3
L3 vertebra detection model . . . 4
Segmentation example . . . 4

3. Test Data **4**
Candidate pool . . . 4
Training data isolation . . . 6
200-case cohort selection . . . 6
Inclusion and exclusion criteria . . . 9

4. Reference Standard **9**

5. Acceptance Criteria **9**
Aggregate performance . . . 9
L3 localization accuracy . . . 10
Per-subgroup performance . . . 10
Secondary endpoints . . . 11

6. Statistical Analysis **11**

7. Study Conduct **12**

8. Ethical Considerations **12**

9. Known Limitations and Risks **12**

References **13**

1. Purpose

This document describes the design of a standalone performance evaluation for Soma, a software device that (1) automatically segments four body composition tissue classes from CT scans (skeletal muscle, subcutaneous adipose tissue/SAT, visceral adipose tissue/VAT, and intramuscular adipose tissue/IMAT) and (2) identifies the axial slice corresponding to the L3 vertebral level for standardized body composition measurement [1, 2]. The study will compare Soma's output against expert radiologist annotations on a curated dataset of 200 CT scans drawn from six independent, publicly available collections.

The goal is to demonstrate that Soma produces clinically acceptable segmentation accuracy across a representative range of patient demographics, scanner hardware, and clinical contexts. Performance must meet pre-specified acceptance thresholds not only in aggregate but within every demographic and technical subgroup of sufficient size, so that no population is left inadequately validated.

Intended use

Soma is intended to provide automated segmentation and quantification of body composition tissues (skeletal muscle, subcutaneous adipose tissue, visceral adipose tissue, and intramuscular adipose tissue) from CT images, to assist clinicians in the assessment of sarcopenia, cachexia, and metabolic risk.

2. Device Description

Soma accepts axial CT images as input and performs two functions. First, it produces a voxel-level segmentation map assigning each pixel to one of five classes: background (0), skeletal muscle (1), SAT (2), VAT (3), or IMAT (4). Second, it automatically identifies the axial slice corresponding to the L3 vertebral level. Together, these enable computation of downstream quantities such as tissue cross-sectional area at L3, total tissue volume, and (when patient height is available) the skeletal muscle index (SMI = L3 muscle area / height^2) for sarcopenia assessment [1, 3].

Segmentation model

The segmentation component is a U-Net convolutional neural network [4] with a four-block encoder (1->64->128->256->512 channels), a 1024-channel bottleneck, and a four-block decoder with skip connections. Input is a single-channel 480×480 axial CT image with Hounsfield unit values clipped to $[-190, 150]$ and normalized using z-score standardization (mean -115.78, std 103.15). Output is a five-class segmentation map. The model was trained on labeled data from the SAROS [5] and TotalSegmentator [6] collections using a combined cross-entropy and Dice loss [7] (0.5 each), with tissue-specific HU post-processing to enforce physiologically plausible attenuation ranges [1, 8].

