

AI-based multimodal behavioral phenotyping for sarcopenia and muscle health assessment in older adults: study protocol

Insuk Oh

Yeungnam University

Chan-Su Lee

chansu@ynu.ac.kr

Yeungnam University

Study protocol

Keywords: Sarcopenia, Behavioral phenotyping, Digital biomarkers, Facilitator-mediated assessment, Embodied conversational agents (ECA), Older adults, Non-invasive screening

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10857313/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Sarcopenia, characterized by age-related declines in muscle mass and strength, is associated with functional decline in older adults. Objective assessment may be impractical for frequent community monitoring, while brief self-report tools such as the SARC-F have limited sensitivity. Facial dynamics, vocal characteristics, and movement may carry information outside conscious control, yet few studies have tested whether such phenotypes distinguish older adults with normal muscle mass and strength from those with low muscle mass. This protocol therefore frames community assessment in three linked layers: standard-instrument responses, a mediating layer accommodating varying digital familiarity, and multimodal behavioral signals generated throughout. The mediating layer is currently performed by a trained facilitator whose interaction patterns can inform the design of future embodied conversational agents (ECAs).

Methods

This study is conducted within an ongoing interdisciplinary sarcopenia research program at Yeungnam University. Up to 120 community-dwelling adults aged ≥ 65 years are recruited independently of the parent program. Each participant completes a single session of under 30 minutes. The session is mediated throughout by a trained facilitator, and facilitator–participant interaction patterns are systematically documented for descriptive analysis. The session comprises warmup conversation, a read-aloud passage, facial expression tasks, open-ended emotion elicitation, STAI-6, SGDS-K, SARC-F, and selected SarQoL-K, with continuous audiovisual recording. Verbal and nonverbal features, grounded in behavioral and psychosocial theory, are analyzed using interpretable machine-learning models and exploratory multimodal fusion to examine their potential for human-centered AI assessment of muscle health. For group comparison, appendicular skeletal muscle mass index (ASMI) by bioelectrical impedance analysis and handgrip strength (HGS) by dynamometry are used to classify participants according to AWGS 2019 thresholds.

Discussion

This protocol serves a dual purpose. First, it evaluates whether multimodal behavioral phenotypes are comparable to objective muscle-strength and body-composition measures. It also examines whether these phenotypes provide information complementary to self-report screening and may inform the development of non-invasive approaches to community muscle-health assessment. Second, the facilitator-mediated design documents interaction patterns that may inform future ECAs for older adults. The framework provides a structured approach to multimodal data collection across varying levels of digital familiarity.

Trial registration:

Not applicable (observational study).

Background

Sarcopenia, the age-related progressive loss of skeletal muscle mass, strength, and function, is a major geriatric syndrome associated with physical disability, increased fall risk, and elevated mortality [1, 2]. The Asian Working Group for Sarcopenia (AWGS) 2019 [3] and the European Working Group on Sarcopenia in Older People (EWGSOP2) [4] have established standardized diagnostic criteria; however, implementation in routine clinical settings remains limited by the cost and logistical requirements of current diagnostic modalities. The recent AWGS 2025 consensus update further reframes sarcopenia management through a life-course approach to muscle health promotion, expanding diagnosis to middle-aged adults and simplifying the diagnostic algorithm [5]. These evolving standards underscore the growing need for scalable, accessible screening tools.

Despite a growing global prevalence estimated at 10–27% among community-dwelling older adults [2, 6], sarcopenia remains underdiagnosed. Objective assessment of muscle mass commonly relies on modalities such as dual-energy X-ray absorptiometry (DXA) and bioelectrical impedance analysis (BIA), with computed tomography (CT) providing additional body-composition information in selected settings; these approaches require equipment and dedicated assessment workflows that limit their use as ubiquitous community self-screening tools [7, 8]. Simple screening tools such as SARC-F offer accessibility but demonstrate limited sensitivity (29–55%), often failing to identify individuals in the early stages of muscle decline [9]. This diagnostic gap is particularly relevant in rapidly aging societies such as South Korea, where the proportion of adults aged 65 years and older is projected to reach 46.4% by 2070 [10].

A fundamental limitation of self-report screening tools is their dependence on the respondent's subjective awareness and willingness to report functional decline [11, 12]. Older adults in the early stages of sarcopenia may not yet perceive their muscle loss as problematic, or may underreport difficulties due to gradual adaptation to declining function, social desirability, or cultural norms. Behavioral phenotyping offers a complementary approach: facial muscle dynamics, vocal characteristics, linguistic content, and upper-body kinematics may capture aspects of behavior that are not fully represented in conscious self-report and may be associated with underlying neuromuscular status. This use of 'behavior' is conceptually distinct from the movement- and performance-based indicators commonly emphasized in sarcopenia research, such as gait speed, chair-rise time, or physical performance batteries, which quantify functional capacity during structured motor tasks. The present protocol instead focuses on spontaneous expressive behavior that emerges during naturalistic face-to-face interaction rather than during an explicit physical task. If validated, such markers could provide information complementary to existing screening pathways by capturing muscle-health-related variation that may not be reflected in self-report instruments, without replacing established diagnostic criteria.

Beyond initial screening, validated non-invasive behavioral markers may also have potential longitudinal applications, such as monitoring change following exercise or nutritional interventions and supporting periodic health assessment. Repeated objective assessment using equipment-based methods may be impractical in some community or longitudinal monitoring settings, motivating interest in complementary low-burden approaches. The behavioral phenotyping approach proposed here is designed with such scalable, repeated-use applications in mind, although the present protocol focuses on cross-sectional discriminant validity as a necessary first step.

To address these limitations, digital phenotyping has emerged as a promising paradigm, leveraging artificial intelligence (AI) to quantify human behavior as a proxy for health status [13]. In the visual domain, age-related differences in facial emotion recognition are well established, and the magnitude of these effects varies with task and stimulus characteristics [14]. Facial-expression research in older adults also highlights age-related modeling challenges [15], while computer-analyzed facial data demonstrate altered muscle recruitment and Action Unit (AU) intensity with age [16]. Recent work on mild cognitive impairment (MCI) detection has demonstrated that spatial-to-temporal attention models applied to facial video can capture clinically meaningful behavioral phenotypes [17].

In the acoustic domain, vocal biomarkers have shown promise for detecting frailty and related conditions. Lin et al. demonstrated that specific vocal parameters discriminate between frailty phenotypes [18], while subsequent work extended these findings using deep learning-based voice embeddings [19]. Rosen-Lang et al. achieved classification accuracy above 80% using voice features for frailty categorization [20], and Kim et al. validated machine learning-based vocal biomarkers for community-based frailty identification [21]. A recent scoping review concluded that voice biomarkers represent a feasible and emerging tool for frailty assessment [22], supporting the foundation for their application to sarcopenia.

In the kinematic domain, sensor-based upper-extremity function tests have demonstrated that arm movement parameters such as speed, range of motion, and smoothness discriminate between frailty phenotypes in older adults [23]. However, these studies employed structured motor tasks rather than naturalistic seated interaction. Whether spontaneous upper-body movements during conversational assessment, such as postural shifts, gesture frequency, and reach patterns, carry discriminative information for sarcopenia remains unexplored.

These converging lines of evidence suggest that multimodal behavioral data, encompassing facial expressions, speech characteristics, linguistic content, and body kinematics, may serve as informative digital biomarkers. Indeed, Takeshige-Amano et al. demonstrated that combined facial and vocal features extracted during chatbot interactions can differentiate Alzheimer's disease patients from controls [24], illustrating the potential of multimodal approaches. Evidence from aging, cognitive impairment, and frailty therefore provides biological and methodological plausibility rather than direct evidence for sarcopenia. Evidence directly examining whether multimodal behavioral phenotypes

distinguish older adults with normal muscle mass and preserved strength from those with low muscle mass, or vary with handgrip strength, remains limited. This evidence gap motivates the present protocol.

Even when behavioral markers are shown to exist under controlled laboratory conditions, it remains uncertain whether they can be reliably captured from community-dwelling older adults during real-world health assessments. Older adults may have difficulty following explicit expression-posing instructions [25], and closed-ended questionnaire formats inherently constrain the temporal and expressive space for naturalistic behavioral expression [26]. These challenges motivate a multi-method elicitation strategy combining structured expression tasks, naturalistic open-ended narrative, and facilitator-mediated clinical instruments. The design rationale for each component is detailed in the Methods section.

