Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Self-report and objective measures of swallowing capture different constructs in community-dwelling older people: a cross-classification study across two surveys

**Jie Yu**
the Fourth Affiliated Hospital of School of Medicine, International Institutes of Medicine, Zhejiang University

**Siyuan Wu**
the Fourth Affiliated Hospital of School of Medicine, International Institutes of Medicine, Zhejiang University

**Mengxue Fu**
the Fourth Affiliated Hospital of School of Medicine, International Institutes of Medicine, Zhejiang University

**Wenhao Qi**
the Fourth Affiliated Hospital of School of Medicine, International Institutes of Medicine, Zhejiang University

**Lijuan Quan**
the Fourth Affiliated Hospital of School of Medicine, International Institutes of Medicine, Zhejiang University

**Changxin Xu**
the Fourth Affiliated Hospital of School of Medicine, International Institutes of Medicine, Zhejiang University

**Xinran Tian**
the Fourth Affiliated Hospital of School of Medicine, International Institutes of Medicine, Zhejiang University

**Huafang Zhang**
wuyizhangyichi@zju.edu.cn

the Fourth Affiliated Hospital of School of Medicine, International Institutes of Medicine, Zhejiang University

---

**Research Article**

**Keywords:** Aged, Cognition, Deglutition Disorders, Mass Screening, Nursing Assessment, Reproducibility of Results, Surveys and Questionnaires

**Posted Date:** November 19th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11530411/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Self-report and objective measures of swallowing capture different constructs in community-dwelling older people: a cross-classification study across two surveys

Jie Yu, MSN[1], Siyuan Wu, PhD[1], Mengxue Fu, MSN[1], Wenhao Qi, MSN[1], Lijuan Quan, MSN[1], Changxin Xu, MSN[1], Xinran Tian, MSN[1], Huafang Zhang, MSN[1]

**Affiliation:** [1]Department of Nursing, the Fourth Affiliated Hospital of School of Medicine, and International School of Medicine, International Institutes of Medicine, Zhejiang University, Yiwu, China, 322000

**Corresponding author:** Huafang Zhang, Department of Nursing, the Fourth Affiliated Hospital of School of Medicine, and International School of Medicine, International Institutes of Medicine, Zhejiang University, Yiwu, China, 322000. Email: wuyizhangyichi@zju.edu.cn

**Abstract**

**Background:** Self-report questionnaires such as the Eating Assessment Tool are widely recommended for community dysphagia screening, but evidence for them comes largely from clinical populations, and the abbreviated version was validated against its parent questionnaire. Whether they measure the same thing as an observed swallowing task is unresolved. We examined agreement between two self-report instruments and an objective water swallow test, tested whether cognitive impairment accounts for discordance, and assessed whether objective measures carry information the questionnaires do not.

**Methods:** Cross-classification analysis of two consecutive community surveys in eastern China: 2024 (2,886 participants, 121 positive tests) derived a model; 2025 (2,765 participants, 230 positive tests) validated it. The comparator was a positive 30 mL water swallow test, grade II or above. Because the comparator is a bedside observation, not an instrumental reference standard, agreement was reported as positive and negative percent agreement and Cohen's kappa. Agreement was examined by cognitive stratum in both surveys, and discrimination was compared with DeLong tests.

**Results:** Agreement between the instruments was slight in both surveys (kappa 0.022 and 0.101; positive percent agreement 15.7% and 27.0%). Agreement did not fall with cognitive impairment: it was higher in the impaired group in 2025 (kappa 0.280 versus 0.082), driven by a higher prevalence of positive tests (15.2% versus 7.9%) rather than closer classification, but this was not reproduced in 2024 (-0.011 versus 0.025). A five-variable model discriminated the comparator outcome better than the questionnaire (0.661 versus 0.595, P = 0.010), but tongue pressure, calf circumference and cognitive score added nothing to age and grip strength (0.659; P = 0.971). Calibration remained acceptable when prevalence rose from 4.2% to 8.3%.

**Conclusions:** A self-report dysphagia questionnaire and an objective swallow test classify the same older people differently, and cognitive impairment does not explain this. They should be treated as

measuring different constructs rather than as interchangeable screens, and pathways should state which construct they intend to capture.

**Clinical trial number:** Not applicable.

**Keywords:** Aged; Cognition; Deglutition Disorders; Mass Screening; Nursing Assessment; Reproducibility of Results; Surveys and Questionnaires

## Background

Dysphagia is common in later life, but estimates of how common differ more than twofold depending on how it is measured: a meta-analysis of community studies reported a prevalence of 12.1% when dysphagia was defined by a water swallow test and 30.5% when defined by a validated symptom questionnaire [1]. That discrepancy is itself a measurement finding, and it frames the question this paper addresses. Unrecognised dysphagia contributes to malnutrition, dehydration, aspiration pneumonia and hospital admission [2,3], and is now regarded as a geriatric syndrome in its own right [4,5]. Identifying those at risk matters most where formal instrumental assessment by videofluoroscopy or fibreoptic endoscopic evaluation of swallowing is neither available nor feasible at population scale, so a brief screening step is needed to decide who should be assessed in greater depth.

The most widely used brief screen is the Eating Assessment Tool, a ten-item self-report questionnaire recommended in several national guidelines [6] and the subject of a systematic review of diagnostic accuracy [7]. Its appeal is that it takes under two minutes and requires no instrumentation. A two-item abbreviated version has been proposed as a still shorter alternative. A previous survey of 5,090 community-dwelling older people compared the Eating Assessment Tool, the two-item version and a 30 mL water swallow test [8]. The two self-report instruments agreed substantially with each other (kappa 0.64 at the conventional cut-off), but neither agreed with the water swallow test (kappa 0.11 and 0.09). The authors concluded that the questionnaire and the swallow test should be used together rather than interchangeably.

That survey established a discrepancy; it did not establish what the discrepancy means. Two readings are possible, and they lead to opposite recommendations. If the questionnaire is a less precise version of the same measurement, its poor agreement is a sensitivity problem to be solved by a better questionnaire. If it measures a different construct, the experienced difficulty of eating rather than the mechanics of swallowing, then low agreement is not a defect, and substituting one instrument for the other is a category error. Accuracy statistics cannot separate these readings, because accuracy requires a reference standard that this literature does not have: the water swallow test is itself a bedside screening observation, not an instrumental criterion.

Cognitive status is the axis on which the two readings diverge most sharply. The intuitive expectation is that self-report fails in people with cognitive impairment, because they may not recognise, recall or attribute a developing swallowing difficulty. If that were the mechanism, agreement should fall as cognition falls, and the practical response would be to withhold the questionnaire from that group. If instead the two instruments measure different constructs, agreement should be unrelated to cognition, and withholding the questionnaire would remove a source of information without replacing it. Community surveys can distinguish these predictions: they generate large samples, record both cognitive and swallowing measures, and can be repeated in the same population.

