ORIGINAL ARTICLE

Journal of Cachexia, Sarcopenia and Muscle 2023; **14**: 418–428
Published online 1 December 2022 in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/jcsm.13144

Chest X-ray-based opportunistic screening of sarcopenia using deep learning

Jin Ryu[1], Sujeong Eom[2], Hyeon Chang Kim[2,3], Chang Oh Kim[4], Yumie Rhee[1], Seng Chan You[2,3]* & Namki Hong[1,3]*

[1]*Department of Internal Medicine, Severance Hospital, Endocrine Research Institute, Yonsei University College of Medicine, Seoul, South Korea;* [2]*Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, South Korea;* [3]*Institute for Innovation in Digital Healthcare, Yonsei University, Seoul, South Korea;* [4]*Division of Geriatrics, Department of Internal Medicine, Yonsei University College of Medicine, Seoul, South Korea*

Abstract

Background Early detection and management of sarcopenia is of clinical importance. We aimed to develop a chest X-ray-based deep learning model to predict presence of sarcopenia.

Methods Data of participants who visited osteoporosis clinic at Severance Hospital, Seoul, South Korea, between January 2020 and June 2021 were used as derivation cohort as split to train, validation and test set (65:15:20). A community-based older adults cohort (KURE) was used as external test set. Sarcopenia was defined based on Asian Working Group 2019 guideline. A deep learning model was trained to predict appendicular lean mass (ALM), handgrip strength (HGS) and chair rise test performance from chest X-ray images; then the machine learning model (SARC-CXR score) was built using the age, sex, body mass index and chest X-ray predicted muscle parameters along with estimation uncertainty values.

Results Mean age of the derivation cohort ($n = 926$; women $n = 700$, 76%; sarcopenia $n = 141$, 15%) and the external test ($n = 149$; women $n = 95$, 64%; sarcopenia $n = 18$, 12%) cohort was 61.4 and 71.6 years, respectively. In the internal test set (a hold-out set, $n = 189$, from the derivation cohort) and the external test set ($n = 149$), the concordance correlation coefficient for ALM prediction was 0.80 and 0.76, with an average difference of 0.18 ± 2.71 and 0.21 ± 2.28, respectively. Gradient-weight class activation mapping for deep neural network models to predict ALM and HGS commonly showed highly weight pixel values at bilateral lung fields and part of the cardiac contour. SARC-CXR score showed good discriminatory performance for sarcopenia in both internal test set [area under the receiver-operating characteristics curve (AUROC) 0.813, area under the precision-recall curve (AUPRC) 0.380, sensitivity 0.844, specificity 0.739, F1-score 0.540] and external test set (AUROC 0.780, AUPRC 0.440, sensitivity 0.611, specificity 0.855, F1-score 0.458). Among SARC-CXR model features, predicted low ALM from chest X-ray was the most important predictor of sarcopenia based on SHapley Additive exPlanations values. Higher estimation uncertainty of HGS contributed to elevate the predicted risk of sarcopenia. In internal test set, SARC-CXR score showed better discriminatory performance than SARC-F score (AUROC 0.813 vs. 0.691, $P = 0.029$).

Conclusions Chest X-ray-based deep leaning model improved detection of sarcopenia, which merits further investigation.

Keywords Sarcopenia; Chest X-ray-based deep learning model; Appendicular lean mass; Artificial intelligence; Chest radiograph

Received: 19 June 2022; Revised: 26 October 2022; Accepted: 31 October 2022

**Correspondence to: Seng Chan You, M.D., Ph.D., and Namki Hong, M.D., M.P.H., Ph.D., Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, South Korea. Institute for Innovation in Digital Healthcare, Yonsei University, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea. Email: chandryou@yuhs.ac;nkhong84@yuhs.ac*

Introduction

Sarcopenia is a progressive, generalized skeletal muscle disorder that is characterized by decreased muscle function and mass. The prevalence of sarcopenia varies from 10% to 40% depending on the operational definitions in community-dwelling older adults.[1] Individuals with sarcopenia are associated with twofold elevated pooled risk of mortality according to a meta-analysis using 57 studies with 42 108 participants, independent of population and sarcopenia definitions.[2] Sarcopenia also predisposes individuals to elevated risk of falls, osteoporosis, fractures and disability.[3–5] Given the substantial burden of sarcopenia, current clinical practice guidelines emphasize the importance of the case-finding steps, including the SARC-F questionnaire to detect individuals at high risk of sarcopenia.[6,7] However, sarcopenia often remains underdetected in routine clinical practice, which suggests the need for an effective and pragmatic opportunistic screening strategy to improve sarcopenia detection.

Chest radiography is one of the most frequently and widely used diagnostic imaging modalities that accounts for about one-quarter of the annual total number of diagnostic imaging procedures in developed countries.[8] Recent progress in deep neural network algorithms has enabled the quantification of clinically useful latent features from chest radiographs, with promising results in computer-aided diagnosis of lung disease, mortality prediction and opportunistic screening of metabolic diseases such as osteoporosis.[9–11] Nonetheless, limited explainability and robustness of these artificial intelligence approaches has challenged active applications.

In this study, we aimed to develop an explainable artificial intelligence model, called SARC-CXR (chest X-ray), using chest radiographs and basic clinical parameters to predict sarcopenia, and to validate the model using the community-based prospective cohort.

Methods

Study subjects

Derivation set (hospital-based osteoporosis clinic cohort)

The patients who visited the osteoporosis clinic at Severance Hospital, Seoul, South Korea, between January 2020 to June 2021 were screened (*Figure* 1). A total of 1076 patients underwent chest X-ray scans within 3 months from the visit date to the osteoporosis clinic for various reasons including routine health examinations, evaluation for related symptoms or follow-up for underlying cardiopulmonary diseases. After excluding individuals with age under 40 (*n* = 58) or those who did not undergo any muscle function measurements (*n* = 92), 926 patients remained in the derivation cohort. To build train, validation and internal test set, we applied commonly recommended 80:20 (20 for test set) split ratio.[12] Within the remaining 80% dataset, we applied same ratio to split train and validation set, which yielded 65% train set and 15% validation set. The presence of osteoporosis, previous fractures, hypertension and diabetes was defined as the presence of related diagnosis codes and medication prescriptions in the electronic health records for each disease in the derivation set. The SARC-F questionnaire score was collected for all subjects in the derivation cohort.[13]

This study was approved by the Institutional Review Board of Severance Hospital (IRB No. 4-2021-1466).