Despite the technological potential, significant methodological barriers remain. Older adults invited to an unfamiliar laboratory setting for a structured assessment may experience situational stress that extends well beyond technostress, the anxiety induced by interacting with complex technology, [27, 28]. More broadly, systematic-review evidence indicates that older adults' adoption of health technologies is shaped by interacting health-related, technology-related, dispositional, socioeconomic, and social factors [29]. Unfamiliarity with the environment, the presence of recording equipment, and the perceived evaluative nature of the session can all alter the behavioral markers this protocol aims to measure. An AI model trained on data affected by such situational artifacts may learn stress-related rather than muscle-health-related behavioral phenotypes. This motivates the Digital Facilitator Framework described in the Methods section: a trained human facilitator accompanies the participant throughout the session to help reduce situational and technological stress and support more naturalistic behavioral expression. Importantly, the facilitator's interaction patterns, including when and how assistance is provided, are systematically documented to generate empirical observations that may inform design requirements for future autonomous screening interfaces, such as adaptive applications or embodied conversational agents (ECAs), that could replicate the facilitator's supportive role without requiring human presence.

This protocol aims to: (1) establish whether multimodal behavioral markers, including facial expression dynamics, speech and vocal features, verbal content, and upper-body kinematics, discriminate between older adults with normal muscle mass and preserved strength and those with low muscle mass, with secondary characterization by handgrip strength; (2) develop and validate a facilitator-mediated single-session data collection framework optimized for digital phenotyping in older adult populations, while systematically documenting facilitator-participant interaction patterns to inform the future design of autonomous screening interfaces such as ECAs; and (3) develop and internally evaluate a non-invasive sarcopenia-related risk probability score as an exploratory step toward future muscle-health screening.

We hypothesize that the low-muscle-mass group, relative to the reference group, will exhibit distinct behavioral profiles: (H1) altered facial expression patterns, including diminished expression diversity, lower Action Unit intensity, or differential AU activation profiles; (H2) reduced vocal richness, slower speech rate, and more monotone prosody; and (H3) reduced lexical diversity and shorter utterance

length. Additionally, exploratory analyses will examine whether upper-body kinematic features (gesture frequency, range of motion, postural dynamics) differ between groups and vary with HGS.

Methods

Study design and setting

This study is designed as an observational behavioral phenotyping protocol conducted within an ongoing interdisciplinary sarcopenia research program at Yeungnam University, Republic of Korea (IRB No.: 7002016-A-2026-178; task number: 2025-R-0606-005). The broader program includes age-stratified assessment of physical function and nutritional intake; however, eligibility for and participation in the behavioral phenotyping study do not depend on participation in any parent or companion study. Community-dwelling adults aged 65 years and older may be recruited directly for the present study. Muscle-health classification for this protocol uses bioelectrical impedance analysis (BIA; InBody380, InBody Co., Ltd., Cheonan, Republic of Korea) to derive appendicular skeletal muscle mass index (ASMI) and handgrip dynamometry (HGS; IB-HGS, InBody Co., Ltd., Cheonan, Republic of Korea), applying the sex-specific thresholds specified in AWGS 2019. Behavioral phenotyping data collection commenced in August 2026 and is ongoing within the approved research period, from August 2026 to July 2027 (Fig. 1).

The behavioral phenotyping study is implemented as an operationally independent module: all stages of the data collection session are administered primarily by the principal investigator in a dedicated laboratory (Computer Vision & Human Perception Laboratory) within Yeungnam University. Recruitment and participation are independent of involvement in other studies within the broader sarcopenia research program. All behavioral data collection procedures are conducted under the present approved protocol, including the Digital Facilitator Framework, multimodal behavioral assessment tasks, structured questionnaire administration, continuous audiovisual recording, BIA-derived ASMI assessment, and handgrip strength assessment. This design is intended to allow the behavioral phenotyping protocol to be replicated as a standalone study in future multi-site contexts without modification. The protocol was developed with reference to relevant STROBE reporting principles for observational studies [30]. Ethical approval has been obtained from the Institutional Review Board of Yeungnam University (IRB No.: 7002016-A-2026-178).

Participants

Up to 120 community-dwelling adults aged 65 years and older will be recruited for the behavioral phenotyping study. Participation is independent of enrollment in any other study within the broader research program. Participants will be classified for the primary analysis using ASMI measured by BIA and HGS, with sex-specific thresholds drawn from AWGS 2019 [3]. The study groups are operationally defined using these two objective measures rather than presented as complete AWGS diagnostic categories. The two primary groups are defined as follows:

Group 1: Reference group. Defined as normal appendicular skeletal muscle mass index (ASMI ≥ 7.0 kg/m² for men and ≥ 5.7 kg/m² for women by BIA) together with preserved handgrip strength (HGS ≥ 28 kg for men and ≥ 18 kg for women).

Group 2: Low-muscle-mass group. Defined as low appendicular skeletal muscle mass index (ASMI < 7.0 kg/m² for men and < 5.7 kg/m² for women by BIA), irrespective of HGS. For secondary analyses, this group will be stratified into participants with preserved HGS and those with low HGS (< 28 kg for men and < 18 kg for women). Participants with both low ASMI and low HGS meet the AWGS 2019 muscle-mass and muscle-strength criteria for sarcopenia. Participants with normal ASMI but low HGS will not be included in the primary two-group inferential comparison; consistent with the AWGS 2019 concept of possible sarcopenia, they are explicitly retained as an exploratory subgroup and examined in descriptive and continuous-outcome analyses. Given the anticipated low prevalence of this profile among community-dwelling volunteers, a fully powered three-group inferential comparison is not planned. This operational grouping avoids assigning a complete AWGS 2019 diagnostic status without a physical-performance criterion.

Table 1
Operational group stratification using AWGS 2019 ASMI and HGS thresholds

Criterion	Group 1: Reference	Group 2: Low-muscle-mass	Exploratory subgroup: Possible sarcopenia
Muscle Mass (ASMI)	Normal (≥ 7.0 kg/m ² men; ≥ 5.7 kg/m ² women)	Low (< 7.0 kg/m ² men; < 5.7 kg/m ² women)	Normal (≥ 7.0 kg/m ² men; ≥ 5.7 kg/m ² women)
Muscle Strength (HGS)	Preserved (≥ 28 kg men; ≥ 18 kg women)	Preserved or low (< 28 kg men; < 18 kg women); used for secondary stratification	Low (< 28 kg men; < 18 kg women)
Operational classification	Normal ASMI + preserved HGS	Low ASMI, stratified by preserved vs. low HGS	Normal ASMI + low HGS; excluded from primary two-group inferential comparison; retained in descriptive and continuous-outcome analyses
Note: Group assignment is operational and based only on BIA-derived ASMI and HGS. Participants with low ASMI and low HGS meet the AWGS 2019 muscle-mass and muscle-strength criteria for sarcopenia; participants with normal ASMI but low HGS are retained for continuous-outcome analyses but excluded from the primary two-group comparison.			

Rationale for reference vs. low-muscle-mass design

This two-group design positions older adults with normal ASMI and preserved HGS as the reference comparator, enabling the AI pipeline to learn behavioral contrasts associated with low muscle mass while retaining HGS as an objective strength measure. This approach offers three methodological advantages. First, ASMI provides a direct body-composition anchor for the primary contrast, while HGS

permits clinically meaningful secondary stratification. Second, community-based recruitment yields a natural range of muscle mass and strength without requiring prior diagnostic labeling, maximizing recruitment feasibility. Third, the resulting model addresses a practical screening need: identifying individuals whose behavioral profile may warrant formal muscle-health evaluation with established objective measures. Participants with low ASMI and low HGS can be identified as meeting AWGS 2019 muscle-mass and muscle-strength criteria for sarcopenia, whereas other combinations are retained without imposing a full diagnostic label.

Muscle-mass and handgrip thresholds: AWGS 2019 with reference to AWGS 2025

Although the AWGS 2025 consensus update has recently been published [5], the present protocol uses the established AWGS 2019 sex-specific thresholds for BIA-derived ASMI and HGS [3] to maintain consistency with the approved research framework and accumulated data. These thresholds remain clinically interpretable for adults aged 65 years and older. The study therefore describes ASMI/HGS-based groups operationally rather than assigning complete consensus diagnostic categories. Sensitivity analyses may examine classification implications under updated consensus recommendations where applicable.

Age, sex distribution and group composition

Recruitment will remain community-based and will not use diagnostic quotas or clinical enrichment. The principal investigator will monitor age, sex, and muscle-health distributions during recruitment to characterize the achieved sample composition. Because the prevalence of low muscle mass and low HGS may be limited in a community sample, a predominance of participants with relatively preserved muscle health is anticipated as a plausible recruitment outcome rather than treated as a protocol deviation. Formal post-hoc matching is not applied, as it would further reduce the already modest sample size; instead, age and sex will be incorporated as pre-specified covariates in all statistical and machine learning analyses (see Statistical analysis plan).

Eligibility criteria

Inclusion criteria: (1) age ≥ 65 years; (2) ability to communicate verbally and follow basic instructions. Exclusion criteria: (1) severe cognitive impairment that precludes comprehension of study procedures, as determined by the facilitator during the session; (2) observable motor or neurological conditions that severely affect facial expression or speech control, as determined by the facilitator during the session; (3) history of facial surgery or facial paralysis, as reported by the participant; (4) acute infection or hospitalization within two weeks prior to assessment; (5) severe hearing or visual impairment precluding communication. These criteria are designed to exclude conditions that independently alter the

behavioral markers targeted in this protocol, which would confound the interpretation of sarcopenia-specific behavioral signatures.

