Review

Predictive Modeling of Sarcopenia Using Machine Learning: Systematic Review and Meta-Analysis

Yinying Peng[1*], MM; Zhexue Ren[1*], MS; Lina Xue[1], MM; Mingshao Zhao[2], MM; Ce Bian[1], MM; Qi Shen[1], MM; Mengjun Zhou[3], MM; Bo Li[3*], MD; Xiaodong Jin[2*], MD

[1]School of Clinical Medicine, Shandong Second Medical University, Weifang, Shandong, China

[2]Department of Geriatrics, Central Hospital of Zibo, Zibo, Shandong, China

[3]Department of Cardiology, Central Hospital of Zibo, Zibo, Shandong, China

[*]these authors contributed equally

Corresponding Author:
Xiaodong Jin, MD
Department of Geriatrics
Central Hospital of Zibo
No. 54, Gongqingtuan West Road
Zibo, Shandong, 255000
China
Phone: 86 1 867 818 6557
Email: jinxiaodongdxj@sina.com

Abstract

Background: Sarcopenia is an age-related syndrome characterized by muscle loss and functional decline, increasing fall risks. Early identification is crucial, yet convenient and accurate screening tools remain limited in routine practice. Machine learning (ML) models show promise for early sarcopenia detection, but their diagnostic performance and methodological quality have not been systematically quantified.

Objective: This review aimed to quantitatively evaluate the diagnostic performance of ML models for sarcopenia prediction, compare performance across ML algorithms, and identify key sources of heterogeneity and methodological limitations.

Methods: PubMed, Web of Science, Embase, and the Cochrane Library were systematically searched from inception to April 15, 2026. Eligible studies developed or validated at least one ML model for sarcopenia and provided sufficient data to reconstruct a 2×2 contingency table. Risk of bias was assessed using PROBAST (Prediction Model Risk of Bias Assessment Tool) developed by Moons et al. The primary analysis used a bivariate random-effects (Reitsma) model to jointly synthesize sensitivity and specificity. Area under the receiver operating characteristic curve (AUROC) values were pooled using a univariate random-effects model with the Hartung-Knapp-Sidik-Jonkman (HKSJ) method, and 95% prediction intervals (PIs) were estimated. A secondary analysis incorporated all models using cluster-robust variance estimation (RVE). Subgroup analyses and meta-regression explored heterogeneity. The certainty of evidence was assessed using the Grading of Recommendations Assessment, Development and Evaluation (GRADE).

Results: Twenty studies (20 prespecified index models) were included. The pooled AUROC was 0.89 (95% CI 0.82-0.93; 95% PI 0.54-0.98), with pooled sensitivity of 0.77 (95% CI 0.68-0.83; 95% PI 0.37-0.95) and specificity of 0.88 (95% CI 0.83-0.91; 95% PI 0.62-0.97). Gradient boosting had the highest pooled AUROC (0.92), but CIs overlapped across algorithms. Externally validated models had a pooled AUROC of 0.85, lower than that of internal ones (0.91). The secondary analysis of all 115 models yielded consistent results. Meta-regression identified no significant covariates. PROBAST assessment revealed that half of the studies were at high risk of bias.

Conclusions: This review provides the first quantitative synthesis of ML-based sarcopenia prediction models, with several innovations: (1) it provides algorithm-stratified pooled estimates of diagnostic accuracy, (2) it formally quantifies performance across algorithms and addresses the unit-of-analysis issue through a 2-tier approach, and (3) it quantifies the performance gap between externally and internally validated models. By identifying specific weaknesses—insufficient external validation, poor calibration reporting, and high risk of bias—this review offers a roadmap for improving future studies. Practically, these models may serve as cost-effective prescreening tools. However, GRADE certainty was very low, with substantial heterogeneity, insufficient external validation, and high risk of bias, indicating that current models are not yet ready for routine deployment.

XSL•FO
RenderX

Future research should prioritize rigorous external validation, transparent reporting, calibration assessment, and clinically implementable design.

Trial Registration: PROSPERO CRD420251267598; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251267598

(J Med Internet Res 2026;28:e91297) doi: 10.2196/91297

KEYWORDS

sarcopenia; machine learning; predictive modeling; diagnostic accuracy; risk stratification; systematic review; meta-analysis; clinical decision-making

Introduction

Sarcopenia is a skeletal muscle disease associated with aging, primarily characterized by progressive, systemic loss of muscle mass, decline in muscle strength, and deterioration of physical function [1]. In the context of the increasingly significant global trend of population aging, sarcopenia has become an important geriatric syndrome that significantly increases the risk of falls, functional impairment, hospitalization, and mortality among the older adults. It not only severely affects the quality of life of patients but also places a heavy burden on health care systems [2,3]. Therefore, achieving early and accurate risk identification and prediction of sarcopenia is a key prerequisite for improving health outcomes in older adults and implementing targeted interventions.

The clinical diagnosis of sarcopenia currently relies primarily on international consensus standards such as those established by the European Working Group on Sarcopenia in Older People (EWGSOP), the Asian Working Group on Sarcopenia 2019 (AWGS 2019), the International Working Group on Sarcopenia (IWGS), and the Foundation for the National Institute of Health (FNIH) [1,4-6]. However, these standards face dual challenges in practical application: on one hand, the gold standard for diagnosis (such as dual-energy X-ray absorptiometry [DXA] and computed tomography [CT]) is difficult to popularize in community and primary health care settings due to the high cost of equipment and complexity of operation; on the other hand, while simpler screening tools (such as the SARC-F [strength, assistance in walking, rising from a chair, climbing stairs, and falls] questionnaire) are easy to implement, their sensitivity is limited, particularly in identifying early or "possible" sarcopenia, which poses a risk of missing high-risk populations [7]. Therefore, developing an objective, efficient, and easily disseminated early prediction tool is of significant clinical importance. In recent years, machine learning (ML), as a core branch of AI, has demonstrated considerable potential in the fields of medical prediction and diagnosis [8]. Its core advantage lies in its ability to automatically learn patterns from complex data through algorithms, thereby enhancing predictive performance [9]. Compared to traditional statistical models (such as logistic regression [LR]), advanced ML algorithms such as random forest (RF), gradient boosting (GB) machines, and artificial neural networks (NNs) are better at handling high-dimensional, nonlinear relationships and can integrate heterogeneous data from multiple sources [10,11]. In the field of sarcopenia research, scholars have attempted to use ML techniques to construct diagnostic and risk prediction models by combining demographic, anthropometric, laboratory indicators, and imaging features, and have preliminarily shown good performance [12-14].

However, research in this field remains fragmented, with substantial heterogeneity in methodological quality and reporting standards. A recent systematic review by Zhang et al [15] provided a valuable qualitative synthesis of diagnostic prediction models for sarcopenia and assessed their risk of bias using the PROBAST (Prediction Model Risk of Bias Assessment Tool), revealing that over half (56.25%) of the 16 studies included in that review exhibited a high risk of bias, primarily due to insufficient sample sizes, improper handling of predictors, and deficiencies in statistical analysis. While this review offered important insights into the methodological landscape, it did not perform a quantitative synthesis of model performance, nor did it systematically compare the discriminative ability of different ML algorithms. To address this gap, this study extends the work of Zhang et al [15] by conducting a comprehensive quantitative meta-analysis focused specifically on ML-based prediction models for sarcopenia. Unlike prior qualitative reviews, our study provides pooled estimates of key diagnostic accuracy metrics—including area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, diagnostic odds ratio (DOR), positive likelihood ratio (PLR), and negative likelihood ratio (NLR)—stratified by algorithm type. Subgroup analyses and meta-regression are used to explore sources of heterogeneity and to identify study-level factors that influence model performance.

Accordingly, this systematic review and meta-analysis was conducted with the following objectives: (1) to quantitatively synthesize the overall diagnostic performance of ML models for sarcopenia prediction, as measured by pooled AUROC, sensitivity, specificity, DOR, PLR, and NLR; (2) to compare the discriminative performance of commonly used ML algorithms—including LR, RF, GB, support vector machine, and NNs—through subgroup analyses; and (3) to explore potential sources of between-study heterogeneity using meta-regression. The risk of bias of the included studies was assessed with PROBAST to evaluate its possible influence on performance estimates. We hypothesized that the diagnostic performance of ML models for sarcopenia would vary across studies, and that this variability could be partially explained by algorithm type, sample size, study design, and validation strategy.

XSL•FO
RenderX

Methods

Protocol and Reporting Standards

This systematic review and meta-analysis was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [16]. The completed PRISMA 2020 checklist is provided in Multimedia Appendix 1. The abstract was prepared in accordance with the PRISMA 2020 for Abstracts checklist (Multimedia Appendix 2). The study protocol was prospectively registered with PROSPERO (CRD420251267598).

Literature Search Strategy

A systematic literature search was performed in PubMed, Web of Science, Embase, and the Cochrane Library from database inception to April 15, 2026. The search strategy was developed and reported in adherence to the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) guidelines [17]. The complete search strategy for each database, including all subject headings and free-text keywords combined with Boolean operators, is provided in Multimedia Appendix 3, and the completed PRISMA-S checklist is provided in Multimedia Appendix 4. The search was initially conducted on December 1, 2025, and was updated on April 15, 2026, prior to revision; 4 additional eligible studies were identified and included [18-21]. Preprints, conference proceedings, and gray literature were not systematically searched. Citation chasing of included studies and relevant reviews was performed to identify any additional eligible articles.

Eligibility Criteria

Studies were eligible if they (1) were published in English; (2) included any human population with sarcopenia diagnosed by established consensus criteria, including the European Working Group on Sarcopenia in Older People (EWGSOP and EWGSOP2) [1,22], the AWGS 2019 [4], the IWGS [5], the FNIH criteria [6], or the Sarcopenia Definitions and Outcomes Consortium (SDOC) criteria [23]; (3) developed or updated at least one ML-based prediction model for sarcopenia; and (4) provided sufficient data to reconstruct a 2×2 contingency table (true positives, false positives, false negatives, and true negatives), either reported directly or derivable from reported sensitivity, specificity, and sample size or disease prevalence.

Studies were excluded if they (1) were not published in English; (2) were case reports, review articles, conference abstracts, or editorials; (3) did not use ML algorithms; or (4) reported only discrimination metrics (eg, AUROC) without threshold-specific operating points, such that a 2×2 contingency table could not be reconstructed. The number of studies excluded due to insufficient data for 2×2 table reconstruction is documented in the PRISMA flow diagram.

