

<https://doi.org/10.1038/s41746-026-02379-x>

Neck-to-knee dixon MRI thigh volume as a superior mass biomarker for Sarcopenia: evidence from the UK biobank

Check for updates

Hyeon Su Kim^{1,2}, Hyunwoo Park³, Junseok Kang⁴, Hyunbin Kim^{1,2}, Bonsang Gu², Bhola Shivam⁵ & Jun-Il Yoo^{5,6} ✉

Sarcopenia assessment requires biomarkers capturing muscle-specific strength beyond single-slice measurements. We developed an automated MRI framework segmenting 27 pelvic–thigh musculoskeletal structures to investigate muscle distribution as functional biomarkers. Among 37,004 UK Biobank participants (64.5 ± 7.9 years), transformer-based segmentation achieved Dice similarity coefficient of 0.896. Dixon MRI-derived thigh muscle volume showed exceptional DEXA concordance ($r = 0.936$). Posterior/anterior (P/A) muscle ratio independently predicted adverse outcomes: weak grip strength (OR 1.60, 95%CI 1.45–1.77), sarcopenia (OR 1.42, 95%CI 1.13–1.78), mortality (OR 1.49, 95%CI 1.23–1.81), and falls (OR 1.12, 95%CI 1.05–1.20), all $p < 0.005$, while left/right asymmetry showed no associations. Automated MRI phenotyping reveals muscle distribution patterns, particularly reduced anterior compartment volume, predict functional decline independent of total muscle mass, supporting evolution toward composition-aware sarcopenia criteria.

Sarcopenia—the age-associated decline in skeletal muscle quantity and function—impairs balance and gait in older adults and is strongly linked to falls and hip fractures, with downstream risks of institutionalization and mortality. Consensus statements from European and Asian working groups provide operational criteria, yet comparative evaluations show that definitions differ in predicting falls^{1–3}. Sarcopenia frequently coexists with osteoporosis (osteosarcopenia) and is enriched among older adults with a history of falling^{4,5}. With ageing populations, standardized and scalable measures of muscle status are needed to support epidemiologic phenotyping and cross-cohort harmonization.

Multiple expert groups have proposed diagnostic frameworks, most prominently the European Working Group on Sarcopenia in Older People (EWGSOP2) and the Asian Working Group for Sarcopenia (AWGS)^{1,2}. Despite broad agreement that muscle strength is central and that muscle mass (typically by DEXA or bioimpedance) provides complementary information, differences remain in operational cut-offs, measurement modalities, and staging schemes³. In practice, this heterogeneity impairs comparability across studies and limits the transportability of algorithms, underlining the need for mass metrics that are both physiologically meaningful and technically reproducible at scale.

The global leadership initiative in sarcopenia (GLIS) recently advanced the field by articulating a conceptual definition that disentangles components (muscle mass, muscle strength, and muscle-specific strength) from downstream outcomes such as physical performance⁶. In the literature exploring GLIS-inspired operationalizations, muscle-specific strength has often been computed as the ratio of strength to mass, with the latter approximated by abdominal CT at the third lumbar vertebra (L3)^{7–9}. This “CT-SMI” approach—using a single axial slice to estimate cross-sectional muscle area—has gained traction because it correlates with whole-body lean mass and is opportunistically available in many clinical settings¹⁰.

However, reliance on a single L3 slice has inherent limitations. First, it samples trunk musculature rather than the muscle groups that drive locomotion; selective atrophy of the thighs—highly relevant to gait and balance—may therefore be under-represented^{11,12}. Second, a one-slice measurement is susceptible to variability from slice positioning and body habitus, and to bias in the presence of regional disproportionality⁹. Third, cross-sectional area alone cannot capture the three-dimensional distribution of muscle bulk across the length of the limb⁸. These constraints motivate volumetric strategies that better reflect true muscle burden in the lower extremities.

¹Program in Biomedical Science and Engineering, Inha University, Incheon, South Korea. ²Department of Biomedical Research Institute, Inha University Hospital, Incheon, South Korea. ³Department of Internal Medicine, Hallym University Medical Center, Hallym University College of Medicine, Anyang, South Korea. ⁴Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. ⁵Department of Biomedical Sciences, College of Medicine, and Program in Biomedical Science & Engineering, Inha University, Incheon, South Korea. ⁶Department of Orthopedic Surgery, Inha University Hospital, Inha University College of Medicine, Incheon, South Korea. ✉e-mail: furim@hanmail.net

Manual delineation of individual muscles across many slices is prohibitively time-consuming. However, recent advances in medical image analysis, enable accurate, automated, and scalable segmentation of musculoskeletal structures. Deep-learning pipelines can recover consistent three-dimensional masks from neck-to-knee acquisitions and yield volumetric readouts within minutes, supporting population-scale quantification with quality control procedures to ensure reliability^{13,14}. These automated approaches are particularly valuable for studying muscle changes associated with mobility and aging.

Mobility limitations, falls, and the clinical construct of locomotive syndrome are largely mediated by lower-limb dysfunction⁵. To understand these conditions better, prior epidemiologic and biomechanical studies have linked measures of thigh muscle size and architecture with walking speed, balance impairment, and hip-fracture risk^{15,16}. In this context, the thigh muscle region merits greater emphasis, and lower-limb volumetry represents a biologically relevant mass metric to be evaluated alongside established approaches.

To address this need for accurate thigh muscle quantification, our group has previously developed fully automated segmentation models for 27 pelvic–thigh musculoskeletal structures on Whole Thigh CT (from L3 to knee joint), achieving a mean Dice similarity coefficient of 0.89 at the individual-muscle level across all 27 labels¹⁷. Building on this per-muscle performance, we now adapt these methods to Dixon MRI. With cross-modality pipelines now mature, this approach enables consistent thigh volumetry at scale.

To validate and apply this approach in a large-scale epidemiological context, we leverage the UK Biobank (UKB) dataset. The UK Biobank is a prospective cohort of ~500,000 adults with extensive baseline phenotyping and linkage to health records¹⁸. For traits relevant to muscle and function, UKB includes dynamometer-based grip strength, anthropometry (height, weight, waist/hip), bioimpedance analysis (whole-body composition) at baseline, blood and urine biomarkers. Its imaging study includes standardized neck-to-knee Dixon MRI in a large imaging subset (tens of thousands to date) and whole-body DEXA in another subset^{19,20}.

Using these rich multimodal data, we aim to develop a functionally oriented muscle-mass biomarker—muscle-specific muscle volume (MSMV)—derived from neck-to-knee Dixon MRI via automated segmentation of 27 anatomically defined thigh muscles, addressing aspects not captured by conventional muscle mass or muscle-specific strength indices.

Results

The study population was drawn from the UK Biobank, a large-scale prospective cohort. As shown in the Table 1, we analyzed cross-sectional data from 37,004 participants selected from an initial pool of 64,524 individuals based on age (> 50 years) and the availability of complete datasets for grip strength, Dixon MRI, and DEXA scans.

The cohort comprised 17,920 men and 19,084 women, with a mean age of 64.5 ± 7.9 years. As shown in Table 1, significant sex differences were observed in physical and muscular characteristics. Men were, on average, taller (176.0 vs. 162.7 cm) and heavier (83.7 vs. 69.0 kg), and exhibited greater maximum grip strength (39.3 vs. 24.1 kg) and a higher ALMI (8.18 vs. 6.53 kg/m²) compared to women. Based on the EWGSOP2 criteria, 1.13% of the total sample had confirmed sarcopenia, with a slightly higher prevalence in men (1.20%) than in women (1.06%).

Figures 1, 2 present the performance of our automated MRI muscle segmentation model using UNETR architecture, which achieved an overall mean DSC of 0.896 across 27 distinct pelvic–thigh musculoskeletal structures classes. Large muscle groups demonstrated the highest segmentation accuracy: gluteus maximus (Dice=0.956), multifidus (0.937), abdominal oblique (0.936), and vastus lateralis (0.935). The quadriceps complex (vastus lateralis, medialis, and intermedius) maintained DSC above 0.899, while the hamstring group (biceps femoris, semimembranosus, and semitendinosus) achieved scores exceeding 0.910. Smaller, deeper muscles showed lower performance, with obturator externus representing the most challenging

Table 1 | Demographic information of inclusion criteria in UK Biobank

Variable	Total	Men	Women
N	37,004	17,920	19,084
Age	64.5 ± 7.9	65.2 ± 7.9	63.9 ± 7.7
BMI (Kg/m ²)	26.5 ± 4.4	27.0 ± 3.9	26.1 ± 4.8
Height (cm)	169.2 ± 9.3	176.0 ± 6.7	162.7 ± 6.3
Weight (Kg)	76.1 ± 15.2	83.7 ± 13.4	69.0 ± 13.1
MAX Grip Strength (Kg)	31.5 ± 10.7	39.3 ± 8.9	24.1 ± 6.1
TMV (cm ³)	9321.7 ± 2276.5	11,166.4 ± 1689.8	7586.2 ± 1076.1
TVI (cm ³ /m ²)	0.32 ± 0.06	0.36 ± 0.05	0.29 ± 0.04
ALMI (kg/m ²)	7.33 ± 1.21	8.18 ± 0.95	6.53 ± 0.84
P/A Ratio	0.45 ± 0.06	0.45 ± 0.05	0.44 ± 0.06
Confirmed Sarcopenia	417 (1.13%)	215 (1.20%)	202 (1.06%)
Probable Sarcopenia	2165 (5.85%)	1023 (5.71%)	1142 (5.99%)
Non-Sarcopenia	34,422 (93.02%)	16,682 (93.09%)	17,740 (92.95%)

Data are presented as mean ± standard deviation (SD) for continuous variables and number (%) for categorical variables.