We therefore analysed two consecutive community surveys of older people in the same district of eastern China, in 2024 and 2025, with three aims: to quantify agreement between two self-report instruments and an objective water swallow test, and to test whether cognitive impairment accounts for discordance; to examine whether routinely measured objective variables discriminate the comparator outcome better than the questionnaire, and how much of any advantage is attributable to individual variables; and to assess whether a model built in one survey remains calibrated in the next, when prevalence changes. We did not set out to estimate diagnostic accuracy, and we do not claim to.

The comparison is treated here as a problem of construct validity rather than of respondent accuracy. Validity is a property of the inferences drawn from scores, not of the instrument in isolation [9], and rests on evidence assembled for the intended interpretation [10]. Three consequences follow: two questionnaires asking the same question can be expected to agree for reasons unconnected to dysphagia; agreement between self-report and an observed task is a cross-method question that requires evidence independent of both; and evidence for a screening recommendation is specific to the population in which it was gathered.

## Methods

### Design and reporting

This was a cross-sectional analysis of two consecutive community surveys. The 2024 survey provided the derivation cohort and the 2025 survey the temporal validation cohort, so that validation used data collected strictly after the model had been specified. Reporting follows the TRIPOD+AI statement [11]; the completed checklist is submitted as a supplementary file. Within that framework this is a Type 3 study: development with external validation. The study was observational with no health care intervention, so prospective registration was not applicable. Because the study estimates agreement and model performance rather than diagnostic accuracy against an instrumental criterion, no diagnostic accuracy checklist was completed and no sensitivity or specificity is reported; the comparator, and the limits this places on interpretation, are set out under Comparator measurement in the Methods.

## Data sources and ethics

Both surveys were conducted in the same community setting in a city in eastern China, with the same eligibility criteria and protocol, from March to July 2024 and March to July 2025; the second round began eight months after the first ended, holding seasonal variation constant while capturing change over time. Both covered a single, largely urban district and recruited through community health service points; people in institutions and those unable to attend were not included. Educational attainment varied widely and, because the Hasegawa Dementia Scale-Revised score was not adjusted for schooling, some participants classified as impaired on the prespecified threshold may have had limited education rather than cognitive impairment. The two rounds form part of one research team's programme of work on community dysphagia screening, and earlier publications from that programme report other analyses in this population [8,12]. Because each round was de-identified and the two datasets carried no shared participant identifier, the two waves could not be linked at the individual level; the 2025 cohort is therefore treated as a later, independently collected sample from the same source population rather than as a wholly distinct one. Each survey round was approved by the human research ethics committee of the coordinating hospital: the 2024 round under approval K2023186 (issued 12 December 2023) and the 2025 round under approval K2025041 (issued 14 March 2025). All participants provided written informed consent; where cognitive impairment made written consent questionable, a family member also consented on their behalf. The study followed the Declaration of Helsinki.

## Participants

Eligibility was age 65 or older, residence in the study district, ability to understand spoken questions, and ability to complete the swallowing screening procedures. Cognitive status was assessed with the Hasegawa Dementia Scale-Revised [13,14], a 30-point scale on which 20 or below conventionally indicates impairment; this threshold was fixed a priori in the survey design, not derived from these data. Because the study question concerned community screening rather than screening within a cognitively impaired subgroup, the analytic population comprised all participants aged 65 or older with a recorded water swallow test result and complete predictor data. Cognitive status was retained as a candidate predictor and used to define prespecified subgroups. Because this is the axis along which a self-report screen might perform unequally, performance was examined separately in participants with and without impairment, which we treat as the fairness assessment for the model (TRIPOD+AI item 14).

## Comparator measurement

The comparator was a positive water swallow test, grade II or above on a standardised 30 mL test performed by trained community nurses [15,16]. Grade III or above, conventionally regarded as clearly abnormal, occurred in too few 2024 participants to serve as the outcome (four); grade II or above was

therefore used, so that a swallow graded as suspected difficulty was counted positive. This matches the definition used in previous community surveys [17,18]. The test is a bedside screening observation, not an instrumental criterion: it is feasible at community scale where videofluoroscopy is not, but it cannot confirm aspiration or identify silent aspiration. It is therefore treated throughout as a second, different measurement of swallowing, and every agreement statistic is interpreted in that light. Assessors were trained in administration and grading before each survey round using a standardised protocol, and were unaware of participants' questionnaire responses at the time of testing.

The Eating Assessment Tool total score (range 0 to 40) was recorded at the same visit. A score of 2 or above was taken as positive, the conventional threshold being 3. The lower threshold was used because it captures more symptomatic people, and because a more inclusive threshold works against our own argument: disagreement at a cut-off of 2 or above cannot be attributed to a stringent definition.

**Candidate predictors**

Candidate predictors were selected a priori from the literature on sarcopenia, oral motor function and falls rather than by data-driven selection: age, sex, Hasegawa Dementia Scale-Revised score, maximum tongue pressure, maximum grip strength, calf circumference, number of natural teeth, subjective xerostomia, ability to eat hard food, falls in the past year and pneumonia in the past year (univariable discrimination in Supplementary Table S3). Oral-health variables were included because poor oral health and oral dryness have each been associated with dysphagia in community-dwelling older adults [17,18]; an objective oral moisture value was recorded in the 2024 round only and is examined in a post hoc sensitivity analysis in the Results. Maximum tongue pressure was measured with a JMS device (JMS, Hiroshima, Japan) and maximum grip strength with a Jamar dynamometer (Sammons Preston, USA). Smoking and alcohol status were recorded as descriptive variables only.

The same tongue pressure device was used in both survey years. Maximum grip strength was the higher of two handgrip dynamometer readings. Falls in the past year were collapsed from five categories into three (none, one, two or more) and educational attainment from six into three, because the original extreme categories contained fewer than five participants, which makes coefficient estimates unstable. Values outside physiologically plausible ranges (for example a calf circumference below 15 cm) were set to missing rather than winsorised; nine such values were identified in the 2025 survey.

**Sample size**

Sample size followed the framework of Riley and colleagues [19]. With an anticipated outcome proportion of 4.2%, eleven candidate predictors, an anticipated C-statistic of 0.75 and a conservative Nagelkerke R-squared of 0.27, the minimum was 2,790 participants with about 118 events; the

derivation cohort exceeded this (n = 2,886; 121 events). The final model retained five predictors, giving 24.2 events per predictor and comfortably exceeding the conventional minimum of ten.

**Missing data**

Missingness among candidate predictors ranged from 0.1% to 14.0%: the cognitive score was most often missing, in 13.1% of the 2024 and 14.0% of the 2025 sample, and maximum tongue pressure next, in 2.2% and 5.8%; other predictors were missing in under 3.5%. Complete-case analysis was therefore used, retaining 2,886 of 3,402 participants with a recorded water swallow test in 2024 (121 positive tests) and 2,765 of 3,498 in 2025 (230). Multiple imputation by chained equations with 20 imputations was a pre-specified sensitivity analysis and gave materially identical estimates.