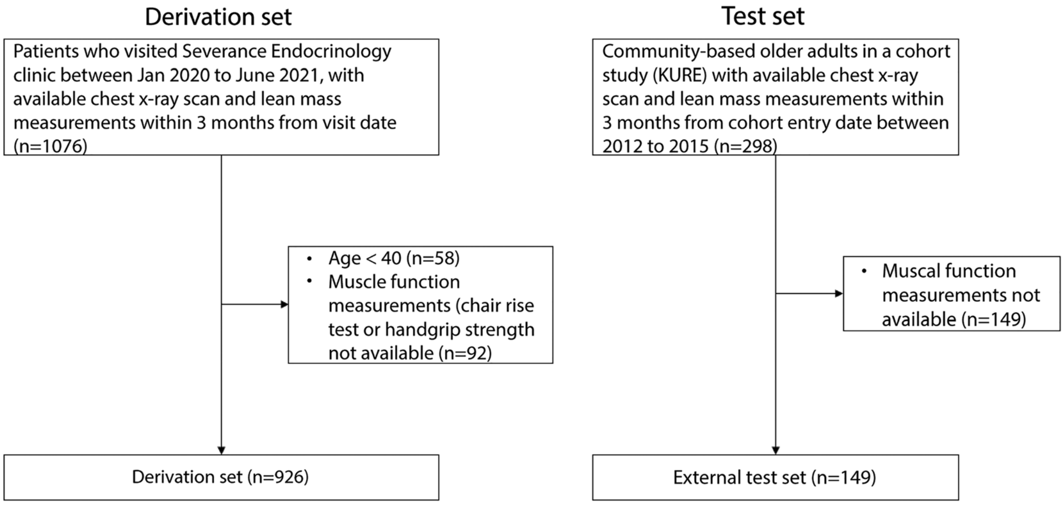
Derivation set
Patients who visited Severance Endocrinology clinic between Jan 2020 to June 2021, with available chest x-ray scan and lean mass measurements within 3 months from visit date (n=1076)
• Age < 40 (n=58)
• Muscle function measurements (chair rise test or handgrip strength not available (n=92)
Derivation set (n=926)
Test set
Community-based older adults in a cohort study (KURE) with available chest x-ray scan and lean mass measurements within 3 months from cohort entry date between 2012 to 2015 (n=298)
• Muscal function measurements not available (n=149)
External test set (n=149)

Figure 1 Study flow.

DOI: 10.1002/jcsm.13144

External test set (community-based prospective cohort)

To test the model performance in an external setting, we used a subset of a prospective community-based cohort dataset. Briefly, the Korean Urban Rural Elderly (KURE) cohort was built to investigate newly emerging risk factors in community-dwelling Korean older adults, mainly related to cardiovascular and musculoskeletal disorders.[14,15] Baseline recruitment for the cohort was performed between 2012 and 2015 (n = 3517), and all examinations for this cohort were performed at Severance Hospital with individual written permission (IRB No. 4-2012-0172). Among them, chest X-ray images are available for 298 patients within 3 months from cohort entry (n = 298; *Figure* 1). After excluding individuals without complete ALM data and muscle function assessment at the cohort entry, data of 149 individuals were analysed as the final external test set. The presence of osteoporosis, previous fractures, hypertension and diabetes was defined as any history of physician-made diagnosis of each disease with current medication use obtained from an interviewer-assisted questionnaire according to the cohort assessment protocol.

Sarcopenia assessment

The presence of sarcopenia was defined according to the 2019 AWGS consensus update.[7] ALM was estimated using multi-frequency BIA (InBody 720, Biospace Co., Ltd., Seoul, Korea), measuring the impedance with frequencies of 1, 5, 50, 250 and 500 kHz and 1 MHz at five locations—the right arm, left arm, trunk, right leg and left leg—in both the derivation set and external test set.[14] As muscle function assessments, HGS (testing twice for both hands using a handheld dynamometer and then selecting the highest force value in kilograms) and chair rise test (CRT) performance (measuring time to perform five repetitions) were measured using the same digital equipment (Leonardo Mechanography Ground Reaction Force Platform; Leonardo software version 4.4, Novotec Medical GmbH, Pforzheim, Germany) in both cohorts.[15] Sarcopenia was defined as the presence of both low muscle function [low HGS (men, below 28 kg; women, below 18 kg) or low CRT performance (12 s or longer)] and low muscle mass (low ALM index, under 7.0 kg/m^2 in men and under 5.7 kg/m^2 in women).

Model architecture

To develop a deep learning-based model using chest radiographs to predict sarcopenia, we used a two-step approach (*Figure* 2). In step 1, the deep neural network algorithm was trained to develop models to predict ALM, HGS and CRT performance. To quantify predictive uncertainty as well as values related with sarcopenia, we employed deep en-

