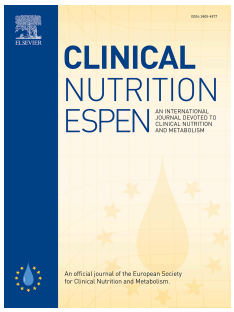
Journal Pre-proof



Use of automated assessment for determining associations of low muscle mass and muscle loss with overall survival in patients with colorectal cancer – a validation study

Karel C. Smit, Jeroen W.G. Derksen, Sophie A. Kurk, Pim Moeskops, Miriam Koopman, Wouter B. Veldhuis, Anne M. May

PII: S2405-4577(24)00203-1

DOI: https://doi.org/10.1016/j.clnesp.2024.07.001

Reference: CLNESP 2160

To appear in: *Clinical Nutrition ESPEN*

Received Date: 18 June 2024

Revised Date: 1 July 2024

Accepted Date: 2 July 2024

Please cite this article as: Smit KC, Derksen JWG, Kurk SA, Moeskops P, Koopman M, Veldhuis WB, May AM, Use of automated assessment for determining associations of low muscle mass and muscle loss with overall survival in patients with colorectal cancer – a validation study, *Clinical Nutrition ESPEN*, https://doi.org/10.1016/j.clnesp.2024.07.001.

Use of automated assessment for determining associations of low muscle mass and muscle loss with overall survival in patients with colorectal cancer – a validation study

Karel C. Smit[1,2], Jeroen W.G. Derksen[1], Sophie A. Kurk[1,2], Pim Moeskops[3], Miriam Koopman[2] Wouter B. Veldhuis[4*] & Anne M. May [1*]

[1] Department of Epidemiology and Health Economics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, PO Box 85500, 3508 GA, Utrecht, The Netherlands.

[2] Department of Medical Oncology, University Medical Center Utrecht, Utrecht University, PO Box 85500, 3508 GA, Utrecht, The Netherlands.

[3] Quantib, Westblaak 130, 3012 KM Rotterdam, The Netherlands

[4] Department of Radiology, University Medical Center Utrecht, Utrecht University, PO Box 85500, 3508 GA, Utrecht, The Netherlands.

[*] Anne M. May and Wouter B. Veldhuis share last authorship

Corresponding author

Professor Anne M. May, PhD

E-mail: a.m.may@umcutrecht.nl

Tel. +31 88 75 681 81

Use of automated assessment for determining associations of low muscle mass and muscle loss with overall survival in patients with colorectal cancer – a validation study

Karel C. Smit[1,2], Jeroen W.G. Derksen[1], Sophie A. Kurk[1,2], Pim Moeskops[3], Miriam Koopman[2] Wouter B. Veldhuis[4*] & Anne M. May [1*]

[1] Department of Epidemiology and Health Economics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, PO Box 85500, 3508 GA, Utrecht, The Netherlands.

[2] Department of Medical Oncology, University Medical Center Utrecht, Utrecht University, PO Box 85500, 3508 GA, Utrecht, The Netherlands.

[3] Quantib, Westblaak 130, 3012 KM Rotterdam, The Netherlands

[4] Department of Radiology, University Medical Center Utrecht, Utrecht University, PO Box 85500, 3508 GA, Utrecht, The Netherlands.

[*] Anne M. May and Wouter B. Veldhuis share last authorship

Corresponding author

Professor Anne M. May, PhD

E-mail: a.m.may@umcutrecht.nl

Tel. +31 88 75 681 81

Abstract

Background: Low muscle mass and skeletal muscle mass (SMM) loss are associated with adverse patient outcomes, but the time-consuming nature of manual SMM quantification prohibits implementation of this metric in clinical practice. Therefore, we assessed the feasibility of automated SMM quantification compared to manual quantification. We evaluated both diagnostic accuracy for low muscle mass and associations of SMM (change) with survival in colorectal cancer (CRC) patients.

Methods: Computed tomography (CT) images from CRC patients enrolled in two clinical studies were analyzed. We compared i) manual vs. automated segmentation of preselected slices at the third lumbar [L3] vertebra ("semi-automated"), and ii) manual L3-slice-selection + manual segmentation vs. automated L3-slice-selection + automated segmentation ("fully-automated"). Automated L3-selection and automated segmentation was performed with Quantib Body Composition v0.2.1. Bland-Altman analyses, within-subject coefficients of variation (WSCVs) and Intraclass Correlation Coefficients (ICCs) were used to evaluate the agreement between manual and automatic segmentation. Diagnostic accuracy for low muscle mass (defined by an established sarcopenia cut-off) was calculated with manual assessment as the “gold standard”. Using either manual or automated assessment, Cox proportional hazard ratios (HRs) were used to study the association between changes in SMM (>5% decrease yes/no) during first-line metastatic CRC treatment and mortality adjusted for prognostic factors. SMM change was also assessed separately in weight-stable (<5%, i.e. occult SMM loss) patients.

Results: In total, 1580 CT scans were analyzed, while a subset of 307 scans were analyzed in the fully-automated comparison. Included patients (n=553) had a mean age of 63±9 years and 39% were female. The semi-automated comparison revealed a bias of -2.41 cm^2, 95% limits of agreement [-9.02 to 4.20], a WSCV of 2.25%, and an ICC of 0.99 (95% confidence intervals (CI) 0.97 to 1.00). The fully-

automated comparison method revealed a bias of -0.08 cm^2 [-10.91 to 10.75], a WSCV of 2.85% and an ICC of 0.98 (95% CI 0.98 to 0.99). Sensitivity and specificity for low muscle mass were 0.99 and 0.89 for the semi-automated comparison and 0.96 and 0.90 for the fully-automated comparison. SMM decrease was associated with shorter survival in both manual and automated assessment (n=78/280, HR 1.36 [95% CI 1.03 to 1.80] and n=89/280, HR 1.38 [95% CI 1.05 to 1.81]). Occult SMM loss was associated with shorter survival in manual assessment, but not significantly in automated assessment (n=44/263, HR 1.43 [95% CI 1.01 to 2.03] and n=51/2639, HR 1.23 [95% CI 0.87 to 1.74]).