Hasegawa Dementia Scale-Revised scores were missing for 541 of 3,531 participants in the 2024 survey and 550 of 3,674 in the 2025 survey, because participants were unable or unwilling to complete the assessment. These participants were excluded. They were older, with weaker grip and tongue pressure, and far more likely to have an abnormal water swallow test: 12.1% versus 4.4% in 2024 and 14.9% versus 8.6% in 2025 (Supplementary Table S8). The analysis population is therefore healthier than the source population, and the prevalence reported here is correspondingly an underestimate; multiple imputation cannot repair this, because missingness is plausibly related to the unobserved cognitive score itself.

**Model development**

The primary model was a logistic regression specified before any model fitting. Although the event count is adequate for more flexible methods, selecting among algorithms on the same data would bias the reported performance, and logistic regression remains the reference standard for clinical prediction models. Five approaches were prespecified, namely logistic regression, lasso regression, random forest, gradient boosting and a support vector machine, each assessed by stratified ten-fold cross-validation in the derivation cohort. Logistic regression and lasso regression performed almost identically (area under the curve 0.730 and 0.726) and both exceeded the three non-linear methods (0.682, 0.669 and 0.607), so the choice of primary model was supported by the comparison rather than assumed (Supplementary Table S2).

In that comparison the cross-validated area under the curve was 0.730 for logistic regression and 0.726 for lasso regression. The two differed by less than the resampling error, so logistic regression was retained without a formal test of difference.

Logistic regression was therefore retained as the primary model, and the comparison is reported in full rather than as a search for the best performer. The final model retained five predictors, selected a priori as objective and routinely collectable measures rather than by data-driven selection: maximum tongue

pressure, maximum grip strength, calf circumference, age and Hasegawa Dementia Scale-Revised score. No variable selection was performed on these data, so the reported performance is not inflated by predictor search.

**Internal and temporal validation**

Apparent performance was corrected for optimism by 1,000 bootstrap resamples of the entire modelling process [20,21]. The fitted model was then applied unchanged to the 2025 cohort, treated as a temporally separate wave of the same population; because the two waves shared no participant identifier, individual-level overlap could not be determined. Discrimination was assessed by the area under the receiver operating characteristic curve with a DeLong variance estimate [22]; calibration was assessed by the calibration-in-the-large, the calibration slope and the Brier score, each estimated on the linear-predictor scale. Logistic recalibration was performed by regressing the observed outcome on the linear predictor in the validation cohort and applying the resulting intercept and slope; because this is a monotone transformation it leaves discrimination unchanged and is reported alongside the un-recalibrated values.

**Agreement and comparison with self-report instruments**

Agreement between instruments was quantified with Cohen's kappa and with positive and negative percent agreement: the proportion of participants positive, or negative, on the comparator who were also positive, or negative, on the questionnaire. Because kappa depends on the prevalence of the outcome [23], observed agreement and both percent agreement values are reported alongside it. Percent agreement rather than sensitivity and specificity is used throughout, because the comparator is not a reference standard: the proportions it yields describe concordance between two measurements, not the probability of correctly classifying disease.

The two-item version comprised items 4 and 8, which define the published version [8] and had the two highest positive response rates here (14.2% and 11.7%). Its agreement with the full questionnaire (positive percent agreement 92.8%, negative 95.9%) was consistent with the 96.3% reported previously. Agreement with the comparator was examined in both surveys and across cognitive strata: the prespecified contrast of a Hasegawa Dementia Scale-Revised score of 20 or below versus 21 or above, and, as exploratory analyses, three and five bands (Supplementary Table S5), with differences between strata assessed by bootstrap resampling with 5,000 replicates. Discrimination of the comparator outcome was compared with DeLong tests for paired curves. Reclassification relative to the questionnaire total score was quantified by the continuous net reclassification improvement and the integrated discrimination improvement [24], with bootstrap confidence intervals from 2,000 resamples.

### Exploratory decision analysis

As a secondary and explicitly exploratory analysis, we computed net benefit across threshold probabilities of 5% to 30% using the water swallow test as the outcome [25]. The analysis inherits the limitation of the comparator: it describes the consequences of acting on a bedside observation, not on confirmed aspiration. Results are given in Supplementary Table S4 and are not used to support any claim of clinical utility.

### Software and reproducibility

All analyses were performed in R version 4.6.1, using pmsampsize for sample size, rms for modelling and validation, mice for multiple imputation and dcurves for decision curves; package versions are given in the supplementary material. The complete analysis is reproducible from the analysis scripts, which carry a fixed random seed and are available from the corresponding author on reasonable request.

Ethical approval and consent procedures are given in the Methods and in the Declarations.

## Results

### Participants

The 2024 survey yielded 3,531 records, of which 3,402 had a recorded water swallow test result and 2,886 had complete predictor data and formed the derivation sample, with 121 events (4.2%); 129 records lacked a test result and 516 lacked predictor data. The 2025 survey yielded 3,674 records, of which 3,498 had a recorded test result and 2,765 had complete predictor data and formed the validation sample, with 230 events (8.3%); 176 lacked a test result and 733 lacked predictor data. Baseline characteristics of all participants with a recorded cognitive score, stratified by cognitive status, are given in Table 1 and in full in Supplementary Table S1. Participants with impairment were older, more often female, and had lower tongue pressure, grip strength and calf circumference, in both surveys.

### Agreement between self-report and objective measurement

All 2,886 participants in the derivation cohort had both measurements. Agreement there was slight (kappa 0.022; positive percent agreement 15.7%, negative percent agreement 88.4%), and in the 2,765 validation participants it was slightly higher (kappa 0.101; positive percent agreement 27.0%, negative percent agreement 86.6%). In the validation cohort, of 230 participants with a positive water swallow test, 62 also screened positive on the questionnaire; of 2,535 with a negative test, 2,196 also screened negative.

Stratification by cognitive status gave a different picture in the two surveys. In 2024 agreement was, if anything, slightly lower among participants with cognitive impairment (kappa -0.011, n = 251, 23 positive tests) than among those with intact cognition (kappa 0.025, n = 2,635, 98 positive tests); the difference was -0.036 (95% confidence interval -0.146 to 0.097). In 2025 it was higher (kappa 0.280, n

= 164, versus 0.082, n = 2,601; difference 0.199, 95% confidence interval 0.013 to 0.379). The excess observed in 2025 was therefore not reproduced. What both surveys showed is the absence of the predicted gradient: agreement did not fall as cognition fell.

In 2025 the higher kappa in the impaired group did not reflect closer classification, because raw observed agreement was slightly lower there (78.7% versus 81.8%; difference -3.2%, 95% confidence interval -9.8% to 3.1%). It arose from the marginal distribution: the impaired group had both a higher prevalence of positive tests (15.2% versus 7.9%) and a higher positive percent agreement (48.0% versus 24.4%) at a similar negative percent agreement, and these move kappa in the same direction. In 2024 only the first of those two components was present: prevalence was higher in the impaired group (9.2% versus 3.7%) but positive percent agreement was not (13.0% versus 16.3%), which is why kappa did not rise. Five cognitive bands in 2025 confirmed that the pattern was not a gradient: kappa was 0.280, 0.134, 0.089, 0.063 and 0.084 for scores of 20 or below, 21 to 22, 23 to 24, 25 to 26 and 27 or above, and the two highest bands did not differ (difference -0.005, 95% confidence interval -0.095 to 0.088).