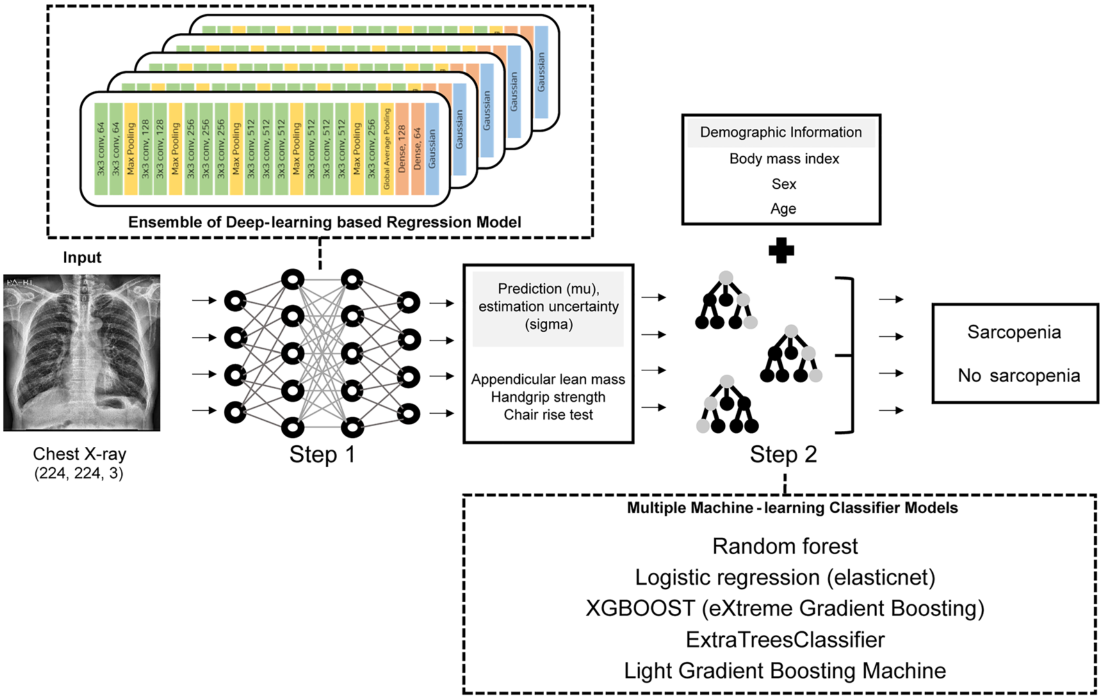

Figure 2 Architecture of the sarcopenia prediction model using chest X-ray images with clinical variables.

sembles. Unlike typical machine learning models, which output a single prediction value, deep ensembles yield two values, corresponding to predicted values (μ) and variance or estimates for predictive uncertainty (σ).[16] Three deep ensembles were developed to predict ALM, HGS and CRT performance and quantify the degrees of uncertainty. In step 2, machine learning models to predict the presence of sarcopenia were developed using the predicted muscle parameters and estimates for uncertainty derived from step 1 along with basic clinical variables including age, sex and BMI.

Step 1: Deep neural network to predict muscle parameters from chest radiograph

For each patient's raw Digital Imaging and Communications in Medicine (DICOM) image, pixel values were extracted using the PYDICOM library in Python. Images were resized to 224 × 224; then, the contrast-limited adaptive histogram equalization library was applied to increase the contrast of the original image with histogram equalization. A VGG16 model pre-trained in ImageNet was used as the main algorithm.[17] Feature maps were flattened using global average pooling and dense layers. Data augmentation was performed on the training set using random shift, rotation, horizontal flip, zoom, brightness and random multiplication five times. For the final training, we ran the step 1 model for 100 epochs with early stopping (stop training when a monitored metric has stopped improving to avoid overfitting; patience = 15), batch size 32 (total number of training examples present in a single batch) and cosine annealing learning rate (set learning rate of each parameter group using a cosine annealing schedule to improve model performance). The Adam optimizer (a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments) was used to minimize the Gaussian loss function (negative log-likelihood loss following Gaussian distributions).[16] A deep ensemble model approach from a previous study was implemented to measure the degree of uncertainty of predicted values (σ).[16] A higher σ value indicates higher uncertainty of the predicted value, whereas a lower σ value indicates that the model is more confident regarding the predicted results. Model performance was assessed using R^2 values. A GRAD-CAM (gradient-weighted class activation mapping) heat map (a localized map highlighting the important regions in the image for predicting the outcome to enhance model interpretability) was used to visualize which parts of the chest radiograph are most important for the prediction of muscle parameters.[18]

Step 2: Machine learning model to predict sarcopenia from muscle parameters

In step 2, to determine an optimal sarcopenia prediction model by aggregating predicted ALM, HGS and CRT performance values from the chest radiograph along with basic clinical features [age, sex, body mass index (BMI)], prediction models were developed using various machine learning algorithms, including random forest, regularized logistic regression, extreme gradient boosting, extra tree classifier and light gradient boosting machine (*Figure* 2). The estimated uncertainty (σ) for each muscle parameter prediction in the deep learning models was also used in the models. Among the five different algorithms trained and fine-tuned with the three-fold five-repeat cross-validation in the training set, the best-performing model, random forest, was selected according to model performance metrics (mainly AUROC and F1 score) in the internal test set (a hold-out set of the derivation cohort). Model calibration was performed using sigmoid methods.

Statistical analysis

Data were presented as mean ± standard deviation, median (interquartile range) or number (%) as appropriate. Categorical variables were compared using the independent two-sample *t*-test, Wilcoxon rank-sum test or chi-square test between groups with or without sarcopenia. The concordance correlation coefficient (Lin's concordance coefficient) and Bland–Altman plot were used to assess the agreement between predicted muscle parameters from step 1 and reference values.[19,20] To derive the 95% confidence interval of the step 2 model performance metrics in the internal and external test sets, the bootstrap method with 2000 repetitions was used. A calibration plot was created using the PMCALPLOT command of STATA software version 16.1 (StataCorp, College Station, TX, USA). In the internal test set, the AUROC of the chest radiograph-based machine learning model score and SARC-F were compared using the DeLong method.[21] Statistical significance was set at two-sided $P < 0.05$. The reporting of this study adheres to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) reporting guideline.[22]

Results

Clinical characteristics

The mean age of the derivation cohort (n = 926, women 76%) and the external test cohort (n = 149, women 64%) was 61.4 and 71.6 years, respectively (*Table* 1). The prevalence of sarcopenia was 15% in the derivation set and 12% in the external test set. In both cohorts, individuals with sarcopenia were older and had lower BMI, handgrip strength (HGS), CRT performance and appendicular lean