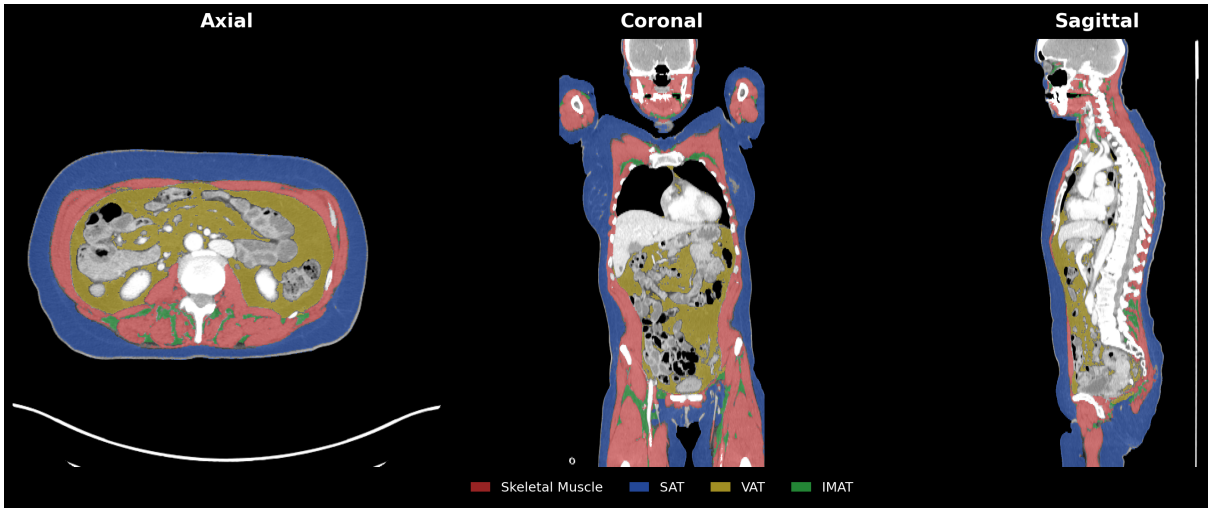

Figure 1: Soma automated segmentation in axial, coronal, and sagittal views. Tissue classes: skeletal muscle (red), SAT (blue), VAT (yellow), IMAT (green).

L3 vertebra detection model

The L3 detection component is a two-stage pipeline. Stage 1 is an EfficientNet-Lite0 CNN [9] (~3.4M parameters) that classifies each axial slice independently using three-channel input (target slice ± 1 adjacent neighbor for anatomical context) at 224×224 resolution. It produces a 26-class probability distribution over vertebral levels (C1–C7, T1–T12, L1–L5, sacrum, background) and a 1280-dimensional feature embedding per slice. Stage 2 is a two-layer bidirectional LSTM (BiLSTM; 256 hidden units per direction) that processes the ordered sequence of CNN embeddings for all slices in a scan, capturing anatomical ordering constraints inaccessible to the per-slice CNN. The BiLSTM was trained with a custom hit-rate loss that directly optimizes for each vertebra's peak probability falling within the correct anatomical range. L3 is identified as the slice with maximum L3-class probability, using a top-2 midpoint inference strategy (average of the two highest L3-probability slice indices). Both stages were trained on 2,186 subjects from TotalSegmentator [6] and SAROS [5].

Segmentation example

Figure 1 shows Soma's automated segmentation of a representative whole-body CT scan in axial, coronal, and sagittal views. Tissue compartments are color-coded: skeletal muscle (red), subcutaneous adipose tissue (blue), visceral adipose tissue (yellow), and intramuscular adipose tissue (green). The axial view is taken at the abdominal level where all four tissue compartments are visible.

3. Test Data

Candidate pool

The evaluation draws from a pool of 2,066 CT scans across six publicly available datasets, none of which were used during Soma training. The datasets were selected to span a range of

body regions (whole-body and abdomen-only), scanner manufacturers (Siemens, GE, Philips, Toshiba), geographic origins (Germany, China, USA, Italy, and 14 additional countries via RATIC), and clinical contexts (healthy controls, oncology staging, surgical planning).

autoPET (900 scans) is the largest contributor. These are whole-body FDG-PET/CT attenuation correction CTs acquired on a Siemens Biograph128 mCT at the University Hospital Tübingen, published by Gatidis et al. [10]. The collection includes 426 cancer-free negative controls and 474 patients with melanoma, lung cancer, or lymphoma. Full demographics are available for every subject: age (mean 59.7 ± 15.9, range 11–95), sex (500 male, 396 female), and diagnosis. BMI is available for a subset of 130 subjects where unique matching to the source demographic registry was possible.

AMOS (500 scans) provides multi-scanner abdominal CTs from two Chinese hospitals, acquired on eight scanner models spanning all four major manufacturers [11]. Age, sex, and scanner metadata are available for a 20-subject subset selected for an earlier validation phase. The remaining 480 subjects contribute scanner and population diversity without individual demographics.

MSD Pancreas (420 scans) consists of portal venous phase abdominal CTs from Memorial Sloan Kettering Cancer Center at a standardized 120 kVp protocol [12]. No per-subject demographics are available, but the standardized single-center acquisition provides a controlled technical baseline.

CT-ORG (140 scans) is an international multi-center collection with variable body coverage from abdomen to whole-body, including a PET/CT subset from Stanford [13]. No per-subject demographics are available.