Sarcopenia status was defined according to the European Working Group on Sarcopenia in Older People 2 (EWGSOP2) guidelines.

TMV, Total Muscle Volume.

TVI, Total Muscle Volume/Height².

ALMI, Appendicular Lean Mass/Height².

Ratio of posterior to anterior thigh muscle volume.

structure (Dice=0.797). The mean IoU across all classes was 0.814, with precision and recall metrics both averaging above 0.90.

Figure 3 validates our automated MRI muscle segmentation model against DEXA gold standard measurements in 37,004 participants. Figure 3a (TMV vs DEXA Legs Lean Mass) demonstrates exceptional concordance TMV and DEXA legs lean mass ($r = 0.936, p < 0.001$), while Fig. 3b (TVI vs ALMI) shows robust correlation between height-adjusted TVI and ALMI ($r = 0.885, p < 0.001$).

Figure 4 presents age- and sex-stratified distributions of muscle mass metrics. Figure 4a (DEXA Appendicular Lean Mass by Age and Sex) demonstrates ALMI declining from median values of 8.5 kg/m² (men, 40–49 y) to 7.5 kg/m² (men, 80+) and 6.8 kg/m² (women, 40–49 y) to 6.4 kg/m² (women, 80+), representing declines of 0.25 and 0.13 kg/m²/decade, respectively. The proportion below sarcopenia cutoffs (7.0 kg/m² men, 5.5 kg/m² women) increased from <5% to 30% in men and 10% to 35% in women across the age span, with a critical transition at age 70 where >20% fell below thresholds.

Figure 4b (TMV by Age and Sex) reveals more pronounced absolute TMV losses, with median values decreasing from 12,158 cm³ to 9,550 cm³ in men (21% decline) and 8,358 cm³ to 6,894 cm³ in women (19% decline). Men maintained 40–50% higher absolute TMV throughout aging, though this gap narrowed due to greater absolute losses. Figure 4c shows that height-adjusted TVI partially normalized sex differences, with men maintaining only 15–20% higher values (0.38–0.34 cm³/m² vs 0.30–0.26 cm³/m²). Both sexes demonstrated similar relative TVI declines of ~13% over 40 years, suggesting height normalization accounts for much of the apparent sex difference in muscle loss.

The complementary nature of MRI-derived thigh metrics (Fig. 4b, c) to traditional DEXA measurements (Fig. 4a) provides enhanced granularity for sarcopenia assessment. Increasing variance with age, particularly evident in 70+ groups across all panels, suggests heterogeneous aging trajectories that may distinguish successful from pathological muscle aging, with implications for targeted intervention strategies.

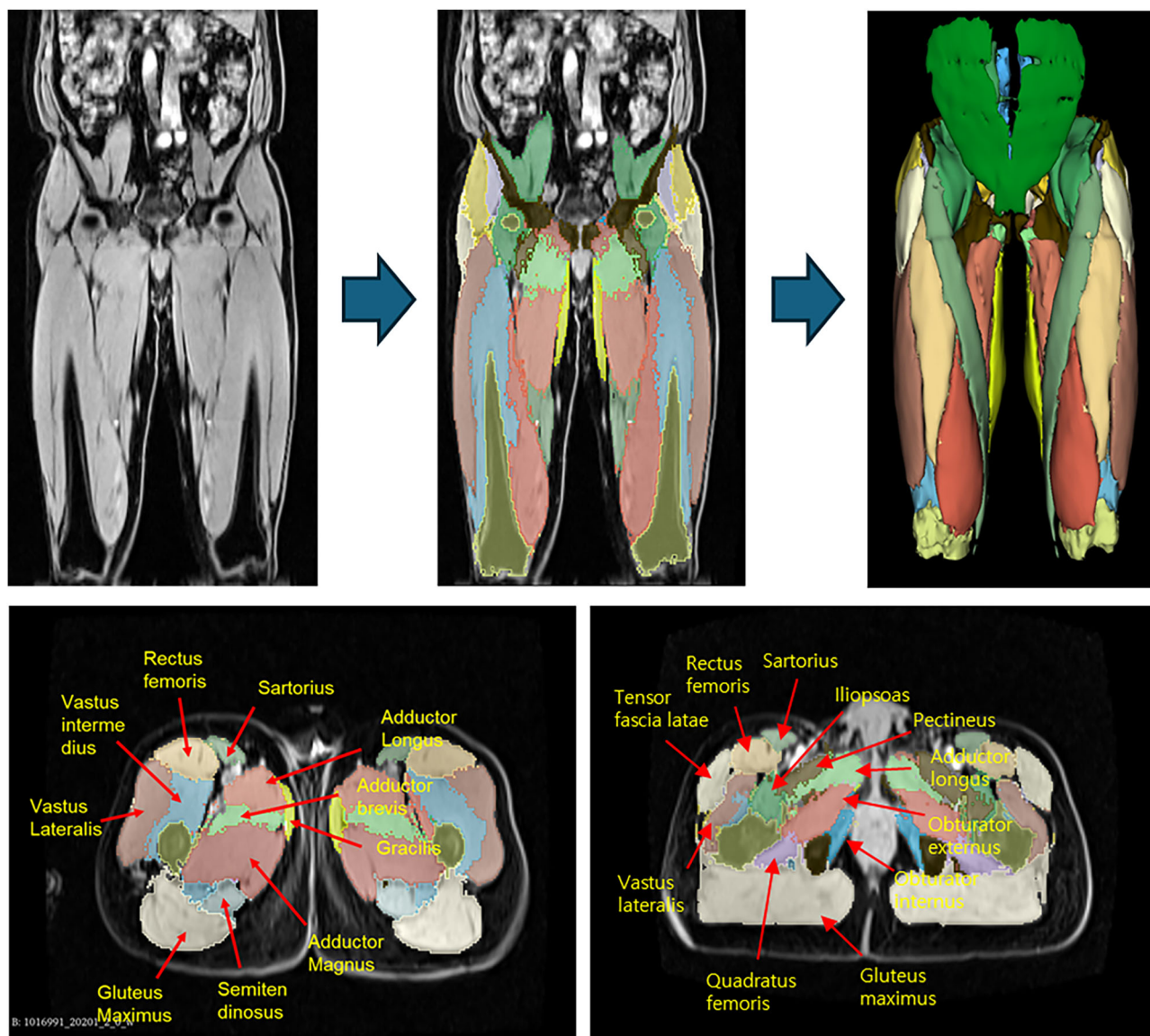


Fig. 1 | Representative examples of automated pelvic-thigh musculoskeletal structures segmentation on Dixon MRI. Three-dimensional visualization of automated muscle segmentation results from the UNETR model. The model

achieved a mean DSC of 0.896 across all muscle classes, with particularly high accuracy for larger muscle groups (Dice >0.93).

Figure 5 and Table 2 presents the associations between muscle balance patterns and clinical outcomes through tertile analysis of posterior-anterior (P/A) ratio and left-right (L/R) asymmetry in 24,670 participants. Figure 5a (P/A Ratio) demonstrates that participants in the highest P/A ratio tertile (> 0.48) compared to the lowest tertile (< 0.41) had significantly increased odds of adverse outcomes. The highest tertile showed increased odds of weak grip strength (OR 1.60, 95% CI 1.45–1.77, $p < 0.001$; 1,083 vs 698 events), sarcopenia (OR 1.42, 95% CI 1.13–1.78, $p = 0.003$; 179 vs 127 events), all-cause mortality (OR 1.49, 95% CI 1.23–1.81, $p < 0.001$; 257 vs 174 deaths), and falls in the last year (OR 1.12, 95% CI 1.05–1.20, $p = 0.001$; 2,373 vs 2,158 events). Low muscle mass defined by DEXA ALMI showed no association with P/A ratio (OR 1.02, 95% CI 0.93–1.11, $p = 0.68$; 1,096 vs 1,076 events).

Figure 5b (L/R Asymmetry) reveals that L/R asymmetry measured by DEXA showed no significant associations with any clinical outcome. OR for high versus low asymmetry tertiles were 0.95 (95% CI 0.87–1.04) for low muscle mass (1,057 vs 1,106 events), 0.93 (95% CI 0.85–1.03) for weak grip strength (838 vs 893 events), 0.86 (95% CI 0.68–1.08) for sarcopenia (133 vs 155 events), 0.91 (95% CI 0.75–1.11) for mortality (196 vs 215 deaths), and

1.02 (95% CI 0.95–1.09) for falls (2,240 vs 2,207 events). The median L/R asymmetry was 3.1% (IQR 1.4–4.9%), with 20% of participants exceeding 5% asymmetry. Adjustment for age, sex, BMI, and physical activity level did not alter these associations. The differential associations between P/A ratio and clinical outcomes, contrasted with the null findings for L/R asymmetry, indicate that sagittal plane muscle balance represents a distinct risk factor for functional decline independent of coronal plane symmetry.

In sex-stratified models, the associations were directionally similar in men (adjusted OR 0.76, 95% CI 0.66–0.89) and women (0.84, 0.73–0.96), with no significant sex interaction ($p = 0.339$), suggesting consistent effects across sexes.