Health status documentation

During the warmup conversation, participants are asked about their general health status and any ongoing medical conditions. If a participant reports a health concern, the facilitator inquires about current medications and diagnoses, which are noted in the session record. This naturalistic approach to health documentation is consistent with the facilitator-mediated framework and avoids imposing a formal medical intake procedure that could increase assessment-related stress. Common conditions reported in this age group (e.g., hypertension, diabetes) and associated medications are recorded when disclosed. Medication status is reported descriptively and used qualitatively when interpreting results, but it is not entered as a covariate in any model, as systematic pharmacological profiling is beyond the scope of this behavioral phenotyping protocol and polypharmacy is near-universal in this population.

Pre-study protocol optimization

A six-participant pre-study pilot was conducted solely to optimize procedures for the definitive observational study described in this protocol. Key findings included: (a) participants had significant difficulty operating a PC with a mouse, prompting a switch to touchscreen administration, which participants handled more comfortably due to widespread smartphone familiarity; (b) participants' preference for self-administered versus facilitator-mediated interaction was recorded at the beginning of the questionnaire phase, with the majority (5 of 6) preferring facilitator-mediated administration, supporting the Digital Facilitator Framework design; (c) to maximize session efficiency, only the SARC-F (5 items) was administered via touchscreen self-response, with screen interaction recorded for subsequent analysis of digital interaction behavior; and (d) structured closed-ended questionnaire items elicited limited observable behavioral expression, motivating the addition of dedicated behavioral elicitation tasks (facial expression recognition and production, open-ended emotion narrative) as a separate phase preceding questionnaire administration. The two-phase protocol was feasible within a single session of less than 30 minutes.

Sample size justification

Up to 120 community-dwelling older adults aged 65 years and older will be recruited directly for the behavioral phenotyping study, with primary group assignment based on ASMI and HGS as defined above. Recruitment is intentionally community-based rather than clinically enriched; therefore, the observed muscle-health distribution may be skewed toward participants with preserved ASMI and HGS. As group sizes depend on the naturally occurring distribution in the recruited sample, a formal a priori power analysis assuming equal groups is reported only as a reference benchmark. Using G*Power 3.1, a two-

sided independent two-group comparison with Cohen's $d = 0.7$ (medium-to-large effect), $\alpha = 0.05$, and power $(1 - \beta) = 0.80$ requires approximately 34 participants per group.

Because participants are recruited from the community, the number meeting low-muscle-mass criteria cannot be guaranteed in advance, and the achieved group ratio may be substantially unbalanced. A recruitment contingency is therefore pre-specified. If the low-muscle-mass group falls below approximately 34 participants at the planned close of recruitment, the dichotomous between-group comparison will be treated as secondary and analytical emphasis will shift to continuous-outcome models evaluating associations between behavioral features and ASMI and HGS across the full sample. This strategy preserves information across the observed muscle-health continuum and remains informative even if most participants have relatively preserved muscle health. Such a distribution would also characterize the range of behavioral variation encountered in the community population for which low-burden muscle-health assessment is ultimately intended, rather than representing a recruitment failure.

Regarding the machine learning analysis pipeline, our analytical strategy follows a staged approach. Step 1 (Primary analysis): interpretable, classical machine learning models (Elastic Net, Random Forest, SVM) will serve as the primary analytical framework. Step 2 (Exploratory analysis): deep learning fusion architectures (LSTM, attention-based transformers) will be applied through feature-level transfer from established open-source toolkits and pre-trained representations (wav2vec 2.0 for speech embeddings, OpenFace 2.0 for facial AU and landmark features), with regularization (dropout, weight decay, early stopping). Data augmentation strategies (e.g., temporal jittering, noise injection for audio, frame-level perturbation for video) will be applied where appropriate to increase effective training sample diversity. Vabalas et al. [31] demonstrated that nested cross-validation provides unbiased performance estimates even with modest sample sizes under these conditions.

The Digital Facilitator Framework

To support high-quality, ecologically valid data collection from digitally inexperienced older adults, this protocol introduces the Digital Facilitator Framework. A trained researcher (the principal investigator or a trained research assistant following a standardized SOP) accompanies the participant throughout the entire session, serving as both technical support and a social presence that reduces assessment-related stress. The framework addresses three key challenges: stress mitigation, by providing a supportive human presence that reduces both technological and situational stress; ecological validity, by allowing participants to respond naturally so that captured behavioral markers reflect their health status rather than assessment-related anxiety; and standardization, by providing a replicable procedure for facilitator-mediated data collection that can be adopted across research sites.

The facilitator's role varies across the two phases of the protocol. In Phase 1 (Face-to-Face Behavioral Assessment), the facilitator directly guides all tasks: initiating warmup conversation to establish rapport and normalize the recording environment (Stage 1), presenting a read-aloud passage for vocal baseline (Stage 2), administering facial expression recognition and production tasks using FACS-referenced

stimuli (Stage 3), and eliciting open-ended personal narratives in positive and negative valences (Stage 4). Throughout Phase 1, the facilitator provides minimal supportive responses (e.g., acknowledgment, nodding) without directing the content of participant responses, allowing naturalistic behavioral expression to unfold.

In Phase 2 (Facilitator-Mediated Questionnaire Interaction), the facilitator first assesses the participant's familiarity with touchscreen interaction and records their preferred administration mode (self-administered or facilitator-mediated) as a categorical variable. For STAI-6 and SGDS-K (Stages 5 and 6), the facilitator reads items aloud and records responses. For SARC-F (Stage 7), the participant is encouraged to respond independently via touchscreen, with the facilitator providing assistance only when needed; screen interaction is recorded for subsequent analysis of digital interaction behavior. For SarQoL-K selected domains (Stage 8), the facilitator reads each item aloud through face-to-face dialogue, allowing participants to respond freely and elaborate on their experiences.

Importantly, the facilitator's interaction patterns throughout both phases — including when and how assistance is provided, participant response behaviors, and digital literacy observations — are systematically documented for future analysis. This documented interaction data provides empirical design requirements for future autonomous screening interfaces, such as adaptive applications or ECAs, that could replicate the facilitator's supportive role.

Experimental procedure

To maximize recording condition consistency and data quality for AI model training, behavioral data will be collected within a single, controlled session conducted in the designated laboratory. This approach minimizes inter-environment variability (lighting, acoustics, camera positioning) that would arise from multi-team, multi-location recording, and is intended to ensure that all participants are assessed under comparable technical conditions. Phase 1 guides participants through a deliberate progression from conversational ease to affective disclosure, while Phase 2 transitions to structured questionnaire administration.

Structured interaction protocol and burden management

The session will be administered by the principal investigator and a trained research assistant following the standardized SOP. Recording equipment is configured prior to the session and runs continuously throughout all stages. The session follows a two-phase sequence, each targeting a distinct modality of data collection and a specific cognitive-affective state.

The session comprises two phases: Phase 1 (face-to-face behavioral assessment session) and Phase 2 (facilitator-mediated questionnaire interaction session), for a total session duration of less than 30 minutes. Phase 1 is positioned first to capture behavioral data while the participant is cognitively fresh and unaffected by assessment fatigue. Breaks are offered between phases or between stages at the facilitator's discretion.

Phase 1: face-to-face behavioral assessment

The session begins with face-to-face behavioral data collection while the participant is cognitively fresh and unaffected by assessment fatigue.

Stage 1: Warmup conversation (3 items). The facilitator initiates a brief naturalistic conversation with the participant. Three open-ended conversational prompts address daily-life topics covering weather, directions traveled to the assessment site, and general wellbeing. This stage will serve to establish rapport, normalize the recording environment, and capture a conversational behavioral baseline. It can enable intra-session comparison of facial and vocal expression dynamics, linguistic content, and gesture patterns.

Stage 2: Vocal baseline (1 passage). The participant reads aloud a phonetically balanced Korean passage ('Gaeul' [Autumn]), a standardized speech sample widely used in Korean voice and speech research [32], for extraction of fundamental frequency, jitter, shimmer, speech rate, and pause patterns.

Stage 3: Facial expression recognition and production (7 expressions). Prototypical facial expression stimuli depicting six basic emotions (happiness, sadness, anger, surprise, disgust, fear) and a neutral face were selected from the Facial Action Coding System (FACS) manual [33] by a FACS-certified investigator, with one male and one female exemplar per expression. For each stimulus, the participant first identifies the emotion displayed (recognition phase), then reproduces the target expression while viewing the reference image, maintaining it for approximately 5–10 seconds, as tolerated (production phase). This dual-phase design yields an emotion recognition accuracy measure and a participant-specific AU activation range for subsequent within-subject normalization of spontaneous expressions.