**Model performance**

The apparent area under the curve in the derivation cohort was 0.751 and the bootstrap-corrected value 0.744. In temporal validation the area under the curve was 0.661 (95% CI 0.620 to 0.702). Calibration was acceptable: calibration-in-the-large 0.064 (95% bootstrap CI -0.54 to 0.66) and slope 0.796 (0.61 to 1.00), and both intervals include their ideal values, so the data are compatible with good calibration and with moderate miscalibration. Logistic recalibration improved the Brier score only modestly, from 0.074 to 0.073.

Discrimination was also examined within cognitive strata in the validation cohort: the model reached an area under the curve of 0.741 (95% CI 0.636 to 0.846) in the 164 participants with cognitive impairment (25 events) and 0.645 (95% CI 0.602 to 0.689) in the 2,601 with intact cognition (205 events); the difference fell short of significance (DeLong P = 0.100) and is treated as exploratory.

The model explained little variance in the comparator outcome: the Brier skill score relative to the null model was 2.7% and Nagelkerke's R-squared 0.007. A low value is expected when an outcome is uncommon and largely determined by factors outside the model, but it sets the scale for what follows: the model orders people usefully without accounting for much of why they differ. Grip strength was the strongest single predictor in the derivation cohort (univariable area under the curve 0.697), followed by age (0.694) and maximum tongue pressure (0.686); the full ranking is given in Supplementary Table S3.

**Discrimination of the comparator outcome**

In the validation cohort (n = 2,765; 230 positive tests), the model discriminated the comparator outcome better than the questionnaire total score (area under the curve 0.661, 95% CI 0.620 to 0.702, versus

0.595; DeLong P = 0.010), the questionnaire at its cut-off (0.568, P < 0.001) and the two-item version (0.571, P < 0.001) (Table 2, Figure 1). It also improved reclassification (continuous net reclassification improvement 0.019, 95% CI 0.009 to 0.030) and discrimination (integrated discrimination improvement 0.006, 0.003 to 0.008) relative to the total score.

That advantage was not evenly distributed across the five variables. In the validation cohort, age alone reached an area under the curve of 0.643 and grip strength 0.620; a two-variable model of age and grip strength reached 0.659, and adding tongue pressure, calf circumference and cognitive score did not improve it (0.659; DeLong P = 0.971). Against age alone the full model was not significantly better (P = 0.206). The incremental information is therefore carried almost entirely by two variables that are already recorded in routine community assessment.

**Exploratory analysis: the objective oral moisture value**

An objective oral moisture value was recorded in 2024 (median 31.4, interquartile range 30.0 to 32.9) but not 2025, so it could be examined only in the derivation cohort. Among the 2,884 participants with complete data it had no useful univariable discrimination (area under the curve 0.542, 95% confidence interval 0.486 to 0.598), and adding it to the five-predictor model improved neither discrimination (0.750 versus 0.753; DeLong P = 0.677) nor optimism-corrected discrimination (0.729 in both), despite a small adjusted coefficient (P = 0.013). It was not retained (Supplementary Table S6).

## Discussion

### Principal findings

Self-report and objective measurement of swallowing disagreed substantially in both surveys (kappa 0.022 and 0.101), and neither survey showed the degradation with cognitive impairment that the intuitive account predicts. The 2025 survey showed an excess in impaired participants rather than a deficit, but 2024 did not reproduce it, and in 2025 that excess reflected a higher prevalence of positive tests rather than closer classification. A five-variable objective model discriminated the comparator outcome better than the questionnaire, but the advantage rested on age and grip strength, and the three remaining variables added nothing measurable. The model nevertheless transported to the later survey with acceptable calibration when prevalence roughly doubled.

### What the disagreement does and does not mean

The most consequential finding concerns how these instruments have been validated. In previous work the two-item version showed 96.3% agreement with the ten-item version, and this was read as evidence that a two-item screen could replace a ten-item one [8]. Two instruments that ask the same question in the same way measure the same construct; agreement between them is close to inevitable and says little about swallowing. Against an observed task, the two-item version discriminated only weakly (area

under the curve 0.571). Oropharyngeal dysphagia is readily concealed within what is taken to be normal ageing [26], which is what one would expect if the score indexes experienced difficulty rather than measured swallowing capacity.

The same caution applies to the ten-item version. Estimates supporting it come largely from clinical populations with overt disease [7], and a recent Chinese consensus recommends it as the preferred initial screen in community health examinations on the strength of that evidence [27]. In our community sample the questionnaire and the swallow test classified the same people differently in about one case in five. That is what one would expect if the two instruments measure different things, and it is not evidence that the questionnaire is defective: a person who reports no difficulty but coughs on a timed swallow has given two pieces of information, not one true and one false.

This is a problem of construct validity rather than of instrument quality [28,9,10]. A two-item version validated against its parent questionnaire, and an accuracy estimate carried over from a clinical population, both draw conclusions that their evidence cannot support: the first cannot show that anything is detected, only that another instrument of the same kind is reproduced, and the second is specific to the population in which it was gathered. Screening programmes inherit inferences of this kind without testing them in the population they serve.

**Cognitive impairment is not the mechanism**

The intuitive explanation for poor agreement is that people with cognitive impairment cannot report symptoms accurately. Our data do not support that explanation, but they do not invert it either. In 2025 positive percent agreement was substantially higher in participants with cognitive impairment (48.0%) than in those with intact cognition (24.4%), at a similar negative percent agreement, and the model's point estimate was also higher there, though that difference fell short of significance (DeLong P = 0.100) and is treated as exploratory. In 2024 the direction was the opposite and the difference negligible: positive percent agreement was 13.0% versus 16.3% and kappa -0.011 versus 0.025, with a confidence interval spanning zero. Both impaired subgroups are small (251 and 164 participants), so neither year can settle the direction. What the two surveys do agree on is that agreement does not fall as cognition falls, which is the observation the recall-and-insight account requires.

A more plausible explanation for the overall pattern is construct divergence. The questionnaire asks about the subjective experience of swallowing, including effort, discomfort and worry; the water swallow test observes whether a timed swallow is completed safely. These are different things and need not agree even in people with intact cognition: a person may complete a test swallow safely yet struggle with a normal meal, or report no difficulty while aspirating silently. Poor agreement should not be read as a failure of the respondent: the two instruments answer different questions, and a screening pathway

must decide which question it is asking. Chinese expert guidance takes the same position, describing the questionnaire as reflecting subjective symptoms rather than objective impairment [29].