Table 1 Clinical characteristics of study subjects

	Derivation set			External test set		
	Sarcopenia (n = 141, 15%)	No sarcopenia (n = 785, 85%)	P value	Sarcopenia (n = 18, 12%)	No sarcopenia (n = 131, 88%)	P value
Age, year	65.6 ± 11.5	60.7 ± 10.2	<0.001	74.7 ± 6.2	71.2 ± 4.0	0.001
Women, n (%)	116 (82)	584 (74)	0.045	15 (83)	80 (61)	0.065
BMI, kg/m^2	20.8 ± 2.9	23.2 ± 3.6	<0.001	16.6 ± 1.5	19.3 ± 2.6	<0.001
HGS, kg	17.7 ± 4.4	25.1 ± 7.3	<0.001	21.5 ± 4.5	27.1 ± 6.7	<0.001
Low HGS, n (%)	110 (78)	123 (16)	<0.001	7 (39)	11 (8)	<0.002
CRT, sec	12.7 [10.3–14.9]	9.1 [7.4–11.1]	<0.001	13.0 [12.1–14.8]	9.3 [7.8–12.1]	<0.001
Low CRT, n (%)	83 (59)	140 (18)	<0.001	15 (83)	34 (26)	<0.001
Low muscle function, n (%)	141 (100)	214 (27)	<0.001	18 (100)	40 (31)	<0.001
ALMi, kg/m^2	5.4 ± 0.6	6.6 ± 1.2	<0.001	5.5 ± 0.6	6.8 ± 0.8	<0.001
Low ALMi, n (%)	141 (100)	145 (9)	<0.001	18 (100)	20 (15)	<0.001
Osteoporosis, n (%)[a]	109 (77)	467 (59)	<0.001	10 (55)	39 (30)	0.029
Previous fracture, n (%)	16 (11)	42 (5)	0.007	6 (33)	23 (18)	0.113
Hypertension[a], n (%)	76 (54)	398 (51)	0.484	5 (28)	36 (27)	0.979
Diabetes[a], n (%)	38 (27)	192 (24)	0.532	2 (11)	24 (18)	0.450

ALMi, appendicular lean mass index; BMI, body mass index; CRT, chair rise test; HGS, handgrip strength.

[a] Diagnosis of osteoporosis, hypertension and diabetes mellitus was defined as presence of diagnosis codes and corresponding medication use for each disease in derivation set.

mass (ALM) index and higher prevalence of osteoporosis ($P < 0.05$ for all).

Chest radiograph-based prediction of muscle parameters

In the internal test set (a hold-out set, n = 189, from the derivation cohort) and the external test set (n = 149), the concordance correlation coefficient (rho) for ALM prediction was 0.80 and 0.76, with an average difference of 0.18 ± 2.71 and 0.21 ± 2.28, respectively, in the Bland–Altman plot (*Figure* 3; *Figure* S1). Compared with ALM, the predicted HGS and CRT performance values showed weak to modest concordance correlation with observed values in the internal test set (0.70 for HGS, 0.31 for CRT) and external test set (0.50 for HGS, 0.11 for CRT). For the ALM, HGS and CRT parameters, most observation values (ground truth) were within the estimation uncertainty boundaries for given predicted values (*Figure* 3).

Class activation map for deep neural network models

The gradient-weight class activation mapping (GRAD-CAM) for deep neural network models to predict muscle parameters (ALM, HGS and CRT performance) is shown in *Figure* 4. In both men and women with or without sarcopenia, prediction models for ALM and HGS had highly weighted pixel values at bilateral lung lesions and part of the cardiac contour in common. However, GRAD-CAM for the CRT performance prediction model showed a distinctive pattern where pixel values at upper abdomen lesions contributed most to the prediction of CRT performance.

Performance of sarcopenia prediction models

On the basis of muscle parameters derived from deep neural network models using chest radiographs, machine learning models to predict the presence of sarcopenia were developed and tested. Among all trained machine learning models (*Table* S1), the random forest model [area under the receiver-operating characteristics curve (AUROC) 0.813 in the internal test set] was chosen as the best representative model owing to its highest F1 score (0.540) for the internal test set (*Table* 2; sensitivity 0.844, specificity 0.739). Similar model performance was observed in the external test set (AUROC 0.780, F1 score 0.458, sensitivity 0.611, specificity 0.855). When the model was re-trained using individuals with age older than 60 according to AWGS 2019 age threshold in train set, performance of retrained model was similar to the original model in internal test set (AUROC 0.777 vs. 0.813, P = 0.137) or inferior (AUROC 0.708 vs. 0.780, P = 0.018) in external set (*Figure* S2). The score derived from the representative model (SARC-CXR score) showed good calibration in both the internal and external test sets (*Figure* S3; Brier score 0.124 and 0.098, respectively). The SHapley Additive exPlanations (SHAP) summary plot for the sarcopenia prediction model revealed that lower predicted ALM values based on chest X-ray scans were associated with higher risk of sarcopenia with the highest feature importance, followed by BMI, lower predicted HGS and lower predicted CRT performance (longer time to complete the test; *Figure* 5). Higher estimation uncertainty (estimation uncertainty/