Stage 4: Open narrative (2 questions, ≤ 3 minutes each). Two open-ended emotion-elicitation questions are administered: the participant is first asked to describe a recent positive experience, then a recent negative experience. Unlike the preceding structured tasks, these questions elicit self-paced naturalistic affective expression in two contrasting valences within a single participant, providing a within-subject contrast for facial action unit dynamics and vocal affect features under conditions of minimal response constraint.

Phase 2: facilitator-mediated questionnaires interaction

Following behavioral data collection, the session transitions to structured questionnaire administration. The facilitator first assesses the participant's familiarity with touchscreen interaction and records their preferred administration mode (self-administered or facilitator-mediated) as a categorical variable.

Stage 5: STAI-6 (6 items). Participants complete the six-item short form of the Spielberger State Anxiety Inventory (STAI-6) [34]. The STAI-6 comprises three anxiety-present and three anxiety-absent items from the original 20-item State scale, with reported correlations exceeding $r = .90$ with the full form.

Administering the STAI-6 at this early stage captures baseline state anxiety before any sarcopenia-specific questionnaires, serving as a pre-elicitation covariate and enabling statistical separation of anxiety-driven behavioral variance from sarcopenia-related variance. The STAI-6 items are a direct

subset of the validated Korean STAI; the absence of an independently validated Korean short form is acknowledged in the Discussion.

Stage 6: SGDS-K (15 items). The Short-Form Geriatric Depression Scale, Korean version (SGDS-K) [35, 36] is administered by the facilitator, who reads each item aloud and records the participant's verbal response on screen. The SGDS-K score serves a dual purpose: it provides a validated measure of depressive symptomatology for use as a key statistical covariate, and the affective self-disclosure during its administration constitutes independently analyzable behavioral data.

Stage 7: SARC-F (5 items). The SARC-F questionnaire [37] is presented on a touchscreen, and the participant is encouraged to respond independently. The facilitator provides assistance only if requested. This stage serves a dual purpose: sarcopenia risk screening and evaluation of unassisted screen-based interaction feasibility in this population.

Stage 8: SarQoL-K selected domains (18 items). To balance assessment burden with behavioral signal richness, this protocol administers the validated 14-item SF-SarQoL [38] supplemented with four items from Domain 7 (Fears), selected for their affective elicitation potential. The SF-SarQoL, a validated short form of the Sarcopenia Quality of Life questionnaire [39], comprises items from six domains (Physical and Mental Health, Locomotion, Body Composition, Functionality, Activities of Daily Living, and Leisure Activities) shown to retain the discriminative properties of the full 55-item instrument [38]. The four supplementary Fears items address anxiety about pain, functional loss, post-activity fatigue, and fear of falling; although this domain demonstrates lower QoL score differences between groups, these items are expected to elicit strong affective responses during face-to-face administration. The facilitator reads each item aloud and allows participants to respond freely. When participants elaborate on their experiences, the facilitator provides minimal supportive responses (e.g., acknowledgment, nodding) without directing the conversation, allowing naturalistic affective expression to unfold. The recall of locomotion difficulties, fear of falling, and loss of independence is expected to elicit naturalistic affective responses, including frustration, resignation, and anxiety, that manifest as analyzable facial expressions, vocal affect features, and gesture patterns.

Data acquisition system

To maximize accessibility and minimize participant burden, no wearable sensors are employed. The non-invasive recording setup consists of video, audio, and screen capture channels operating continuously throughout both phases.

Video capture: A primary camera positioned at a standardized distance and angle captures the participant's face and upper body for facial AU and kinematic analysis. A secondary camera captures the facilitator-participant interaction throughout the session to document interaction patterns. Screen activity is recorded simultaneously to enable event-locked alignment of questionnaire items with behavioral responses.

Audio capture: A dedicated microphone captures participant speech, supplemented by camera-embedded audio as a secondary source. Microphone functionality and ambient noise levels are checked prior to each session. Post-hoc audio quality assessment is conducted before acoustic feature extraction, and sessions with insufficient audio quality are flagged.

Environment: A dedicated assessment room with controlled lighting to prevent facial shadows and minimized background noise. Equipment positioning (camera angles, lighting conditions, microphone placement) follows a standardized technical protocol intended to provide consistent recording conditions across participants; where unavoidable technical constraints require equipment substitution or adjustment during the study period, any resulting deviations will be documented and considered in downstream data quality assessment.

Due to the extended data collection period, minor equipment substitutions (e.g., camera or microphone replacement) may be necessary. All equipment changes are logged in the session record, and any potential impact on data comparability is addressed during preprocessing and reported in the results.

AI analysis pipeline (Fig. 2)

Facial analysis (OpenFace 2.0 and MediaPipe)

Google MediaPipe Face Mesh will extract 468 3D facial landmarks for robust tracking of head pose (pitch, yaw, roll). OpenFace 2.0 [40] will quantify AU presence (binary occurrence) and intensity (0–5 scale) based on the Facial Action Coding System (FACS) [33, 41]. Key AUs of interest include AU4 (Brow Lowerer), AU6 (Cheek Raiser), AU12 (Lip Corner Puller), AU15 (Lip Corner Depressor), and AU20 (Lip Stretcher). Temporal dynamics of AU activations, including transition frequency, co-occurrence patterns, and asymmetry indices, will be computed to capture the complexity and diversity of emotional expression.

Speech and audio analysis (wav2vec 2.0 and OpenSMILE)

wav2vec 2.0 [42], a self-supervised speech representation learning framework, will extract deep contextualized speech embeddings. A Korean-adapted model fine-tuned on Korean speech corpora will be employed where available; otherwise, language-agnostic acoustic features will be extracted from the pre-trained model's intermediate layers. OpenSMILE [43] will extract interpretable acoustic features including fundamental frequency (F0), jitter, shimmer, speech rate, and Harmonics-to-Noise Ratio (HNR). Special attention will be given to pause frequency, filled pauses, and response latency as potential markers of cognitive-motor processing efficiency.

Linguistic analysis

Linguistic analysis, including lexical diversity and utterance length, will be conducted as a primary analysis to assess whether verbal content features complement acoustic and visual markers.

Kinematic analysis (exploratory, MediaPipe Pose)

Pose estimation algorithms will detect 2D/3D keypoints of the upper body. Motion features including velocity, acceleration, and range of motion (ROM) of hand gestures will be computed. Reduced gesture frequency and fluidity will be explored as potential markers of upper-limb strength decline associated with sarcopenia [44].

Multi-stream feature fusion and prediction

The analysis pipeline employs a multi-stream feature fusion approach following the staged analytical strategy described above. In Step 1, interpretable regularized models (Elastic Net, Random Forest, SVM) will generate baseline performance metrics and identify the most discriminative individual features and modality combinations. In Step 2, deep learning fusion architectures (LSTM-based temporal fusion, attention-based multimodal transformers) will be applied via transfer learning with regularization (dropout, weight decay, early stopping). The final output is a continuous sarcopenia-related risk probability score (0–1), reflecting the probability of membership in the low-muscle-mass group under the primary operational definition. To translate this score into a screening threshold, optimal classification thresholds will be determined using Youden’s Index from the receiver operating characteristic (ROC) curve [45].

Given the sample size (up to $N = 120$), nested stratified Group k -fold cross-validation will be employed to ensure subject independence between training and test sets and prevent data leakage. Performance metrics (AUC-ROC, F1-score, sensitivity, specificity) will be audited across sex (male vs. female) and age subgroups (65–74 vs. ≥ 75 years) to detect and mitigate potential algorithmic bias [46]. Age and sex will be included as covariates in all ANCOVA models and as input features in all machine learning models, irrespective of group balance at recruitment, to ensure that between-group behavioral differences are not confounded by demographic variation. External validation will be planned as a future multi-center extension.

Table 2
Multimodal behavioral features and hypothesized muscle-health markers

Modality	Tools	Key Features	Hypothesized Muscle-Health Markers
Facial	OpenFace 2.0, MediaPipe Face Mesh	AU presence and intensity (0–5), head pose, expression diversity, temporal dynamics	Reduced AU intensity, differential AU activation profile, diminished expression diversity, slower transitions
Speech / Vocal	wav2vec 2.0, OpenSMILE	F0, jitter, shimmer, speech rate, HNR, pause patterns	Slower speech rate, monotone prosody, increased pause frequency
Linguistic	NLP-based analysis	Lexical diversity, utterance length	Reduced lexical diversity, shorter utterances
Kinematic (exploratory)	MediaPipe Pose	Gesture frequency, velocity, acceleration, ROM	Reduced gesture frequency, smaller ROM, slower velocity
Fusion	Step 1: Elastic Net, RF, SVM; Step 2: LSTM, Attention	Sarcopenia-related Risk Probability Score (0–1)	Exploratory classification threshold via Youden’s Index
Note: Kinematic stream is designated as an exploratory analysis. AU, Action Unit; F0, fundamental frequency; HNR, Harmonics-to-Noise Ratio; ROM, Range of motion; RF, Random Forest; SVM, Support Vector Machine; LSTM, Long Short-Term Memory.			