Study Selection

Two reviewers (YP and ZR) independently screened titles, abstracts, and full texts for eligibility. Disagreements were resolved by consensus or by consultation with a third reviewer (LX).

Data Extraction and Items

Data extraction was guided by the CHARMS (Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies) framework [24]: (1) source of data. The data source and participant characteristics (country, setting, sample size, age range, and sex distribution) were extracted. Studies were categorized by data source type as community cohort, hospital cohort, or disease-specific cohort. (2) Participants. The characteristics of participants from each included study, including the inclusion and exclusion criteria applied in the primary studies, were documented. (3) Outcomes to be predicted. The definition of sarcopenia and the diagnostic criteria used in each primary study were extracted. (4) Candidate predictors (or index tests). All candidate predictors considered in each primary study were extracted and categorized into 5 groups: demographic characteristics, anthropometric measurements, laboratory biomarkers, lifestyle and behavioral factors, and clinical functional indicators. (5) Sample size. The sample size of each included study, including the total number of participants and the number of cases and noncases, was extracted. (6) Missing data. Information on missing data handling in the primary studies was extracted when available, and categorized as advanced methods (eg, multiple imputation by chained equations [MICE]), simple methods (eg, mean imputation and complete-case analysis), or not reported. (7) Model development. For each included model, the type of ML algorithm, feature selection method, data balancing technique, and internal validation strategy were extracted. When studies reported multiple models, all eligible models were initially extracted. (8) Model performance. For each included model, the reported AUROC, sensitivity, specificity, and, if available, the 2×2 contingency table were extracted. When multiple thresholds were reported, the operating point selected by the study authors or the threshold corresponding to the maximum Youden index was recorded. For models evaluated on both internal and external validation datasets, the validation type from which performance metrics were extracted was documented. (9) Model evaluation. For each included model, the type of validation (internal or external) was extracted. External validation was defined as validation on a dataset independent from the development cohort, including temporally or geographically distinct samples.

Assessment of Bias

The methodological quality and risk of bias of the included studies were assessed using the PROBAST. PROBAST covers four domains: (1) participants, (2) predictors, (3) outcomes, and (4) analysis. Each domain was rated using the original signaling questions without adaptation. Applicability and risk of bias were assessed separately. Each domain was classified as low, high, or unclear risk of bias. Overall risk of bias for each study was determined as follows: low if all domains were low; high if any domain was high; unclear if at least one domain was unclear and none were high.

XSL•FO
RenderX

Grading of Recommendations Assessment, Development and Evaluation Assessment of Evidence Quality

The certainty of evidence was assessed using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach for diagnostic test accuracy studies. Evidence started at high certainty and was downgraded based on 5 domains: risk of bias, indirectness, inconsistency, imprecision, and publication bias. Risk of bias was judged using PROBAST results; indirectness considered the applicability of study populations and diagnostic criteria; inconsistency was assessed using the I^2 statistic; imprecision was evaluated using clinical decision thresholds for likelihood ratios (PLR≥2.0; NLR≤0.5); and publication bias was assessed using Deeks funnel plot asymmetry test. The GRADE assessment was based on the 18 studies [12-14,19-21,25-36] that contributed to the bivariate meta-analysis of sensitivity and specificity. A summary of findings table was generated using GRADEpro GDT (McMaster University and Evidence Prime).

Statistical Analysis

Multiple entries from the same study represented distinct prediction models or validation methods, which created potential within-study dependence. To address this issue, a 2-tiered analytical approach was used throughout the analyses. All statistical analyses were performed using R software (version 4.4.1; R Core Team) with the *mada*, *metafor*, *pimeta*, and *clubSandwich* packages.

Several included studies contributed more than one prediction model, resulting in multiple model-level performance estimates from the same study. To address potential within-study dependence and avoid double-counting participants, the primary analysis was based on one prespecified index model per study. The index model was selected according to the following hierarchical rules: first, the model evaluated in an external validation dataset was selected when available; second, among models evaluated at the same validation level, the model with the highest reported AUROC was selected; third, when AUROC was not available or when model selection remained unclear, the model explicitly identified by the original study authors as the primary or recommended model was selected.

For models evaluated on both internal and external validation datasets, performance estimates from the external validation dataset were preferentially extracted. When no external validation dataset was available, estimates from the internal validation dataset, test set, or validation cohort were used according to the information reported in the original study.

For threshold-specific diagnostic accuracy, we used a bivariate random-effects model, commonly referred to as the Reitsma model, to jointly synthesize sensitivity and specificity, with model parameters estimated using restricted maximum likelihood (REML) [37]. Summary receiver operating characteristic (SROC) curves were generated based on the bivariate model. The bivariate meta-analysis included only index models with performance estimates derived from a clearly defined validation dataset or validation cohort and with sufficient information to reconstruct a 2×2 contingency table, including true positives, false positives, false negatives, and true negatives.

Studies that reported sensitivity and specificity only as averages across repeated random train-test splits, bootstrap resampling, or cross-validation, without sufficient information to reconstruct a single validation-set-specific 2×2 table, were not included in the bivariate meta-analysis. These studies were retained in the systematic review and were included in the AUROC meta-analysis when AUROC estimates and corresponding uncertainty measures could be obtained or approximated.

When studies reported performance at multiple thresholds, a single operating point was selected according to the following hierarchy: first, the threshold corresponding to the maximum Youden index, if available; second, the threshold recommended by the original authors for clinical use; third, the default probability threshold of 0.5 for models outputting predicted probabilities; and fourth, the only reported threshold if no alternative threshold was provided.

For descriptive and supplementary analyses of threshold-specific accuracy, sensitivity and specificity were synthesized using univariate random-effects models on the logit scale. The DOR, PLR, and NLR were derived from the pooled sensitivity and specificity estimates of the bivariate (Reitsma) model, and their 95% CIs were calculated using the delta method. This approach avoids potential statistical inflation caused by extreme values in individual studies. Between-study inconsistency was described using I^2 statistics from the corresponding models. Between-study heterogeneity was assessed using the 95% prediction intervals (PIs), which illustrate the expected dispersion of sensitivity and specificity across future settings.

For the meta-analysis of AUROC values, AUROC estimates were logit-transformed before pooling. A univariate random-effects model was fitted using REML, and the Hartung-Knapp-Sidik-Jonkman (HKSJ) method was applied to estimate pooled effects with 95% CIs, providing more robust inference when the number of studies is limited [38,39]. Pooled estimates were back-transformed to the original AUROC scale for reporting.

When the standard error of AUROC was not directly reported, it was derived from the reported 95% CI when available, using the formula:

$$SE = [\text{logit (upper limit)} - \text{logit (lower limit)}]/(2 \times 1.96)$$

For studies reporting AUROC without a CI or SE, the SE was approximated using a large-sample method based on the reported AUROC and the numbers of cases and noncases in the validation dataset [40]. The approximated standard error on the original AUROC scale was then converted to the logit scale using the delta method. Studies without sufficient information to estimate the SE were excluded from the AUROC meta-analysis.

In addition to 95% CIs, 95% PIs were calculated for pooled AUROC estimates using the Nagashima confidence distribution approach with bootstrap calibration [41]. The width of the PI was used to describe the expected range of true AUROC values across different clinical settings and to quantify the practical implications of between-study heterogeneity [42].

XSL•FO
RenderX

As a secondary analysis, all eligible model-validation records with reconstructable 2×2 contingency tables were included to assess the robustness of the primary findings when multiple models from the same study were considered. Random-effects meta-analyses with cluster-robust variance estimation (RVE) were performed separately for logit-transformed sensitivity and logit-transformed specificity, with clustering at the study level to account for within-study dependence among multiple models from the same study [43]. The pooled DOR, positive PLR, and NLR in this secondary analysis were not synthesized using separate RVE models. Instead, they were derived from the RVE-pooled sensitivity and specificity estimates, and their 95% CIs were approximated using the delta method.

Subgroup analyses were performed on the prespecified index models from the primary analysis to compare diagnostic performance across different ML algorithm types and between externally validated models versus those validated only internally. For algorithm-type subgroups, GB was defined to include all boosting variants: extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), CatBoost, gradient boosting classifier (GBC), gradient boosting decision tree (GBDT), and AdaBoost. RF included RF and extra trees. NNs included multilayer perceptron (MLP) and Wide & Deep (W&D) architectures. Decision tree (DT) included classification and regression trees (CART). For each algorithm subgroup, if at least 3 models had complete 2×2 tables available for bivariate modeling, sensitivity and specificity were pooled using a bivariate random-effects (Reitsma) model, and the DOR, PLR, and NLR were derived. If at least 3 models reported AUROC with calculable standard errors, AUROC was pooled using a univariate random-effects model. Algorithm subgroups with fewer than 3 contributing models were described narratively without pooled estimates.

For validation-type subgroups, models were categorized as externally validated or internally validated. The same meta-analytic methods described above were applied to each validation subgroup.

To explore potential sources of heterogeneity, we performed univariate meta-regression based on study-level covariates. Given the limited number of studies (k=18 with complete 2×2 data) [12-14,19-21,25-36], this analysis was prespecified as exploratory, intended for hypothesis generation rather than confirmatory inference [44,45]. Covariates included study design (multicenter vs single-center), validation type (external vs internal-only validation), sarcopenia definition (AWGS 2019 vs other criteria), and sample size (log10-transformed). The dependent variable was the log-transformed DOR, and study-specific sampling variances were calculated from the 2×2 tables. REML estimation was used.

To assess whether the pooled results were unduly influenced by any single study, a leave-one-out sensitivity analysis was performed, in which the bivariate meta-analysis was repeated after omitting each index model in turn, and the changes in pooled sensitivity and specificity were examined.

Potential small-study effects were assessed using Deeks funnel plot asymmetry test. Funnel plot asymmetry may indicate small-study effects, of which publication bias is one potential cause [46]. A *P* value less than .05 was considered statistically significant, unless otherwise specified.

Ethical Considerations

This meta-analysis did not involve primary human participants or experimental procedures, so ethical approval was not required. All included studies have been approved by their respective institutional review boards, and informed consent was obtained from participants where applicable.