Conclusion: Deep-learning based assessment of SMM at L3 shows reliable performance, enabling the use of CT measures to guide clinical decision making. Implementation in clinical practice helps to identify patients with low muscle mass or (occult) SMM loss who may benefit from lifestyle interventions.

Keywords: colorectal cancer, body composition, muscle mass, survival analysis, deep-learning

Introduction

Cross-sectional skeletal muscle area (SMA), commonly measured through computed tomography (CT) imaging at the level of the 3rd lumbar vertebra, serves as a reliable indicator of overall skeletal muscle mass (SMM) (1,2). SMA assessment has garnered substantial attention due to its association with various adverse clinical outcomes. Both low muscle mass and a decline in muscle mass over time are associated with increased morbidity and mortality among diverse patient populations, but especially among the elderly and cancer patients (3–9). While manual assessment of SMA currently remains the gold standard, it is time-consuming and labor-intensive. This limits its feasibility in routine clinical practice and large-scale studies. Several studies have already explored automated segmentation techniques on a single axial CT slice, showcasing promising outcomes in accurately delineating SMA (10). However, studies focusing on a fully-automated approach, combining automated L3 localization with automated segmentation, are less prevalent, although also reporting promising results (11–15). Importantly, research focusing on reproducibility and validity of these methods through repeated measures within the same patient is sparse. This poses a limitation in the translation of these techniques into clinical practice and longitudinal research. Additionally, in clinical oncology practice there is a need for a diagnostic tool to identify sarcopenic and pre-cachectic patients that – from the outside – seem to be weight stable and in good condition. In a retrospective cohort of colorectal cancer (CRC) patients, it was found that incident sarcopenia occurs in 8.5% of patients during a mean follow-up of 15 months and was associated with a higher risk of death (16). Such patients easily remain unnoticed, while early interventions could prevent muscle depletion and may improve clinical outcome.(16)

Therefore, the primary objective of this study was to evaluate and compare the accuracy of (fully-)automated SMA assessment at the third lumbar level against manual assessment in a clinically relevant context. This included: i) analyzing the diagnostic accuracy of automated SMA assessment compared to manual assessment for identifying low muscle mass and tracking muscle loss over time using repeated measures within patients; and ii) comparing associations for changes in SMM and

88 overall survival (OS) between automated and manual assessment in patients with CRC, with

89 additional comparisons for weight stable and weight decreased patients separately.

Methods

Study population

CT scans assessed in this study were from patients with CRC enrolled in two clinical studies. The first is the Dutch Colorectal Cancer Group CAIRO3 study (17). This phase III study investigated the effect of maintenance treatment with low dose capecitabine plus bevacizumab (CAP-B) versus observation on overall survival in previously untreated metastatic CRC (mCRC) patients not progressing during first-line systemic treatment with six cycles of treatment with capecitabine plus oxaliplatin and bevacizumab (CAPOX-B). Patients were recruited from 64 Dutch hospitals between May 30, 2007 and Oct 15, 2012. CT scans were available from 55/64 hospitals. CT scans were collected at four timepoints: (i) before start of initial six cycles of CAPOX-B; (ii) at randomization between CAP-B and observation; (iii) at first disease progression (PD1), before start of CAPOX-B or other reintroduction treatment; (iv) at second disease progression (PD2), after progression on any systemic reintroduction treatment. Depending on availability of the CT scans, up to four CT scans per patient were used.

The second clinical study is PLCRC PROTECT-Plus, which is a substudy of the Prospective Dutch Colorectal Cancer (PLCRC) cohort (18,19). PLCRC PROTECT-Plus is an observational study investigating the association between body composition and treatment toxicity among patients with stage II-III colon cancer receiving adjuvant systemic therapy. Patients were recruited from three Dutch hospitals from 15 Jun, 2017 onwards. CT scans were collected at three timepoints: (i) at first CRC diagnosis; (ii) prior to the fourth adjuvant systemic therapy cycle; (iii) six months after surgery of primary tumor. Up to three CT scans per patient were used, depending on availability of the CT scans. Survival data is not yet available for PLCRC PROTECT-Plus.

A schematic overview of study characteristics and performed data analyses in both studies can be found in **Supplement 1**. All patients of both studies provided informed consent. Protocols were in accordance with the declaration of Helsinki.

Manual segmentation

All CT scans were manually segmented by trained researchers using the Slice-o-matic (version 5.0, Tomovision) software package using pre-specified tresholds for Hounsfield units (HU; -29 to 150) to identify and demarcate the muscle compartments. A single slice evaluation was performed at the level of the third lumbar vertebra (L3). SMA was calculated in square centimeters by adding the surface values of the psoas, abdominal and long spine muscle. SMA is highly correlated with total SMM, which can be calculated based on SMA by linear transformation (20–22). Results from manual assessment of skeletal muscle from the CAIRO3 studies have been previously published (5,23–27).

Automated segmentation

Automated segmentation was performed by the deep learning-based software Quantib Body Composition (version 0.2.1, Quantib, Rotterdam, Netherlands) (28). The Quantib Body Composition software is available online for testing through research.quantib.com. The automatic method consists of two steps: first, the slice at the mid-L3 level is automatically selected from the full image using a convolutional neural network. Second, 5 slices centered on the the selected mid-L3-level are automatically segmented into visceral fat, subcutaneous fat, psoas muscle, abdominal muscle and long spine muscle, using a second convolutional neural network. Before the analysis, the software resamples CT scans to a standard slice thickness of 5 mm. Segmentation results are averaged over the five segmented slices, i.e. over a standard 25 mm range centered on mid-L3.

The body composition tool was evaluated with two approaches:

1. The single L3 slices as selected during manual assessment were used as input , i.e., to evaluate the segmentation independently (semi-automated comparison).
2. Manually assessed L3 slices were compared with subsequent automated L3 localization and segmentation, (i.e., fully-automated comparison).

Differences between manual and automated L3 localization were calculated based on the middle of the five selected L3 slices (mid-L3) in automated L3 localization. The fully-automated comparison was performed with the subset of PLCRC PROTECT-Plus CT scans, for which full abdominal CT scans were

available. Results of fully-automated assessment of the psoas muscle at L3 compared to manual segmentation have been published previously and yielded a Dice index of 0.93 (±0.04 SD) (12). Four random examples as well as the four biggest outliers based on absolute SMA differences will be shown and possible causes for discrepancies between manual and automated assessment will be discussed.