Two qualifications keep this from becoming too tidy. The comparator is a bedside observation rather than an imaging criterion, and it too requires holding a cup, controlling a swallow and following instructions, so part of what it records is cognitive as well as pharyngeal. And where pharyngeal sensation is severely reduced, as in advanced bulbar or sensorimotor disease, a person may neither perceive difficulty nor cough during the test: the two methods then fail together, and a pathway relying on a positive response to either will miss that subgroup. Construct divergence explains the general pattern; it does not license relying on self-report where sensation itself is the problem.

Why positive percent agreement was higher in the impaired group in 2025, and not in 2024, can be proposed but not tested here. More advanced impairment may mean more overt difficulty, easier to notice and to report, and a greater likelihood of a positive test, so questionnaire and test move together. In clinically diagnosed Alzheimer's disease, Eating Assessment Tool scores rise with dementia stage and correlate with penetration-aspiration scores [30].

**What the objective measures add, and how little**

Grip strength and age carried almost all of the model's discrimination. That is a smaller claim than a five-variable model implies, and it should be stated plainly: adding tongue pressure, calf circumference and cognitive score to age and grip strength produced no measurable improvement ($P = 0.971$), and against age alone the full model was not significantly better ($P = 0.206$). The practical implication is not that a five-variable model should be deployed, but that most of the objective information worth having is already collected in routine community assessment.

The model predicted the comparator better for a reason that supports rather than contradicts the argument above: it is built from observed physical measurements, as the comparator is, so the two share a method. That is the same principle that made the two questionnaires agree with each other, applied on the objective side. It is evidence for construct divergence, not a claim that one instrument is more accurate than the other.

Grip strength and calf circumference are not merely proxies for age. Both are sarcopenia markers, and their association with swallowing muscle function is recognised [31,32,33]; Chinese guidance recommends calf circumference below 34 cm in men and 33 cm in women as a field measure [29]. An analogous pattern has been reported for oral dryness, where an instrument-measured index identified dysphagia more often than the corresponding self-reported symptom [18]. Tongue pressure captures oropharyngeal motor function more directly, but it needs a calibrated device and was the most often

missing objective measurement, so its marginal contribution may be underestimated by complete-case analysis.

The model explained little variance in the comparator outcome (Nagelkerke R-squared 0.007). That is expected when an outcome is uncommon and largely determined by factors outside the model, and it is the main reason we do not propose the model as a replacement for clinical assessment.

**Calibration held despite the change in prevalence**

A model carried from one survey to the next is exposed to any shift between them, and the shift here was substantial: prevalence rose from 4.2% to 8.3%. Serious miscalibration would have been reasonable to expect. We did not find it: calibration-in-the-large (0.064) and slope (0.796) were close to their ideal values, with the slope slightly below 1 as expected from optimism carried over.

Two caveats apply. The confidence intervals are wide, so the data are compatible with moderate miscalibration as well as with good calibration. A single year-on-year comparison in one district cannot establish behaviour under a larger or more sustained change in prevalence, and services adopting the model should still monitor calibration.

**Implications for practice and research**

The implication for services is narrower than a recommendation to replace questionnaires with models. It is that self-report and objective measurement answer different questions, and a pathway that records only one of them will classify people differently from a pathway that records the other, without either classification being demonstrably correct.

Two practical consequences follow. First, the two should not be treated as substitutes. Guidance that pairs a positive questionnaire with a bedside test is sequential, and a second test reached only through a positive first test cannot reclassify those the first test classified as negative. In our data the questionnaire pathway screened 14.5% of participants and identified 27.0% of positive tests; running both in parallel screened 24.7% and identified 49.1% (Supplementary Table S7). Second, if a service can record only one measure, the choice should be made explicitly rather than by default, because the two are not interchangeable.

For nursing practice this shifts part of screening from an administered questionnaire to an observed task, with implications for training and workload. Tongue pressure is the demanding measure: it needs a calibrated device, consumable probes, infection control, and a repeated maximal-effort instruction that frail or cognitively impaired people may not sustain. A service that cannot meet those conditions should not treat it as a screening requirement; the water swallow test alone, using a cup and a stopwatch, generated the objective information in this study. If a measurement is unobtainable at a visit it should be

left missing and the visit repeated rather than imputed, because imputation at the point of use substitutes a population average for the measurement the model depends on.

For research, the priority is not another comparison of one self-report instrument with another, but prospective work establishing what either classification predicts: whether people classified differently by the two methods differ in pneumonia, malnutrition or hospital admission. Until that is known, the choice between them rests on construct and cost, not on accuracy.

**Strengths and limitations**

Strengths include the sample sizes, the prespecified primary model with a full report of alternative algorithms, and an objective physiological measure rarely available in community surveys.

The most important limitation is that no instrumental assessment was performed. We did not use videofluoroscopy or fibreoptic endoscopic evaluation, so we cannot estimate how well either instrument detects aspiration or identifies the underlying disorder. This is a design boundary rather than an oversight, because instrumental assessment is not feasible at community scale; but it means that every number reported here describes agreement between two bedside measurements, and none can be read as diagnostic accuracy. Where pharyngeal sensation is reduced, both measurements may miss the same people, and this design cannot detect that.

Three further limitations qualify the model. First, most of its discrimination came from age and grip strength and the remaining variables added nothing measurable, so it should be read as a small extension of routine assessment rather than a new instrument. Second, the threshold used to report classification was selected within the validation cohort; thresholds transferred from the derivation cohort and thresholds selected within the validation cohort are both given in Supplementary Table S9, and neither is proposed as an operating point for practice. Third, an exploratory decision analysis (Supplementary Table S4) suggested that at a prevalence of 8.3% universal testing would yield higher net benefit than any selective strategy, so the model should not be described as improving on universal objective testing.

Other limitations are as follows. Within the comparator most positive results were grade II, suspected difficulty; grade III or above was too rare to analyse separately, so the outcome should be read as screen-detected risk. Both surveys were conducted in one district of one province, and the observed shift in prevalence may be setting specific; transport elsewhere would require recalibration rather than reuse of the coefficients, because prevalence differs, sarcopenia thresholds are population-specific, and the cognitive score requires adjustment for education. The cognitive score was missing for about 15% of each survey (see Missing data in the Methods), which selects a healthier analysis population and therefore understates community prevalence. Cognitive status was defined by a prespecified threshold of 20 without adjustment for education (see Methods). The two waves could not be linked at the individual

level, so temporal validation reflects performance in the same population rather than a fully independent sample. The agreement analysis by cognitive stratum was exploratory and the impaired subgroups were small (n = 251 and 164, with 23 and 25 positive tests), so the direction of the difference between strata could not be established and the 2025 excess was not reproduced in 2024. Finally, the study is cross-sectional: it addresses how instruments classify people, not what either classification predicts.

**Conclusions**

Self-report and objective measurement of swallowing classify the same community-dwelling older people differently, and cognitive impairment does not explain the difference. The two should be treated as measuring different constructs rather than as interchangeable screens. Services should decide explicitly which construct their screening pathway is intended to capture, record both where feasible, and not substitute a questionnaire for an observed swallowing task on the assumption that the two are measuring the same thing.