Results

Study Selection and Characteristics

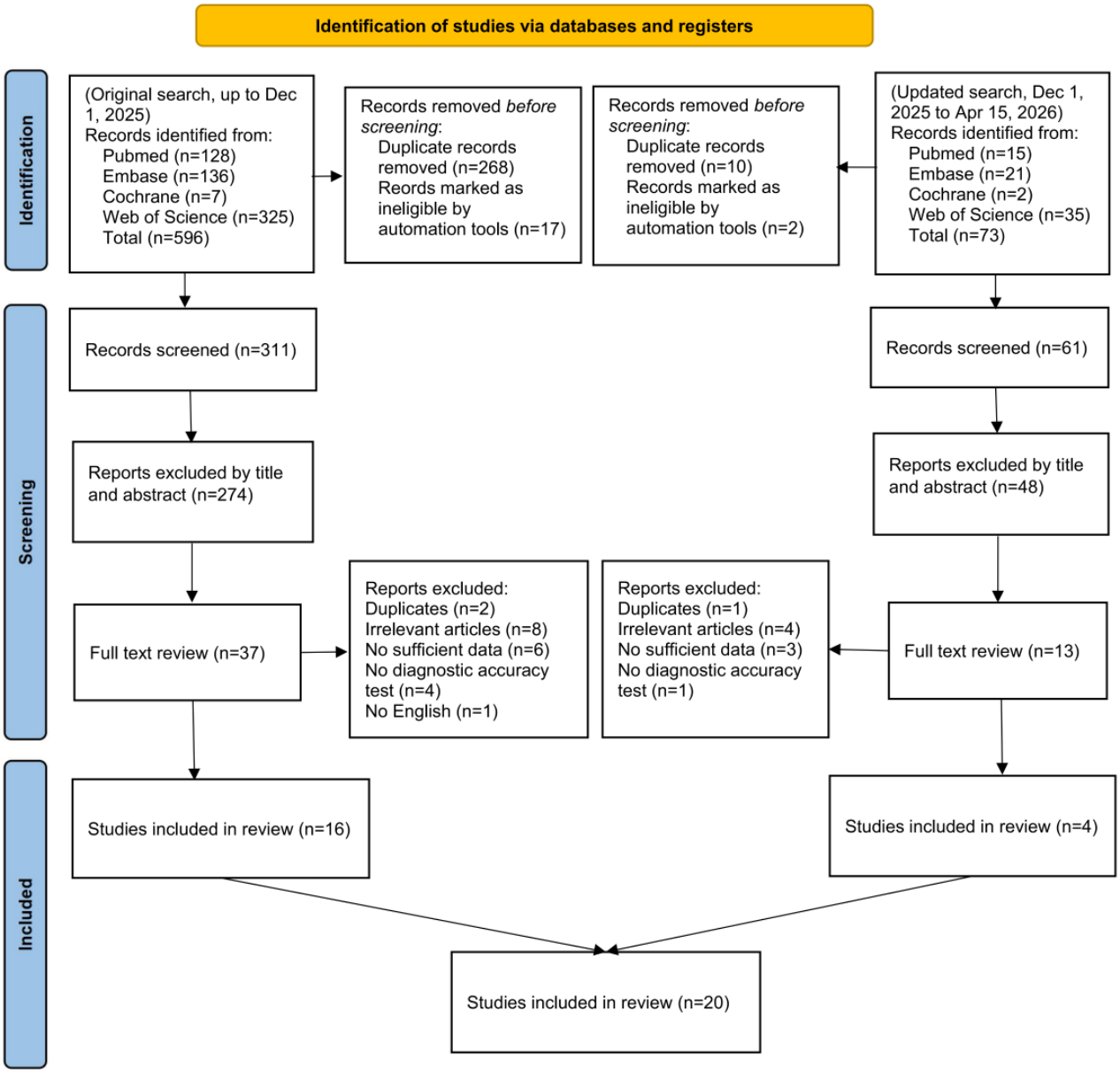
As shown in Table S1 in Multimedia Appendix 5 [12-14,18-21,25-36,47] and Figure 1, the systematic search identified 669 records, of which 372 remained after removing duplicates and automated exclusions for title/abstract screening; 322 were excluded as irrelevant. Full-text assessment of 50 articles led to the exclusion of 30 studies, primarily due to duplicate publications, nondiagnostic accuracy design, or non-English language. Ultimately, 20 studies were included (16 from the original search, 4 from the updated search), encompassing approximately 44,167 participants, with sample sizes ranging from 31 to 10,356. A total of 115 ML models and their validation records were extracted for secondary analyses, while the primary analysis was based on 20 prespecified index models (one per study).

Key characteristics of the included studies are summarized in Table S1 in Multimedia Appendix 5. Most studies applied the AWGS 2019 criteria (n=17) [12-14,19,21,25,27-36,47], with the remainder using EWGSOP2 or FNIH standards. Ten studies [12,19,20,25,30,31,33-36] included external validation, while the other 10 relied on internal validation data [13,14,18,21,26-29,32,47]. Reporting of thresholds varied: 6 studies [20,26,30,31,35,47] used data-driven thresholds based on the Youden index, 1 study [34] adopted an author-recommended clinical threshold, and the remaining 13 studies [12-14,18,19,21,25,27-29,32,33,36] applied a default probability cutoff of 0.5 or reported only a single operating point without further specification (Table S2 in Multimedia Appendix 5).

Candidate predictors were diverse, including demographic characteristics (eg, age, sex, height, weight, and BMI), anthropometric measures (eg, waist, calf, and upper-arm circumference), laboratory markers (eg, lipids and inflammatory indicators), lifestyle and behavioral factors (eg, physical activity, smoking, alcohol consumption, and diet), and clinical functional or health indicators (eg, grip strength, cognitive function, activities of daily living, and comorbidities). Age and BMI were the most frequently included predictors, followed by height, weight, and waist circumference. Functional measures, such as grip strength and cognitive/activities of daily living assessments, were also commonly incorporated. Differences in predictor selection and importance across studies likely reflect variation in study design, data sources, and modeling approaches.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for study selection.



Risk of Bias Assessment

The risk of bias of the 20 included studies is summarized in Multimedia Appendix 5 (Figure 2) [12-14,18-21,25-36,47]. In the participants' domain, all studies (100%) were rated low risk; in the predictors domain, 18 (90%) studies were low risk [13,14,18-21,25,26,28-36,47]; and in the outcomes domain, 17 (85%) studies were low risk [13,18-21,26-36,47].

The analysis domain showed the highest risk, with only 8 (40%) studies rated low [20,25,29-34]. High-risk issues included insufficient sample size relative to candidate predictors (9, 45% studies) [13,14,18,19,27-29,34,47], inadequate handling of missing data (12, 60% studies) [12-14,25-33], reliance on univariable predictor selection (2, 10% studies) [13,29], and insufficient control of overfitting or optimism (among 10 studies without external validation [13,14,18,21,26-29,32,47], 4 used complex ML algorithms with only a simple random split for internal validation [14,27-29]). Lack of model calibration assessment was common. Only 11 (55%) studies [12,13,18,20,25,26,30,31,33,35,47] reported some form of calibration, of which 8 provided calibration curves [12,20,25,26,30,31,33,35]. Overall, 7 (35%) studies were low risk [20,29-34], 10 (50%) studies were high risk [12-14,18,19,21,25-27,36], and 3 (15%) studies [28,35,47] were unclear, mainly due to limitations in the analysis domain.

XSL•FO
RenderX

Figure 2. Risk of bias assessment results (Prediction Model Risk of Bias Assessment Tool) [12-14,31,56-70].



Primary Analysis: Diagnostic Performance of Index Models

Twenty prespecified index models (one per study) were included in the primary analysis. Performance metrics were extracted preferentially from external validation datasets; if unavailable, internal validation or test set results were used. Eighteen models provided sufficient data to reconstruct 2×2 tables and were analyzed using a bivariate random-effects model (Reitsma model, 0.5 continuity correction). Two studies (Yu et al [47] and Xiang et al [18]) reported only mean sensitivity and specificity from repeated random train-test splits and were excluded from the bivariate model but retained for AUROC analysis.

The SROC curve and 95% prediction region are shown in Figure 3. Pooled sensitivity was 0.77 (95% CI 0.68-0.83; 95% PI 0.37-0.95; I^2=94.4%) and pooled specificity 0.88 (95% CI 0.83-0.91; 95% PI 0.62-0.97;I^2=97.4%; Figures 4 and 5). DOR, PLR, and NLR were derived from the bivariate model via the delta method: DOR 23.91 (95% CI 13.42-42.61; 95% PI 2.24-301.44; Figure 6), PLR 6.38 (95% CI 4.51-9.03; 95% PI 1.58-24.41), NLR 0.27 (95% CI 0.19-0.37; 95% PI 0.09-0.88; Figures S1 and S2 in Multimedia Appendix 5).

AUROC meta-analysis included 17 models. Logit-transformed AUROC values were fitted using a random-effects model with the HKSJ method, yielding a pooled AUROC of 0.89 (95% CI 0.82-0.93; 95% PI 0.54-0.98, I^2=97.8%; Figure 7). Most SEs were derived from reported 95% CIs, with 2 approximated from validation sample sizes.

XSL•FO
RenderX

Figure 3. The bivariate boxplot and overall combined area under the receiver operating characteristic curve for machine learning models used in sarcopenia prediction. SENS: sensitivity; SPEC: specificity.