CT scans were excluded from comparison when noise-level or acquisition artifacts precluded manual or automated segmentation, when manual segmentation was performed unilaterally due to the abdominal being partly missing and/or abnormalities in the abdominal wall or when the original DICOM source data could not be retrieved. Additionally, CT scans from PLCRC PROTECT-Plus were excluded when the L3 distance from the most caudal slide could not be calculated based on the available metadata.

Statistical analyses

SMA of manual and automated assessment was compared using Bland-Altman analyses with 95% limits of agreement (LoA) for both the semi-automated and fully-automated comparison, with adjustment for multiple observations per individual (29,30). Within-subject coefficients of variation (WSCVs) were also calculated, considering the broad normal range of SMA values and the possibility that within-subject variation is not uniform, but proportional to the magnitude of the measurement. The root mean squared method was used for the calculation of WSCVs (31). Two-way mixed Intraclass Correlation Coefficients (ICCs) were calculated as well. The same analyses were performed to compare L3 slice location for manual and fully-automated assessment. For percentual SMA changes between subsequent CT scans, WSCV will not be calculated, considering most patients will have minor SMM changes. This results in possible large relative differences, despite small absolute differences between manual and automated assessment, leading to large but uninformative WSCV values.

Several stratified analyses were conducted to identify potential subgroups in which automated assessment may be particularly (in)accurate. This was performed for sex (male/female), BMI (<18.5 / 18.5-25 / 25-30 / ≥30), and muscle mass (low muscle mass/normal muscle mass). Low muscle mass was defined based on the skeletal muscle index (SMI) and BMI of patients, using an established cut-off for sarcopenia, in accordance with prior research (5,32). SMI is calculated by dividing the SMA by squared height in meters. Cut-offs for low muscle mass were a SMI <43 and a BMI <25 or a SMI <53 and a BMI ≥25 for males, and a SMI <41 for females. These tresholds were established based on a large cohort of cancer patients by using optimal stratification to best separate patients based on their survival time. Results of the two clinical studies were compared as well, since PLCRC PROTECT-Plus is a more recent study and imaging quality might have improved thereby possibly influencing accuracy. WSCV differences between subgroups were calculated with Welch's t-test for subgroups with 2 categories, and with a one-way ANOVA with post-hoc Tukey-Kramer tests for subgroups with >2 categories.

Diagnostic accuracy of possible clinical implementations was assessed for low muscle mass, and for SMA change between subsequent CT scans. To improve generalizability, supplemental content will be provided of the same using two alternative sarcopenia cut-offs which are used in prior research, i.e. a SMI <55 for males and a SMI <39 for females (15,33), and a SMI <52.4 for males and a SMI <38.5 for females (14,34). SMA change was categorized as a >5% decrease (yes/no) in accordance with prior research (16). Manual assessment was used as gold standard. For low muscle mass, subgroup analyses were performed for sex, BMI, and study cohort. For SMA decrease, subgroup analyses were performed for sex and weight change. Weight change was categorized as >5% decrease (yes/no), in accordance with the cancer cachexia criteria by Fearon et al., and used in prior research (16,34). Diagnostic accuracy terms (sensitivity, specificity, negative predictive value [NPV], and positive predictive value [PPV]) were classified based on predefined thresholds. Values >0.7 were categorized as moderate, >0.8 as good, and >0.9 as excellent.

SMA – OS analysis (CAIRO3 study)

For these analyses, we used CAIRO3 data, since survival data for PROTECT-Plus are not yet available. To analyze associations between SMA and OS, Cox proportional hazards models were employed, comparing manual and automated SMM assessments. For these analyses, we selected the first two CT scans acquired before and after first-line treatment with six cycles of CAPOX-B. Median interval between the two CT scans was 4.6 months (IQR 4.1 to 5.2). Survival time was calculated from the time of randomization which can occur up to a few weeks after CT scan 2. It is worth noting that inclusion in CAIRO3 required stable disease or better post-first-line treatment, meaning we effectively evaluated associations with survival in mCRC patients who exhibited at least stable disease after six cycles of CAPOX-B. Analyses were performed for SMA decrease (yes/no) and the combination of weight decrease and SMA decrease (both stable/weight stable, SMA decrease >5% (i.e. 'occult SMM loss')/weight decrease >5%, SMA stable/weight & SMA decrease >5%). Furthermore, we conducted analyses for muscle mass (low muscle mass/normal muscle mass) for both CT scans separately, and for changes between the two CT scans (normal muscle mass both CT scans/new low muscle mass/persistent low muscle mass/remission of low muscle mass), in accordance with prior research (16). Similar to our approach evaluating diagnostic accuracy for low muscle mass, analyses will be repeated for two other commonly used sarcopenia cut-offs and provided as supplemental content. All models were adjusted for age (continuously), sex (male/female), WHO performance status (0/1), primary tumor site (colon/sigmoid/rectum), resection primary tumor (yes/no), response to initial treatment (stable disease/partial or complete response), lactate dehydrogenase (LDH) level at randomization (normal/elevated), metastasis patteren (synchronous/metachronous), dose reductions during initial treatment (yes/no), and treatment arm (CAP-B/observation).

All data were analyzed using R statistical software version 4.0.3 (35). All statistical tests were two-sided with an alpha level of 0.05.

Results

Study population

In total, 553 patients (39% female) were included in the full study population including CAIRO3 and PLCRC PROTECT-Plus patients. Baseline characteristics of both study populations can be found in **table 1**. After exclusion of scans that were unsuitable for comparison, 1580 and 307 CT scans remained for semi-automated and fully-automated comparison (**figure 1**). For the semi-automated comparison, average SMA was 139.9 ± 31.6 cm^2 for manual segmentation and 137 ± 31.5 cm^2 for automated segmentation, and 139.1 ± 33.2 cm^2 and 139.1 ± 31.8 cm^2 for the fully-automated comparison (**table 2**).