Statistical analysis plan

Between-group comparisons (reference group vs. low-muscle-mass group) of demographic and clinical characteristics will be performed using independent-samples t-tests or Mann–Whitney U tests (continuous variables) and chi-square tests (categorical variables). Multimodal behavioral features will be compared between groups using analysis of covariance (ANCOVA), with age and sex as the only pre-specified covariates in the primary model. Body mass index (BMI), self-reported medication use, SGDS-K score and STAI-6 score are reported as baseline characteristics and compared between groups descriptively, but are not entered into the primary model. BMI is deliberately excluded because ASMI, which defines the primary group assignment, is itself a function of body composition and height, so BMI is strongly related to the exposure; adjusting for it could remove variance attributable to group membership rather than control confounding. Medication use is likewise not entered into any model. It is ascertained only through voluntary disclosure during the warmup conversation, without systematic pharmacological profiling, and polypharmacy is common among community-dwelling older adults, leaving limited interpretability for adjustment from incomplete data. Depressive symptoms and anxiety are instead examined in pre-specified sensitivity analyses, because the behavioral markers targeted in this protocol (facial expression dynamics, vocal characteristics, linguistic content, and gesture patterns) are known to be influenced by affective states [47, 48]. In the sensitivity analyses, SGDS-K and STAI-6 scores are added to the primary model to test whether between-group differences persist after

adjustment. This staged approach avoids over-adjustment in the primary analysis while allowing us to examine whether observed differences are explained by comorbid depression or anxiety. In parallel with the group-based comparison, behavioral features will also be evaluated in relation to ASMI and HGS as continuous muscle-health outcomes. These models retain within-group variation lost to dichotomization and remain informative when one operational group is small.

For machine learning analyses, primary models use behavioral markers with age and sex only; SGDS-K and STAI-6 scores are added in secondary models to quantify their incremental contribution. Feature importance analyses will quantify the relative contribution of affect-related and muscle-health-related features to model predictions. Subgroup analyses stratified by SGDS-K severity (no depression: 0–4; mild depression: 5–9; moderate-to-severe: ≥ 10) will assess model robustness across affective states. To address potential confounding by age and sex, both of which independently influence facial AU dynamics, vocal characteristics, linguistic content, and upper-body kinematics, age and sex are pre-specified covariates in all ANCOVA models and will be included as features in all machine learning pipelines. Algorithmic bias audits will be stratified by both sex (male/female) and age band (65–74 vs. ≥ 75 years) to detect and mitigate differential model performance across subgroups.

Within the low-muscle-mass group, a pre-specified exploratory analysis will compare participants with preserved versus low HGS to examine whether different behavioral streams show differential sensitivity to muscle-strength impairment. This differential modality sensitivity is one rationale for a multimodal rather than single-modality approach: no single stream is expected to capture the full behavioral spectrum, but their fusion may compensate for the limitations of individual modalities. These subgroup analyses are exploratory and the study may not be adequately powered to confirm within-group differences; they serve to generate hypotheses for subsequent studies.

In addition to the primary muscle-health behavioral analyses, facilitator-participant interaction patterns observed during the Digital Facilitator Framework, including the type and frequency of facilitator assistance, participant response latency to facilitator prompts, and touchscreen versus facilitator-mediated administration preference, are logged and coded descriptively rather than entered into the primary statistical or machine-learning models. Two trained coders will independently review a subset of sessions to establish inter-rater agreement for the assistance-type categories, with discrepancies resolved by consensus. These descriptive summaries are intended to characterize facilitator-participant interaction patterns for future methodological reference, including the design of autonomous screening interfaces, rather than to test interaction patterns as muscle-health biomarkers within the present protocol.

Data management and privacy protection

Given the sensitive nature of facial and voice data, security measures are implemented throughout the data lifecycle. Raw video and audio files are stored on password-protected research devices and encrypted external storage with access restricted to authorized research personnel. For academic dissemination and publication, appropriate de-identification procedures will be applied to protect

participant privacy while preserving the behavioral features necessary for analysis. De-identified feature-level data (AU intensities, acoustic parameters, kinematic coordinates) may be shared upon reasonable request. The data management process complies with the Personal Information Protection Act (PIPA) of the Republic of Korea and adheres to the ethical guidelines of the Declaration of Helsinki.

Use of artificial intelligence

An artificial intelligence language model tool was used to assist with language refinement, reference formatting. All scientific content, study design, data analysis plans, and intellectual contributions, including the conceptualization of the Digital Facilitator Framework, the two-phase session architecture, and the multimodal AI analysis pipeline, are the sole responsibility of the authors.

Discussion

This protocol addresses a critical gap in geriatric AI research by providing a systematic framework for collecting standardized, high-fidelity multimodal behavioral data from community-dwelling older adults for multimodal muscle-health phenotyping and screening. Several methodological contributions merit discussion.

A central contribution is the Digital Facilitator Framework. While previous studies have utilized digital phenotyping for cognitive decline [17, 24] and frailty detection [19–21], these approaches typically assume that participants can independently operate digital devices or wear sensors, assumptions that do not hold for a substantial proportion of older adults. The pre-study protocol optimization confirmed this: most participants preferred or required facilitator-mediated administration over self-administered screen interaction. Despite the modest pilot sample, the facilitator-mediated approach preserves the benefits of standardized, quantifiable assessment while acknowledging the sociotechnical realities of geriatric populations. The AWGS 2025 expansion of screening to middle-aged adults [5] further increases the demand for scalable, non-invasive screening tools such as the behavioral phenotyping approach proposed here.

The two-phase session architecture represents a deliberate methodological choice informed by pilot observations. Structured closed-ended questionnaire items elicited limited observable behavioral expression, motivating the separation of dedicated behavioral elicitation tasks (facial expression recognition and production, open-ended emotion narrative) into Phase 1, positioned before questionnaire administration while participants are cognitively fresh. Phase 2 then administers standardized questionnaires through facilitator-mediated interaction, with SARC-F administered via touchscreen to evaluate unassisted screen interaction feasibility. This design enables comparison of behavioral signal richness across distinct elicitation contexts, including structured expression tasks, naturalistic narrative, and questionnaire-based interaction, within the same participant.

A further methodological contribution is the controlled single-session design. Unlike multi-team, multi-environment recording paradigms that introduce inter-session variability, this approach ensures that

camera positioning, lighting, acoustic environment, and facilitator interaction style remain consistent across all participants. This consistency is an important prerequisite for training reliable AI models: behavioral signal differences between groups must reflect genuine muscle-health-related phenotypic variation rather than environmental artifacts.

Compared to recent behavioral marker-based approaches using ecological momentary assessment and actigraphy [49], this protocol employs contactless multimodal AI to extract behavioral biomarkers without wearable sensors or additional participant burden. Beyond motor function, converging evidence links sarcopenia to psychological and affective changes, including depressive symptoms and related affective concerns [47, 48]. By capturing facial expression dynamics, vocal affect, and movement quality simultaneously, while statistically controlling for depressive symptoms and anxiety, this protocol moves beyond traditional motor-only assessments to characterize affective correlates of muscle health and sarcopenia.

The facilitator-mediated design also generates a secondary but important data stream: systematic documentation of interaction patterns, including when and how assistance is provided, participant response behaviors, and digital literacy barriers. These documented patterns provide empirical design requirements for future autonomous screening interfaces, such as adaptive applications or ECAs, tailored to older adult populations.

Current computer vision and affective computing datasets are predominantly skewed towards younger, Western populations [15, 46]. The geriatric-specific behavioral dataset generated through this protocol may help address this representational gap by providing data from an underrepresented demographic. Each session yields time-synchronized video and audio across four standardized elicitation contexts, together with four validated instruments and AWGS 2019-referenced ASMI and HGS indices, for every participant. This dense within-participant phenotyping may support exploratory secondary analyses beyond the primary aim despite the modest cohort size, including studies of affective expression in later life, voice and speech ageing, human-facilitator interaction behavior, and digital literacy in questionnaire administration. The dataset is therefore intended to support both the primary analysis and hypothesis-generating secondary work.

Limitations

Several methodological limitations should be noted. A single facilitator will administer all sessions. While this ensures procedural consistency and benefits from the facilitator's expertise in verbal and nonverbal communication research, it limits generalizability to other facilitators; the SOP is designed to enable future multi-facilitator replication. Additionally, participants' awareness of video recording may theoretically influence behavioral expression, though pilot observations indicated that older adult participants showed minimal awareness of recording equipment and generally engaged with the session as an enjoyable interpersonal interaction. The warmup stage is additionally designed to normalize the recording environment.