ENHANCE.PET (56 scans) comprises whole-body PET/CT attenuation CTs from Leipzig, Germany (Siemens) and Florence, Italy (Philips), with full demographics including age (mean 67.0, range 26–91), sex, weight, height, and BMI (mean 25.4) for all subjects [14].

RATIC (50 scans) is an injury-negative subset of the Radiology AI Test Image Collection, drawn from 23 institutions across 14 countries [15]. Age (mean 62.9, range 18–90) and sex (28 male, 22 female) are known for all subjects. RATIC provides the strongest scanner diversity of any single source in the pool.

The following table summarizes the key characteristics of each dataset:

Dataset	N	Body Region	Demographics	Scanner(s)	Geography
autoPET	900	Whole-body	Age, sex; BMI for 130	Siemens Biograph128 mCT	Tübingen, Germany
AMOS	500	Abdomen/pelvis	20 subjects: age, sex	8 models (Siemens/GE/Philips/Toshiba)	China
MSD Pancreas	420	Abdomen	None	Single scanner, 120 kVp	USA (MSKCC)

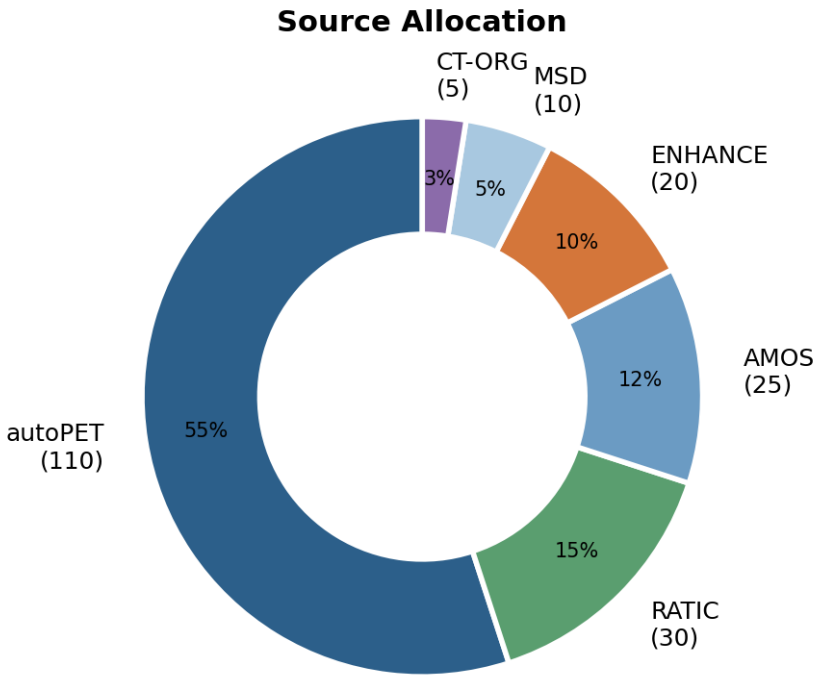

Figure 2: Source allocation

Dataset	N	Body Region	Demographics	Scanner(s)	Geography
CT-ORG	140	Abdomen–whole body	None	Multiple	International (8 centers)
ENHANCE.PET	56	Whole-body	Age, sex, BMI	Siemens + Philips	Germany, Italy
RATIC	50	Abdomen/pelvis	Age, sex	23+ models	14 countries

Training data isolation

None of the six datasets were used in Soma training. Soma was trained on the SAROS [5] and TotalSegmentator [6] datasets, both of which draw exclusively from distinct source institutions. All six evaluation datasets are independently published collections with no overlap with either training source.

200-case cohort selection

From the 2,066-scan pool, 200 cases will be selected using stratified sampling designed to ensure that every clinically relevant demographic subgroup contains at least 20 subjects. This minimum is grounded in the Central Limit Theorem, which requires approximately 20–30 observations for sample means to approximate normality and support valid parametric inference [16], and in the power requirements of non-parametric tests such as the Mann-Whitney U, which achieves ~80% power to detect a large effect ($d \approx 0.8$) at $\alpha = 0.05$ with $N = 20$ per group [17].