Discussion

This study presents a comprehensive automated segmentation framework for quantifying TMV from Dixon MRI in 37,004 UK Biobank participants, achieving exceptional performance (mean DSC = 0.896) across 27 distinct muscle groups. Our findings demonstrate strong concordance with DEXA-derived measurements while revealing novel insights into muscle composition patterns that predict functional decline and mortality. These results

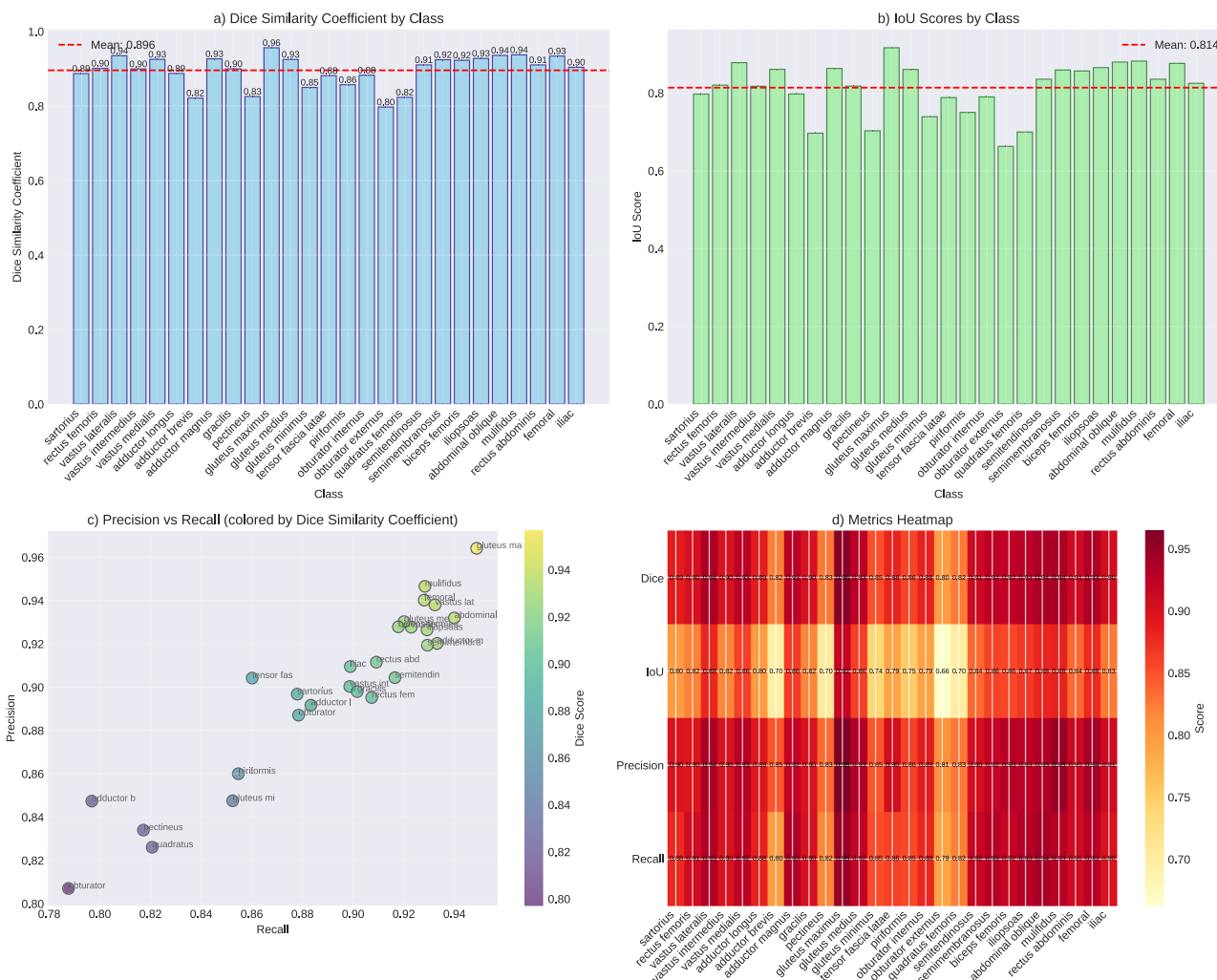


Fig. 2 | Segmentation performance metrics across individual thigh muscles. Quantitative evaluation of the UNETR segmentation model performance for each of the 27 pelvic–thigh musculoskeletal structures. **a** Per-class Dice similarity coefficient (DSC) values; the red dashed line indicates the mean DSC (0.896). **b** Per-class Intersection over Union (IoU) values for the same structures; the red dashed line indicates the mean IoU (0.814). **c** Precision–recall scatter plot by class, with point

color denoting the corresponding DSC, illustrating the trade-off between over- and under-segmentation; larger superficial muscles cluster at higher precision/recall, whereas smaller deep muscles show comparatively lower values. **d** Heatmap summarizing per-class DSC, IoU, precision, and recall, providing an integrated view of model performance across muscle and bony classes (including femur and iliac).

address critical limitations in current sarcopenia assessment methods and provide a scalable solution for population-level muscle phenotyping.

To contextualize these results against current clinical entry-point measures, DEXA-derived appendicular lean mass remains the most practical entry-point measure for sarcopenia: scanners are inexpensive, whole-body scans take only a few minutes, radiation dose is low, and guideline-based cut-offs are well established. However, as a two-dimensional projection, DEXA cannot separate anterior and posterior thigh compartments, is insensitive to sagittal-plane imbalances in muscle distribution, and mixes muscle with adjacent soft tissues along the X-ray path. In contrast, the neck-to-knee Dixon MRI protocol used here acquires co-registered water and fat volumes over the entire thigh in approximately 6–7 min without ionizing radiation, and our automated three-dimensional segmentation converts these into anatomically resolved metrics (TMV, TVI, and P/A ratio) that, in this cohort, captured compartment-specific risk for weakness, sarcopenia, mortality, and falls beyond that provided by ALMI alone.

Beyond DEXA, it is also important to consider widely used opportunistic imaging surrogates. The prevailing CT-based skeletal muscle index (SMI) measured at the L3 vertebral level has gained widespread adoption due to its opportunistic availability and correlation with whole-body muscle mass^{8,9}. However, our findings highlight fundamental limitations of this

single-slice approach that may compromise its utility for sarcopenia assessment and mobility prediction. First, the L3-SMI samples trunk musculature rather than the locomotor muscles that directly mediate gait, balance, and fall risk. As Goodpaster et al.²¹ demonstrated, preferential atrophy of thigh muscles precedes functional decline by several years, yet this critical regional loss remains invisible to L3-based measurements. Our data showing stronger associations between thigh muscle patterns and functional outcomes support this locomotor-centric view of sarcopenia pathophysiology.

Second, the reliance on cross-sectional area from a single axial slice introduces substantial measurement variability. Slice positioning errors of even 1–2 cm can alter SMI values by 5–10% due to the tapering anatomy of paraspinal muscles²². Our volumetric approach, integrating information across the entire thigh length, eliminates this positional dependency while capturing the full three-dimensional muscle distribution. This methodological advance is particularly relevant given the heterogeneous patterns of muscle loss we observed, where some individuals showed preferential proximal versus distal atrophy patterns that would be missed by any single-slice approach.

Third, and perhaps most critically, the L3-SMI approach conflates anatomically and functionally distinct muscle groups into a single metric.

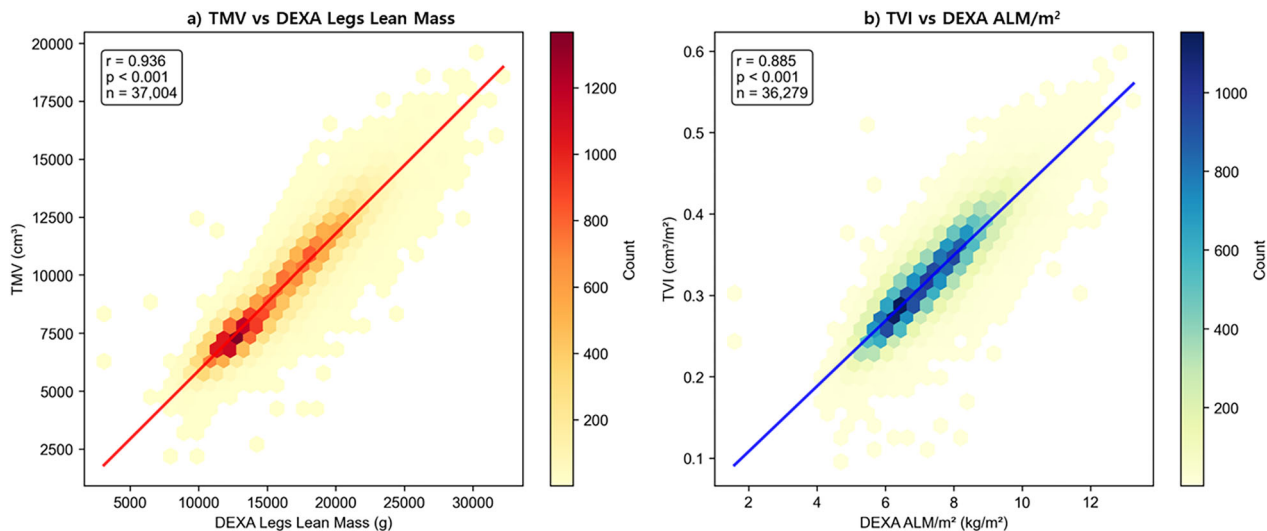


Fig. 3 | Validation of MRI-derived muscle measurements against DEXA gold standard. Concordance analysis between automated MRI segmentation and DEXA measurements in 37,004 participants. **a** Scatter plot showing correlation between thigh muscle volume (TMV) from MRI and DEXA legs lean mass (lean soft tissue of

the entire lower extremities) ($r = 0.936, p < 0.001$). **b** Correlation between height-adjusted thigh volume index (TVI) and DEXA appendicular lean mass index (ALMI) ($r = 0.885, p < 0.001$).