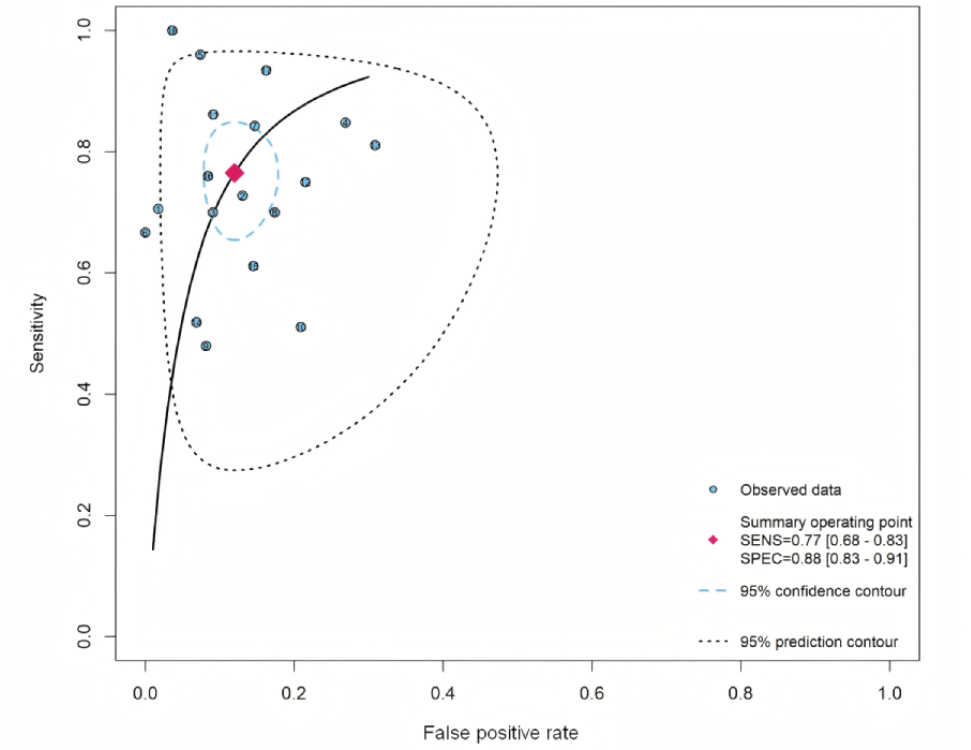
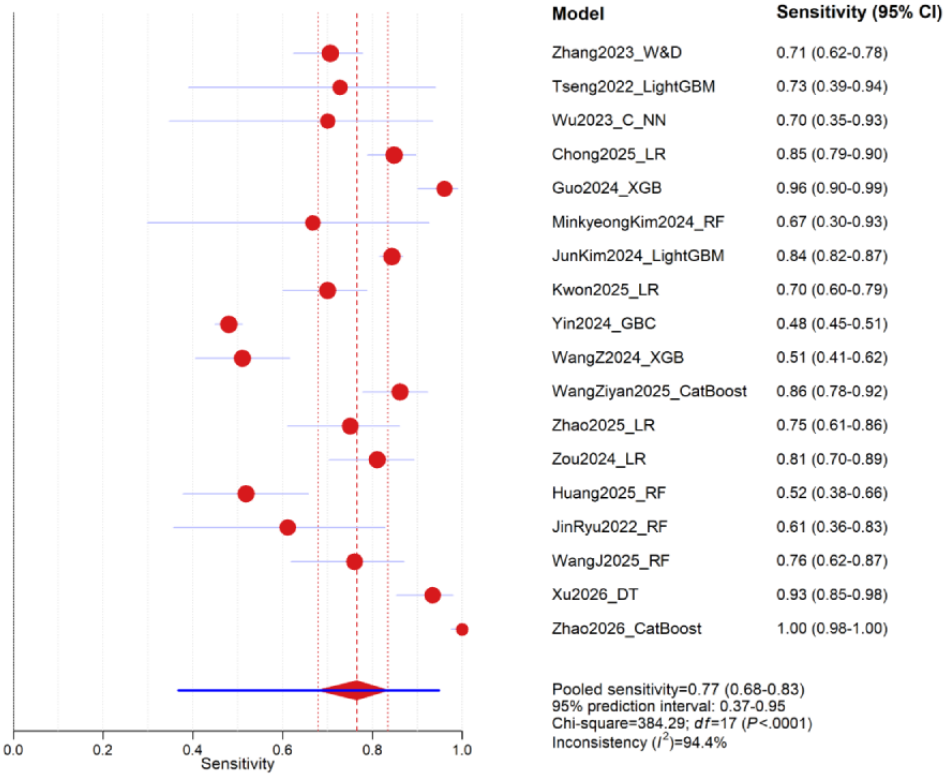


Figure 4. Forest plot of pooled sensitivity for the 18 index models with complete 2×2 contingency tables [12-14,56-70].



XSL•FO
RenderX

Figure 5. Forest plot of pooled specificity for the 18 index models with complete 2×2 contingency tables [12-14,56-70].

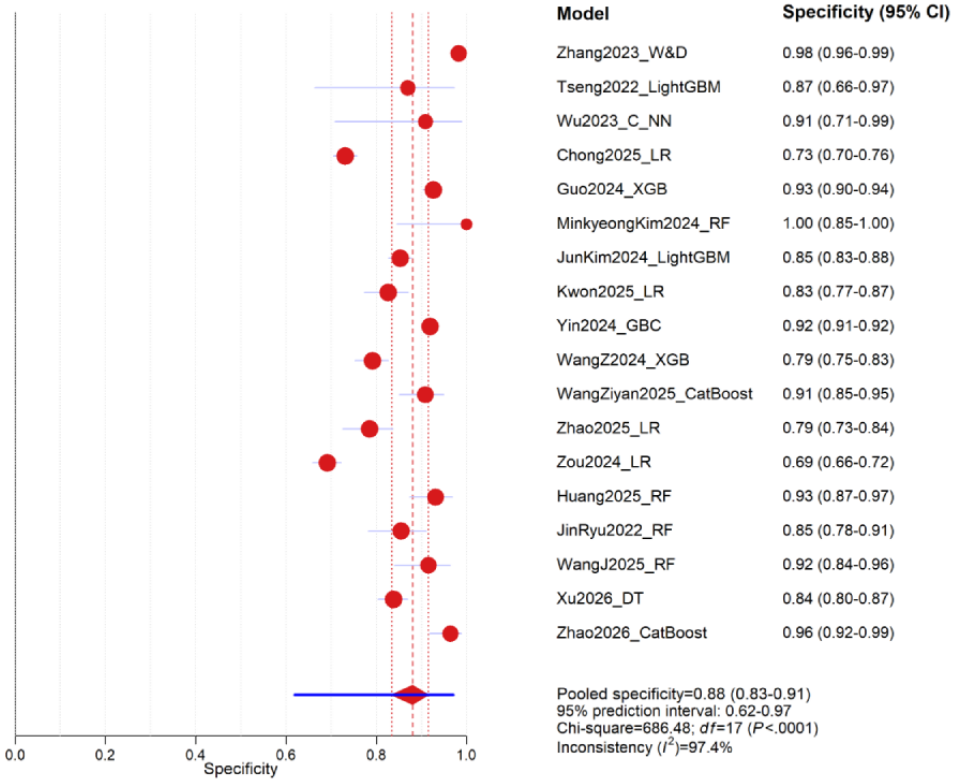
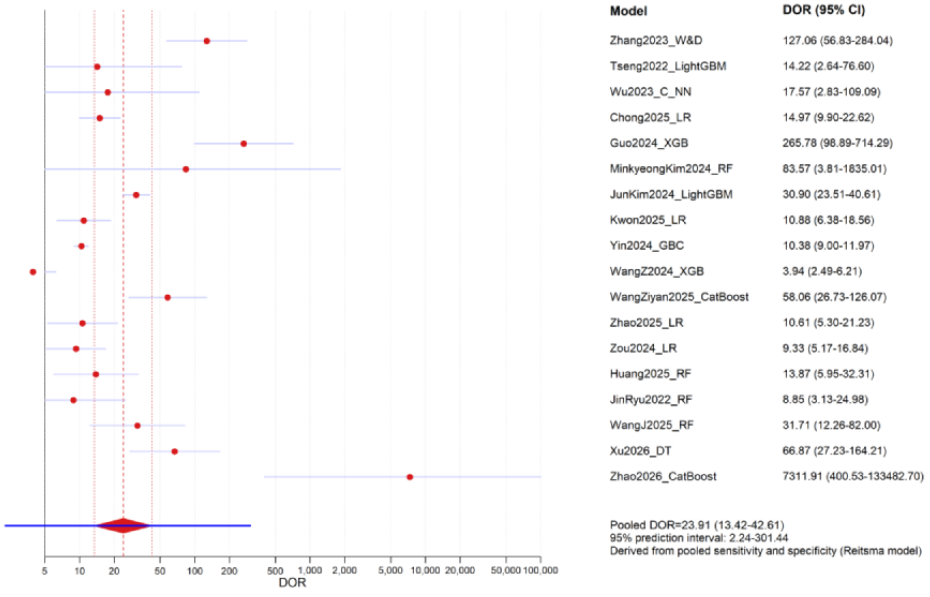
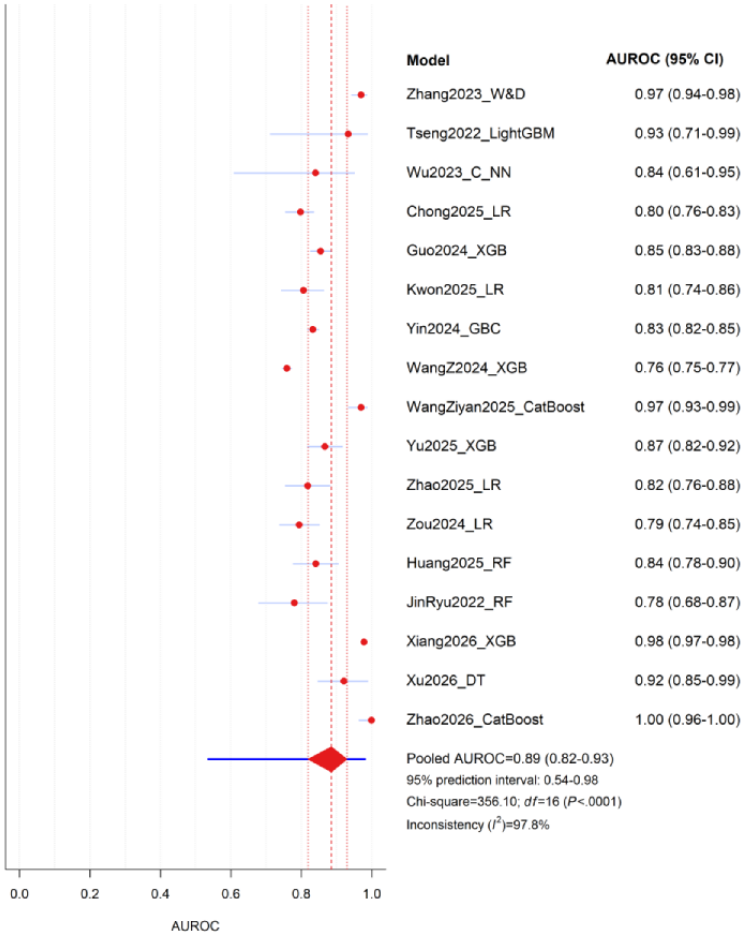


Figure 6. Forest plot of pooled diagnostic odds ratio (DOR) for the 18 index models included in the bivariate meta-analysis [12-14,56-70].