Semi-automated comparison

Median difference between manual and automated assessment was -2.40 cm^2 [IQR: -3.80 to -1.10]. Bland-Altman analyses displayed a non-significant small negative bias of -2.41 cm^2 [95% LoA: -9.02 to 4.20] for the entire dataset (n=1580) with a WSCV of 2.25% and an ICC of 0.99 (95% CI 0.97 to 1.00). (**figure 2, table 2**). Biases for subgroup analyses ranged from -1.81 cm^2 [95% LoA -8.90 to 5.27] to -2.80 cm^2 [95% LoA -9.99 to 4.40] cm^2. The degree of bias was not evidently correlated with SMA magnitude. WSCV values in subgroup analyses ranged from 1.81% to 3.78% with a statistically significant difference for PROTECT-Plus CT scans compared to CAIRO3 CT scans (1.81 vs 2.23%) and for BMI 25-30 compared to ≤18.5 (1.91 vs 3.78%). ICC values ranged from 0.94 to 0.99, with the lowest ICC for the n=26 BMI ≤18.5 subgroup. Four random examples and the four largest outliers based on absolute SMA difference between manual and automated assessment can be found in **figure 3**

Fully-automated comparison

L3 slice selection

Median difference between manual and automated mid-L3 slice selection was -0.50 mm [IQR -4.10 to 4.65]. Bland-Altman analyses showed a non-significant bias of 1.04 mm [95% LoA -22.17 to 24.25], a WSCV of 3.18% and an ICC of 0.85 (95% CI 0.82 to 0.88). Based on the height of an average L3 corpus of 32 mm (36), 82.1% of the mid-L3 slices are within 16 mm below and 16 mm above the manual level and are therefore expected to be on the same vertebra (**Figure 4**). No significant differences between subgroups were observed. ICC values ranged from 0.39 to 0.92 with the lowest ICC for the n=6 BMI ≤18.5 subgroup.

SMA assessment

Median difference between manual and automated assessment was 0.27 cm^2 [IQR -2.01 to 2.89]. Bland-Altman analyses displayed a non-significant small negative bias of -0.08 cm^2 [95% LoA -10.91 to 10.75] for the entire dataset (n=307) with a WSCV of 2.85% and an ICC of 0.98 (95% CI 0.98 to 0.99). (**figure 2, table 2**). Biases for subgroup analyses ranged from 1.26 cm^2 [95% LoA -10.63 to 13.15] to -1.51 cm^2 [95% LoA -14.11 to 11.08]. In fully-automated SMA comparison, the degree of bias was also not evidently correlated with SMA magnitude. WSCV values in subgroup analyses ranged from 2.58% to 3.10%. No significant different differences between subgroups were observed. ICC values ranged from 0.96 to 0.99, with the lowest ICC for both the n=157 male and n=150 female subgroups.

Validation of SMA change between subsequent CT scans

In the semi-automated comparison, SMA changes between subsequent CT scans could be calculated for n=1027 changes. Median change between two subsequent CT scans was small with -0.62% (IQR –5.08 to 3.57%) in manual assessment and -0.39% (IQR -5.48 to 4.37) in automated assessment, with a median difference between manual and automated assessment of 0.18% (IQR -0.88 to 1.26). Bland-Altman analyses displayed a non-significant bias 0.20% [95% LoA -6.13 to 6.53] for the entire dataset. In the fully-automated comparison, SMA changes between subsequent CT scans could be calculated for n=186 changes. Median change between two subsequent CT scans was small with 0.84% (IQR -

3.34 to 4.66) in manual assessment and 0.51% (IQR –3.07 to 4.36) in automated assessment, with a median difference between manual and automated assessment of -0.11% (IQR -1.51 to 1.59). Bland-Altman analyses displayed a non-significant bias 0.08% [95% LoA -6.48 to 6.64] for the entire dataset (**figure 5, table 3**).

Diagnostic accuracy for low muscle mass and SMA decrease between subsequent CT scans

A comprehensive overview of the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for both semi- and fully-automated comparisons for low muscle mass detection can be found in **table 4**. This same overview can be found for two other commonly used sarcopenia cut-offs in **supplement 2a and 2b.** In the semi-automated comparison, low muscle mass occurred in 680/1262 (54%) of CT scans. Overall sensitivity and NPV were excellent (0.99 and 0.98), with minor differences between subgroups. In the fully-automated comparison, low muscle mass occurred in 105/187 (57%) of CT scans, with excellent sensitivity and NPV as well (0.96 and 0.95). Compared to semi-automated assessment, greater differences across subgroups were seen, with sensitivity and NPV being notably lower in males compared to females and for BMI >30 compared to the other BMI subgroups.

The diagnostic accuracy for both semi- and fully-automated comparisons for SMA decrease can be found in **table 5**. In the semi-automated analysis, SMA decrease occurred in 259/1027 (25%) of subsequent CT scans. Overall sensitivity and NPV were excellent (0.90 and 0.97). Weight values were available for n=732 changes between subsequent CT scans. SMA decrease occurred in 139/624 (22%) patients with stable weight, and in 61/108 (56%) of patients with weight decrease. In the fully-automated analysis, SMA decrease occurred in 33/186 (18%) of subsequent scans. Overall sensitivity was moderate (0.73), with an excellent NPV (0.94). Weight values were available for n=68 changes between subsequent CT scans. SMA decrease occurred in 20/61 (33%) of patients with stable weight, and in 5/7 (71%) of patients with weight decrease. In general, diagnostic misclassification occurred

mainly for patients with a small SMA change in cm^2, especially in the semi-automated comparison (**Supplement 3**).

Survival analyses

Median follow-up time in the CAIRO3 study population (n=432) was 20.3 months (IQR: 12.0 to 31.4), during which a total number of 414 (96%) of patients died. CT scan and weight information was accessible for 308 patients for the initial scan and for 341 for the subsequent scan. 280 patients had scan data available for both scans, and 263 patients had both weight and CT scan data available at both timepoints. Detailed patient characteristics can be found in **supplement 4**. Low muscle mass both before and after completion of first-line CAPOX-B was not associated with shorter survival, nor were muscle mass changes during first-line treatment in both manual and automated assessment (**figure 6, supplement 5**). A comparison of the multivariate analyses with two other commonly used sarcopenia cut-offs can be found in **supplement 6**. SMA decrease during this time was associated with shorter survival in both manual (adjusted HR [HR_{adj}]) 1.36 [1.03 to 1.80]) and semi-automated assessment (HR_{adj} 1.38 [1.05 to 1.81]). No significant associations with early death were observed among patients with weight decrease and stable SMA compared to weight and SMA stable patients. SMA decrease among weight stable patients compared to both weight and SMA stable patients was associated with significantly shorter survival in manual assessment (HR_{adj} 1.42 [1.01 to 2.03]), but not in automated assessment (HR_{adj} 1.23 [0.87 to 1.74]). A weight and SMA decrease compared to weight and SMA stable patients was not associated with death in manual assessment (HR_{adj} 1.48 [0.98 to 2.22]), but with shorter survival in automated assessment (HR_{adj} 1.72 [1.15 to 2.58]).