Our compartment-specific analysis reveals that posterior-to-anterior muscle ratio independently predicts adverse outcomes. Specifically, individuals in the highest P/A ratio tertile (indicating relatively preserved posterior muscles but diminished anterior muscles) demonstrated a 60% increased odds of weak grip strength (OR 1.60, $p < 0.001$), 42% increased odds of sarcopenia (OR 1.42, $p < 0.001$), 49% increased odds of mortality (OR 1.49, $p < 0.001$), and 12% increased odds of falls (OR 1.12, $p = 0.001$) compared to those in the lowest tertile. This means that for every 100 individuals with balanced muscle distribution who develop weakness, 160 individuals with high P/A ratios will develop weakness—a clinically meaningful difference.

Paradoxically, this increased risk occurs despite slight association with total muscle mass, suggesting that muscle distribution may be more important than absolute quantity (OR 1.02, $p = 0.68$). This finding aligns with biomechanical principles, as the anterior compartment muscles (quadriceps) are crucial for knee stability, stair climbing, and rising from chairs—activities directly tested in grip strength and functional assessments. While posterior muscles (hamstrings and gluteals) contribute to forward propulsion, their relative preservation in the setting of quadriceps atrophy may indicate a maladaptive compensation pattern that ultimately compromises overall function¹⁶. The inability of L3-SMI to capture these functionally relevant distribution patterns may explain its modest predictive value for mobility outcomes in recent validation studies.

Against this backdrop, the clinical implications of distribution- and composition-aware phenotyping become clearer. The differential associations between muscle distribution patterns and clinical outcomes underscore the importance of moving beyond simple quantity metrics toward composition-aware assessments. Our finding that high P/A ratios predict adverse outcomes independent of total muscle mass has immediate implications for sarcopenia phenotyping and intervention targeting. As Kirk et al.⁴ emphasized in their comprehensive review of osteosarcopenia, the syndrome encompasses not merely muscle loss but dysregulated muscle quality and distribution. Our data provide quantitative support for this conceptual framework by demonstrating that individuals with preserved total muscle mass but altered distribution patterns exhibit functional impairments comparable to those with overt sarcopenia.

This composition-centric view aligns with findings from Linge et al., who analyzed 40,178 UK Biobank participants using the AMRA platform and demonstrated that adverse muscle composition—defined by the combination of low fat-tissue free muscle volume and high muscle fat infiltration—was a strong independent predictor of all-cause mortality, with hazard ratios exceeding 2.0 even after adjustment for grip strength and BMI.

While their approach focused on intramuscular fat as a marker of metabolic dysfunction, our P/A ratio captures a complementary dimension: the selective vulnerability of anterior compartment muscles that directly mediate locomotor function²³. Together, these findings suggest that sarcopenia phenotyping should incorporate both metabolic (fat infiltration) and biomechanical (compartment distribution) dimensions of muscle quality.

In line with this biomechanical dimension, the stronger predictive value of sagittal plane imbalance (P/A ratio) compared to coronal plane asymmetry (L/R ratio) offers mechanistic insights into fall pathophysiology. Falls in older adults predominantly occur in the sagittal plane during activities like sit-to-stand transitions, stair navigation, and recovery from forward perturbations⁵. The posterior muscle groups, particularly the gluteus maximus and hamstrings, generate the hip extension torque necessary for these activities. Our observation that individuals in the highest P/A ratio tertile have 49% higher mortality risk suggests this pattern may reflect broader neuromuscular dysfunction beyond simple disuse atrophy.

Importantly, the preservation of predictive associations after adjustment for physical activity levels indicates that muscle distribution patterns capture intrinsic biological aging processes not fully modifiable by exercise^{24,25}. This interpretation is consistent with histological and imaging studies showing selective age-related atrophy of type II fibers and heterogeneous, muscle-specific trajectories of muscle loss across the lower limb. Our findings add an imaging-based, whole-limb perspective to this literature, while deliberately avoiding specific inferences about fiber-type composition of individual muscles from Dixon MRI alone.

Building on these clinical and mechanistic considerations, our volumetric approach enables several advances in sarcopenia phenotyping that address limitations identified in recent consensus statements^{1,6}. First, the strong correlation between MRI-derived TMV and ALMI ($r = 0.885$) validates our method against the current gold standard while providing substantially greater anatomical detail. The ability to parse 27 individual muscles allows investigation of differential atrophy patterns that may define sarcopenia subtypes with distinct etiologies and therapeutic responses.

The validity of our MRI-based sarcopenia associations is further supported by comparison with large-scale BIA studies. Jauffret et al.²⁶ examined 387,025 UK Biobank participants using bioimpedance-derived skeletal muscle index and reported that both pre-sarcopenic and sarcopenic participants had significantly elevated fracture risk (adjusted HR 1.20–1.30) independent of heel ultrasound parameters. Our sarcopenia-related odds ratios (OR 1.42 for sarcopenia, OR 1.12 for falls) show comparable magnitude, validating cross-modality consistency despite fundamentally

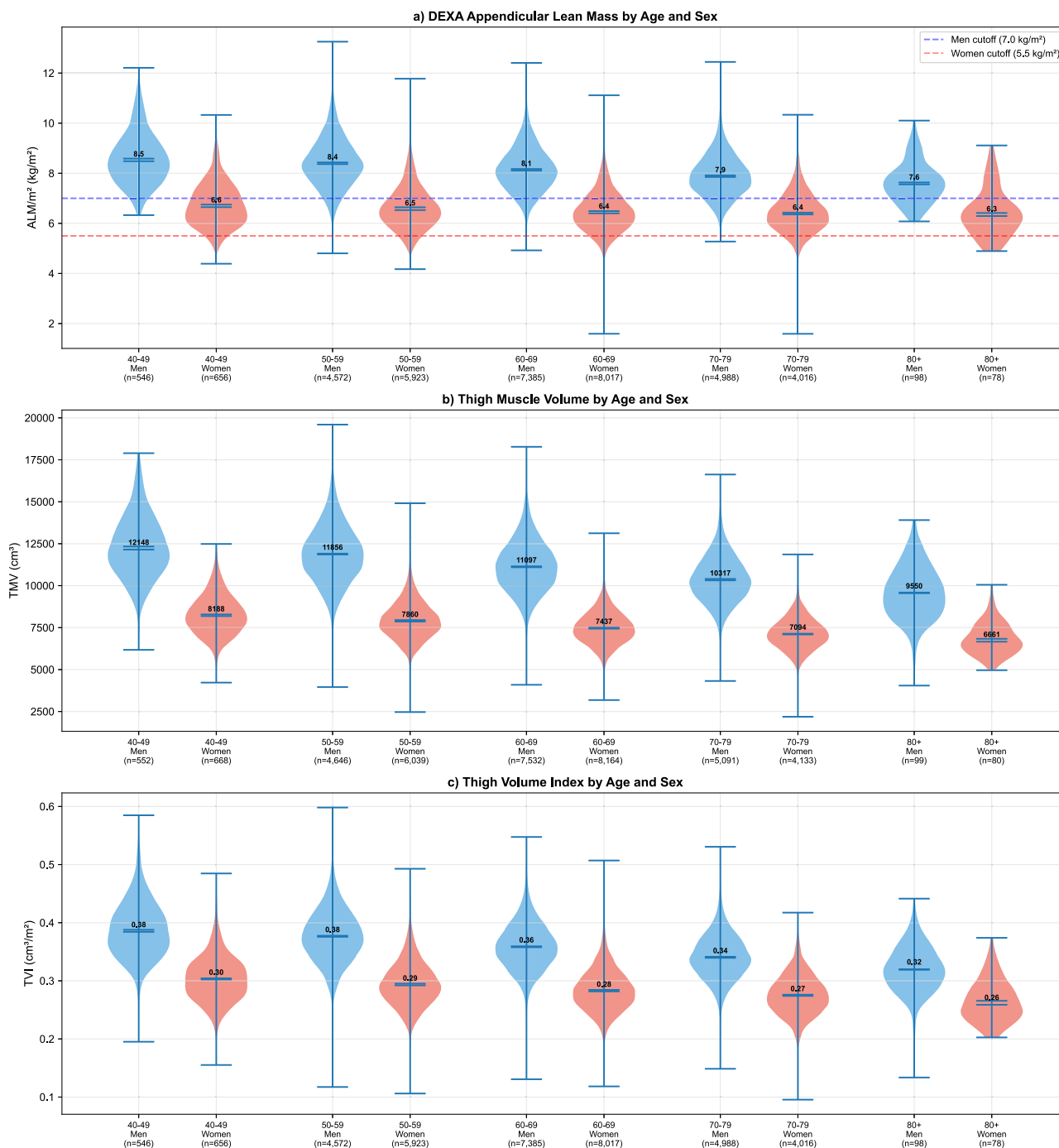


Fig. 4 | Age- and sex-stratified distributions of muscle mass metrics. Box plots showing muscle mass distributions across age groups for men (blue) and women (red). **a** DEXA-derived appendicular lean mass index (ALMI) with EWGSOP2 sarcopenia cutoff thresholds indicated by dashed lines (7.0 kg/m² for men, 5.5 kg/m² for women). **b** Absolute thigh muscle volume (TMV) from MRI segmentation.

c Height-adjusted thigh volume index (TVI). Box plots display median (center line), interquartile range (IQR; box limits showing 25th-75th percentiles), whiskers extending to 1.5×IQR, and outliers as individual points. Numbers below each box indicate sample size. All metrics show progressive decline with aging, with increasing variance in older age groups.

different measurement approaches. However, while BIA-based studies establish epidemiological associations, they cannot explain why sarcopenia increases fall risk. Our compartment-specific analysis addresses this mechanistic gap: selective quadriceps atrophy—reflected in elevated P/A ratios—directly compromises the knee extension torque required for sit-to-stand transitions and perturbation recovery, providing an anatomical basis for targeted rehabilitation interventions.