XSL•FO
RenderX

Figure 7. Forest plot of pooled area under the receiver operating characteristic curve (AUROC) pooled using a univariate random-effects model with the Hartung-Knapp-Sidik-Jonkman (HKSJ) method [12-14,31,32,56,57,60-67,69,70].



Secondary Analysis: All Models With RVE

The secondary analysis included all 115 model-validation records with reconstructable 2×2 contingency tables (from 18 studies [12-14,19-21,25-36]). Since multiple models from the same study are not independent, random-effects models with cluster-RVE were applied, clustering at the study level to account for within-study dependence. RVE analyses of logit-transformed sensitivity and specificity (Figures S3 and S4 in Multimedia Appendix 5) gave a back-transformed pooled sensitivity of 0.75 (95% CI 0.57-0.88) and a pooled specificity of 0.88 (95% CI 0.83-0.92). The corresponding DOR, PLR, and NLR were derived from the RVE-pooled sensitivity and specificity estimates using the delta method, yielding values of 22.64 (95% CI 9.49-54.01), 6.34 (95% CI 4.23-9.50), and 0.28 (95% CI 0.16-0.50), respectively. These estimates were highly consistent with the primary analysis, confirming that the main findings are robust to the inclusion of multiple models per study. Detailed pooled estimates are provided in Table S3 in Multimedia Appendix 5.

Subgroup Analyses

By Algorithm Type

Subgroup analyses were performed on the index models from the primary analysis. Pooled estimates for each algorithm subgroup based on the 18 index models with complete 2×2 tables are summarized in Table 1. Among algorithm subgroups with at least 3 models, the GB subgroup (n=7) had a pooled AUROC of 0.92 (95% CI 0.79-0.97), pooled sensitivity of 0.84 (95% CI 0.60-0.95), and pooled specificity of 0.90 (95% CI 0.85-0.93). The LR subgroup (n=4) had a pooled AUROC of 0.80 (95% CI 0.79-0.82), pooled sensitivity of 0.79 (95% CI 0.71-0.86), and pooled specificity of 0.76 (95% CI 0.70-0.82). The RF subgroup (n=4) had a pooled sensitivity of 0.63 (95% CI 0.50-0.75) and pooled specificity of 0.90 (95% CI 0.85-0.94); AUROC was not pooled for RF because only 2 models in this subgroup reported AUROC values. CIs for the different algorithm subgroups overlapped. The patterns for DOR, PLR, and NLR were generally consistent with the AUROC findings. For algorithms with fewer than 3 models (NN and DT), no pooled estimates were calculated. AUROC analyses were based on the 17 index models that reported AUROC values; the corresponding subgroup-level AUROC results are also presented in Table 1.

XSL•FO
RenderX

Table 1. Subgroup analyses by algorithm type and validation type.

Subgroup	N models[a] (2X2/AUROC[b])	AUROC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	DOR[c] (95% CI)	PLR[d] (95% CI)	NLR[e] (95% CI)
Algorithm type							
GB[f]	7/8	0.92 (0.79-0.97)	0.84 (0.60-0.95)	0.90 (0.85-0.93)	46.21 (18.78-113.70)	8.23 (6.18-8.98)	0.18 (0.05-0.48)
LR[g]	4/4	0.80 (0.79-0.82)	0.79 (0.71-0.85)	0.76 (0.70-0.82)	11.93 (9.12-15.59)	3.29 (2.80-3.82)	0.28 (0.180-0.419)
RF[h]	4/2	—[i]	0.63 (0.50-0.75)	0.90 (0.85-0.94)	15.86 (13.71-18.36)	6.45 (4.87-8.13)	0.41 (0.27-0.59)
Validation type							
External[j]	8/7	0.85 (0.74-0.92)	0.70 (0.59-0.79)	0.88 (0.78-0.94)	17.11 (16.35-17.91)	5.83 (3.66-9.32)	0.34 (0.22-0.52)
Internal[k]	10/10	0.91 (0.81-0.96)	0.84 (0.70-0.92)	0.87 (0.82-0.90)	34.74 (19.55-61.74)	6.32 (5.19-7.13)	0.18 (0.08-0.37)

[a]N models:(2X2/AUROC): number of models with a complete 2×2 table (used for bivariate Reitsma model)/AUROC and calculable standard error (used for AUROC meta-analysis).

[b]AUROC: area under the receiver operating characteristic curve.

[c]DOR: diagnostic odds ratio.

[d]PLR: positive likelihood ratio.

[e]NLR: negative likelihood ratio.

[f]GB: gradient boosting (including XGBoost, LightGBM, CatBoost, GBC, GBDT, and AdaBoost).

[g]LR: logistic regression (including nomogram based on LR).

[h]RF: random forest (including its variant extra trees).

[i]Not pooled (RF subgroup had only 2 models reporting AUROC).

[j]External: model evaluated on datasets independent from the development cohort (temporal or geographically distinct).

[k]Internal: model validated only using random split sample, cross-validation, or bootstrapping within the development cohort.

By Validation Type

Subgroup analyses compared externally validated models (n=8) with internally validated models (n=10). Pooled estimates for all metrics are shown in Table 1. Externally validated models had a pooled AUROC of 0.85 (95% CI 0.74-0.92), pooled sensitivity of 0.70 (95% CI 0.59-0.79), and pooled specificity of 0.88 (95% CI 0.78-0.94). Internally validated models had a pooled AUROC of 0.91 (95% CI 0.81-0.96), pooled sensitivity of 0.84 (95% CI 0.70-0.92), and pooled specificity of 0.87 (95% CI 0.82-0.90).

Meta-Regression Analysis

Univariable random-effects meta-regression was performed on the 18 index models with complete 2×2 contingency table data to explore associations between study-level characteristics and the log-transformed DOR (Table S4 in Multimedia Appendix 5). None of the examined covariates reached conventional statistical significance. External validation showed the strongest signal among the covariates examined (coefficient=−1.14, 95% CI −2.34 to 0.05; *P*=.06), with a trend toward a lower DOR. AWGS 2019 criteria (coefficient=−1.25, *P*=.11), multicenter design (coefficient=−0.51, *P*=.53), and validation sample size (coefficient=−0.15, *P*=.79) showed no significant associations. Residual heterogeneity remained high across all univariable models (I^2>93%), indicating that the included study-level characteristics explained little of the between-study variance. Given the small number of index models (k=18) [12-14,19-21,25-36], statistical power was limited; these findings should therefore be interpreted as exploratory.

Sensitivity Analyses

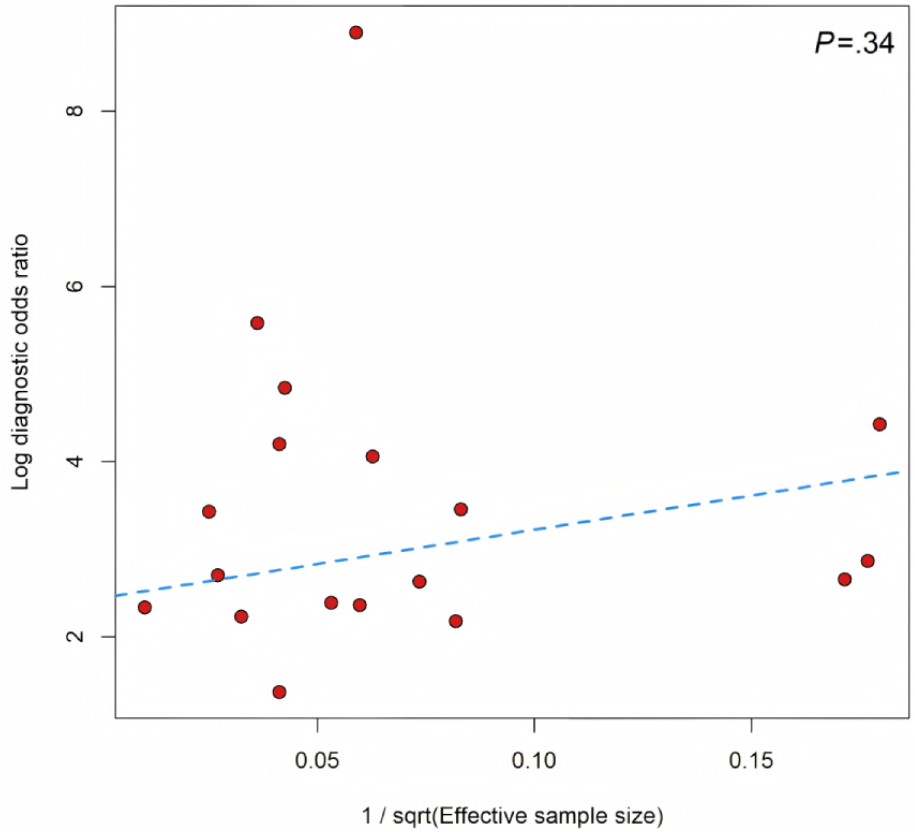
The leave-one-out analysis showed that after sequentially omitting each index model, the pooled sensitivity ranged from 0.74 to 0.78 and the pooled specificity ranged from 0.87 to 0.89. All leave-one-out estimates fell within the 95% CIs of the primary estimates (sensitivity 0.77, specificity 0.88), indicating that no single model had a disproportionate influence on the pooled results (Figure S5 in Multimedia Appendix 5).

Small-Study Effects

Deeks funnel plot asymmetry test did not suggest marked small-study effects (*P*=.34) in Figure 8.

XSL•FO
RenderX

Figure 8. Deeks funnel plot for assessment of small-study effects.



GRADE Assessment of Evidence Quality

The GRADE assessment rated the certainty of evidence as very low across all outcomes (sensitivity, specificity, AUROC, DOR, PLR, and NLR). This rating was primarily driven by high risk of bias (50% of studies at high risk, mainly in the analysis domain), indirectness (most studies conducted in Asian populations with heterogeneous diagnostic criteria), and substantial inconsistency (I^2>94% for sensitivity and specificity, I^2=97.8% for AUROC). Imprecision was not a concern, as the 95% CIs for PLR (4.51-9.03) and NLR (0.19-0.37) did not cross the clinical decision thresholds (2.0 and 0.5, respectively). Publication bias was not detected (Deeks test *P*=.34). The very low certainty indicates that the pooled estimates should be interpreted as hypothesis-generating, and the models are not yet ready for routine clinical deployment. Detailed results are provided in Table S5 in Multimedia Appendix 5.