Discussion

In this analysis, we found that deep-learning based assessment of skeletal muscle mass at the third lumbar vertebra shows reliable performance. Furthermore, diagnostic accuracy was excellent for ruling out low muscle mass and good for detecting SMM decrease above a pre-established threshold, with semi-automated assessment performing notably better compared to fully-automated assessment for the latter. Lastly, we found an association between (occult) SMM loss and shorter survival during first-line treatment in patients with mCRC, with comparable hazard ratios between manual and automated assessment of SMM.

Manual vs automatic SMA assessment

Our validation results are in line with an increasing number of studies showing reliable performance of automated assessment of SMA (10). Generally we observed a small underestimation of skeletal muscle compared to the current gold standard of manual segmentation, which has been reported by others as well (37). Differences in performance among subgroups are small, with automated assessment generally performing better in subgroups with higher SMA. Several explanations have been given for this, including that small underestimations have less impact for patients with larger muscle areas, and that in case of severe muscle wasting, the abdominal wall can be compromised, leading to segmentation errors (12,37,38). Performance in the PROTECT-Plus cohort was slightly better compared to the CAIRO3 cohort, which might be explained by improvements in imaging modalities over time, since PROTECT-Plus is a more recent study. Detailed analyses comparing e.g. different CT scan brands could not be performed, since this data was not available. Furthermore, we found that automated L3 slice selection was accurate in the large majority of cases. Small changes in L3 slice selection are not expected to lead to large differences, given that no substantial changes occur in skeletal muscle area at the L3 level (39). This is indeed consistent with our observation that biases, WSCV and ICC values were comparable between semi-automated and fully-automated assessment in our analyses. Additionally, based on the visual inspection of outliers in our analyses, it

342 is important to exercise caution when applying automated assessment to patients with unusual intra-
343 abdominal anatomy. This is demonstrated by the four largest outliers in **Figure 3**. These outliers
344 include patients who exhibit unusual HU-densities in anatomical structures adjacent to the skeletal
345 muscle due to underlying pathological conditions, resulting in a less distinguishable abdominal wall.
346
347 **Using automated SMA assessment as a diagnostic tool**
348 To our knowledge, no other validation study has shown diagnostic accuracy of automated
349 assessment for changes of low muscle mass and skeletal muscle changes over time; i.e. analyzing
350 repeated assessments of the same patients, and limited research has been conducted on low muscle
351 mass (13–15). We found excellent diagnostic accuracy for low muscle mass in both semi-automated
352 and fully-automated assessment, with its main strength being its ability to identify true negative
353 cases. Our diagnostic results fit clinical implementation as a screening tool, where automated SMM
354 assessment could help identify patients who might benefit from exercise or dietary interventions,
355 and thereby also providing additional means to evaluate such interventions. Minor overdiagnosis is
356 not expected to lead to adverse consequences, since these interventions are recognized as safe and
357 feasible for patients with CRC, and potentially improve various health-related outcomes, benefitting
358 not only sarcopenic patients but CRC patients in general (40–42).
359 Total diagnostic accuracy for SMA decrease was excellent in both semi-automated and fully-
360 automated assessment as well, although sensitivity, specificity, PPV and NPV values showed larger
361 variety. In this context especially, it should be noted that using manual assessment as gold standard
362 has some limitations. Manual assessment of SMA is not perfectly accurate, and our reported
363 variability of a few percent between manual and automated assessment is in fact comparable with
364 interobserver variability of manual assessment performed by trained researchers (37,43). In other
365 words, a current limitation of both manual and automated assessment is an expected error rate of a
366 few percent per CT scan, resulting in increased diagnostic misclassification when assessing SMM
367 changes between subsequent CT scans. This is especially important for patients with low SMM,

illustrated by **supplement 2**, showing that most disagreement between automated and manual assessment occurred with relatively small absolute SMM changes. Additionally, the generally lower sensitivity values in fully-automated assessment compared to semi-automated assessment might reflect minor differences in selection of the L3 slice compared to manual selection, resulting in a subsequent small difference besides the inter-observer/intermodality segmentation error rate. Automated assessment however, provides the opportunity to segment multiple CT slices, which eventually might improve accuracy and enable future studies to focus on evaluating predictive value of skeletal muscle volume, instead of SMA. Automated assessment also enables the possibility to assess muscle mass without using a HU threshold, so that changes in intramuscular fat can also be assessed, thereby more directly targeting muscle mass changes, which might prove to be an alternative predictive marker (44).

Survival analyses

Whilst the primary focus for our survival analyses was to validate automated assessment of SMA, the results of manual assessment of different SMM variables should be put into context first. An increasing body of research shows associations between low muscle mass and shorter survival in cancer patients (4,9). In our study however, and probably due to the study population, we did not observe any associations between low muscle mass (change) and survival, which is in line with a previous analysis with the CAIRO3 study population. CAIRO3 consists of mCRC patients with a relatively good prognosis, because only patients with stable disease or better after 6 cycles of CAPOX-B are included, and patients with disease progression and/or inacceptable toxicity during first-line CAPOX-B were eligible for inclusion (5). Additionally, low muscle mass prevalence in our study exceeded 50%, which is higher than other reported prevalences in CRC patients of around 40%, further indicating our study population might be a specific CRC subgroup (9,45). Besides this, low muscle mass prevalence differs widely between cancers while certain types of cancer do not show significant associations, which might indicate that cancer type-specific cutoff values for low muscle

mass might be warranted (46). This hypothesis would also fit the differences we found on diagnostic accuracy compared to Gu et al., who reported a lower sensitivity (0.718) compared to specificity (0.876) in a group of females in China who underwent DXA examinations. After calculating an optimal cutoff value based on their data instead of the pre-established cut-off by Prado et al., they reported a higher AUC with higher sensitivity (0.875) compared to specificity (0.778) (15,33). The impact of applying different sarcopenia cut-offs is also illustrated by the varying effect sizes between those cut-offs as shown in **supplement 6.** Also, contrary to prior research (16), we did not find significant associations with shorter survival for patients with newly diagnosed low muscle mass; although the referenced study was conducted conducted in stage I-III CRC patients and might thus not be directly comparable.