Second, the height-normalized TVI shows remarkably consistent age-related decline rates between sexes (~13% over 40 years), contrasting with

the apparent sex disparity in absolute muscle loss. This finding suggests that much of the reported sex difference in sarcopenia prevalence may reflect anthropometric scaling rather than differential biological aging.

Third, the Dixon fat–water separation that underlies our segmentation also yields quantitative fat-fraction maps, which could address a critical gap in current sarcopenia definitions focused solely on lean mass^{11,12}. While we focused on lean muscle volume and distribution in this initial validation, the same acquisition is well suited for future derivation of intramuscular fat and other fat-fraction-based indices. Previous studies have reported that

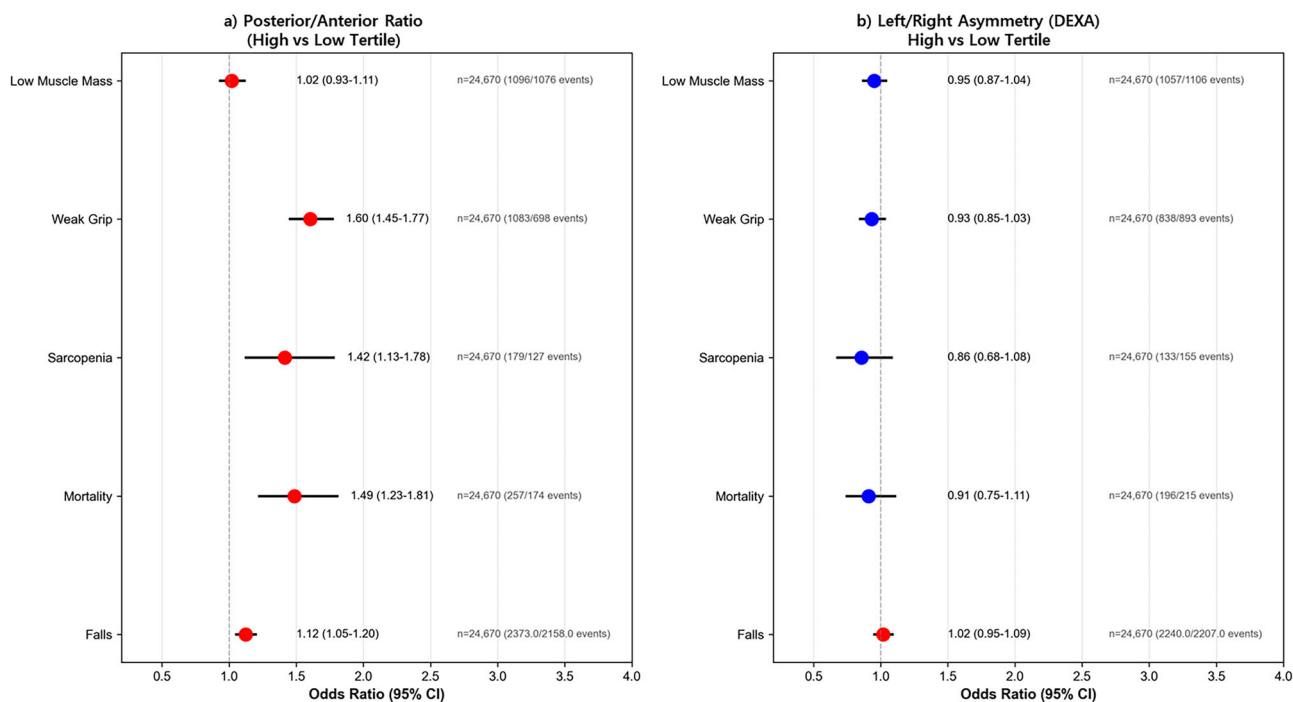


Fig. 5 | Association between muscle distribution patterns and clinical outcomes. Forest plots showing odds ratios (OR) with 95% confidence intervals for adverse outcomes comparing high versus low tertiles of muscle balance metrics. **a** Posterior/anterior (P/A) muscle ratio associations, where high P/A ratio indicates relatively preserved posterior muscles but diminished anterior muscles. **b** Left/right (L/R)

asymmetry associations based on DEXA measurements. ORs are adjusted for age, sex, BMI, and physical activity level. Numbers indicate events in high/low tertile groups. The vertical dashed line represents OR = 1.0 (no association). P/A ratio showed significant associations with all outcomes except low muscle mass, while L/R asymmetry showed no significant associations.

Table 2 | Association Between Muscle Balance Patterns and Clinical Outcomes in UK Biobank

	Outcome	Number (low & high)	Odds Ratio	CI Lower	CI Upper	Events High	Events Low
Posterior/Anterior Ratio (High vs Low Tertile)	Falls	12,335	1.12	1.05	1.20	2373	2158
	Mortality	12,335	1.49	1.23	1.81	257	174
	Sarcopenia	12,335	1.42	1.13	1.78	179	127
	Weak Grip	12,335	1.60	1.45	1.77	1083	698
	Low Muscle Mass	12,335	1.02	0.93	1.11	1096	1076
Left/Right Asymmetry (DEXA) (High vs Low Tertile)	Falls	12,335	1.02	0.95	1.09	2240	2207
	Mortality	12,335	0.91	0.75	1.11	196	215
	Sarcopenia	12,335	0.86	0.68	1.08	133	155
	Weak Grip	12,335	0.93	0.85	1.03	838	893
	Low Muscle Mass	12,335	0.95	0.87	1.04	1057	1106

Odds ratios (OR) with 95% confidence intervals (CI) for clinical outcomes comparing high vs low tertiles of posterior/anterior (P/A) ratio and left/right (L/R) asymmetry. P/A ratio represents the ratio of hamstring to quadriceps muscle volume, where values < 1 indicate quadriceps dominance. L/R asymmetry represents the absolute difference between left and right leg lean mass normalized to the larger side. All analyses are based on 37,004 participants with complete DEXA and MRI data after outlier removal. High and low tertiles each contain approximately 12,335 participants (middle tertile excluded). Events indicate the number of participants experiencing each outcome within the respective tertile groups. P/A ratio tertile cutoffs: low (< 0.41), medium (0.41–0.48), high (> 0.48). L/R asymmetry tertile cutoffs: low (< 0.017), medium (0.017–0.041), high (> 0.041).

individuals with high intramuscular fat despite preserved muscle volume exhibit functional impairments comparable to those with low muscle mass, supporting the inclusion of muscle quality metrics in next-generation sarcopenia criteria. These phenotyping advances also align with emerging conceptual frameworks. The GLIS framework emphasizes muscle-specific strength—the ratio of strength to muscle mass—as a key pathophysiological metric⁶. Our findings reveal important considerations for operationalizing this concept. Using grip strength as the numerator and thigh muscle volume as the denominator may seem anatomically mismatched, yet we observed strong associations between this ratio and functional outcomes. This apparent paradox likely reflects grip strength’s role as a biomarker of global neuromuscular function rather than isolated forearm capacity.

More fundamentally, our discovery that muscle distribution predicts outcomes independent of total mass challenges the assumption that muscle-specific strength can be reduced to a simple ratio. Consider two individuals with identical grip strength and total thigh volume but different P/A ratios: our data suggest the individual with higher P/A ratio (relatively less posterior muscle) will have worse functional outcomes despite equivalent muscle-specific strength by conventional calculation. This finding indicates that the denominator of muscle-specific strength equations must incorporate distribution information to achieve optimal predictive validity.

The compartment-specific approach also enables anatomically matched strength-mass ratios. Future studies could pair our thigh muscle volumes with lower extremity strength measures (knee extension/flexion,

hip abduction) to derive true regional muscle-specific strength metrics. Such measurements would better reflect the mechanical coupling between muscle tissue and force generation, potentially improving sensitivity for detecting pre-clinical sarcopenia.

From an implementation and deployment standpoint, automated segmentation using transformer-based architectures (UNETR) represents a methodological advance with immediate translational potential. The mean DSC of 0.896 across 27 muscles exceeds the 0.85 threshold considered clinically acceptable for treatment planning in radiation oncology, suggesting sufficient accuracy for phenotyping applications¹⁴. The superior performance on larger muscles (DSC > 0.93 for quadriceps, gluteals) that contribute most to total volume ensures robust total muscle quantification even if smaller muscle boundaries are imperfectly delineated.

The computational efficiency of our pipeline—processing a complete thigh volume in under 2 min—enables population-scale deployment. Applied to the full UK Biobank imaging cohort, this approach could generate muscle phenotypes for >100,000 individuals, creating unprecedented opportunities for genetic and epidemiological discovery. The method's reliance on Dixon MRI, now standard in population imaging protocols, ensures broad applicability without specialized sequences.

Despite these strengths, several limitations merit consideration. First, our cross-sectional design precludes causal inference regarding muscle patterns and outcomes. Longitudinal analysis of repeat imaging visits will establish whether distribution changes precede or follow functional decline. Second, while we validated against DEXA-derived lean mass, comparison with muscle biopsy findings would strengthen claims about muscle quality assessment. Third, our UK Biobank cohort, while large, may not represent sarcopenia patterns in non-European populations or clinical samples with advanced frailty. Fourth, our segmentation pipeline is restricted to the pelvic–thigh region and does not include lower-leg musculature, despite the recognized relevance of calf muscle mass in ambulation and its incorporation into EWGSOP2 and AWGS2019 criteria. This limitation reflects the UK Biobank neck-to-knee Dixon MRI protocol, which lacks continuous knee-to-ankle coverage. Future extensions of our framework to whole-body or dedicated lower-leg Dixon acquisitions would enable quantification of gastrocnemius and soleus volumes and allow direct comparison with calf-based indices recommended by international consensus groups.