The modest sample size (up to $N = 120$) is intended to support the primary between-group comparison and interpretable machine learning models, though deep learning fusion architectures are applied as exploratory analyses using transfer learning with regularization [31]. As a single-center study, generalizability to other ethnic and cultural populations requires future multi-center validation. Because recruitment is community-based and not clinically enriched, the final sample may contain substantially more participants with preserved muscle health than with low muscle mass or low HGS. This may limit power for categorical group comparisons, but the pre-specified continuous analyses of ASMI and HGS allow the study to characterize behavioral variation across the observed muscle-health spectrum. The operational grouping in this protocol is based exclusively on BIA-derived ASMI and HGS and therefore should be interpreted as a muscle-mass/strength classification rather than a complete clinical sarcopenia diagnosis.

The STAI-6 items are administered as a direct subset of the validated Korean STAI; no independently validated Korean short form exists, and its psychometric properties in this population have not been separately established. STAI-6 scores are therefore used as a covariate rather than as a diagnostic measure of state anxiety.

The 18-item composite comprises the 14 validated SF-SarQoL items plus four supplementary Fears-domain items selected for their affective elicitation potential. Because the SF-SarQoL has not been independently validated in Korean and the four additional items fall outside the validated short form, the composite is employed here as an elicitation instrument rather than as a scored quality-of-life outcome.

Future directions

A two-phase research roadmap governs the long-term development of this behavioral phenotyping program. Phase 1 (the present protocol) characterizes behavioral phenotypes associated with low muscle mass and their relationship to handgrip strength. If Phase 1 establishes feasibility, Phase 2 will examine more clinically specific transitions across muscle-health states using a larger, more homogeneous dataset and the Digital Facilitator Framework refined in the present protocol. Future studies with larger samples may incorporate additional clinical assessment components when complete consensus-based sarcopenia classification is required and may explore whether depressive symptoms and anxiety mediate, rather than merely confound, the relationship between muscle health and behavioral markers. Additionally, the feasibility of remote facilitator-mediated administration, in which a trained facilitator guides the session via video link rather than in-person presence, warrants investigation as a pathway to broader community deployment. The facilitator interaction patterns documented in this protocol provide empirical grounding for designing adaptive screening interfaces and may inform the development of ECAs for autonomous screening contexts.

Abbreviations

AI
Artificial Intelligence

ANCOVA
Analysis of Covariance
ASMI
Appendicular Skeletal Muscle Mass Index
AU
Action Unit
AUC-ROC
Area Under the Receiver Operating Characteristic Curve
AWGS
Asian Working Group for Sarcopenia
BIA
Bioelectrical Impedance Analysis
BMI
Body Mass Index
CT
Computed Tomography
DXA
Dual-energy X-ray Absorptiometry
ECA
Embodied Conversational Agent
EWGSOP2
European Working Group on Sarcopenia in Older People 2
F0
Fundamental Frequency
FACS
Facial Action Coding System
HGS
Handgrip Strength
HNR
Harmonics-to-Noise Ratio
IRB
Institutional Review Board
LSTM
Long Short-Term Memory
MCI
Mild Cognitive Impairment
NLP
Natural Language Processing
PIPA
Personal Information Protection Act

QoL
Quality of Life
ROM
Range of Motion
ROC
Receiver Operating Characteristic
SARC-F
Strength, Assistance in walking, Rise from a chair, Climb stairs, Falls
SarQoL-K
Sarcopenia Quality of Life Questionnaire, Korean Version
SF-SarQoL
Short-Form Sarcopenia Quality of Life Questionnaire
SGDS-K
Short-Form Geriatric Depression Scale, Korean Version
SOP
Standard Operating Procedure
STAI-6
Six-Item Short Form of the Spielberger State Anxiety Inventory
STROBE
Strengthening the Reporting of Observational Studies in Epidemiology
SVM
Support Vector Machine

Declarations

Ethics approval and consent to participate

This study has received approval from the Institutional Review Board of Yeungnam University (IRB No.: 7002016-A-2026-178) under the broader sarcopenia research program (task number: 2025-R-0606-005), including approval for non-invasive audiovisual recording and AI-based behavioral analysis. Participation in the behavioral phenotyping study is independent of participation in other studies within the program. All participants will provide written informed consent prior to participation. The study adheres to the principles of the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (grant no. RS-2021-NR060125). The overarching research project under which this protocol was developed underwent independent external peer review by expert panels convened by the NRF as part of the competitive grant approval process. The specific study protocol described in this manuscript did not undergo separate peer review by the funding body but received formal ethical approval from the Institutional Review Board of Yeungnam University, as detailed in the Ethics approval and consent to participate section. The funding body had no role in the design of the study; the collection, analysis, or interpretation of data; or the writing of the manuscript.

Author Contribution

IO conceived and designed the study, including the multimodal AI analysis pipeline, the Digital Facilitator Framework, and the statistical analysis plan; drafted the manuscript; and is responsible for data collection and analysis. CL, as principal investigator of the research laboratory, secured funding for the study, provided valuable supervision and will give critical intellectual input on the computer vision and deep learning methodology and analysis. Both authors read and approved the final manuscript.

Acknowledgements

The authors wish to thank the research team of the broader sarcopenia research program at Yeungnam University for logistical and scientific support. The authors also thank all participants for their contribution to this research.

Data Availability

No datasets were generated or analyzed for this protocol manuscript. De-identified feature-level data generated during the study will be available from the author on reasonable request. Raw video and audio recordings contain identifiable personal information and cannot be shared publicly under the Personal Information Protection Act (PIPA) of the Republic of Korea; access may be granted through a formal data-sharing agreement with the Computer Vision and Human Perception Laboratory, Yeungnam University, subject to institutional review board approval and the terms of participant consent.

References

1. Cruz-Jentoft AJ, Sayer AA. Sarcopenia. *Lancet* 2019, 393(10191):2636–2646.

2. Petermann-Rocha F, Balntzi V, Gray SR, Lara J, Ho FK, Pell JP, Celis-Morales C. Global prevalence of sarcopenia and severe sarcopenia: a systematic review and meta-analysis. *J Cachexia Sarcopenia Muscle*. 2022;13(1):86–99.
3. Chen LK, Woo J, Assantachai P, Auyeung TW, Chou MY, Iijima K, Jang HC, Kang L, Kim M, Kim S, et al. Asian Working Group for Sarcopenia: 2019 Consensus Update on Sarcopenia Diagnosis and Treatment. *J Am Med Dir Assoc*. 2020;21(3):300–7. e302.
4. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyere O, Cederholm T, Cooper C, Landi F, Rolland Y, Sayer AA, et al. Sarcopenia: revised European consensus on definition and diagnosis. *Age Ageing*. 2019;48(1):16–31.
5. Chen LK, Hsiao FY, Akishita M, Assantachai P, Lee WJ, Lim WS, Muangpaisan W, Kim M, Merchant RA, Peng LN, et al. A focus shift from sarcopenia to muscle health in the Asian Working Group for Sarcopenia 2025 Consensus Update. *Nat Aging*. 2025;5(11):2164–75.
6. Shafiee G, Keshtkar A, Soltani A, Ahadi Z, Larijani B, Heshmat R. Prevalence of sarcopenia in the world: a systematic review and meta- analysis of general population studies. *J Diabetes Metab Disord*. 2017;16:21.
7. Turimov Mustapoevich D, Kim W. Machine Learning Applications in Sarcopenia Detection and Management: A Comprehensive Survey. *Healthc (Basel)*. 2023;11(18):2483.
8. Dent E, Morley JE, Cruz-Jentoft AJ, Arai H, Kritchevsky SB, Guralnik J, Bauer JM, Pahor M, Clark BC, Cesari M, et al. International Clinical Practice Guidelines for Sarcopenia (ICFSR): Screening, Diagnosis and Management. *J Nutr Health Aging*. 2018;22(10):1148–61.
9. Voelker SN, Michalopoulos N, Maier AB, Reijnierse EM. Reliability and Concurrent Validity of the SARC-F and Its Modified Versions: A Systematic Review and Meta-Analysis. *J Am Med Dir Assoc*. 2021;22(9):1864–e187616.
10. Statistics Korea. Population projection for Korea: 2022–2072. In. Daejeon: Statistics Korea; 2023.
11. Kera T, Kawai H, Hirano H, Kojima M, Watanabe Y, Motokawa K, Fujiwara Y, Osuka Y, Kojima N, Kim H. Limitations of SARC-F in the diagnosis of sarcopenia in community-dwelling older adults. *Arch Gerontol Geriatr*. 2020;87:103959.
12. Lim JY, Low NA, Merchant RA. Prevalence of sarcopenia in pre-frail community dwelling older adult and utility of SARC-F, SARC-CalF and calf circumference in case finding. *J frailty sarcopenia falls*. 2020;5(3):53.
13. Torous J, Kiang MV, Lorme J, Onnela JP. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *JMIR Ment Health*. 2016;3(2):e16.
14. Hayes GS, McLennan SN, Henry JD, Phillips LH, Terrett G, Rendell PG, Pelly RM, Labuschagne I. Task characteristics influence facial emotion recognition age-effects: A meta-analytic review. *Psychol Aging*. 2020;35(2):295–315.
15. Gaya-Morey FX, Buades-Rubio JM, Palanque P, Lacuesta R, Manresa-Yee C. Deep learning-based facial expression recognition for the elderly: a systematic review. *arXiv preprint* 2025,

arXiv:2502.02618.