In line with another study examing SMM decrease during first-line systemic therapy, we observed a significant association with shorter survival for patients with SMM decrease (47). Results are also similar compared to other CRC domains and/or other SMM cut-offs (48,49). Morever, we found a significant association with shorter survival for occult SMM loss, not for a decrease of weight only, nor - unexpectedly - for a decrease of both, although numbers were small in both those groups.

Comparing manual assessment with automated assessment, we observed agreement in statistical significance for most SMM variables. In general, automated assessment labels more patients as having low muscle mass and/or decrease in SMM, thereby possibly overdiagnosing a subset of patients. This might explain that most HRs of automated assessment are attenuated towards one compared to manual assessment. The only exception here is the subgroup of patients who lost both body weight and SMM, where automated assessment was significantly associated with shorter survival, but manual assessment showed no significant association with death. This change was based on 4/31 patients in automated assessment that were classified differently compared to manual assessment (data not shown), indicating that subgroups categorized by weight/SMM lack optimally stabilized hazard ratios in our study because of their small sample sizes.

Regardless, our results indicate that weight decrease might not always be indicative of SMM decrease. Additionally, individuals experiencing weight loss alone appear to exhibit a lower risk profile compared to those with SMM loss. This underscores the importance of evaluating SMM in clinical practice—an aspect that is not commonly implemented, likely owing to time constraints, for which automated assessment provides a viable solution. Future studies should concentrate on establishing more robust criteria for SMM loss and further evaluate whether SMM decrease can be stopped or even reversed.

Strengths and limitations

There are several strengths to this study. We analyzed a large sample of patients with longitudinal data, providing the opportunity to validate performance to assess patterns of body composition; an aspect that has received limited attention in prior research. Moreover, validation of fully-automated assessment of SMA, which includes automated localization of L3, remains an area that has not been thoroughly investigated in current literature. Furthermore, we provided a comprehensive overview by both validating SMA assessment, and subsequently evaluating the performance of automated SMM assessment in possible clinical contexts, while also focusing on low muscle mass and SMM changes in weight stable patients, which has not been extensively evaluated.

Limitations are that we were only able to evaluate fully-automated SMM in a subset of our study, due to missing full abdominal CT scan data, which resulted in small subsets for some stratified analyses in the fully-automated comparison, especially for BMI subcategories. Furthermore, despite having excellent agreement in most cases, automated assessment does not always produce optimal segmentation, especially among patients with low BMI. We believe that the reported accuracy is adequate for scientific research purposes, optionally supported by quality checks in selected subgroups with poorer performance, thereby enabling and facilitating large scale studies. Based on our results, we advise such quality checks for patients with low muscle mass and/or uncommon intro-abdominal anatomy. These checks are especially important when assessing SMM change. For

445 application in clinical context, based on the current data, a quick visual inspection is advisable to
446 identify rare automated segmentation failures.
447
448 **Conclusion**
449 Deep-learning based assessment of skeletal muscle mass at the third lumbar vertebra shows reliable
450 performance, enabling the use of CT measures to guide clinical decision making. Our data suggests
451 implementation as a screening tool to help identify patients with low muscle mass or (occult) SMM
452 loss for whom lifestyle interventions might improve their outcomes.

Acknowledgement section

Acknowledgements

The authors would like to thank all patients and staff at each of the study centres.

Funding statement

This research was supported by a grant from the Dutch national and local government (Regio Deal Foodvalley, grant nr 162135). The funders had no role in the design, conduct, collection, management, analysis, or interpretation of the data, nor drafting, reviewing, approving the manuscript, or deciding to publish this study.

Conflict of interest

MK reports institutional financial interests with Amgen, Bayer, BMS, Merck-Serono, Nordic Pharma, Roche, Servier, Sirtex, and Sanofi-Aventis. Miriam Koopman reports the following non-financial interests: an advisory role for ZON-MW, membership of the scientific board of the Dutch Cancer Society (KWF), chairmanship of the Dutch Colorectal Cancer Group (DCCG), and principal investigator (PI) of the Prospective Dutch Colorectal Cancer (PLCRC) cohort.

WV was a co-founder of Quantib-U.

PM was employed by Quantib-U.

All other authors declare no conflicts of interest.

Author contribution

All authors contributed to conceptualization, methodology and writing – review & editing of the research. KS, JD, SK, PM, and WV contributed to data curation. KS performed formal analysis, visualization and wrote the first draft under supervision of JD, MK, AM, and WV. All authors read and approved the final version of the manuscript.