Future work should explore several promising directions. Integration of fat fraction data could yield composite metrics incorporating both quantity and quality dimensions. Machine learning approaches might identify optimal combinations of muscle volumes that maximize outcome prediction. Genome-wide association studies (GWAS) of muscle-specific phenotypes could reveal susceptibility loci that remain undetected when whole-body lean mass is used as the primary trait. Most importantly, intervention studies should test whether exercise programs targeting posterior chain muscles can normalize P/A ratios and reduce fall risk.

Overall, this study establishes automated MRI-based thigh muscle segmentation as a powerful tool for sarcopenia research and clinical assessment. By moving beyond single-slice, single-number metrics to comprehensive volumetric phenotyping, we reveal that muscle distribution patterns predict functional decline and mortality independent of total muscle mass. These findings challenge current sarcopenia definitions focused solely on quantity and support evolution toward composition-aware criteria. As population imaging cohorts expand globally, the methods presented here offer a scalable pathway to precision medicine in sarcopenia, enabling risk stratification and treatment selection based on individual muscle phenotypes rather than population averages.

Methods

Study design and participants

This cross-sectional analysis utilized baseline imaging data from the UK Biobank, a population-based prospective cohort study. UK Biobank recruited 502,492 adults aged 40–69 years from 22 assessment centers across

the United Kingdom between 2006 and 2010. Participants underwent comprehensive baseline assessments including sociodemographic questionnaires, physical measurements, and biological sampling.

From 2014 onwards, a subset of participants was invited for multi-modal imaging based on geographic proximity to imaging centers and willingness to travel. By 2023, approximately 85,000 participants had completed at least one imaging visit. Our analysis focused on participants who underwent neck-to-knee Dixon MRI as part of the standardized imaging protocol. Inclusion criteria were: (1) age ≥ 50 years at imaging to ensure adequate representation across the aging spectrum, (2) completed neck-to-knee Dixon MRI, (3) DEXA scan performed within 2 years of MRI, and (4) grip strength measurement available from baseline or imaging visit. We excluded participants with incomplete imaging coverage, motion artifacts preventing accurate segmentation, or missing key covariates.

From 64,524 participants with Dixon MRI, we restricted the sample to those meeting the above criteria and with complete DEXA, grip strength, and demographic data, excluding 27,521 individuals in total. The final analytical cohort comprised 37,004 participants (17,920 men, 19,084 women).

All measurements were obtained during single imaging visits except DEXA scans, which were performed separately but typically within 6 months of MRI. This study was conducted using data from the UK Biobank resource under Application ID 622629. UK Biobank has ethical approval as a Research Tissue Bank from the North West Multi-centre Research Ethics Committee (REC reference: 21/NW/0157; IRAS project ID: 299116), which permits the use of stored data and samples for health-related research in the public interest. All UK Biobank participants provided written informed consent at recruitment for the use of their data and linkage to health records in approved research studies. Under this framework, the present analysis of de-identified UK Biobank data did not require additional project-specific ethics approval or consent.

Sarcopenia classification

Sarcopenia was defined according to EWGSOP2 consensus as low grip strength (< 27 kg for men, < 16 kg for women) and low appendicular lean mass index (ALMI < 7.0 kg/m² for men, < 5.5 kg/m² for women)¹. Grip strength was measured twice per hand with a calibrated Jamar dynamometer, and the highest value was retained. Appendicular lean mass (ALM, kg) was obtained directly from whole-body dual-energy X-ray absorptiometry (DEXA) scans performed in the UK Biobank imaging centers on GE-Lunar iDXA systems (enCORE software v17). The lean soft-tissue masses of both arms and both legs (fields 23263–23266) were summed to yield ALM. In this DEXA implementation, “legs lean mass” denotes the total lean soft tissue of the entire lower extremities (thigh, lower leg, and foot), not isolated thigh muscle volume. Quality assurance included daily calibration with a stepped phantom and quarterly cross-calibration across scanners. Standing height (m), recorded via UK Biobank field 50, was used to calculate ALMI. Participants who met neither EWGSOP2 criterion were classified as non-sarcopenic; those with low strength but normal ALMI were designated “probable sarcopenia” and included only in sensitivity analyses.

Whole Thigh MRI acquisition

Dixon MRI exploits the chemical-shift difference (~ 3.5 ppm at 1.5 T) between fat and water protons to generate co-registered water, fat and proton-density-fat-fraction (PDFF) volumes. We used the UK Biobank Dixon protocol on 1.5 T Siemens MAGNETOM Aera scanners at four imaging centers. The sequence was a three-dimensional spoiled gradient-echo acquisition with a repetition time of 6.53 ms and dual echo times of 2.39 ms (in-phase) and 4.77 ms (opposed-phase). A 10° flip angle and a receiver bandwidth of 960 Hz pixel⁻¹ were employed^{15,27,28}.

Images were reconstructed on a 224 × 224 matrix, giving a native in-plane resolution of 2.0 mm across a 448 × 448 mm field of view; 320 contiguous 3.0 mm slices were collected, spanning from the head to below the knees. Parallel imaging with GRAPPA (acceleration factor 2) kept the total acquisition time to approximately 6.5 min per participant.

Each of six consecutive table positions (chest, abdomen, pelvis, hip, mid-thigh, and thigh-to-knee), the scanner produced four spatially aligned magnitude images—opposed-phase, in-phase, water-only, and fat-only—resulting in 24 compressed DICOM series per subject.

For the analysis of pelvic–thigh musculoskeletal structures estimation model, a continuous three-dimensional (3D) volumes was computationally generated by fusing four separate, overlapping axial Dixon MRI stations from the UK Biobank dataset. These stations collectively spanned the region from the pelvic to the knee, encompassing the pelvic, hip, mid-thigh, and thigh-to-knee areas. A custom Python-based algorithm systematically reconstructed and merged these stations to produce a single, cohesive image for each participant.

The process began by reconstructing each of the four stations into a distinct 3D volume. Within each station, the individual 2D DICOM slices were sorted anatomically in the inferior-to-superior direction based on their z-axis coordinate, as specified in the “ImagePositionPatient” metadata tag. These sorted slices were then stacked to form a 3D numpy array. Voxel spacing was defined using the “PixelSpacing” tag for the in-plane (X, Y) dimensions, while the through-plane (Z) spacing was robustly calculated as the median distance between consecutive slices to ensure accuracy. Following reconstruction, the 3D stations were sorted in superior-to-inferior anatomical order. A consistency check was performed to resolve any spatially duplicated stations, retaining only the most recently acquired data based on the “SeriesTime” tag.

The core fusion process involved creating a single high-resolution global grid that encompassed the entire spatial extent of all four validated stations. Each station was then precisely resampled onto this common grid using trilinear interpolation.

The value P at a target point with normalized coordinates (x_d, y_d, z_d) within a voxel is estimated using trilinear interpolation, defined as the weighted average of the values C_{ijk} at the eight surrounding corner points (i, j, k) of the voxel, where $i, j, k \in \{0, 1\}$.

The formula is:

$$P(x_d, y_d, z_d) = \sum_{i=0}^1 \sum_{j=0}^1 \sum_{k=0}^1 C_{ijk} w_i(x_d) w_j(y_d) w_k(z_d) \quad (1)$$

Where the weights for each axis are calculated by linear interpolation:

$$w_i(d) = d^i (1 - d)^{1-i} \quad (2)$$

This can be expanded into the full summation form

$$P(x_d, y_d, z_d) = \sum_{i=0}^1 \sum_{j=0}^1 \sum_{k=0}^1 C_{ijk} x_d^i (1 - x_d)^{1-i} y_d^j (1 - y_d)^{1-j} z_d^k (1 - z_d)^{1-k} \quad (3)$$

$P(x_d, y_d, z_d)$ is the interpolated value at the target point.

C_{ijk} is the known value at the corner (i, j, k) of the voxel.

x_d, y_d, z_d are the normalized distances (ranging from 0 to 1) of the target point from the corner $(0, 0, 0)$ along each respective axis.

In the overlapping regions between adjacent stations, voxel intensities were blended using a simple average to guarantee a seamless transition. This entire pipeline was implemented in Python, leveraging libraries such as pydicom and numpy. To significantly reduce processing time, the computationally demanding interpolation step was accelerated using NVIDIA CUDA via the numba library. The final continuous 3D volume was exported in the nearly Raw Raster Data (NRRD) format, which preserves the image data along with its complete spatial information (origin, spacing and orientation) and metadata in a single file.

Ground Truth Annotation for MRI Muscle Segmentation

Ground truth annotations for the pelvic–thigh musculoskeletal structures segmentation model were meticulously created from the fused Dixon MRI volumes. Expert anatomists manually delineated 27 distinct pelvic–thigh musculoskeletal structures segments using 3D Slicer software (version 5.2.2), a widely adopted platform for medical image annotation²⁹. The annotation protocol followed standardized anatomical guidelines to ensure consistency across raters.