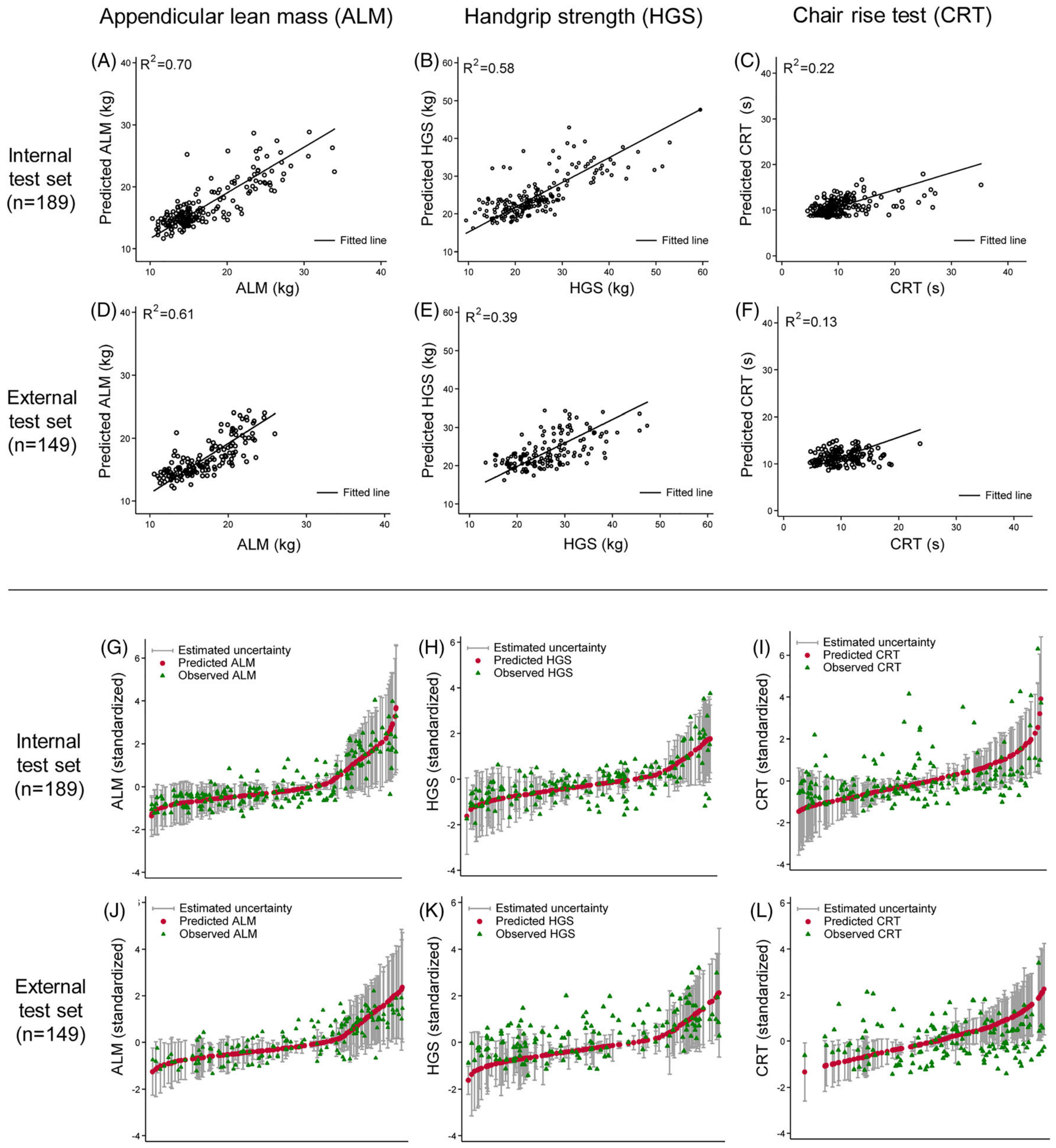

Figure 3 Prediction of appendicular lean mass and muscle function measurements based on chest X-ray scans using deep learning algorithm. Upper panel: Concordance correlation plot between ground truth and predicted values for appendicular lean mass (ALM), handgrip strength (HGS) and chair rise test (CRT). Lower panel: For ALM, HGS and CRT performance, data were plotted after standardization using mean and standard deviation values from the training set in the derivation cohort. Red dots indicate predicted values. Green dots indicate observed truth values corresponding to predicted values. The grey spike plot indicates the estimation uncertainty for a given predicted value.

predicted value, %) for HGS based on chest X-ray scans was associated with higher risk of sarcopenia in the prediction model, with relatively lower feature importance than prediction values.

The SARC-F score was available for the internal test set. When the SARC-CXR score performance was compared with that of the SARC-F questionnaire, the SARC-CXR score showed better discriminatory performance compared with SARC-F (AUROC 0.813 vs. 0.691, P = 0.029; *Figure* 6).

Discussion

For sarcopenia prediction, we developed the SARC-CXR ensemble machine learning model equipped with uncertainty-aware deep learning for predicting muscle mass and function from chest radiographs. The model's performance was externally validated in a community-based prospective cohort, for which the AUROC and F1 score were 0.773 and 0.444, respectively. The SHAP value revealed that

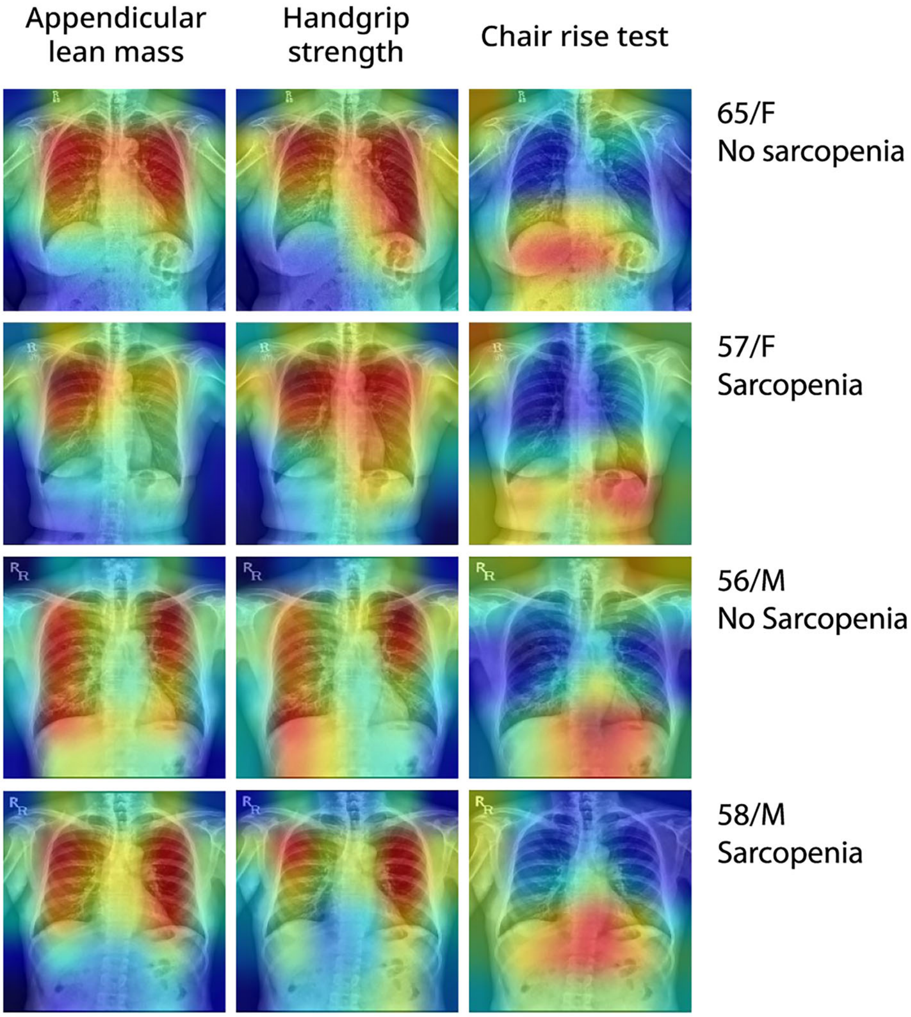

Figure 4 Gradient-weight class activation mapping for deep neural network models to predict components of sarcopenia including appendicular lean mass, handgrip strength and chair rise test performance. Red regions indicate regions with high importance pixel values for the classification of sarcopenia in the deep neural network model. F, woman; M, man.