The allocation across sources is as follows: 110 from autoPET (the only source with sufficient BMI diversity to fill all four BMI categories), 30 from RATIC (scanner diversity), 25 from AMOS (manufacturer diversity and Chinese population), 20 from ENHANCE.PET (additional whole-body coverage with two scanner models), 10 from MSD Pancreas, and 5 from CT-ORG. This yields 160 subjects with known age and sex (80%) and approximately 118 with known BMI (59%).

The 110 autoPET subjects are selected with deliberate oversampling of rare subgroups. All 18 underweight subjects (BMI < 18.5) with known BMI are included exhaustively, since this is the smallest BMI category. Twenty subjects aged 80 or older are drawn from the available pool. Twenty-five obese subjects (BMI ≥ 30) are drawn from the 166 available. Fifteen subjects aged 18–39 ensure young adult representation. The remaining slots are filled with sex-balanced random selection, targeting approximately equal male and female representation within the autoPET allocation. The split between negative controls (~50) and cancer patients (~60) emerges naturally from this stratification.

The resulting cohort achieves the following demographic coverage, with every subgroup exceeding the N ≥ 20 minimum:

Subgroup	Categories	Expected N	Meets N ≥ 20
BMI	Underweight (< 18.5)	~22	Yes
	Normal (18.5–24.9)	~37	Yes
	Overweight (25–29.9)	~30	Yes
	Obese (≥ 30)	~29	Yes
Age	18–39	~33	Yes
	40–59	~38	Yes
	60–79	~62	Yes
	≥ 80	~27	Yes
Sex	Male	~86	Yes
	Female	~74	Yes
Clinical context	Healthy / negative	~95	Yes
	Oncology	~65	Yes
Body region	Whole-body	~130	Yes
	Abdomen-only	~70	Yes

Figures 3 and 4 illustrate one representative realization of the stratified selection from the candidate pool. The exact subject allocation will vary depending on the final randomization, but the stratification constraints (exhaustive inclusion of all underweight subjects, fixed quotas for age ≥ 80 and obese, sex balancing) guarantee that every subgroup exceeds the N ≥ 20 minimum regardless of the specific draw.

Scanner diversity is achieved through AMOS (four manufacturers across eight scanner models), RATIC (23 global institutions), and ENHANCE.PET (Siemens and Philips models). The 40 subjects without individual demographics (from AMOS, MSD Pancreas, and CT-ORG) are