Anterior compartment muscles ($n = 5$): sartorius, rectus femoris, vastus lateralis, vastus medialis, and vastus intermedius. Gluteal region muscles ($n = 8$): gluteus maximus, gluteus medius, gluteus minimus, piriformis, obturator internus, obturator externus, pectineus, and tensor fasciae latae. Medial compartment muscles ($n = 5$): adductor magnus, adductor longus, adductor brevis, gracilis, and quadratus femoris. Posterior compartment muscles ($n = 3$): semimembranosus, semitendinosus, and biceps femoris. Core muscles ($n = 4$): multifidus, iliopsoas, abdominal oblique, and rectus abdominis. Bone structures ($n = 2$): iliac bone and femur. Additionally, subcutaneous and intermuscular adipose tissue can be annotated within the same framework to enable future analyses of fat-related muscle quality, although these measures were not analyzed in the present work. The annotation masks were saved in NRRD format, preserving spatial information including voxel spacing, orientation matrices, and origin coordinates essential for accurate volumetric quantification.

Image Pre-processing

The preprocessing pipeline was implemented following standardized medical imaging protocols to ensure robust and reproducible segmentation results. Several critical transformations were applied systematically to both training and validation datasets.

Intensity Normalization: Intensity windowing specific to soft tissue visualization was applied, mapping voxel intensities from the range [80, 450] to normalized values [0.0, 1.0]. This windowing, optimized for muscle tissue contrast enhancement. This approach aligns with established practices in medical image preprocessing^{14,30}.

Spatial Resampling: All volumes underwent resampling to achieve consistent voxel spacing of (1.5, 1.5, 2.0). Trilinear interpolation was employed for image volumes to preserve intensity continuity, while nearest-neighbor interpolation was utilized for segmentation masks to maintain label integrity. This standardization protocol ensures uniform spatial resolution across heterogeneous MRI acquisitions, as recommended by recent medical imaging benchmarks³¹.

Patch-based Sampling Strategy: The training pipeline utilized the RandCropByPosNegLabel transformation from the MONAI framework to extract three-dimensional patches of dimensions (96, 96, 96) voxels. Four patches were sampled per volume during each training iteration, maintaining a 1:1 ratio between positive samples (containing target muscle tissue) and negative samples (background regions). This balanced sampling strategy addresses the class imbalance inherent in medical image segmentation tasks³².

UNETR Architecture for 3D Muscle Segmentation

We employed the UNETR (U-Net Transformers) architecture for 3D muscle segmentation, leveraging its hybrid design that combines Vision Transformers for global context modeling with convolutional decoders for precise spatial localization³³. The transformer encoder processes 3D patches of size (96, 96, 96) voxels through 12 self-attention layers with hidden dimension 768, capturing long-range anatomical dependencies crucial for distinguishing morphologically similar muscle groups. The CNN decoder follows a U-Net-like architecture with skip connections from transformer layers at multiple resolutions (1/2, 1/4, 1/8, and 1/16), employing instance normalization and residual blocks for stable training. The model outputs 29 channels corresponding to 27 distinct pelvic–thigh musculoskeletal structures segments plus background, enabling comprehensive multi-class segmentation.

Training configuration

Model training was conducted on NVIDIA RTX 4090 and A6000 GPUs using PyTorch 2.0 and MONAI 1.2.0 frameworks. We employed a hybrid loss function combining Dice and Cross-Entropy losses ($L_{total} = L_{Dice} + L_{CE}$) to address class imbalance in multi-muscle segmentation. The AdamW optimizer was configured with learning rate 2×10^{-5} , weight decay 1×10^{-5} , and batch size 2, with hyperparameters determined through systematic grid search. Training proceeded for 12,000 epochs with validation every 2 epochs and early stopping based on validation performance. Mixed precision training with automatic mixed precision (AMP) was utilized to accelerate computation and enable gradient accumulation, effectively increasing the batch size despite memory constraints imposed by 3D volumetric data.

Performance evaluation metrics of AI segmentation model

To comprehensively assess segmentation performance, we employed multiple complementary metrics that capture different aspects of segmentation quality, following established evaluation protocols in medical image segmentation^{34,35}. Each metric was computed on a per-class basis and aggregated across all 27 muscle segments.

The dice similarity coefficient (DSC) and Intersection over Union (IoU) were calculated to quantify volumetric overlap between predicted and ground truth segmentations. For each muscle class c :

$$DSC_c = \frac{|P_c \cap G_c|}{|P_c + G_c|} \times 2 \tag{4}$$

$$IoU_c = \frac{|P_c \cap G_c|}{|P_c \cup G_c|} \tag{5}$$

Here P_c and G_c represent the sets of voxels predicted and labeled as class c , respectively. These metrics are related by: $DSC = 2 \times IoU / (1 + IoU)$. The mean values across all muscle classes were computed as:

$$Mean\ DSC = \left(\frac{1}{n}\right) \times \sum (Dice_c) \text{ and } Mean\ IoU = \left(\frac{1}{n}\right) \times \sum (IoU_c) \tag{6}$$

where $n = 27$ muscle classes. Classes absent from the ground truth ($tp = 0, fn = 0$) were excluded from averaging to prevent artificial inflation of performance metrics, as recommended by recent segmentation challenges³⁶.

The Precision and Recall were computed to evaluate the model's discriminative capability:

$$Precision_c = \frac{TP_c}{(TP_c + FP_c)} \tag{7}$$

$$Recall_c = \frac{TP_c}{(TP_c + FN_c)} \tag{8}$$

Where TP_c , FP_c , and FN_c represent true positives, false positives, and false negatives for class c , respectively. These metrics provide complementary insights into the model's tendency toward over- or under-segmentation of specific muscle groups³⁷.

Statistical analysis

The concordance between MRI-derived muscle measurements and DEXA gold standard assessments was evaluated using multiple statistical approaches to ensure robust validation of our automated segmentation method. Pearson correlation coefficients were calculated to assess the linear relationships between paired measurements, specifically examining thigh muscle volume (TMV) versus DEXA legs lean mass ($n = 37,004$) and thigh volume index (TVI) versus ALMI per height squared ($n = 37,004$).

The correlation coefficient was computed as $r = \Sigma[(x_i - \bar{x})(y_i - \bar{y})] / \sqrt{[\Sigma(x_i - \bar{x})^2 \times \Sigma(y_i - \bar{y})^2]}$, where x_i and y_i represent paired MRI and DEXA measurements for participant i , and \bar{x} and \bar{y} denote the respective sample means. To quantify the uncertainty in our correlation estimates, 95% confidence intervals were constructed using Fisher's z -transformation: $z = 0.5 \times \ln[(1 + r)/(1 - r)]$, with standard error $SE(z) = 1 / \sqrt{(n - 3)}$.

The confidence intervals were then back-transformed to the correlation scale. Correlations exceeding 0.8 were interpreted as strong, indicating excellent concordance between measurement modalities.

To characterize age-related patterns of muscle loss, participants were stratified into decade-based age groups (50–59, 60–69, 70–79, and ≥ 80 years); an additional 40–49 year stratum was included only for descriptive visualization in Fig. 4. This stratification enabled examination of both linear trends and potential non-linear patterns in muscle changes across the aging spectrum. Within each age stratum, descriptive statistics including median, interquartile range, and distribution parameters were calculated separately for men and women to account for known sex differences in muscle mass and aging trajectories. The rate of muscle decline was quantified using linear regression models fitted separately for each sex: $Y_{ij} = \beta_0 + \beta_1 \times Age_{ij} + \epsilon_{ij}$, where Y_{ij} represents the muscle metric (ALMI, TMV, or TVI) for participant i of sex j , β_0 represents the intercept, β_1 represents the annual rate of change, and ϵ_{ij} represents the random error term assumed to follow a normal distribution with mean zero and constant variance.

The decline rate per decade was calculated as $10 \times \beta_1$, providing clinically interpretable estimates of muscle loss over ten-year periods. The 95% confidence intervals for these estimates were derived from the standard error of the regression coefficient: $CI = 10 \times \beta_1 \pm 1.96 \times 10 \times SE(\beta_1)$. Additionally, to examine the proportion of participants meeting sarcopenia criteria across age groups, we calculated the percentage falling below established EWGSOP2 thresholds (ALMI < 7.0 kg/m² for men, < 5.5 kg/m² for women) within each age stratum. The Cochran-Armitage test for trend was applied to assess whether the proportion with sarcopenia increased linearly with age category, with the test statistic calculated as $Z = \Sigma w_i(p_i - \bar{p}) / \sqrt{[\bar{p}(1 - \bar{p})\Sigma w_i^2/n_i]}$, where w_i represents the weight for age group i , p_i represents the proportion with sarcopenia in group i , and \bar{p} represents the overall proportion.

The associations between muscle balance patterns and adverse clinical outcomes were investigated using multivariable logistic regression models, with participants stratified into tertiles based on the distribution of posterior/anterior (P/A) ratio and left/right (L/R) asymmetry. Tertile cutpoints were determined using the 33rd and 67th percentiles of each distribution, with P/A ratio tertiles defined as low (< 0.41), middle (0.41–0.48), and high (> 0.48). For the primary analysis, we compared the highest versus lowest tertile, excluding the middle tertile to maximize contrast between groups and increase statistical power to detect associations. The logistic regression model for each clinical outcome was specified as:

$$\text{logit}(p_i) = \ln \left[\frac{p_i}{1 - p_i} \right] = \beta_0 + \beta_1 \times I_{\text{high},i} + \beta_2 \times Age_i + \beta_3 \times Sex_i \tag{9}$$

where p_i represents the probability of the outcome for participant i , $I_{\text{high},i}$ is an indicator variable coded as 1 for high tertile membership and 0 for low tertile (reference category), Age_i represents age in years (continuous), Sex_i is coded as 1 for male and 0 for female, BMI_i represents body mass index in kg/m² (continuous), and PA_i represents physical activity level in MET-hours/week (categorical).