16. Ko H, Kim K, Bae M, Seo MG, Nam G, Park S, Park S, Ihm J, Lee JY. Changes in Computer-Analyzed Facial Expressions with Age. *Sens (Basel)*. 2021;21(14):4858.
17. Alsuhaibani M, Dodge HH, Mahoor MH. Mild cognitive impairment detection from facial video interviews by applying spatial-to-temporal attention module. *Expert Syst Appl*. 2024;252(Pt B):124185.
18. Lin YC, Yan HT, Lin CH, Chang HH. Predicting frailty in older adults using vocal biomarkers: a cross-sectional study. *BMC Geriatr*. 2022;22(1):549.
19. Lin YC, Yan HT, Lin CH, Chang HH. Identifying and Estimating Frailty Phenotypes by Vocal Biomarkers: Cross-Sectional Study. *J Med Internet Res*. 2024;26:e58466.
20. Rosen-Lang Y, Zoubi S, Cialic R, Orenstein T. Using voice biomarkers for frailty classification. *Geroscience*. 2024;46(1):1175–9.
21. Kim T, Choi JY, Ko MJ, Kim KI. Development and Validation of a Machine Learning Method Using Vocal Biomarkers for Identifying Frailty in Community-Dwelling Older Adults: Cross-Sectional Study. *JMIR Med Inf*. 2025;13:e57298.
22. Chen P, Chuang M, Tan EC. The application of voice biomarkers in frailty: a scoping review. *J Am Med Dir Assoc*. 2025;26(2):105447.
23. Toosizadeh N, Mohler J, Najafi B. Assessing Upper Extremity Motion: An Innovative Method to Identify Frailty. *J Am Geriatr Soc*. 2015;63(6):1181–6.
24. Takeshige-Amano H, Oyama G, Ogawa M, Fusegi K, Kambe T, Shiina K, Ueno SI, Okuzumi A, Hatano T, Motoi Y, et al. Digital detection of Alzheimer's disease using smiles and conversations with a chatbot. *Sci Rep*. 2024;14(1):26309.
25. Kim H, Kwak S, Yoo SY, Lee EC, Park S, Ko H, Bae M, Seo M, Nam G, Lee J-Y. Facial expressions track depressive symptoms in old age. *Sensors*. 2023;23(16):7080.
26. Cohn JF, De la Torre F. Automated face analysis for affective computing. *Oxf Handb Affect Comput* 2014:131.
27. Jiang D, Yan L, Mayrand F. Emotion expressions and cognitive impairments in the elderly: review of the contactless detection approach. *Front Digit Health*. 2024;6:1335289.
28. Ragu-Nathan TS, Tarafdar M, Ragu-Nathan BS, Tu Q. The Consequences of Technostress for End Users in Organizations: Conceptual Development and Empirical Validation. *Inform Syst Res*. 2008;19(4):417–33.
29. Bertolazzi A, Quaglia V, Bongelli R. Barriers and facilitators to health technology adoption by older adults with chronic diseases: an integrative systematic review. *BMC Public Health*. 2024;24:506.
30. von Elm E, Altman DG, Egger M, Pocock SJ, Gotsche PC, Vandenbroucke JP, Initiative S. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med*. 2007;4(10):e296.

31. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PLoS ONE*. 2019;14(11):e0224365.
32. Choi SH. Development of Korean standardized sentences on voice quality evaluation for dysphonia. *Audiol Speech Res*. 2018;14(2):128–42.
33. Ekman P, Friesen W, Hager J. Facial Action Coding System. Manual and Investigator's Guide, Research Nexus. 2002.
34. Marteau TM, Bekker H. The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI). *Br J Clin Psychol*. 1992;31(3):301–6.
35. Bae JN, Cho MJ. Development of the Korean version of the Geriatric Depression Scale and its short form among elderly psychiatric patients. *J Psychosom Res*. 2004;57(3):297–305.
36. Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, Leirer VO. Development and validation of a geriatric depression screening scale: a preliminary report. *J Psychiatr Res*. 1982;17(1):37–49.
37. Malmstrom TK, Morley JE. SARC-F: a simple questionnaire to rapidly diagnose sarcopenia. *J Am Med Dir Assoc*. 2013;14(8):531–2.
38. Geerinck A, Beaudart C, Reginster J-Y, Locquet M, Monseur C, Gillain S, Bruyère O. Development and validation of a short version of the Sarcopenia Quality of Life questionnaire: the SF-SarQoL. *Qual Life Res*. 2021;30(8):2349–62.
39. Yoo JI, Ha YC, Kim M, Seo SH, Kim MJ, Lee GY, Seo YM, Sung C, Park KS. Translation and validation of the Korean version of the Sarcopenia Quality of Life (SarQoL-K®) questionnaire and applicability with the SARC-F screening tool. *Qual Life Res*. 2021;30(2):603–11.
40. Baltrušaitis T, Zadeh A, Lim YC, Morency L-P. OpenFace 2.0: facial behavior analysis toolkit. In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition: 2018*. 59–66.
41. Ekman P, Friesen WV. Facial Action Coding System: Investigator's Guide. Palo Alto: Consulting Psychologists; 1978.
42. Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. In: *Advances in Neural Information Processing Systems: 2020*.
43. Eyben F, Wöllmer M, Schuller B. openSMILE: the Munich versatile and fast open-source audio feature extractor. In: *Proceedings of the ACM International Conference on Multimedia: 2010*. 1459–1462.
44. Basile G, Sardella A. From cognitive to motor impairment and from sarcopenia to cognitive impairment: a bidirectional pathway towards frailty and disability. *Aging Clin Exp Res*. 2021;33(2):469–78.
45. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32–5.
46. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: *Proceedings of the Conference on Fairness, Accountability and Transparency: 2018*. 77–91.
47. Gao K, Ma WZ, Huck S, Li BL, Zhang L, Zhu J, Li T, Zhou D. Association Between Sarcopenia and Depressive Symptoms in Chinese Older Adults: Evidence From the China Health and Retirement

Longitudinal Study. Front Med (Lausanne). 2021;8:755705.

48. Kim HJ, Yoo K, Kim M, Won CW, Park YS. Sarcopenia and Depression in Older Adults: 2-Year Findings From the Korean Frailty and Aging Cohort Study. J Am Med Dir Assoc. 2026. 10.1016/j.jamda.2026.106152.

49. Kang B, Ma J, Jeong I, Yoon S, Kim JI, Heo SJ, Oh SS. Behavioral marker-based predictive modeling of functional status for older adults with subjective cognitive decline and mild cognitive impairment: Study protocol. Digit Health. 2024;10:20552076241269555.