Table 2 Discriminatory performance of sarcopenia prediction model using chest X-ray predicted skeletal muscle parameters and clinical variables

Performance metrics	Internal test set	95% CI[a]	External test set	95% CI[a]
AUROC	0.813	0.746–0.878	0.780	0.679–0.873
AUPRC	0.380	0.263–0.566	0.440	0.245–0.628
F1 score	0.540	0.422–0.646	0.458	0.301–0.600
MCC	0.455	0.335–0.573	0.379	0.207–0.533
Sensitivity	0.844	0.719–0.963	0.611	0.412–0.808
Specificity	0.739	0.671–0.805	0.855	0.804–0.906
PPV	0.397	0.288–0.514	0.367	0.222–0.517
NPV	0.959	0.921–0.986	0.941	0.903–0.975
Brier score	0.124	0.093–0.157	0.098	0.077–0.121

AUROC, area under the receiver-operating characteristics curve; AURPC, area under the precision-recall curve; MCC, Matthews correlation coefficient; NPV, negative predictive value; PPV, positive predictive value.
[a]95% CI was calculated using bootstrapping method.

the BMI value and deep learning-predicted muscle mass and function were the important predictors. To our knowledge, this is the first research attempt to screen for sarcopenia on the basis of conventional chest radiographs with patients' baseline demographics. The results demonstrated the potential for opportunistic screening of sarcopenia in broader patient groups.

Clinical artificial intelligence without explainability may predict outcomes depending on confounding variables. Previously, Agniel et al. showed that the presence of a laboratory test order or timing had a much stronger association with survival than laboratory test results did in the electronic healthcare record data.[23] Hence, it would be essential to analyse the predictors in clinical artificial intelligence to avoid developing artificial intelligence depending on confounding variables. In this study, explainability was assessed vigorously.

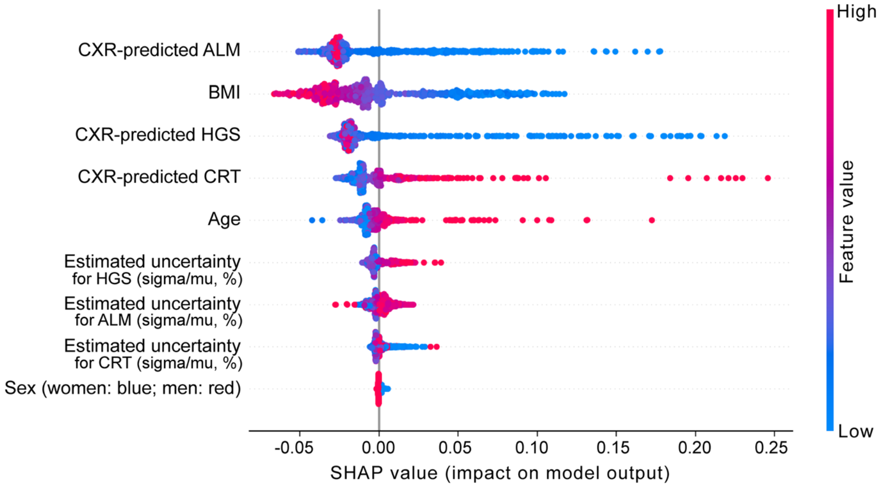

Figure 5 SHapley Additive exPlanations (SHAP) summary plot to visualize the relative additive contribution of each feature for the prediction of sarcopenia in the final machine learning model (SARC-CXR). ALM, appendicular lean mass; BMI, body mass index; CRT, chair rise test; CXR, chest X-ray; HGS, handgrip strength.

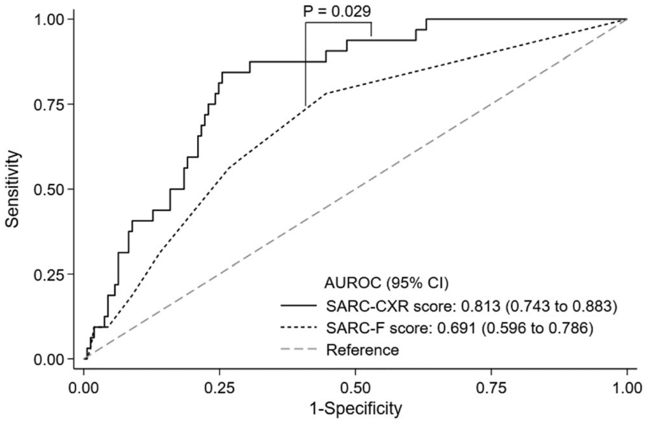

Figure 6 Comparison of discriminatory performance for sarcopenia between chest x-ray based machine learning prediction score (SARC-CXR) and SARC-F in internal test set (AUROC 0.813 vs. 0.691, *P* = 0.029). AUROC, area under the receiver-operating characteristics curve; CXR, chest X-ray; ML, machine learning.

We found that the predicted ALM, HGS and CRT performance derived from chest radiographs using the deep neural network model, along with estimation uncertainties for each predicted value, served as strong predictors in the final model, SARC-CXR.

Until now, the quantification and communication of uncertainty in medical machine learning have been relatively neglected. Kompa et al. argued that medical machine learning should be geared with the ability to say ‘I don’t know’ on the basis of predictive uncertainty estimates to flag physicians for a second opinion.[24] In this study, deep ensembles were leveraged to estimate the uncertainty of predicted surrogate markers of sarcopenia. Interestingly, the performance of deep learning models in predicting ALM was good, whereas the predictive performance for muscle function, particularly for the CRT performance, was not competent. These findings suggest that chest X-ray scans provide relevant information for skeletal muscle quantity, but the inference for muscle function based on chest X-ray images may require further input from other features to enhance the prediction accuracy. Given the possibility of predictive uncertainty as outcome predictors, we utilized estimation uncertainty values for muscle parameters obtained from the chest X-ray-based deep learning model along with predicted values. Notably, the relatively higher degree of estimation uncertainty, particularly for muscle function parameters, substantially contributed to elevating the predicted risk of sarcopenia in the SARC-CXR model.