Discussion

Principal Findings

This systematic review and meta-analysis included 20 studies [12-14,18-21,25-36,47] and quantitatively synthesized the diagnostic performance of ML models for sarcopenia prediction. Overall, ML models showed good average discrimination, with a pooled AUROC of 0.89, a pooled sensitivity of 0.77, and a pooled specificity of 0.88. Two key findings emerged: first, externally validated models exhibited lower apparent performance than internally validated models; second, substantial heterogeneity was observed across studies, and the wide AUROC PI (0.54-0.98) indicates that model discrimination may vary considerably across different clinical settings [48].

Interpretation in Light of GRADE, Risk of Bias, and Heterogeneity

The certainty of evidence was rated as very low across all outcomes using the GRADE approach for diagnostic test accuracy studies [49,50]. This rating was driven by high risk of bias (50% of studies at high risk), indirectness (mostly Asian populations with heterogeneous criteria), and substantial inconsistency (I^2>94%). These factors mean we have little confidence in the effect estimate, and the true effect is likely to differ substantially from the pooled estimate.

A clear distinction must be made between the average effect and the distribution of effects. The pooled estimates reflect the average effect across included studies, as indicated by the 95% CIs (eg, pooled AUROC 0.89, 95% CI 0.82-0.93). However, the 95% PIs were considerably wider (eg, AUROC 95% PI 0.54-0.98), indicating that model performance in a new clinical setting could range from poor to excellent. The PI represents the range within which the true effect of a future study is expected to lie, and thus provides a more clinically relevant estimate of variability than the CI alone [51]. This distinction further supports the very low certainty rating.

Clinicians should therefore be cautious about directly applying these models in routine practice without local validation. Future research should prioritize rigorous external validation, transparent reporting, calibration assessment, and prospective implementation studies to evaluate real-world clinical utility [52].

XSL•FO
RenderX

Interpretation and Comparison With Previous Literature

Compared with previous reviews in the field, especially the recent review by Zhang et al [15], this study provides several important extensions. Earlier reviews mainly provided qualitative summaries of sarcopenia prediction models, whereas we focused specifically on ML-based models and quantitatively synthesized their diagnostic performance. Moreover, we compared performance across algorithm types and explored sources of heterogeneity through subgroup analyses and meta-regression. We also explicitly addressed the unit-of-analysis issue: the primary analysis included one prespecified index model per study, and a secondary analysis using cluster-RVE included all available models, thereby improving methodological rigor [43].

In the algorithm-level subgroup analysis, the GB subgroup had the numerically highest pooled AUROC (0.922), which is conceptually consistent with the ability of ensemble learning methods to improve predictive performance by combining multiple weak learners [53]. However, the CIs for the GB, LR, and RF subgroups overlapped considerably, consistent with the systematic review by Christodoulou et al [54], which concluded that when studies are methodologically rigorous, complex ML methods do not consistently outperform LR. This suggests that model selection should balance performance and clinical interpretability, and current evidence does not support the definitive superiority of any one algorithm. Notably, the RF subgroup showed a lower pooled sensitivity (0.634) and a higher pooled specificity (0.902). This pattern may be related to threshold choice, data imbalance, or between-study heterogeneity, and cannot be attributed solely to inherent algorithm characteristics. In clinical screening, a model with high specificity may carry a substantial risk of missing cases, emphasizing the need to weigh sensitivity and specificity according to the intended use case. Although deep learning has advantages in capturing complex nonlinear relationships, LR and tree-based models remain valuable in clinical practice because of their relative transparency and ease of deployment [55].

An important finding is that externally validated models had lower apparent performance than internally validated models. The validation-type subgroup analysis showed that externally validated models had a pooled AUROC of 0.85 and a pooled sensitivity of 0.70, both lower than those of internally validated models (AUROC 0.91, sensitivity 0.84). This pattern is consistent with the well-recognized problem of optimism bias in prediction modeling [56]. Importantly, this performance drop does not indicate model “failure”; rather, it likely reflects more realistic estimates of generalizability, because external validation typically involves different case-mix, diagnostic criteria, measurement methods, and data distributions. Consequently, externally validated performance is more clinically relevant than internally validated performance [57]. Furthermore, external validation should be viewed as a continuous and necessary process, as a model’s predictive performance can change over time and across different settings [58]. The limited number of externally validated studies also highlights a major barrier to clinical translation in this field. Moreover, repeated validation across diverse populations and settings may be more informative than a single external validation study [59].

Substantial inconsistency was observed across studies (I^2>94%), which is expected in meta-analyses of prediction models, especially those involving ML [40]. This inconsistency indicates that the proportion of total variation across studies due to heterogeneity rather than chance is high. However, it is important to distinguish inconsistency from heterogeneity: the former is measured by I^2, while the latter is reflected by the PIs, which indicate the range of effects expected across different settings [42]. Potential sources of heterogeneity include differences in study populations, sarcopenia diagnostic criteria, predictor definitions, model development strategies, and validation methods. Furthermore, as suggested by our exploratory meta-regression, heterogeneity in this field may arise more from micro-level methodological differences (such as missing data handling, feature selection methods, hyperparameter tuning, and threshold determination) rather than from macro-level design features alone. These micro-level factors were often inadequately reported in the included studies and difficult to quantify in meta-regression [60]. Therefore, our pooled estimates should be interpreted as average performance across heterogeneous contexts and should not be directly transported to any single clinical setting. The wide AUROC PI further illustrates that real-world discrimination performance may vary substantially [48].

PROBAST assessment showed that the risk of bias was generally low in the participant, predictor, and outcome domains, but was prominent in the analysis domain, with only a few studies rated as low risk. Common methodological problems included insufficient sample size relative to the number of candidate predictors, inappropriate handling of missing data [61], reliance on univariable predictor screening, and inadequate control of overfitting and optimism bias (especially in studies using complex ML algorithms without external validation) [62]. Critically, calibration assessment was reported very rarely; only a few studies provided calibration plots or calibration metrics [63]. This deficiency severely limits the ability to judge the accuracy of predicted probabilities (a model with good discrimination but poor calibration may systematically over) or underestimate risk in clinical use, potentially leading to misguided decisions [64], as a well-calibrated model is essential for reliable individualized risk prediction, a perspective that has been consistently emphasized in the methodological literature [65]. Moreover, decision curve analysis (DCA) was reported in only 5 studies [18,26,30,33,34], meaning that the net clinical benefit of these models could not be evaluated, which substantially weakens the willingness and feasibility for clinical translation [64]. These methodological shortcomings collectively indicate that while the current evidence provides a basis for discrimination, major gaps remain in calibration reliability and clinical utility. Future studies should systematically improve reporting quality following the TRIPOD-AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence) guidelines [52,66].

XSL•FO
RenderX

The included models varied considerably in how thresholds were determined: some used data-driven thresholds based on the Youden index, some used author-recommended clinical thresholds, and a few only reported default probability thresholds (eg, 0.5). This inconsistency directly affects the point estimates of sensitivity and specificity and their comparability, and is an important source of between-study heterogeneity. Although our bivariate Reitsma model partially accounted for threshold effects by jointly modeling sensitivity and specificity, it relies on having at least one operable threshold per model. Future studies should follow the TRIPOD-AI guidelines to clearly report the method for threshold determination (prespecified vs data-driven), the specific threshold value, its clinical rationale, and, whenever possible, validate threshold stability across different validation sets. Furthermore, when models are intended for different clinical scenarios (screening vs confirmation), threshold selection should be adapted to the risk tolerance of the setting rather than adhering to a default value [52].

Clinical Translation and Implementation Considerations

From a clinical practice perspective, ML models have potential value for sarcopenia screening and risk stratification. First, in community or primary care settings where gold-standard tools such as DXA or CT are not readily available, ML models based on routine clinical data (eg, age, BMI, grip strength, and calf circumference) could serve as cost-effective prescreening tools to identify high-risk individuals and guide further diagnostic evaluation [1]. Second, in hospital or specialized outpatient settings, ML models could assist clinicians in individualized risk stratification for specific disease populations (eg, patients with chronic kidney disease, diabetes, or a history of stroke) and facilitate timely nutritional support and rehabilitation interventions. However, several obstacles must be overcome before these applications can be realized: real-world model performance is often lower than reported in development (optimism bias), and the lack of standardized thresholds across models makes it difficult for clinicians to interpret outputs directly. Moreover, model interpretability (eg, through Shapley additive explanations [SHAP] values or local interpretable models) is crucial for building clinical trust. Very few of the included studies provided easy-to-use decision tools (eg, web calculators, mobile apps, or electronic health record integration), limiting bedside adoption [67], This reflects a broader implementation gap, where considerations of local performance validation and end user trust are often overlooked in the model development phase [68]. Future research should collaborate with clinicians and health informaticians to develop and validate user-friendly implementation tools and conduct prospective implementation studies to evaluate the cost-effectiveness and net benefit of ML models in real-world clinical pathways [64].

Strengths and Limitations

This study has several strengths. First, to our knowledge, it is the first quantitative meta-analysis focused exclusively on ML-based sarcopenia prediction models. Second, unlike previous reviews that mainly provided narrative summaries, we formally synthesized diagnostic performance and explored sources of heterogeneity at both the algorithm and study-design levels. Third, we explicitly addressed the unit-of-analysis issue: a primary analysis based on one index model per study, supplemented by a secondary RVE analysis that included all models. Fourth, we systematically assessed methodological quality using PROBAST, enabling a more evidence-based interpretation of bias and applicability concerns.