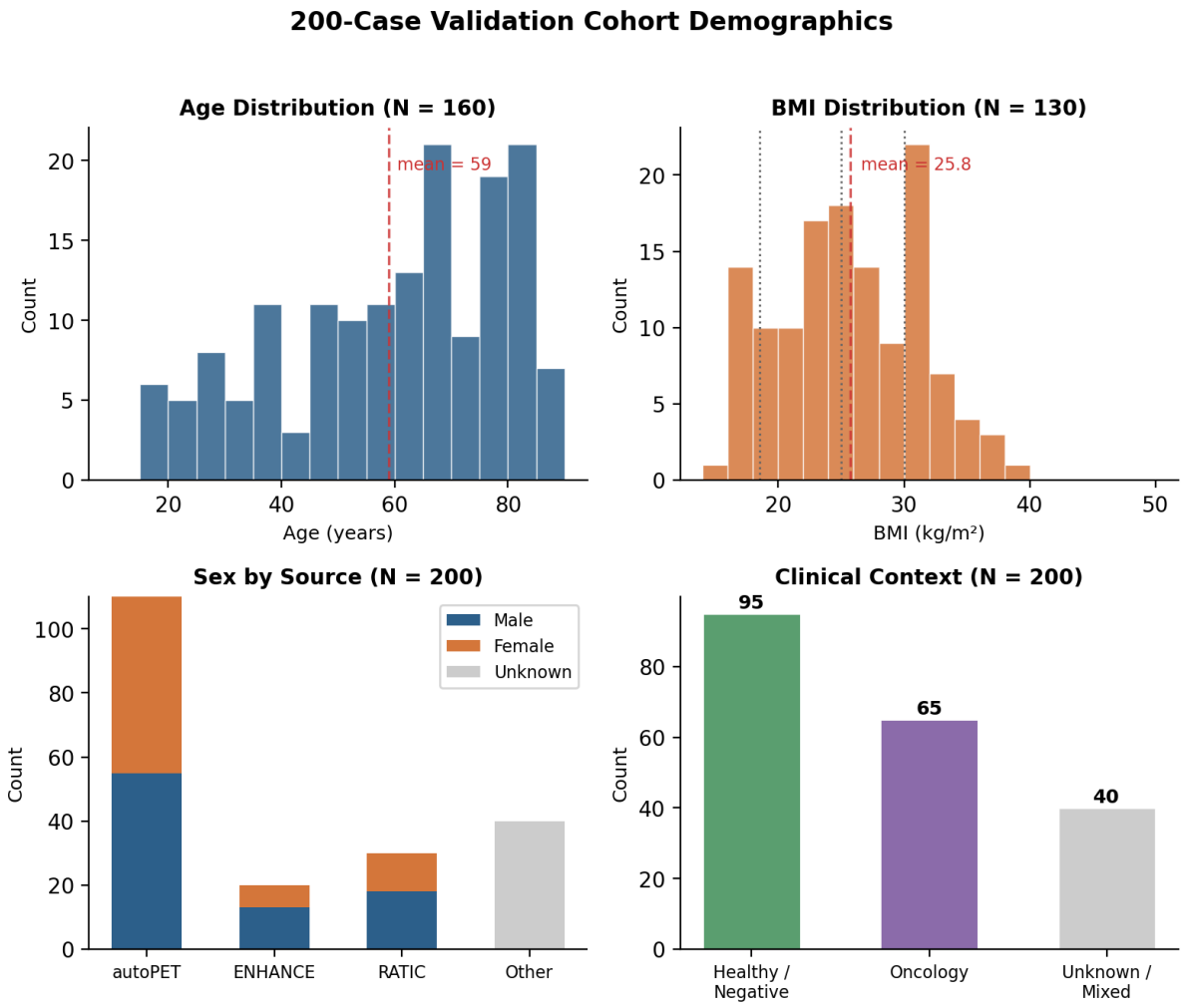

Figure 3: Cohort demographics

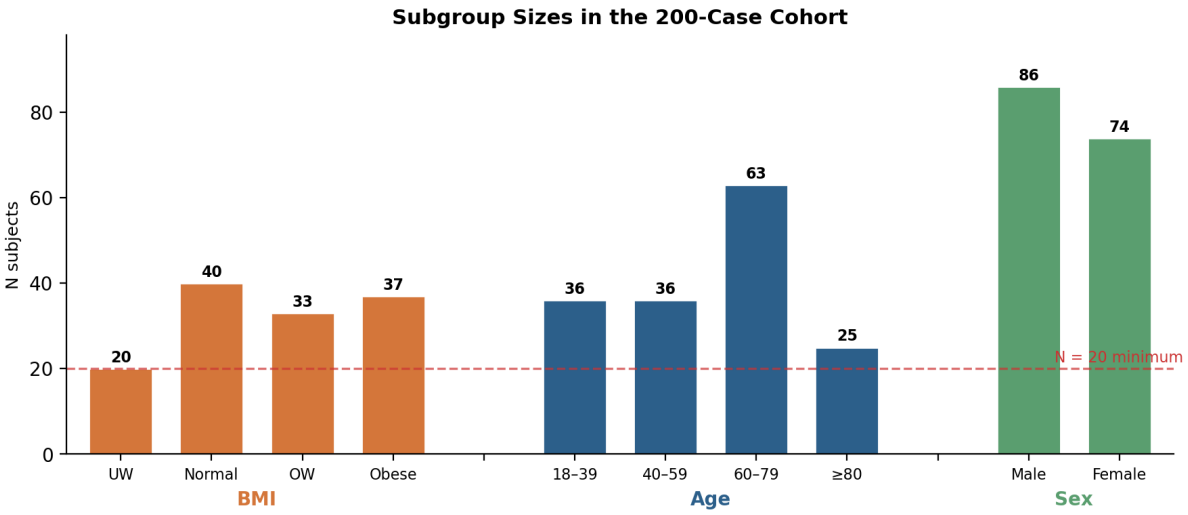

Figure 4: Subgroup adequacy

included in all aggregate analyses and contribute technical diversity, but are excluded from demographic subgroup analyses and reported as an "unknown demographics" group.