Figures

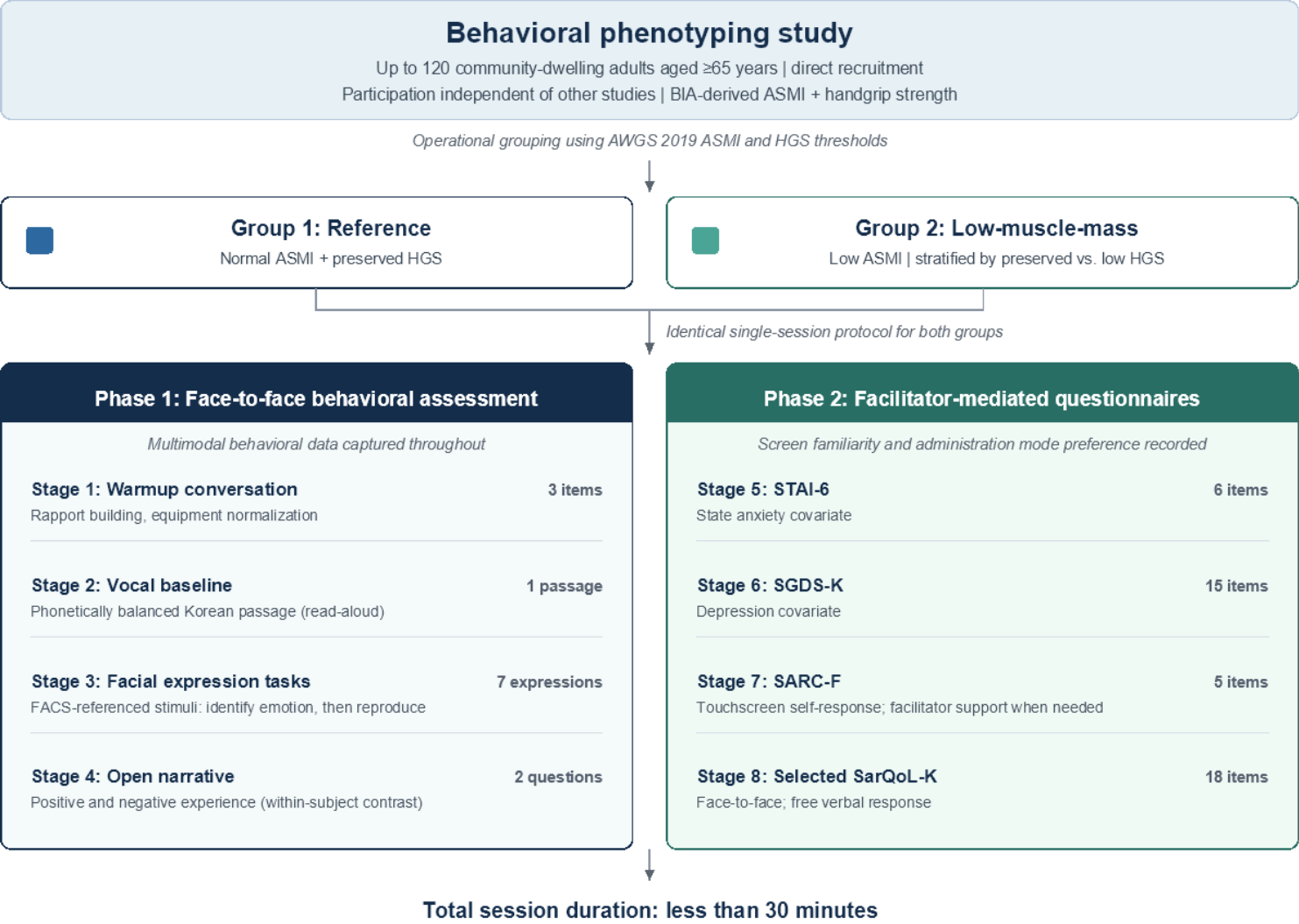


Figure 1

Study flow diagram: participant recruitment, muscle-health group stratification, and the two-phase assessment session. Up to 120 community-dwelling older adults aged ≥65 years are recruited directly for the behavioral phenotyping study; participation does not depend on enrollment in any parent or companion study within the broader sarcopenia research program. Grouping is operational rather than a full AWGS 2019 diagnostic classification and uses two measures: appendicular skeletal muscle mass

index (ASMI) derived from bioelectrical impedance analysis, and handgrip strength (HGS). Group 1 (reference) has normal ASMI with preserved HGS; Group 2 (low-muscle-mass) has low ASMI and is secondarily stratified by preserved versus low HGS. Both groups complete an identical single session of less than 30 minutes comprising two phases. Phase 1 (face-to-face behavioral assessment) elicits multimodal behavioral data through warmup conversation (3 items), a phonetically balanced Korean read-aloud passage, facial expression tasks using FACS-referenced stimuli (7 expressions; identify the emotion, then reproduce it), and open-ended narrative elicitation in two contrasting valences (2 questions). Phase 2 (facilitator-mediated questionnaires) administers the STAI-6 (6 items) and SGDS-K (15 items) as covariates, the SARC-F by touchscreen self-response with facilitator support when needed (5 items), and selected SarQoL-K items face-to-face with free verbal response (18 items); screen familiarity and administration mode preference are recorded. Video and audio are recorded continuously across both phases.

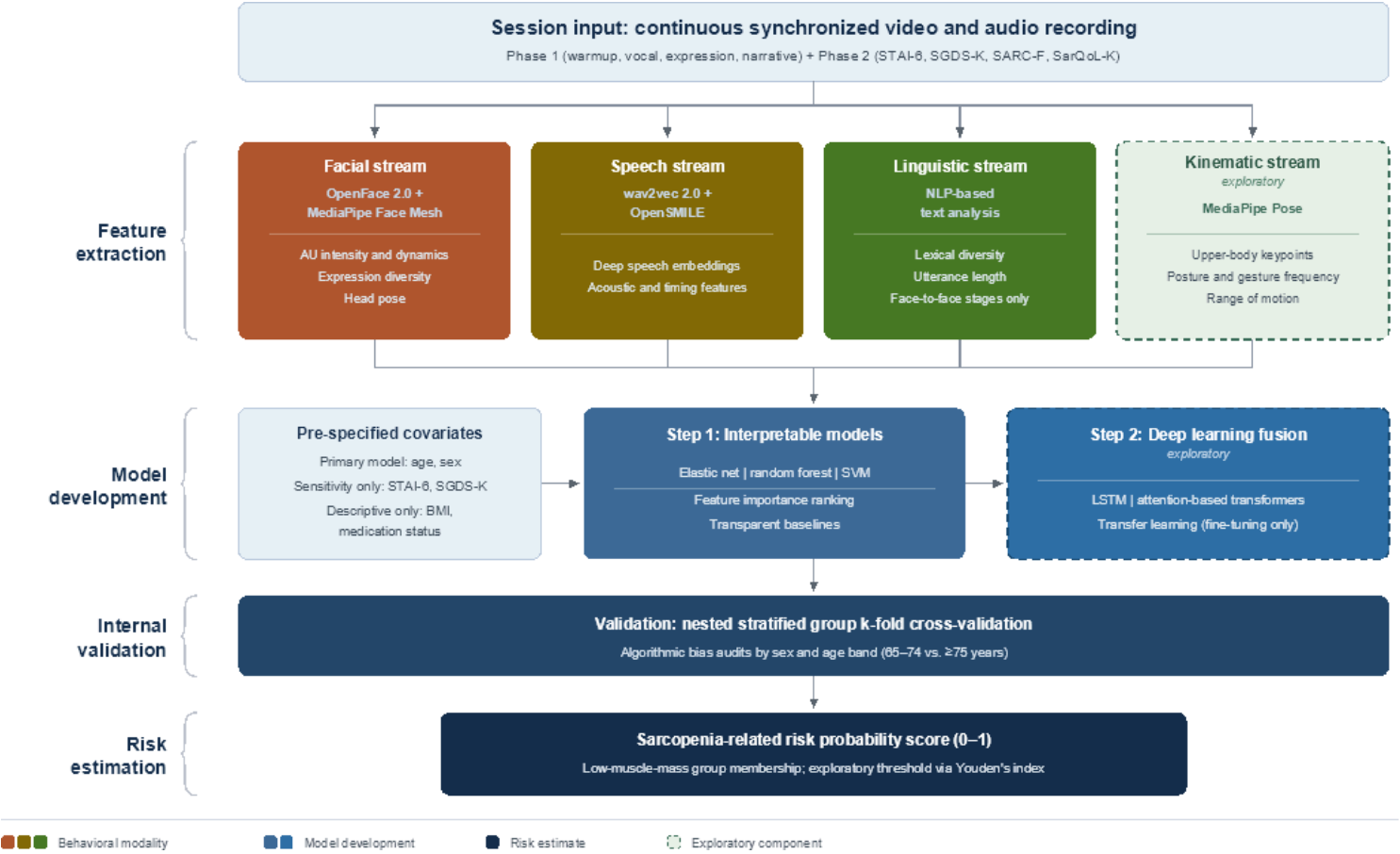


Figure 2

Multimodal analysis pipeline for human-centered AI assessment. Synchronized video and audio from the two-phase session (Phase 1: warmup, vocal, expression and narrative tasks; Phase 2: STAI-6, SGDS-K, SARC-F and SarQoL-K) are processed through four parallel feature-extraction streams. Three primary streams provide the core behavioral features: facial (OpenFace 2.0 and MediaPipe Face Mesh; action unit intensity and dynamics, expression diversity, head pose), speech (wav2vec 2.0 and OpenSMILE; deep speech embeddings, acoustic and timing features), and linguistic (NLP-based text analysis; lexical

diversity and utterance length, face-to-face stages only). One exploratory stream, kinematic (MediaPipe Pose; upper-body keypoints, posture, gesture frequency, range of motion), supports hypothesis-generating analyses. Model development proceeds in two steps with pre-specified covariates: age and sex enter the primary model, STAI-6 and SGDS-K enter sensitivity analyses only, and BMI and medication status are reported descriptively without entering the primary model. Step 1 applies interpretable models (elastic net, random forest, support vector machine) to rank feature importance and establish a transparent performance baseline; Step 2 applies deep learning fusion architectures (LSTM, attention-based transformers) with transfer learning as an exploratory extension. Models are internally validated using nested stratified group k-fold cross-validation with algorithmic bias audits by sex and age band (65–74 versus ≥ 75 years). The output is a continuous sarcopenia-related risk probability score (0–1) for low-muscle-mass group membership, with an exploratory classification threshold determined by Youden's index. Colour encodes function: each behavioral modality has its own colour, blue denotes model development, navy denotes the risk estimate, and dashed borders denote exploratory components.