Five clinical outcomes were examined: (1) low muscle mass defined by ALMI below EWGSOP2 sex-specific cutoffs, (2) weak grip strength using established thresholds (< 27 kg for men, < 16 kg for women), (3) confirmed sarcopenia according to EWGSOP2 criteria requiring both low muscle mass and weak grip strength, (4) all-cause mortality ascertained through linkage with national death registries, and (5) history of falls in the past year derived from UK Biobank Field 2296 where values ≥ 2 indicate fall occurrence. Odds ratios (OR) were calculated as $OR = \exp(\beta_1)$, representing the multiplicative

increase in odds of the outcome for high versus low tertile membership. The 95% confidence intervals were obtained using the profile likelihood method, which provides more accurate coverage than Wald-based intervals, particularly for smaller sample sizes or when OR deviate substantially from 1.0. Statistical significance was assessed using Wald χ^2 tests with the test statistic calculated as $(\beta_1/SE(\beta_1))^2$, where SE represents the standard error. *P*-values less than 0.05 were considered statistically significant. Given the exploratory nature of this analysis and the biological plausibility of the examined associations, no correction for multiple comparisons was applied, though all *p*-values are reported to allow readers to apply their preferred adjustment method if desired.

Code availability

All code related to MRI data processing, including the Dixon MRI station fusion pipeline, is available to researchers holding a valid UK Biobank Application ID. Please contact the corresponding author with proof of UK Biobank approval.

Data availability

All data generated or analyzed during this study are included in the published article and Supplementary Information file. The raw data supporting the results of this study are available from the UK Biobank. UK Biobank data is only available to approved researchers. To access data, researchers need to submit an application to the UK Biobank Access Management System and be approved. Further information on data access is provided at [<https://www.ukbiobank.ac.uk>].

Received: 4 September 2025; Accepted: 15 January 2026;

Published online: 05 February 2026

References

1. Cruz-Jentoft, A. J. et al. Sarcopenia: revised European consensus on definition and diagnosis. *Age Ageing* **48**, 16–31 (2019).
2. Chen, L.-K. et al. Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. *J. Am. Med. Dir. Assoc.* **21**, 300–307.e2 (2020).
3. Bischoff-Ferrari, H. A. et al. Comparative performance of current definitions of sarcopenia against the prospective incidence of falls among community-dwelling seniors age 65 and older. *Osteoporos. Int. J. Establ. Result Coop. Eur. Found. Osteoporos. Natl. Osteoporos. Found. USA* **26**, 2793–2802 (2015).
4. Kirk, B., Zanker, J. & Duque, G. Osteosarcopenia: epidemiology, diagnosis, and treatment-facts and numbers. *J. Cachexia Sarcopenia Muscle* **11**, 609–618 (2020).
5. Huo, Y. R. et al. Phenotype of osteosarcopenia in older individuals with a history of falling. *J. Am. Med. Dir. Assoc.* **16**, 290–295 (2015).
6. Kirk, B. et al. The Conceptual Definition of Sarcopenia: Delphi Consensus from the Global Leadership Initiative in Sarcopenia (GLIS). *Age Ageing* **53**, afae052 (2024).
7. Wu, G.-F. et al. Sarcopenia defined by the global leadership initiative in sarcopenia (GLIS) consensus predicts adverse postoperative outcomes in patients undergoing radical gastrectomy for gastric cancer: analysis from a prospective cohort study. *BMC Cancer* **25**, 679 (2025).
8. Shen, W. et al. Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image. *J. Appl. Physiol. Bethesda Md* **1985** **97**, 2333–2338 (2004).
9. Mourtzakis, M. et al. 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. Physiol. Appl. Nutr. Metab.* **33**, 997–1006 (2008).
10. Pickhardt, P. J. et al. Opportunistic Screening at Abdominal CT: Use of Automated Body Composition Biomarkers for Added Cardiometabolic Value. *Radiogr. Rev. Publ. Radiol. Soc. N. Am. Inc* **41**, 524–542 (2021).

11. Goodpaster, B. H., Kelley, D. E., Thaete, F. L., He, J. & Ross, R. Skeletal muscle attenuation determined by computed tomography is associated with skeletal muscle lipid content. *J. Appl. Physiol. Bethesda Md* **1985** **89**, 104–110 (2000).
12. Delmonico, M. J. et al. Longitudinal study of muscle strength, quality, and adipose tissue infiltration. *Am. J. Clin. Nutr.* **90**, 1579–1585 (2009).
13. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
14. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
15. Schaap, L. A., van Schoor, N. M., Lips, P. & Visser, M. Associations of sarcopenia definitions, and their components, with the incidence of recurrent falling and fractures: the longitudinal aging study Amsterdam. *J. Gerontol. A. Biol. Sci. Med. Sci.* **73**, 1199–1204 (2018).
16. Schaap, L. A., Koster, A. & Visser, M. Adiposity, muscle mass, and muscle strength in relation to functional decline in older persons. *Epidemiol. Rev.* **35**, 51–65 (2013).
17. Kim, H. S. et al. Precise individual muscle segmentation in whole thigh CT scans for sarcopenia assessment using U-net transformer. *Sci. Rep.* **14**, 3301 (2024).
18. Dodds, R. M., Granic, A., Robinson, S. M. & Sayer, A. A. Sarcopenia, long-term conditions, and multimorbidity: findings from UK Biobank participants. *J. Cachexia Sarcopenia Muscle* **11**, 62–68 (2020).
19. Miller, K. L. et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523–1536 (2016).
20. Petersen, S. E. et al. UK Biobank’s cardiovascular magnetic resonance protocol. *J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson.* **18**, 8 (2016).
21. Goodpaster, B. H. et al. Attenuation of skeletal muscle and strength in the elderly: The Health ABC Study. *J. Appl. Physiol. Bethesda Md* **1985** **90**, 2157–2165 (2001).
22. Pickhardt, P. J. et al. Fully automated deep learning tool for sarcopenia assessment on CT: L1 versus L3 vertebral level muscle measurements for opportunistic prediction of adverse clinical outcomes. *AJR Am. J. Roentgenol.* **218**, 124–131 (2022).
23. Linge, J., Petersson, M., Forsgren, M. F., Sanyal, A. J. & Dahlqvist Leinhard, O. Adverse muscle composition predicts all-cause mortality in the UK Biobank imaging study. *J. Cachexia Sarcopenia Muscle* **12**, 1513–1526 (2021).
24. Lexell, J., Taylor, C. C. & Sjöström, M. What is the cause of the ageing atrophy? Total number, size and proportion of different fiber types studied in whole vastus lateralis muscle from 15- to 83-year-old men. *J. Neurol. Sci.* **84**, 275–294 (1988).
25. Frontera, W. R. et al. Aging of skeletal muscle: a 12-yr longitudinal study. *J. Appl. Physiol. Bethesda Md* **1985** **88**, 1321–1326 (2000).
26. Jauffret, C. et al. Association between sarcopenia and fracture risk in a population from the UK biobank database. *J. Bone Miner. Res. Off. J. Am. Soc. Bone Miner. Res.* **38**, 1422–1434 (2023).
27. Starck, S. et al. Using UK Biobank data to establish population-specific atlases from whole body MRI. *Commun. Med.* **4**, 237 (2024).
28. Wachinger, C., Wolf, T. N. & Pölsterl, S. Deep learning for the prediction of type 2 diabetes mellitus from neck-to-knee Dixon MRI in the UK biobank. *Heliyon* **9**, e22239 (2023).
29. Fedorov, A. et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **30**, 1323–1341 (2012).
30. Ma, J. et al. Loss odyssey in medical image segmentation. *Med. Image Anal.* **71**, 102035 (2021).
31. Antonelli, M. et al. The Medical Segmentation Decathlon. *Nat. Commun.* **13**, 4128 (2022).
32. Cardoso, M. J. et al. MONAI: An open-source framework for deep learning in healthcare. Preprint at <https://doi.org/10.48550/arXiv.2211.02701> (2022).

33. Hatamizadeh, A. et al. UNETR: Transformers for 3D Medical Image Segmentation. in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 1748–1758 (2022). <https://doi.org/10.1109/WACV51458.2022.00181>.
34. Taha, A. A. & Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* **15**, 29 (2015).
35. Maier-Hein, L. et al. Metrics reloaded: recommendations for image analysis validation. *Nat. Methods* **21**, 195–212 (2024).
36. Reinke, A. et al. Common Limitations of Image Processing Metrics: A Picture Story. Preprint at <https://doi.org/10.48550/arXiv.2104.05642> (2023).
37. Müller, D., Soto-Rey, I. & Kramer, F. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Res. Notes* **15**, 210 (2022).

Acknowledgements

This research was supported by a grant of Korean ARPA-H Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2024-00507256). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2021-NR060097).

Author contributions

H.S.K. conceived and designed the study, developed the deep learning framework for muscle segmentation, performed all statistical analyses, interpreted the results, created all figures and tables, wrote the manuscript, and managed the UK Biobank data access application. H.P. assisted with data collection and analysis, and provided clinical insights for data interpretation. H.K. contributed to data collection and preprocessing. J.K. provided technical expertise for MRI-based fat infiltration quantification and contributed to the development and training process of the imaging analysis pipeline. B.G. and B.S. assisted with data collection and quality control. J.I.Y. supervised the study, secured the UK Biobank data access application, provided critical revision of the manuscript, and gave final approval for submission. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-026-02379-x>.

Correspondence and requests for materials should be addressed to Jun-Il Yoo.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026