Inclusion and exclusion criteria

Scans presented to annotators are evaluated for suitability prior to segmentation. A scan is included if it contains axial CT slices covering the abdominal or whole-body region with sufficient image quality for tissue delineation. Scans are excluded if they exhibit: (1) poor overall image quality precluding reliable tissue identification, (2) severe artifacts such as those caused by metal prostheses, implanted devices, or beam hardening that substantially obscure tissue boundaries, or (3) incomplete anatomical coverage. Exclusion decisions are documented per case. Excluded scans are replaced by drawing additional cases from the candidate pool according to the same stratification constraints.

4. Reference Standard

Reference segmentations will be generated by three board-certified radiologists with expertise in CT-based body composition analysis. Annotation guidelines and the HU window ([−190, 150], matching Soma's training range) will be standardized across readers.

Annotations will be performed on every fifth axial slice (stride of 5) across the full scan depth, balancing annotation efficiency with volumetric coverage. Dice similarity coefficients will be computed on the annotated slices only. All annotations will be cross-reviewed for quality assurance, with discrepancies resolved by consensus.

For each case, annotators will perform two tasks. The first is tissue segmentation of four compartments: skeletal muscle (SM), comprising all skeletal muscle visible on each axial slice including psoas, paraspinal, rectus abdominis, lateral abdominal wall, and other regional muscles; subcutaneous adipose tissue (SAT), defined as adipose tissue external to the muscular wall and fascial planes; visceral adipose tissue (VAT), defined as adipose tissue within the abdominal and thoracic cavities, including pericardial fat; and intramuscular adipose tissue (IMAT), defined as adipose tissue located within and between skeletal muscle groups. The second is L3 localization: annotators will identify the superior and inferior axial slice boundaries of the L3 vertebral body, defining a slice range.

5. Acceptance Criteria

Aggregate performance

For each tissue class, the primary metric is the Dice similarity coefficient (DSC) computed voxel-wise per case. The acceptance threshold varies by tissue class to reflect clinical importance and inherent segmentation difficulty:

Tissue Class	DSC Threshold	Rationale
Skeletal muscle	≥ 0.90	Critical for sarcopenia assessment; well-defined boundaries
SAT	≥ 0.90	Well-delineated subcutaneous tissue
VAT	≥ 0.90	Clinically important for metabolic risk; complex mesenteric boundaries
IMAT	≥ 0.85	Small, diffuse tissue; threshold reflects inherent annotation difficulty and published inter-rater variability (0.80–0.90) [8, 18]

A tissue class passes the aggregate criterion if the lower bound of a one-sided 95% confidence interval (BCa bootstrap, 10,000 iterations) for the mean DSC exceeds its threshold.

L3 localization accuracy

Soma identifies L3 by selecting the single axial slice that maximizes the output probability of its L3 classifier (using the BiLSTM top-2 midpoint strategy), while annotators define a reference range (superior and inferior boundaries of the L3 vertebral body). The primary metric is the hit rate: the percentage of cases in which Soma's predicted L3 slice falls within the annotator's reference range. For cases where the predicted slice falls outside the reference range, the distance to the nearest range boundary (in mm) is reported as a supplementary metric. L3 localization accuracy is reported descriptively and does not gate validation acceptance; however, its effect on clinical output is captured by the end-to-end L3 body composition analysis (see Secondary Endpoints).

Per-subgroup performance

Critically, the same acceptance thresholds must also be met independently within every demographic subgroup containing at least 20 subjects. This is a hard requirement, not an exploratory analysis. The cohort is deliberately stratified to enable these comparisons, and the rationale is straightforward: because autoPET contributes 110 of the 200 subjects (55%), aggregate metrics would disproportionately reflect Siemens Biograph performance on a German clinical population. Per-subgroup testing ensures that performance generalizes to obese patients, elderly patients, female patients, non-Siemens scanners, and healthy controls separately.

The qualifying subgroups are: four BMI categories (underweight, normal, overweight, obese), four age bands (18–39, 40–59, 60–79, ≥80), two sex categories, two body region categories (whole-body vs. abdomen-only), and two clinical context categories (healthy vs. oncology).

A tissue class passes validation only if it meets its threshold both in aggregate and in every qualifying subgroup. If any subgroup falls below the threshold, the tissue class fails regardless of aggregate performance.