A systematic review of prediction models for the diagnosis and prognosis of coronavirus disease 2019 (COVID-19) found that all 232 reviewed models had high or unclear risk of bias, mainly because of non-representative selection of control patients.[25] We validated the SARC-CXR model externally using a community-based prospective registry with the assessment for calibration. We believe that the robust performance of the SARC-CXR model in the external community-based cohort set observed in this study can provide a proof of concept to design further prospective studies for testing the feasibility of sarcopenia screening based on chest radiographs. The sample size of external test subset from KURE cohort was relatively small limited to the subsamples with available chest X-ray scans. To overcome this limitation, we applied bootstrapping method to address potential uncertainty from the relatively small sample size as possible by calculating 95% confidence interval of model performance metrics. Findings from this study need to be validated in further prospective studies with larger sample size.

Sarcopenia is associated with higher risk of falls, fractures, loss of independence and mortality.[6,7] To diagnose sarcopenia, current clinical guidelines endorse a stepwise approach that begins with case finding based on a simple questionnaire such as SARC-F, followed by measurement of muscle function and mass as subsequent steps.[6,7] Therefore, case finding for sarcopenia is the gate-keeping step to diagnose sarcopenia. Although the SARC-F questionnaire, a five-item self-report questionnaire, is a simple, inexpensive and widely accepted method to identify cases of sarcopenia with high specificity, its low-to-moderate sensitivity to detect sarcopenia remains a challenge. Given the significant improvement of discriminatory ability by applying the SARC-CXR score compared with the SARC-F as observed in this study, and the wide availability of chest X-ray scans in various clinical settings, it is conceivable that chest X-ray-scan-based opportunistic screening may improve sarcopenia detection. Variation of SARC-F with simple measurement such as calf circumference (SARC-CalF) was proposed to have improved discriminatory ability compared to SARC-F. In the study by Bahat et al., performance of SARC-CalF (SARC-F + calf circumference with 33 cm threshold) ranged from 0.68 to 0.83 for AUROC, 0.15 to 0.50 for sensitivity and 0.90 to 0.98 for specificity depending on the definitions of sarcopenia.[26] Although we were not able to directly compare the performance of SARC-CXR and SARC-CalF due to lack of calf circumference data, we observed similar AUROC of SARC-CXR (0.78–0.81) compared with that of SARC-CalF reported in prior literature, with numerically higher sensitivity ranging from 0.61 to 0.84 by SARC-CXR compared with SARC-CalF. Considering that the purpose of the models is to screen individuals at the risk of sarcopenia, better sensitivity (recall) of SARC-CXR model might have advantage for screening sarcopenia; this needs to be tested further using direct comparison of AUPRC or F1-score between SARC-F, SARC-CalF and SARC-CXR scores.

In this study, GRAD-CAM revealed characteristic patterns of regional feature importance for ALM, HGS and CRT. For the prediction of ALM and HGS, the model focused on the bilateral lung lesion and cardiac contour. Prior studies showed that cardiopulmonary function is positively associated with fat-free mass and HGS in individuals without apparent lung disease.[27–29] Muscle wasting was known also an independent predictor of survival in patients with heart failure.[30] These findings and GRAD-CAM results suggest that model might at least partly reflect the association between cardiopulmonary function and muscle mass and HGS, which was indirectly captured by lung size and cardiac contour. Meanwhile, the model mainly utilized pixel values around upper abdomen to estimate CRT performance. Sit-to-stand motion in CRT requires not only lower extremity power but also the balance and coordination of whole body that largely depends on core muscles and hip flexors such as psoas muscle.[31] It is conceivable that silhouette of core muscles visualized at upper abdomen area was utilized to predict CRT performance, although this hypothesis needs to be further tested.

This study has several limitations. Although bioimpedance analysis (BIA) is a widely used method to measure ALM, inaccuracy of BIA-estimated ALM values in patients with dehydration or morbid obesity was reported compared with other clinically available measurements such as dual-energy X-ray absorptiometry testing. In this study, we used the definition of sarcopenia from the Asian Working Group for Sarcopenia (AWGS) 2019 guideline. Although the harmonization of definitions for sarcopenia has not yet reached a consensus, we believe that the two-step approach used in our study (prediction of muscle mass and function parameters as continuous values from chest radiographs using deep neural network regression as step 1, followed by the construction of a machine learning model to classify sarcopenia using the predicted muscle parameters along with basic clinical parameters as step 2) would provide room for improvement to develop better prediction models that fit various definitions of sarcopenia. The SARC-CXR model has not been validated for patients other than those of Korean ethnicity. Applicability of SARC-CXR model can be limited in low-resource setting with barriers to access imaging modalities, although accessibility to plain X-ray is relatively better than other modalities such as computed tomography or magnetic resonance.[32] As observed in high-priority diseases such as tuberculosis, technical solutions such as promotion for free access to artificial intelligence software along with regulatory support have potential to facilitate the utilization of artificial intelligence system in low-resource setting.[32] As we investigated data of apparently healthy, ambulatory individuals in both derivation and external test sets, chest X-ray scans from individuals with acute medical conditions such as pneumonia and pleural or pericardial effusion were not included in this study. Applicability of current model to inpatient setting needs to be validated further.

In conclusion, the chest X-ray-based sarcopenia prediction score, SARC-CXR, improved the detection of sarcopenia, which merits further validation.

Acknowledgements

This study was supported by a Severance Hospital Research fund for Clinical excellence(SHRC; C-2021-0019) and the Korea Disease Control and Prevention Agency (2013-E63007-01,2013-E63007-02, 2019-ER6302-01) and by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT) (Project Number:1711174510, RS-2021-KD000010). This research was supported by a grant of the MD-Phd/Medical Scientist Training Program through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea. We thank Jaehyun Shin, MFA (Research Associate, Department of Biomedical Systems Informatics, Yonsei University College of Medicine), for his graphical assistance and Seung Hyun Chung, BA, MS (Research Associate, Department of Biomedical Systems Informatics, Yonsei University College of Medicine), for her contribution to editing the manuscript.