References

1. Shen W, Punyanitya M, Wang Z, Gallagher D, St.-Onge MP, Albu J, et al. Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image. J Appl Physiol. 2004 Dec;97(6):2333–8.
2. Prado CM, Birdsell LA, Baracos VE. The emerging role of computerized tomography in assessing cancer cachexia. Curr Opin Support Palliat Care. 2009 Dec;3(4):269–75.
3. Veronese N, Demurtas J, Soysal P, Smith L, Torbahn G, Schoene D, et al. Sarcopenia and health-related outcomes: an umbrella review of observational studies. Eur Geriatr Med. 2019 Dec 6;10(6):853–62.
4. Ligibel JA, Schmitz KH, Berger NA. Sarcopenia in aging, obesity, and cancer. Transl Cancer Res. 2020 Sep;9(9):5760–71.
5. Kurk SA, Peeters PHM, Dorresteijn B, de Jong PA, Jourdan M, Creemers GM, et al. Loss of skeletal muscle index and survival in patients with metastatic colorectal cancer: Secondary analysis of the phase 3 CAIRO3 trial. Cancer Med. 2020 Feb 18;9(3):1033–43.
6. Surov A, Strobel A, Borggrefe J, Wienke A. Low skeletal muscle mass predicts treatment response in oncology: a meta-analysis. Eur Radiol. 2023 Mar 16;33(9):6426–37.
7. Pamoukdjian F, Bouillet T, Lévy V, Soussan M, Zelek L, Paillaud E. Prevalence and predictive value of pre-therapeutic sarcopenia in cancer patients: A systematic review. Clinical Nutrition. 2018 Aug;37(4):1101–13.
8. Lopez PD, Nepal P, Akinlonu A, Nekkalapudi D, Kim K, Cativo EH, et al. Low Skeletal Muscle Mass Independently Predicts Mortality in Patients with Chronic Heart Failure after an Acute Hospitalization. Cardiology. 2019;142(1):28–36.
9. Trejo-Avila M, Bozada-Gutiérrez K, Valenzuela-Salazar C, Herrera-Esquivel J, Moreno-Portillo M. Sarcopenia predicts worse postoperative outcomes and decreased survival rates in patients with colorectal cancer: a systematic review and meta-analysis. Int J Colorectal Dis. 2021 Jun 22;36(6):1077–96.

10. Rozynek M, Kucybała I, Urbanik A, Wojciechowski W. Use of artificial intelligence in the imaging of sarcopenia: A narrative review of current status and perspectives. Nutrition. 2021 Sep;89:111227.

11. Dabiri S, Popuri K, Ma C, Chow V, Feliciano EMC, Caan BJ, et al. Deep learning method for localization and segmentation of abdominal CT. Computerized Medical Imaging and Graphics. 2020 Oct;85:101776.

12. Van Erck D, Moeskops P, Schoufour JD, Weijs PJM, Scholte Op Reimer WJM, Van Mourik MS, et al. Evaluation of a Fully Automatic Deep Learning-Based Method for the Measurement of Psoas Muscle Area. Front Nutr. 2022 May 12;9.

13. Pickhardt PJ, Perez AA, Garrett JW, Graffy PM, Zea R, Summers RM. Fully Automated Deep Learning Tool for Sarcopenia Assessment on CT: L1 Versus L3 Vertebral Level Muscle Measurements for Opportunistic Prediction of Adverse Clinical Outcomes. American Journal of Roentgenology. 2022 Jan 1;218(1):124–31.

14. Nandakumar B, Baffour F, Abdallah NH, Kumar SK, Dispenzieri A, Buadi FK, et al. Sarcopenia identified by computed tomography imaging using a deep learning–based segmentation approach impacts survival in patients with newly diagnosed multiple myeloma. Cancer. 2023 Feb 22;129(3):385–92.

15. Gu S, Wang L, Han R, Liu X, Wang Y, Chen T, et al. Detection of sarcopenia using deep learning-based artificial intelligence body part measure system (AIBMS). Front Physiol. 2023 Jan 26;14.

16. Brown JC, Caan BJ, Cespedes Feliciano EM, Xiao J, Weltzien E, Prado CM, et al. Weight stability masks changes in body composition in colorectal cancer: a retrospective cohort study. Am J Clin Nutr. 2021 Jun;113(6):1482–9.

17. Simkens LHJ, Van Tinteren H, May A, Ten Tije AJ, Creemers GJM, Loosveld OJL, et al. Maintenance treatment with capecitabine and bevacizumab in metastatic colorectal cancer (CAIRO3): A phase 3 randomised controlled trial of the Dutch Colorectal Cancer Group. The Lancet. 2015;385(9980):1843–52.

18. Burbach JPM, Kurk SA, Coebergh van den Braak RRJ, Dik VK, May AM, Meijer GA, et al. Prospective Dutch colorectal cancer cohort: an infrastructure for long-term observational, prognostic, predictive and (randomized) intervention research. Acta Oncol (Madr). 2016 Nov 1;55(11):1273–80.

19. Derksen JWG, Vink GR, Elferink MAG, Roodhart JML, Verkooijen HM, van Grevenstein WMU, et al. The Prospective Dutch Colorectal Cancer (PLCRC) cohort: real-world data facilitating research and clinical care. Sci Rep. 2021 Dec 1;11(1).

20. Shen W, Punyanitya M, Wang Z, Gallagher D, St.-Onge MP, Albu J, et al. Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image. J Appl Physiol. 2004 Dec;97(6):2333–8.

21. Mourtzakis M, Prado CMM, Lieffers JR, Reiman T, McCargar LJ, Baracos VE. A practical and precise approach to quantification of body composition in cancer patients using computed tomography images acquired during routine care. Applied Physiology, Nutrition, and Metabolism. 2008 Oct;33(5):997–1006.

22. Ward SR, Lieber RL. Density and hydration of fresh and fixed human skeletal muscle. J Biomech. 2005 Nov;38(11):2317–20.

23. Kurk S, Peeters P, Stellato R, Dorresteijn B, de Jong P, Jourdan M, et al. Skeletal muscle mass loss and dose-limiting toxicities in metastatic colorectal cancer patients. J Cachexia Sarcopenia Muscle. 2019 Aug 15;10(4):803–13.

24. Derksen JWG, Kurk SA, Peeters PHM, Dorresteijn B, Jourdan M, van der Velden AMT, et al. The association between changes in muscle mass and quality of life in patients with metastatic colorectal cancer. J Cachexia Sarcopenia Muscle. 2020 Aug 27;11(4):919–28.

25. Kurk SA, Peeters PHM, Dorresteijn B, de Jong PA, Jourdan M, Kuijf HJ, et al. Impact of different palliative systemic treatments on skeletal muscle mass in metastatic colorectal cancer patients. J Cachexia Sarcopenia Muscle. 2018 Oct 24;9(5):909–19.

26. Kurk SA, Stellato RK, Peeters PH, Dorresteijn B, Jourdan M, Oskam MJ, et al. Trajectory of body mass and skeletal muscle indices and disease progression in metastatic colorectal cancer patients. Am J Clin Nutr. 2019 Dec;110(6):1395–403.

27. Derksen JWG, Kurk SA, Oskam MJ, Peeters PHM, Punt CJA, Koopman M, et al. Factors Contributing to Cancer-Related Muscle Wasting During First-Line Systemic Treatment for Metastatic Colorectal Cancer. JNCI Cancer Spectr. 2019 Jun 1;3(2).