Secondary endpoints

In addition to volumetric DSC, the evaluation includes an end-to-end assessment of Soma's clinical output. Soma reports tissue cross-sectional areas at its predicted L3 slice; the reference is the radiologist tissue area at the annotated L3 midpoint. Bland-Altman analysis [19] (bias and 95% limits of agreement) compares these two quantities per tissue class, directly evaluating the accuracy of the numbers a clinician would see. This end-to-end metric inherits errors from both L3 detection and segmentation, which is intentional: it reflects actual product performance.

Additional secondary metrics include Hausdorff distance at the 95th percentile for boundary accuracy, Pearson correlation for total tissue volumes in milliliters (target $r \geq 0.95$), and L3 vertebra detection hit rate (see above). For the subset of approximately 118 subjects with known height (autoPET and ENHANCE.PET), SMI-based sarcopenia classification concordance against published cutoffs [2] will be reported, comparing Soma's end-to-end SMI (area at predicted L3 / height²) against reference SMI (area at annotated L3 midpoint / height²), with a target Cohen's $\kappa \geq 0.80$.

Contingent on Soma meeting its primary acceptance criteria, a descriptive clinical analysis will be performed using Soma's predictions across all available subjects (training data and the 2,066-scan candidate pool). This analysis will characterize L3 cross-sectional areas and whole-body tissue volumes for skeletal muscle and adipose tissue compartments stratified by age, sex, and BMI. In particular, the correlation between L3 cross-sectional area and total tissue volume will be evaluated for each compartment, testing the assumption, widely used in the body composition literature, that single-slice L3 measurements serve as a reliable proxy for whole-body tissue burden [1]. This assumption has been primarily validated for skeletal muscle and is less well established for adipose tissue compartments; the large and diverse cohort available here provides an opportunity to examine this relationship across demographic subgroups.

All secondary metrics are reported descriptively and do not gate validation.

6. Statistical Analysis

The primary analysis computes the mean DSC per tissue class across all 200 cases with 95% confidence intervals via bias-corrected and accelerated (BCa) bootstrap with 10,000 iterations [20]. A one-sample t-test ($\alpha = 0.05$, one-sided) evaluates whether the mean exceeds the class-specific threshold. Median DSC and interquartile range are reported as supplementary statistics.

The subgroup acceptance analysis repeats this procedure within each qualifying subgroup ($N \geq 20$). Additionally, Kruskal-Wallis tests assess whether DSC differs significantly across multi-level subgroups (BMI categories, age bands), and Mann-Whitney U tests compare two-group subgroups (sex, clinical context). Effect sizes are reported as η^2 for Kruskal-Wallis tests and

rank-biserial correlation (r) for Mann-Whitney U tests. Because the subgroup analyses are pre-specified and confirmatory (not post-hoc), no multiple comparison correction is applied.

Agreement analysis includes Bland-Altman analysis (bias and 95% limits of agreement) for L3 cross-sectional areas, Pearson correlation for tissue volumes, and Hausdorff distance at the 95th percentile for boundary accuracy. For the ~118 subjects with known height, a sarcopenia classification concordance table (sensitivity, specificity, Cohen's κ) will be reported using published SMI cutoffs [2].

Subjects without demographics participate in all aggregate analyses but are excluded from demographic subgroup testing. No imputation of missing data is performed.

7. Study Conduct

The study will proceed in four phases. First, CT scans are presented to radiologist annotators according to the pre-defined stratification. For each scan, the annotator determines whether it is suitable for analysis; suitable cases are segmented using the stride-5 annotation scheme (every fifth axial slice) with cross-review for quality assurance, and the L3 vertebral body range is identified. This process continues until the target cohort size and subgroup minimums are met. Second, Soma inference is run on all accepted cases. Third, statistical analysis is conducted per the pre-specified analysis plan. Fourth, the validation report is prepared.

8. Ethical Considerations

This study uses exclusively publicly available, de-identified CT datasets that have been released for research use by their respective institutions. No new patient data is collected, no patient contact occurs, and no identifiable protected health information is accessed. All six source datasets were published under institutional review and data use agreements permitting secondary research use.

Because the study involves only retrospective analysis of existing de-identified data, it poses no risk to human subjects.