**List of abbreviations**

AUC: area under the receiver operating characteristic curve; CI: confidence interval; EAT-10: Eating Assessment Tool; HDS-R: Hasegawa Dementia Scale-Revised; IDI: integrated discrimination improvement; NRI: net reclassification improvement; PPA: positive percent agreement; NPA: negative percent agreement; TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, artificial intelligence extension; WST: water swallow test.

**Declarations**

**Ethics approval and consent to participate.** Each survey round was approved by the human research ethics committee of the coordinating hospital (approval K2023186, issued 12 December 2023, for the 2024 round; approval K2025041, issued 14 March 2025, for the 2025 round). All participants provided written informed consent before taking part, with additional consent from a family member where cognitive impairment made written consent questionable, in accordance with the approved protocols.

**Consent for publication.** Not applicable; no individual person's data are presented.

**Availability of data and materials.** The datasets analysed in this study contain individual-level health information on older people collected under an approved community screening protocol, and are not publicly available because public release would risk re-identification and is not permitted by the terms of the ethics approval. De-identified data may be requested from the corresponding author, subject to approval by the data governance committee responsible for the surveys. The analysis code is available from the corresponding author on reasonable request, and will be deposited in a public repository upon acceptance.

**Competing interests.** The authors declare that they have no competing interests.

**Funding.** This work was supported by the National Key R&D Program of China (No. 2025YFC3608300; No. 2025YFC3608305), the Medical Scientific Research Foundation of Zhejiang Province, China (No. 2025KY947), and the Scientific Research Fund of the National Health Commission, under the Key Project of the Zhejiang Provincial Health Major Science and Technology Plan (No. WKJ-ZJ-2433). None of the funders had any role in the design of the study, the collection, analysis or interpretation of data, or the writing of the manuscript.

**Authors' contributions.** Jie Yu conceived the study and was responsible for the formal analysis, the investigation, data curation and project administration, and wrote and revised the manuscript. Siyuan Wu contributed to the investigation and to data curation. Mengxue Fu contributed to the investigation and to the formal analysis. Wenhao Qi, Lijuan Quan, Changxin Xu and Xinran Tian contributed to the investigation. Huafang Zhang conceived the study, developed the methodology, validated the results, provided resources and supervision, revised the manuscript, administered the project and acquired the funding. All authors read and approved the final manuscript.

**Acknowledgements.** Not applicable.

**Patient and public involvement.** Patients and the public were not involved in the design of this study, the choice of outcome measures or the analysis plan, because it is a secondary analysis of data collected under an approved community screening programme whose protocol was fixed before the present research question was posed.

**Study registration.** Not registered. This study is a secondary analysis of two existing community surveys conducted in 2024 and 2025; the analysis plan and the specification of the primary model were fixed before any model fitting (see the Supplementary Appendix).

**Declaration of generative AI and AI-assisted technologies in the writing process.** During the preparation of this work the authors used a large language model to assist with language editing and to check the consistency of reported numerical results against the analysis outputs. All scientific content, all interpretations and all numerical values were generated and verified by the authors, who reviewed and edited every section and take full responsibility for the content of the published article. The AI tool does not meet the criteria for authorship.

## References

1. Doan TN, Ho WC, Wang LH, Chang FC, Nhu NT, Chou LW. Prevalence and methods for assessment of oropharyngeal dysphagia in older adults: a systematic review and meta-analysis. J Clin Med. 2022;11:2605. doi:/23390/jcm11092605.

2. Banda KJ, Chu H, Chen R, Kang XL, Jen HJ, Liu D, et al. Prevalence of oropharyngeal dysphagia and risk of pneumonia, malnutrition, and mortality in adults aged 60 years and older: a meta-analysis. Gerontology. 2022;68:841-53. doi:/21159/000520326.

3. Sura L, Madhavan A, Carnaby G, Crary MA. Dysphagia in the elderly: management and nutritional considerations. Clin Interv Aging. 2012;7:287-98. doi:/22147/CIA.S23404.

4. Baijens LWJ, Clavé P, Cras P, et al. European Society for Swallowing Disorders - European Union Geriatric Medicine Society white paper: oropharyngeal dysphagia as a geriatric syndrome. Clin Interv Aging. 2016;11:1403-28. doi:10.2147/CIA.S107750

5. Speyer R, Balaguer M, Cugy E, et al. Expert consensus on clinical decision making in the disease trajectory of oropharyngeal dysphagia in adults: an international Delphi study. J Clin Med. 2023;12:6572. doi:10.3390/jcm12206572

6. Belafsky PC, Mouadeb DA, Rees CJ, Pryor JC, Postma GN, Allen J, Leonard RJ. Validity and reliability of the Eating Assessment Tool (EAT-10). Ann Otol Rhinol Laryngol. 2008;117:919-24. doi:/21177/000348940811701210.

7. Zhang PP, Yuan Y, Li TT, Zhang H, Wang HY, Wang XW, et al. Diagnostic accuracy of the Eating Assessment Tool-10 (EAT-10) in screening dysphagia: a systematic review and meta-analysis. Dysphagia. 2023;38:145-58. doi:10.1007/s00455-022-10486-6

8. Zhang H, Zhang S, Ye C, et al. Cross-sectional study on dysphagia evaluation in community-dwelling older adults using the Eating Assessment Tool (EAT)-10, EAT-2, and Water Swallow Test. Geriatr Nurs. 2023;54:258-63. doi:10.1016/j.gerinurse.2023.10.002

9. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol. 1995;50:741-9. doi:/21037/0003-066X.50.9.741.

10. American Educational Research Association, American Psychological Association, Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 2014.

11. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:/21136/bmj-2023-078378.

12. Lin Y, Qiu Y, Huang Q, et al. The predictive value of tongue pressure for dysphagia in community-dwelling older adults: a cross-sectional study. Geriatr Nurs. 2025;66:103589. doi:10.1016/j.gerinurse.2025.103589

13. Jeong JW, Kim KW, Lee DY, Lee SB, Park JH, Choi EA, et al. A normative study of the Revised Hasegawa Dementia Scale: comparison of demographic influences between the Revised Hasegawa Dementia Scale and the Mini-Mental Status Examination. Dement Geriatr Cogn Disord. 2007;24:288-93. doi:10.1159/000107592

14. Kounnavong S, Vonglokham M, Sayasone S, Savathdy V, Masaki E, Kayano R, et al. Assessment of cognitive function among adults aged 60 years and older using the Revised Hasegawa Dementia Scale: a cross-sectional study, Lao People's Democratic Republic. Health Res Policy Syst. 2022;20:121. doi:10.1186/s12961-022-00919-x

15. Watanabe S, Oh-Shige H, Oh-Iwa I, Miyachi H, Shimozato K, Nagao T. Reconsideration of three screening tests for dysphagia in patients with cerebrovascular disease performed by nurses. Odontology. 2020;108:117-23. doi:10.1007/s10266-019-00431-9

16. Tohara H, Saitoh E, Mays KA, Kuhlemeier K, Palmer JB. Three tests for predicting aspiration without videofluorography. Dysphagia. 2003;18:126-34. doi:/21007/s00455-002-0095-y.