28. Moeskops, P.; de Vos, B.; Veldhuis, W.B.; de Jong, P.A.; Išgum, I.; Leiner T. Automatic quantification of body composition at L3 vertebra level with convolutional neural networks [conference poster]. European Congress of Radiology. 2020;

29. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. J Biopharm Stat. 2007 Jul;17(4):571–82.

30. Bland J, Altman D. Statistical Methods for assessing agreement between two methods of clinical measurement. The Lancet. 1986 Feb;327(8476):307–10.

31. Bland J. How should I calculate a within-subject coefficient of variation? [Internet]. 2006 [cited 2023 Nov 9]. Available from: https://www-users.york.ac.uk/~mb55/meas/cv.htm#:~:text=In%20the%20study%20of%20measurement,within%2Dsubject%20coefficient%20of%20variation

32. Martin L, Birdsell L, MacDonald N, Reiman T, Clandinin MT, McCargar LJ, et al. Cancer Cachexia in the Age of Obesity: Skeletal Muscle Depletion Is a Powerful Prognostic Factor, Independent of Body Mass Index. Journal of Clinical Oncology. 2013 Apr 20;31(12):1539–47.

33. Prado CM, Lieffers JR, McCargar LJ, Reiman T, Sawyer MB, Martin L, 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 Jul;9(7):629–35.

34. Fearon K, Strasser F, Anker SD, Bosaeus I, Bruera E, Fainsinger RL, et al. Definition and classification of cancer cachexia: an international consensus. Lancet Oncol. 2011 May;12(5):489–95.

582 35. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna,
583 Austria: R foundation for Statistical Computing; 2020. Available from: https://www.r-
584 project.org/
585 36. Zhou SH, McCarthy ID, McGregor AH, Coombs RRH, Hughes SPF. Geometrical dimensions of
586 the lower lumbar vertebrae - analysis of data from digitised CT images. European Spine
587 Journal. 2000 Jun 7;9(3):242–8.
588 37. Cespedes Feliciano EM, Popuri K, Cobzas D, Baracos VE, Beg MF, Khan AD, et al. Evaluation of
589 automated computed tomography segmentation to assess body composition and mortality
590 associations in cancer patients. J Cachexia Sarcopenia Muscle. 2020 Oct 20;11(5):1258–69.
591 38. Paris MT, Tandon P, Heyland DK, Furberg H, Premji T, Low G, et al. Automated body
592 composition analysis of clinically acquired computed tomography scans using neural
593 networks. Clinical Nutrition. 2020 Oct;39(10):3049–55.
594 39. Park J, Gil JR, Shin Y, Won SE, Huh J, You MW, et al. Reliable and robust method for abdominal
595 muscle mass quantification using CT/MRI: An explorative study in healthy subjects. PLoS One.
596 2019 Sep 19;14(9):e0222042.
597 40. Singh B, Hayes SC, Spence RR, Steele ML, Millet GY, Gergele L. Exercise and colorectal cancer:
598 a systematic review and meta-analysis of exercise safety, feasibility and effectiveness.
599 International Journal of Behavioral Nutrition and Physical Activity. 2020 Dec 24;17(1):122.
600 41. Campbell KL, Winters-Stone KM, Wiskemann J, May AM, Schwartz AL, Courneya KS, et al.
601 Exercise Guidelines for Cancer Survivors: Consensus Statement from International
602 Multidisciplinary Roundtable. Med Sci Sports Exerc. 2019 Nov 1;51(11):2375–90.
603 42. Rock CL, Doyle C, Demark-Wahnefried W, Meyerhardt J, Courneya KS, Schwartz AL, et al.
604 Nutrition and physical activity guidelines for cancer survivors. CA Cancer J Clin. 2012
605 Jul;62(4):242–74.

606 43. Ozola-Zālīte I, Mark EB, Gudauskas T, Lyadov V, Olesen SS, Drewes AM, et al. Reliability and
607 validity of the new VikingSlice software for computed tomography body composition analysis.
608 Eur J Clin Nutr. 2019 Jan 16;73(1):54–61.
609 44. van Erck D, Moeskops P, Schoufour JD, Weijs PJM, Scholte op Reimer WJM, van Mourik MS, et
610 al. Low muscle quality on a procedural computed tomography scan assessed with deep
611 learning as a practical useful predictor of mortality in patients with severe aortic valve
612 stenosis. Clin Nutr ESPEN. 2024 Jun;
613 45. Surov A, Wienke A. Prevalence of sarcopenia in patients with solid tumors: A meta-analysis
614 based on 81,814 patients. Journal of Parenteral and Enteral Nutrition. 2022 Nov
615 19;46(8):1761–8.
616 46. Li HL, Au PCM, Lee GKY, Li GHY, Chan M, Cheung BMY, et al. Different definitions of
617 sarcopenia and mortality in cancer: A meta-analysis. Osteoporos Sarcopenia. 2021 Mar;7:S34–
618 8.
619 47. Miyamoto Y, Baba Y, Sakamoto Y, Ohuchi M, Tokunaga R, Kurashige J, et al. Negative Impact
620 of Skeletal Muscle Loss after Systemic Chemotherapy in Patients with Unresectable Colorectal
621 Cancer. PLoS One. 2015 Jun 12;10(6):e0129742.
622 48. Brown JC, Caan BJ, Meyerhardt JA, Weltzien E, Xiao J, Cespedes Feliciano EM, et al. The
623 deterioration of muscle mass and radiodensity is prognostic of poor survival in stage I–III
624 colorectal cancer: a population-based cohort study (<scp>C-SCANS</scp>). J Cachexia
625 Sarcopenia Muscle. 2018 Aug 15;9(4):664–72.
626 49. Blauwhoff-Buskermolen S, Versteeg KS, de van der Schueren MAE, den Braver NR, Berkhof J,
627 Langius JAE, et al. Loss of Muscle Mass During Chemotherapy Is Predictive for Poor Survival of
628 Patients With Metastatic Colorectal Cancer. Journal of Clinical Oncology. 2016 Apr
629 20;34(12):1339–44.
630