9. Known Limitations and Risks

The most significant limitation of this evaluation is that 110 of 200 subjects (55%) come from a single scanner at a single institution (Siemens Biograph mCT, Tübingen). This is an unavoidable consequence of autoPET being the only large collection with complete BMI data, which is necessary for the per-subgroup acceptance testing that gives the study its rigor. The risk is mitigated by including 90 subjects from five other sources spanning at least four scanner manufacturers and 14 countries, and by requiring that per-subgroup performance independently meets acceptance thresholds.

Forty of the 200 subjects lack individual demographics entirely (from AMOS, MSD Pancreas, and CT-ORG). These subjects are included for technical diversity (they contribute scanners and populations not represented by autoPET) but cannot participate in demographic subgroup analyses. They are reported as a separate "unknown demographics" group, and any performance degradation in this group is flagged.

Reference annotations are performed on every fifth axial slice (stride of 5) rather than on every slice. While this reduces the total number of evaluated voxels, it provides representative sampling across the full scan depth and is standard practice in large-scale body composition validation studies.

IMAT is the most challenging tissue class due to its small volume and diffuse distribution within muscle. The acceptance threshold of 0.85 reflects the inherent difficulty and published inter-rater variability for this compartment (typically 0.80–0.90) [8, 18]. Volume correlation is reported as a complementary metric in case voxel-level DSC understates clinical agreement on total IMAT burden.

References

[1] Mourtzakis M, Prado CM, Lieffers JR, Reiman T, McCargar LJ, Baracos VE. A practical and precise approach to quantification of body composition in cancer patients using computed tomography images acquired during routine care. *Appl Physiol Nutr Metab.* 2008;33(5):997–1006.

[2] Prado CM, Lieffers JR, McCargar LJ, et al. Prevalence and clinical implications of sarcopenic obesity in patients with solid tumours of the respiratory and gastrointestinal tracts: a population-based study. *Lancet Oncol.* 2008;9(7):629–635.

[3] Cruz-Jentoft AJ, Bahat G, Bauer J, et al. Sarcopenia: revised European consensus on definition and diagnosis. *Age Ageing.* 2019;48(1):16–31.

[4] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *MICCAI 2015.* Springer; 2015:234–241.

[5] 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.* 2021;31(4):1795–1804.

[6] Wasserthal J, Breit HC, Meyer MT, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiol Artif Intell.* 2023;5(5):e230024.

[7] Milletari F, Navab N, Ahmadi SA. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: *3DV 2016.* IEEE; 2016:565–571.

[8] Mitsiopoulos N, Baumgartner RN, Heymsfield SB, et al. Cadaver validation of skeletal muscle measurement by magnetic resonance imaging and computerized tomography. *J Appl Physiol.* 1998;85(1):115–122.

[9] Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: *ICML 2019*. PMLR; 2019:6105–6114.

[10] Gatidis S, Kuestner T, Hepp T, et al. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. *Sci Data.* 2022;9(1):601.

[11] Ji Y, Bai H, Ge C, et al. AMOS: a large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In: *NeurIPS 2022*.

[12] Antonelli M, Reinke A, Bakas S, et al. The Medical Segmentation Decathlon. *Nat Commun.* 2022;13(1):4128.

[13] Rister B, Yi D, Shivakumar K, Nobashi T, Rubin DL. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Sci Data.* 2020;7(1):381.

[14] Gatidis S, et al. ENHANCE.PET dataset. 2025.

[15] Rubin DL, Flanders AE, Kim W, et al. RATIC: Radiology AI Test Image Collection. *Radiology AI.* 2024.

[16] Altman DG. *Practical Statistics for Medical Research.* Chapman and Hall/CRC; 1991.

[17] Lehmann EL, Romano JP. *Testing Statistical Hypotheses.* 3rd ed. Springer; 2006.

[18] Aubrey J, Esfandiari N, Baracos VE, et al. Measurement of skeletal muscle radiation attenuation and estimation of muscle mass composition using computed tomography: a review. *Nutr Metab (Lond).* 2014;11:44.

[19] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;327(8476):307–310.

[20] Efron B, Tibshirani RJ. *An Introduction to the Bootstrap.* Chapman and Hall/CRC; 1994.