17. Huang Q, Zhao L, Chen Y, et al. Relationship between oral health and dysphagia in the Chinese community-dwelling older adults: a cross-sectional study. Dysphagia. 2025;41:329-39. doi:10.1007/s00455-025-10882-8

18. Ye C, Zhao L, He X, et al. Association between oral dryness and dysphagia in community-dwelling older population. J Nutr Health Aging. 2025;29:100533. doi:10.1016/j.jnha.2025.100533

19. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part II - binary and time-to-event outcomes. Stat Med. 2019;38:1276-96. doi:/21002/sim.7992.

20. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87. doi:/21002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

21. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54:774-81. doi:/21016/S0895-4356(01)00341-9.

22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837-45. doi:/22307/2531595.

23. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37-46. doi:/21177/001316446002000104.

588 24. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive
589 ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med.
590 2008;27:157-72. doi:/21002/sim.2929.
591 25. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models.
592 Med Decis Making. 2006;26:565-74. doi:/21177/0272989X06295361.
593 26. Schindler A, de Fátima Lago Alvite M, Robles-Rodriguez WG, Barcons N, Clavé P. History and
594 science behind the Eating Assessment Tool-10 (EAT-10): lessons learned. J Nutr Health Aging.
595 2023;27:597-606. doi:10.1007/s12603-023-1950-9
596 27. Working Group on Oral Hypofunction. Chinese expert consensus on graded screening and
597 functional assessment of oral hypofunction (2025). Geriatr Health Care. 2026;32(1):22-31.
598 doi:10.3969/j.issn.1008-8296.2026.01.007
599 28. Loevinger J. Objective tests as instruments of psychological theory. Psychol Rep. 1957;3:635-94.
600 doi:/22466/pr0.1957.3.3.635.
601 29. Working Group on Dysphagia in Older Adults. Chinese expert consensus on rehabilitation of
602 dysphagia in older adults. Chin J Phys Med Rehabil. 2025;47(12):1057-65.
603 doi:10.3760/cma.j.cn421666-20251021-00956
604 30. Bengisu S, Savaş M, Aygün Üstel D, Kahraman Beğen S, Manay B. The discriminant ability of the
605 Eating Assessment Tool-10 to detect dysphagia and aspiration according to the stage of Alzheimer's
606 disease. J Alzheimers Dis. 2025;107(2). doi:10.1177/13872877251362229
607 31. Fujishima I, Fujiu-Kurachi M, Arai H, Hyodo M, Kagaya H, Maeda K, et al. Sarcopenia and
608 dysphagia: position paper by four professional organizations. Geriatr Gerontol Int. 2019;19:91-7.
609 doi:/21111/ggi.13591.
610 32. Wakabayashi H. Presbyphagia and sarcopenic dysphagia: association between aging, sarcopenia,
611 and deglutition disorders. J Frailty Aging. 2014;3:97-103. doi:/214283/jfa.2014.8.
612 33. Maeda K, Akagi J. Decreased tongue pressure is associated with sarcopenia and sarcopenic
613 dysphagia in the elderly. Dysphagia. 2015;30:80-7. doi:/21007/s00455-014-9577-y.

614 **Tables**

615 **Table 1. Characteristics of participants with a recorded Hasegawa Dementia Scale-Revised score,**
616 **by cognitive status and survey year**

| Characteristic | 2024 survey: no impairment (n = 2,722) | 2024 survey: impairment (n = 268) | 2025 survey: no impairment (n = 2,925) | 2025 survey: impairment (n = 199) |
|---|---|---|---|---|
| Age, years, mean | 72.2 (5.2) | 75.5 (6.8) | 73.4 (5.3) | 77.4 (7.4) |

| | | | | |
|---|---|---|---|---|
| (SD) | | | | |
| Female sex, n (%) | 1,390 (51.1) | 181 (67.5) | 1,489 (50.9) | 135 (67.8) |
| Hasegawa Dementia Scale-Revised score, mean (SD) | 26.7 (2.3) | 15.9 (4.2) | 26.4 (2.0) | 16.8 (3.6) |
| Maximum tongue pressure, kPa, mean (SD) | 26.8 (8.0) | 21.3 (8.4) | 22.8 (6.5) | 20.1 (6.4) |
| Maximum grip strength, kg, mean (SD) | 25.9 (7.9) | 20.9 (6.6) | 26.4 (7.6) | 21.2 (6.3) |
| Calf circumference, cm, mean (SD) | 34.1 (2.9) | 33.3 (2.9) | 33.9 (3.3) | 32.9 (3.0) |
| Unable to eat hard food, n (%) | 711 (26.3) | 112 (42.4) | 833 (28.9) | 100 (51.0) |
| Eating Assessment Tool total score, mean (SD) | 0.45 (1.42) | 0.54 (1.42) | 0.52 (1.24) | 0.76 (1.55) |
| Positive water swallow test, n (%) | 104 (3.9) | 26 (9.9) | 226 (8.0) | 32 (16.7) |

Note. Cognitive impairment is defined as a Hasegawa Dementia Scale-Revised score of 20 or below. EAT-10 = Eating Assessment Tool; HDS-R is used only in this note for brevity. The full baseline table, including smoking, alcohol, dentition, xerostomia, pneumonia, medication count and educational attainment, is given in Supplementary Table S1.

**Table 2. Discrimination of the comparator outcome and calibration of the objective model compared with self-report instruments in the temporal validation cohort**

| Measure | Objective model | Eating Assessment Tool, total score | Eating Assessment Tool at cut-off of 2 or above | Two-item version |
|---|---|---|---|---|
| Participants, n | 2,765 | 2,765 | 2,765 | 2,765 |
| Positive comparator tests, n | 230 | 230 | 230 | 230 |
| Area under the curve (95% CI) | 0.661 (0.620 to 0.702) | 0.595 | 0.568 | 0.571 |
| P value versus the objective model | — | 0.010 | < 0.001 | < 0.001 |

| | | | | |
|---|---|---|---|---|
| Continuous net reclassification improvement (95% CI) | — | 0.019 (0.009 to 0.030) | — | — |
| Integrated discrimination improvement (95% CI) | — | 0.006 (0.003 to 0.008) | — | — |

Note. All P values are from two-sided DeLong tests for paired curves. None of these values is an estimate of diagnostic accuracy: the comparator is a bedside water swallow test, not an instrumental criterion (see Methods). Calibration of the model in the same cohort: calibration-in-the-large 0.064; calibration slope 0.796; Brier score 0.074 before and 0.073 after logistic recalibration. CI = confidence interval.