Conflict of interest

There is no competing interest to declare.

Online supplementary material

Additional supporting information may be found online in the Supporting Information section at the end of the article.

References

1. Mayhew AJ, Amog K, Phillips S, Parise G, McNicholas PD, de Souza RJ, et al. The prevalence of sarcopenia in community-dwelling older adults, an exploration of differences between studies and within definitions: a systematic review and meta-analyses. *Age Ageing* 2019;**48**: 48–56.
2. Xu J, Wan CS, Ktoris K, Reijnierse EM, Maier AB. Sarcopenia is associated with mortality in adults: a systematic review and meta-analysis. *Gerontology* 2022;**68**: 361–376.
3. Hirani V, Blyth F, Naganathan V, Le Couteur DG, Seibel MJ, Waite LM, et al. Sarcopenia is associated with incident disability, institutionalization, and mortality in community-dwelling older men: the Concord health and ageing in men project. *J Am Med Dir Assoc* 2015;**16**:607–613.
4. Petermann-Rocha F, Ferguson LD, Gray SR, Rodríguez-Gómez I, Sattar N, Siebert S, et al. Association of sarcopenia with incident osteoporosis: a prospective study of 168,682 UK biobank participants. *J Cachexia Sarcopenia Muscle* 2021;**12**: 1179–1188.
5. McLean RR, Shardell MD, Alley DE, Cawthon PM, Fragala MS, Harris TB, et al. Criteria for clinically relevant weakness and low lean mass and their longitudinal association with incident mobility impairment and mortality: the foundation for the National Institutes of Health (FNIH) sarcopenia project. *J Gerontol A Biol Sci Med Sci* 2014;**69**:576–583.
6. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyère O, Cederholm T, et al. Sarcopenia: revised European consensus on definition and diagnosis. *Age Ageing* 2019;**48**:16–31.
7. Chen LK, Woo J, Assantachai P, Auyeung TW, Chou MY, Iijima K, et al. Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. *J Am Med Dir Assoc* 2020;**21**: 300–307.e2.
8. Speets AM, van der Graaf Y, Hoes AW, Kalmijn S, Sachs AP, Rutten MJ, et al. Chest radiography in general practice: indications, diagnostic yield and consequences for patient management. *Br J Gen Pract* 2006;**56**:574–578.
9. Willer K, Fingerle AA, Noichl W, De Marco F, Frank M, Urban T, et al. X-ray dark-field chest imaging for detection and quantification of emphysema in patients with chronic obstructive pulmonary disease: a diagnostic accuracy study. *Lancet Digit Health* 2021;**3**:e733–e744.
10. Lu MT, Ivanov A, Mayrhofer T, Hosny A, Aerts HJWL, Hoffmann U. Deep learning to assess long-term mortality from chest radiographs. *JAMA Netw Open* 2019;**2**: e197416.
11. Jang M, Kim M, Bae SJ, Lee SH, Koh JM, Kim N. Opportunistic osteoporosis screening using chest radiographs with deep learning: development and external validation with a Cohort dataset. *J Bone Miner Res* 2022;**37**:369–377.
12. Joseph VR. Optimal ratio for data splitting. *Stat Anal Data Min: ASA Data Sci J* 2022; **15**:531–538.
13. Malmstrom TK, Miller DK, Simonsick EM, Ferrucci L, Morley JE. SARC-F: a symptom score to predict persons with sarcopenia at risk for poor functional outcomes. *J Cachexia Sarcopenia Muscle* 2016;**7**:28–36.
14. Hong N, Kim KJ, Lee SJ, Kim CO, Kim HC, Rhee Y, et al. Cohort profile: Korean urban rural elderly (KURE) study, a prospective cohort on ageing and health in Korea. *BMJ Open* 2019;**9**:e031018.
15. Lee EY, Kim HC, Rhee Y, Youm Y, Kim KM, Lee JM, et al. The Korean urban rural elderly cohort study: study design and protocol. *BMC Geriatr* 2014;**14**:33.
16. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems. 2017;**30**. https://doi.org/10.48550/arXiv.1612.01474
17. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 2014. https://doi.org/10.48550/arXiv.1409.1556
18. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis* 2020;**128**:336–359.
19. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;**45**:255–268.
20. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *J R Stat Soc Ser A Stat Soc Ser D (The Statistician)* 1983;**32**: 307–317.
21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;**44**:837–845.
22. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD):

the TRIPOD statement. *Br J Surg* 2015; **102**:148–158.
23. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* 2018; **361**:k1479.
24. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit Med* 2021;**4**:4.
25. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020;**369**:m1328.
26. Bahat G, Oren MM, Yilmaz O, Kılıç C, Aydin K, Karan MA. Comparing SARC-F with SARC-CalF to screen sarcopenia in community living older adults. *J Nutr Health Aging* 2018;**22**:1034–1038.
27. Santana H, Zoico E, Turcato E, Tosoni P, Bissoli L, Olivieri M, et al. Relation between body composition, fat distribution, and lung function in elderly men. *Am J Clin Nutr* 2001;**73**:827–831.
28. Park CH, Yi Y, Do JG, Lee YT, Yoon KJ. Relationship between skeletal muscle mass and lung function in Korean adults without clinically apparent lung disease. *Medicine (Baltimore)* 2018;**97**:e12281.
29. Chen L, Liu X, Wang Q, Jia L, Song K, Nie S, et al. Better pulmonary function is associated with greater handgrip strength in a healthy Chinese Han population. *BMC Pulm Med* 2020;**20**:114.
30. von Haehling S, Garfias Macedo T, Valentova M, Anker MS, Ebner N, Bekfani T, et al. Muscle wasting as an independent predictor of survival in patients with chronic heart failure. *J Cachexia Sarcopenia Muscle* 2020;**11**:1242–1249.
31. Hardy R, Cooper R, Shah I, Harridge S, Guralnik J, Kuh D. Is chair rise performance a useful measure of leg power? *Aging Clin Exp Res* 2010;**22**:412–418.
32. Frija G, Blažić I, Frush DP, Hierath M, Kawooya M, Donoso-Bach L, et al. How to improve access to medical imaging in low- and middle-income countries ? *EClinicalMedicine* 2021;**38**:101034.