Several limitations should be acknowledged. First, the number of included studies was still relatively limited, and most studies were conducted in Asian populations, which may limit generalizability to other regions or diagnostic definitions [4,6]. Second, high heterogeneity was observed across analyses. We explored it through subgroup analyses and univariate meta-regression, but because of the limited number of available studies (k=18) [12-14,19-21,25-36], the meta-regression was used only for hypothesis generation [44], and residual heterogeneity remained high (I^2>93%), indicating that the included covariates explained only a small fraction of the between-study variance [40]. In addition, the included studies varied markedly in micromethodological details such as missing data handling, feature selection strategies, hyperparameter tuning, and threshold determination; these factors may be important sources of heterogeneity but were not examined in the meta-regression because of incomplete reporting and difficulty in quantitative coding. Future research should follow TRIPOD-AI guidelines to fully report these methodological details, allowing their impact to be identified and quantified in larger evidence syntheses [52]. Third, many studies were judged to be at high or unclear risk of bias, particularly in the analysis domain [66]. Fourth, because the bivariate diagnostic accuracy model requires threshold-specific sensitivity and specificity estimates from a clearly defined validation dataset, studies that reported only averaged sensitivity and specificity from repeated resampling procedures (eg, Yu et al [47] and Xiang et al [18]) could not be included in the bivariate analysis. This may have reduced the number of studies contributing to the synthesis of sensitivity and specificity and introduced some degree of reporting-related selection bias. This limitation highlights the need for future ML prediction studies to report validation-set-specific 2×2 tables or sufficient threshold-specific information to allow reconstruction [69]. Finally, although the Deeks test did not indicate significant small-study effects (P=.34), the possibility of publication bias cannot be completely excluded because of the limited number of studies and the high residual heterogeneity [46,70].

In summary, this is innovative in several respects. First, it is the first quantitative meta-analysis to systematically evaluate ML-based sarcopenia prediction models with algorithm-level stratification. Second, unlike previous qualitative reviews, it provides pooled estimates of diagnostic accuracy across multiple algorithms and explicitly addresses the unit-of-analysis issue through a 2-tier analytical approach. Third, it brings to the field a comprehensive synthesis of diagnostic performance, revealing that externally validated models underperform those validated only internally and that micro-level methodological differences are the primary drivers of heterogeneity. Fourth, from a real-world perspective, these models offer potential value as prescreening tools in community and primary care settings, yet their current performance and methodological limitations

XSL•FO
RenderX

preclude routine clinical adoption. Future research should focus on rigorous external validation, transparent reporting, calibration assessment, and the development of clinically implementable models to support reliable real-world translation.

Conclusion

ML models show promising average performance for predicting sarcopenia, but their current clinical applicability remains limited by substantial heterogeneity, insufficient external validation, and methodological weaknesses in the available studies. Future research should focus on rigorous external validation, transparent reporting, calibration assessment, and clinically implementable model design to support reliable real-world translation.

Acknowledgments

We would like to express our gratitude to all authors involved in this research for their support in providing relevant data.

XJ and BL contributed equally to this work and share corresponding authorship.

Funding

This study was supported by the Taishan Scholars Program of Shandong Province (grant number tsqn202306402) and the Natural Science Foundation of Shandong Province (grant numbers ZR2023MH136 and ZR2020MH025). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

All data generated or analyzed during this study are included in this published paper and its supplementary information files.

Generative AI Use

This manuscript did not use generative AI for content creation, data analysis, or research design.

Authors' Contributions

Conceptualization: YP, ZR, LX, M Zhao, CB
Methodology: LX, M Zhao, CB
Study design: M Zhou, QS
Search strategy: YP, ZR
Study selection: YP, ZR
Data extraction: YP, ZR
Quality assessment: YP, ZR
Supervision: XJ, BL
Writing – review & editing: XJ, BL

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA 2020 checklist.
[DOCX File , 33 KB-Multimedia Appendix 1]

Multimedia Appendix 2

PRISMA 2020 for Abstracts checklist.
[DOCX File , 22 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Search strategy.
[DOCX File , 17 KB-Multimedia Appendix 3]

Multimedia Appendix 4

PRISMA-S checklist.
[DOCX File , 17 KB-Multimedia Appendix 4]

XSL•FO
RenderX

Multimedia Appendix 5

Additional methodological details and results, including study characteristics, threshold selection, secondary analyses, meta-regression, certainty of evidence, and diagnostic performance assessments.
[DOCX File , 3428 KB-Multimedia Appendix 5]

References

1. 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(1):16-31. [FREE Full text] [doi: 10.1093/ageing/afy169] [Medline: 30312372]
2. Petermann-Rocha F, Balntzi V, Gray SR, Lara J, Ho FK, Pell JP, et al. Global prevalence of sarcopenia and severe sarcopenia: a systematic review and meta-analysis. J Cachexia Sarcopenia Muscle. 2022;13(1):86-99. [FREE Full text] [doi: 10.1002/jcsm.12783] [Medline: 34816624]
3. Yeung SS, Reijnierse EM, Pham VK, Trappenburg MC, Lim WK, Meskers CG, et al. Sarcopenia and its association with falls and fractures in older adults: a systematic review and meta-analysis. J Cachexia Sarcopenia Muscle. 2019;10(3):485-500. [FREE Full text] [doi: 10.1002/jcsm.12411] [Medline: 30993881]
4. Chen L, Woo J, Assantachai P, Auyeung T, Chou M, Iijima K, et al. Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. J Am Med Dir Assoc. 2020;21(3):300-307.e2. [doi: 10.1016/j.jamda.2019.12.012] [Medline: 32033882]
5. Cesari M, Fielding RA, Pahor M, Goodpaster B, Hellerstein M, van Kan GA, et al. Biomarkers of sarcopenia in clinical trials-recommendations from the international working group on sarcopenia. J Cachexia Sarcopenia Muscle. 2012;3(3):181-190. [FREE Full text] [doi: 10.1007/s13539-012-0078-2] [Medline: 22865205]
6. Studenski SA, Peters KW, Alley DE, Cawthon PM, McLean RR, Harris TB, et al. The FNIH Sarcopenia Project: rationale, study description, conference recommendations, and final estimates. J Gerontol A Biol Sci Med Sci. 2014;69(5):547-558. [FREE Full text] [doi: 10.1093/gerona/glu010] [Medline: 24737557]
7. Ida S, Kaneko R, Murata K. SARC-F for screening of sarcopenia among older adults: a meta-analysis of screening test accuracy. J Am Med Dir Assoc. 2018;19(8):685-689. [doi: 10.1016/j.jamda.2018.04.001] [Medline: 29778639]
8. Rasool A, Bunterngchit C, Tiejian L, Islam MR, Qu Q, Jiang Q. Improved machine learning-based predictive models for breast cancer diagnosis. Int J Environ Res Public Health. 2022;19(6):3211. [FREE Full text] [doi: 10.3390/ijerph19063211] [Medline: 35328897]
9. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. 2020;9(2):14. [FREE Full text] [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
10. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920-1930. [doi: 10.1161/CIRCULATIONAHA.115.001593] [Medline: 26572668]
11. Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2(3):160. [FREE Full text] [doi: 10.1007/s42979-021-00592-x] [Medline: 33778771]
12. Zhang H, Yin M, Liu Q, Ding F, Hou L, Deng Y, et al. Machine and deep learning-based clinical characteristics and laboratory markers for the prediction of sarcopenia. Chin Med J (Engl). 2023;136(8):967-973. [FREE Full text] [doi: 10.1097/CM9.0000000000002633] [Medline: 37098831]
13. Tseng Y, Mo S, Zeng Y, Zheng W, Song H, Zhong B, et al. Machine learning model in predicting sarcopenia in crohn's disease based on simple clinical and anthropometric measures. Int J Environ Res Public Health. 2022;20(1):656. [FREE Full text] [doi: 10.3390/ijerph20010656] [Medline: 36612977]
14. Wu J, Lin S, Guan J, Wu X, Ding M, Shen S. Prediction of the sarcopenia in peritoneal dialysis using simple clinical information: a machine learning-based model. Semin Dial. 2023;36(5):390-398. [doi: 10.1111/sdi.13131] [Medline: 36890621]
15. Zhang X, Lian R, Tang H, Luo S, Chen X, Lu J, et al. Diagnostic prediction models for sarcopenia: a systematic review and meta-analysis. J Am Med Dir Assoc. 2025;26(8):105698. [doi: 10.1016/j.jamda.2025.105698] [Medline: 40451253]
16. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
17. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev. 2021;10(1):39. [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
18. Xiang W, Zhu H, Liu X, Wu Q. Development and validation of a machine learning-based sarcopenia prediction model using the triglyceride glucose-frailty index. J Int Med Res. 2026;54(1):3000605261417083. [FREE Full text] [doi: 10.1177/03000605261417083] [Medline: 41618756]
19. Wang J, Mu M, Zhang J, Yin X, Gao Y. Development of machine learning-based models for predicting sarcopenia risk in stroke patients and analysis of associated factors. Sci Rep. 2025;16(1):1490. [FREE Full text] [doi: 10.1038/s41598-025-31600-2] [Medline: 41361590]