**Figure legends**

**Figure 1. Receiver operating characteristic curves for the objective model, the Eating Assessment Tool total score, the Eating Assessment Tool at its conventional cut-off and the two-item version in the temporal validation cohort (n = 2,765 participants, 230 events).**

**Supplementary material**

**Supplementary Table S1.** Full baseline characteristics of participants, by cognitive status and survey year.

**Supplementary Table S2.** Comparison of modelling approaches by ten-fold cross-validation in the derivation cohort.

**Supplementary Table S3.** Univariable discrimination of all candidate predictors in the derivation cohort.

**Supplementary Table S4.** Exploratory decision analysis: net benefit, positive tests identified and participants referred for further assessment by threshold probability, for the objective model, the Eating Assessment Tool at its conventional cut-off and the two-item version, with reference strategies of testing everyone and testing nobody, using the water swallow test as the outcome.

**Supplementary Table S5.** Agreement between the Eating Assessment Tool at its cut-off and the water swallow test by cognitive stratum and survey year, with observed agreement, positive and negative percent agreement and Cohen's kappa.

**Supplementary Table S6.** Sensitivity analysis adding the objective oral moisture value to the five-predictor model in the derivation cohort.

**Supplementary Table S7.** Resource use and yield of candidate screening strategies in the temporal validation cohort: the percentage of participants screened, the positive percent agreement achieved and the number screened per positive comparator test identified, for universal testing, questionnaire-only pathways, the objective model alone at a range of thresholds, and the model and questionnaire run in parallel.

**Supplementary Table S8.** Characteristics of participants with and without a recorded Hasegawa Dementia Scale-Revised score, by survey year, in the cohort with a recorded water swallow test result.

**Supplementary Table S9.** Positive and negative percent agreement for the objective model at candidate thresholds, including thresholds transferred from the derivation cohort and thresholds selected within the validation cohort.

**Supplementary Figure S1.** Calibration plot for the temporal validation cohort, by decile of predicted risk, with the fitted calibration slope of 0.796 and calibration-in-the-large of 0.064 shown for reference.

**Supplementary Figure S2.** Nomogram for the five-predictor objective model, with each variable scaled to the range observed in the derivation cohort.

**Supplementary Figure S3.** Exploratory decision curves for the objective model, the Eating Assessment Tool at its cut-off and the two-item version, with reference strategies of testing everyone and testing nobody, using the water swallow test as the outcome. No strategy exceeded testing everyone at any threshold examined.

**Supplementary Appendix.** R package versions, the random seed used, the full model equation, and the calibration and discrimination estimates with bootstrap confidence intervals. The linear predictor of the model is -4.970571 - 0.036224 x maximum tongue pressure (kPa) - 0.051393 x maximum grip strength (kg) - 0.041185 x calf circumference (cm) + 0.076424 x age (years) - 0.015550 x Hasegawa Dementia Scale-Revised score; the predicted probability is 1 / (1 + exp(-linear predictor)).

# Figures

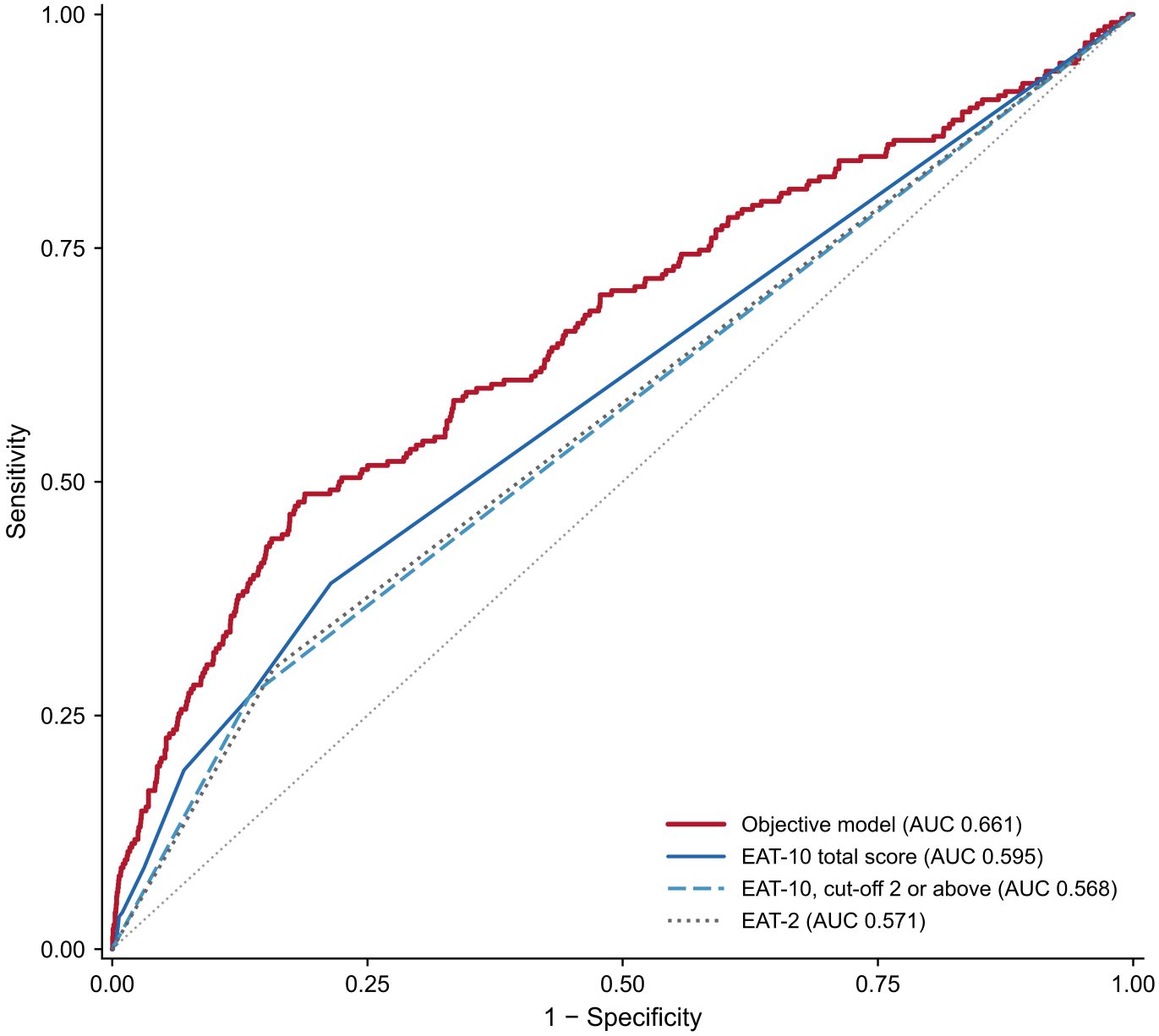



**Figure 1**

**Receiver operating characteristic curves for the objective model, the Eating Assessment Tool total score, the Eating Assessment Tool at its conventional cut-off and the two-item version in the temporal validation cohort (n = 2,765 participants, 230 events).**

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 03Additionalfile1.pdf