XSL•FO
RenderX

20. Xu X. Machine learning-driven personalized risk prediction: developing an explainable sarcopenia model for older european adults with arthritis. J Clin Med. 2026;15(3):1022. [FREE Full text] [doi: 10.3390/jcm15031022] [Medline: 41682702]
21. Zhao X, Yao W, Shen J, Tang X, Zheng J, Guo C, et al. Development and validation of a machine learning-based risk prediction model for sarcopenia in community hospital patients: a retrospective cohort study. Front Aging. 2026;7:1772792. [FREE Full text] [doi: 10.3389/fragi.2026.1772792] [Medline: 41836318]
22. Cruz-Jentoft AJ, Baeyens JP, Bauer JM, Boirie Y, Cederholm T, Landi F, et al. Sarcopenia: European consensus on definition and diagnosis: report of the European Working Group on Sarcopenia in Older People. Age Ageing. 2010;39(4):412-423. [FREE Full text] [doi: 10.1093/ageing/afq034] [Medline: 20392703]
23. Bhasin S, Travison TG, Manini TM, Patel S, Pencina KM, Fielding RA, et al. Sarcopenia definition: the position statements of the sarcopenia definition and outcomes consortium. J Am Geriatr Soc. 2020;68(7):1410-1418. [doi: 10.1111/jgs.16372] [Medline: 32150289]
24. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11(10):e1001744. [FREE Full text] [doi: 10.1371/journal.pmed.1001744] [Medline: 25314315]
25. Chong HJ. A machine learning model for predicting sarcopenia among middle-aged adults: development and external validation. JMIR Med Inform. 2025;13:e75760. [FREE Full text] [doi: 10.2196/75760] [Medline: 40864499]
26. Guo J, He Q, She C, Liu H, Li Y. A machine learning-based online web calculator to aid in the diagnosis of sarcopenia in the US community. Digit Health. 2024;10:20552076241283247. [FREE Full text] [doi: 10.1177/20552076241283247] [Medline: 39360239]
27. Kim J. Machine-learning classifier models for predicting sarcopenia in the elderly based on physical factors. Geriatr Gerontol Int. 2024;24(6):595-602. [doi: 10.1111/ggi.14895] [Medline: 38744528]
28. Kim M, Kim D, Kang H, Park S, Kim S, Yoo J. A machine learning model for prediction of sarcopenia in patients with Parkinson's disease. PLoS One. 2024;19(1):e0296282. [FREE Full text] [doi: 10.1371/journal.pone.0296282] [Medline: 38165980]
29. Kwon S, Kim L, Won CW, Kim N, Chang JY, Kim M, et al. Predictive model development for possible sarcopenia in community-dwelling older adults: a cross-sectional machine learning approach using the Korean frailty and aging cohort study. BMC Geriatr. 2025;25(1):987. [FREE Full text] [doi: 10.1186/s12877-025-06612-2] [Medline: 41327100]
30. Yin L, Zhao J. An artificial intelligence approach for test-free identification of sarcopenia. J Cachexia Sarcopenia Muscle. 2024;15(6):2765-2780. [FREE Full text] [doi: 10.1002/jcsm.13627] [Medline: 39513334]
31. Wang Z, Wu Y, Zhu J, Fang Y. Machine learning-based prediction of sarcopenia in community-dwelling middle-aged and older adults: findings from the CHARLS. Psychogeriatrics. 2025;25(1):e13205. [doi: 10.1111/psyg.13205] [Medline: 39444246]
32. Wang Z, Zhou Y, Zeng X, Zhou Y, Yang T, Hu K. An explainable machine learning-based prediction model for sarcopenia in elderly Chinese people with knee osteoarthritis. Aging Clin Exp Res. 2025;37(1):67. [doi: 10.1007/s40520-025-02931-x] [Medline: 40053240]
33. Zhao Y, Hu Z, Dong H, Zhu L, Xia Q, Lei C, et al. Development and validation of nomogram and machine learning models to predict sarcopenia in patients with chronic kidney disease. Sci Rep. 2025;15(1):42370. [FREE Full text] [doi: 10.1038/s41598-025-26378-2] [Medline: 41309827]
34. Zou M, Shao Z. Construction and evaluation of sarcopenia risk prediction model for patients with diabetes: a study based on the China Health and Retirement Longitudinal Study (CHARLS). Diabetol Metab Syndr. 2024;16(1):230. [FREE Full text] [doi: 10.1186/s13098-024-01467-w] [Medline: 39285494]
35. Huang H, Jiang S, Chen Z, Yu X, Ren K, Zhao Q. A w-ACT model for sarcopenia among community-dwelling older adults based on National Basic Public Health Services: development and validation study. Front Public Health. 2025;13:1522903. [FREE Full text] [doi: 10.3389/fpubh.2025.1522903] [Medline: 40933425]
36. Ryu J, Eom S, Kim HC, Kim CO, Rhee Y, You SC, et al. Chest X-ray-based opportunistic screening of sarcopenia using deep learning. J Cachexia Sarcopenia Muscle. 2023;14(1):418-428. [FREE Full text] [doi: 10.1002/jcsm.13144] [Medline: 36457204]
37. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58(10):982-990. [doi: 10.1016/j.jclinepi.2005.02.022] [Medline: 16168343]
38. IntHout J, Ioannidis JPA, Borm GF. The hartung-knapp-sidik-jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol. 2014;14:25. [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
39. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC Med Res Methodol. 2015;15:99. [FREE Full text] [doi: 10.1186/s12874-015-0091-1] [Medline: 26573817]
40. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460. [FREE Full text] [doi: 10.1136/bmj.i6460] [Medline: 28057641]

XSL•FO
RenderX

41. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res. 2019;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
42. Borenstein M. Avoiding common mistakes in meta-analysis: understanding the distinct roles of Q, I-squared, tau-squared, and the prediction interval in reporting heterogeneity. Res Synth Methods. 2024;15(2):354-368. [doi: 10.1002/jrsm.1678] [Medline: 37940120]
43. Pustejovsky JE, Tipton E. Meta-analysis with robust variance estimation: expanding the range of working models. Prev Sci. 2022;23(3):425-438. [doi: 10.1007/s11121-021-01246-3] [Medline: 33961175]
44. Guolo A, Varin C. Random-effects meta-analysis: the number of studies matters. Stat Methods Med Res. 2017;26(3):1500-1518. [doi: 10.1177/0962280215583568] [Medline: 25953957]
45. Van Lissa CJ. Small sample meta-analyses: exploring heterogeneity using MetaForest. In: Van De Schoot R, Miočević M, editors. Small Sample Size Solutions. London. Routledge; 2020:186-202.
46. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;343(jul22 1):d4002-d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
47. Yu P, Zhang X, Sun G, Zeng P, Zheng C, Wang K. Sarcopenia prediction model based on machine learning and SHAP values for community-based older adults with cardiovascular disease in China. Front Public Health. 2025;13:1527304. [FREE Full text] [doi: 10.3389/fpubh.2025.1527304] [Medline: 40469611]
48. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res. 2023;12(4):101014. [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
49. Schünemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 1. study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy. J Clin Epidemiol. 2020;122:129-141. [doi: 10.1016/j.jclinepi.2019.12.020] [Medline: 32060007]
50. Schünemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 2. test accuracy: inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables. J Clin Epidemiol. 2020;122:142-152. [FREE Full text] [doi: 10.1016/j.jclinepi.2019.12.021] [Medline: 32058069]
51. Murad MH, Morgan RL, Falck-Ytter Y, Mustafa RA, Sultan S, Dahm P, et al. Simultaneous evaluation of the imprecision and inconsistency domains of GRADE can be performed using prediction intervals. J Clin Epidemiol. 2024;175:111543. [doi: 10.1016/j.jclinepi.2024.111543] [Medline: 39322122]
52. Collins GS. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:q902. [FREE Full text] [doi: 10.1136/bmj.q902] [Medline: 38636956]
53. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
54. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
55. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014;14:137. [FREE Full text] [doi: 10.1186/1471-2288-14-137] [Medline: 25532820]
56. Steyerberg EW. Overfitting and optimism in prediction models. In: Clinical Prediction Models. Statistics for Biology and Health. Cham. Springer International Publishing; 2019:95-112.
57. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015;68(1):25-34. [doi: 10.1016/j.jclinepi.2014.09.007] [Medline: 25441703]
58. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024;384:e074820. [FREE Full text] [doi: 10.1136/bmj-2023-074820] [Medline: 38224968]
59. Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. Nat Med. 2023;29(11):2686-2687. [doi: 10.1038/s41591-023-02540-z] [Medline: 37853136]
60. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594. [FREE Full text] [doi: 10.1136/bmj.g7594] [Medline: 25569120]
61. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338(jun29 1):b2393-b2393. [FREE Full text] [doi: 10.1136/bmj.b2393] [Medline: 19564179]
62. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. [doi: 10.1136/bmj.m441] [Medline: 32188600]

XSL•FO
RenderX

63. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic testsprediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. [doi: 10.1186/s12916-019-1466-7] [Medline: 31842878]
64. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574. [FREE Full text] [doi: 10.1177/0272989X06295361] [Medline: 17099194]
65. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc. 2020;27(4):621-633. [FREE Full text] [doi: 10.1093/jamia/ocz228] [Medline: 32106284]
66. Moons KG, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med. 2019;170(1):W1-W33. [doi: 10.7326/m18-1377]
67. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
68. Powis M, Ladak AM, Lakey A, Grant RC, Krzyzanowska MK, Peterson E, et al. Framework for artificial intelligence implementation research in healthcare: synthesizing current evidence on barriers and facilitators. NPJ Digit Med. 2026;9(1):543. [FREE Full text] [doi: 10.1038/s41746-026-02705-3] [Medline: 42104102]
69. Macaskill P, Gatsonis C, Deeks JJ. Handbook for Systematic Reviews of Diagnostic Test Accuracy. Cochrane. 2026. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook-systematic-reviews-diagnostic-test-accuracy [accessed 2026-05-12]
70. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. 2005;58(9):882-893. [doi: 10.1016/j.jclinepi.2005.01.016] [Medline: 16085191]

Abbreviations

AUROC: area under the receiver operating characteristic curve
AWGS 2019: Asian Working Group for Sarcopenia 2019
CART: classification and regression trees
CHARMS: Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies
CT: computed tomography
DCA: decision curve analysis
DOR: diagnostic odds ratio
DT: decision tree
DXA: dual-energy X-ray absorptiometry
EWGSOP: European Working Group on Sarcopenia in Older People
FNIH: Foundation for the National Institutes of Health
GB: gradient boosting
GBC: gradient boosting classifier
GBDT: gradient boosting decision tree
GRADE: Grading of Recommendations Assessment, Development and Evaluation
HKSJ: Hartung-Knapp-Sidik-Jonkman
IWGS: International Working Group on Sarcopenia
LightGBM: light gradient boosting machine
LR: logistic regression
MICE: multiple imputation by chained equations
ML: machine learning
MLP: multilayer perceptron
NLR: negative likelihood ratio
NN: neural network
PI: prediction interval
PLR: positive likelihood ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PROBAST: Prediction Model Risk of Bias Assessment Tool
REML: restricted maximum likelihood
RF: random forest
RVE: robust variance estimation
SARC-F: strength, assistance in walking, rising from a chair, climbing stairs, and falls

XSL•FO
RenderX

SDOC: Sarcopenia Definitions and Outcomes Consortium
SHAP: Shapley additive explanations
SROC: summary receiver operating characteristic
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence
XGBoost: extreme gradient boosting
W&D: Wide & Deep

Edited by S Brini; submitted 13.Jan.2026; peer-reviewed by L Lera, M-H Tsai, T Yusuff, S Kakraba; comments to author 11.Mar.2026; revised version received 14.Oct.2026; accepted 15.Oct.2026; published 23.Nov.2026

Please cite as:
Peng Y, Ren Z, Xue L, Zhao M, Bian C, Shen Q, Zhou M, Li B, Jin X
Predictive Modeling of Sarcopenia Using Machine Learning: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e91297
URL: https://www.jmir.org/2026/1/e91297
doi: 10.2196/91297
PMID:

©Yinying Peng, Zhexue Ren, Lina Xue, Mingshao Zhao, Ce Bian, Qi Shen, Mengjun Zhou, Bo Li, Xiaodong Jin. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX