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.

A machine-learning-benchmarked, ridge-penalised nomogram excluding oral-frailty items for oral-frailty risk stratification in older inpatients with lung cancer: development and internal validation

Zhihua Fan^
Nanning Second People's Hospital
Rongquan Wei^
26466866@qq.com

Guangxi Medical University Cancer Hospital, Guangxi Zhuang Autonomous Region
Guijie Ren^
Nanning Second People's Hospital
Huiling Wang^
Nanning Second People's Hospital
Xuemei Chen^
Nanning Second People's Hospital
Xinghong Liu^
Nanning Second People's Hospital

Research Article

Keywords: oral frailty, lung neoplasms, aged, machine learning, nomograms, risk assessment, decision support techniques

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11181344/v1

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

Additional Declarations: No competing interests reported.

A machine-learning-benchmarked, ridge-penalised nomogram excluding oral-frailty items for oral-frailty risk stratification in older inpatients with lung cancer: development and internal validation

Running title: Internal validation of a non-oral-frailty-item nomogram for oral frailty

Authors: Zhihua Fan^1, Rongquan Wei^2*, Guijie Ren^1, Huiling Wang^1, Xuemei Chen^1, Xinghong Liu^1

^1 Nanning Second People's Hospital, Nanning 530000, Guangxi Zhuang Autonomous Region, China

^2 Guangxi Medical University Cancer Hospital, Nanning 530000, Guangxi Zhuang Autonomous Region, China

^* Correspondence: Rongquan Wei, Guangxi Medical University Cancer Hospital, Nanning 530000, Guangxi Zhuang Autonomous Region, China. Email: 26466866@qq.com

Abstract

Background Oral frailty is common among older inpatients with lung cancer, but published prediction tools embed diagnostic oral-frailty items as predictors, inducing circular reasoning and outcome leakage. This single-centre study develops and internally validates a transparent ridge-penalised nomogram from routine pre-examination admission variables and benchmarks it against six machine-learning classifiers.

Methods Consecutive inpatients aged ≥60 years with histologically confirmed lung cancer were retrospectively enrolled at Guangxi Medical University Cancer Hospital (September 2024–April 2025). Oral frailty was defined as an Oral Frailty Index-8 (OFI-8) score ≥4, and six overlapping oral variables were excluded a priori. The 477 patients were stratified 7:3 (random_state=42) into training (n=333; 236 events) and test (n=144; 102 events) sets. Five-fold cross-validated LASSO over a prespecified 44-term P1 pool selected 11 predictors at λ~1se~=7.543; a ridge-penalised logistic model (C=6.866; λ=0.146) was the primary model, and six further classifiers were fitted without class weights. Discrimination (AUC; 2000 bootstrap replicates), calibration and decision-curve analysis were assessed, with SHAP for XGBoost; three prespecified comparisons used Holm correction, and three sensitivity analyses reintroduced oral, taste/drinking and education variables.

Results Oral frailty was prevalent in 70.9% (338/477). The nomogram, built on pre-examination variables excluding diagnostic items, achieved test AUC 0.888 (95% CI 0.827–0.942; training 0.867), with calibration slope 1.065, intercept 0.295 and Brier 0.122. XGBoost attained the highest AUC (0.959, 95% CI 0.920–0.988), followed by random forest (0.945, 95% CI 0.902–0.978). Both outperformed the nomogram (ΔAUC +0.071 and +0.057; Holm P=0.020 each) but did not differ from each other (ΔAUC +0.014, P=0.156) and were less well calibrated. Reintroducing the six oral variables left XGBoost unchanged (0.950; Δ≈−0.009, ≈−0.9%), whereas the symmetric ridge re-run on the 50-term pool raised ridge AUC to 0.968 (Δ=+0.080; +9.0%) and retained five oral items, quantifying the excluded circularity.

Conclusions A printable, interpretable nomogram built from admission variables available before any oral examination shows good discrimination and a near-ideal calibration slope, supporting nurse-led risk stratification; the higher-AUC ensembles serve as performance

benchmarks. Owing to its positive intercept, absolute-risk estimates require site-specific recalibration, and multicentre external and temporal validation is the priority before bedside deployment.

Keywords: oral frailty; lung neoplasms; aged; machine learning; nomograms; risk assessment; decision support techniques

Background

Lung cancer remains the leading cause of cancer-related death worldwide. GLOBOCAN 2024 ranks it first among cancers for both incidence and mortality, and the burden is concentrated in adults aged 60 years and older [1]. As therapy improves and survival lengthens, more older patients live with chronic cancer and comorbidities. In this population, cancer-related inflammation and cachexia, chemotherapy-induced xerostomia and mucositis, radiation-related oral toxicity, immune-checkpoint-inhibitor adverse events and age-related degenerative change act cumulatively to impair oral function[2-6]. Despite this burden, oral health is seldom assessed systematically during hospitalisation.

Oral frailty denotes a progressive decline in oral functional reserve—spanning mastication, swallowing, salivation and oral hygiene—along the continuum from robust function to severe oral dysfunction[7, 8]. Rooted in the Japanese consensus on oral hypofunction, the construct is operationalised in older adults by instruments such as the Oral Frailty Index-8 (OFI-8)[9]. It accompanies or predicts malnutrition [10], incident physical frailty [11-13], mild cognitive impairment[14], falls [15] and mortality [16]. More broadly, frailty is a recognised prognostic indicator in advanced lung cancer [17, 18]. In hospitalised older patients with cancer, cross-sectional studies likewise report a high prevalence and clusters of modifiable risk factors[19-21].

Three interrelated gaps impede clinical translation.

First, existing oral-frailty prediction tools were developed mainly in community-dwelling older adults[22], and a recent scoping review shows that many enter diagnostic oral-frailty items such as self-reported chewing difficulty, dry mouth, remaining dentition or denture use as predictors [23]. The same review documents marked heterogeneity in development and reporting, few externally validated models and a persistently high risk of bias[23]. When predictors comprise the same items that define the outcome, discrimination is mechanically inflated through circular reasoning (outcome leakage), a source of bias explicitly identified in the TRIPOD+AI guidance[24]. Such models show good apparent performance but cannot be applied before the oral examination they are intended to anticipate.

Second, machine-learning (ML) methods capture non-linear relationships and complex interactions [25, 26], but their "black-box" character limits bedside use and clinical trust [27]. Transparent nomograms, by contrast, support visual point-of-care estimation and, with appropriate penalisation, retain stable calibration[28]. Evidence further indicates that in small clinical datasets ML does not consistently outperform logistic regression[29]. Because decision-curve net benefit is computed at the observed event prevalence, class weights were fixed at unity a priori: re-weighting to equalise group sizes would shift predicted probabilities away from that prevalence and distort the net-benefit scale and downstream threshold decisions[30].

Third, hospital-based oral-frailty models have been confined to chronic-disease cohorts with internal validation only[31]; no lung-cancer-specific tool exists despite the pronounced oral toxicity of its treatment.

To address these gaps, the present study makes three main contributions.

- **Leakage-free by design.** A ridge-penalised logistic nomogram for oral frailty is derived solely from admission variables recorded before any oral examination, with the six diagnostic oral items excluded a priori so that screening is separated from diagnosis and outcome leakage is removed by construction.

- **Circularity quantified, not asserted.** Rather than presuming such bias, a symmetric sensitivity arm (S_oral) reintroduces the excluded oral items to measure directly how far discrimination shifts for the transparent ridge model relative to the non-linear ensemble.

- **Transparent and deployment-ready first.** The printable, interpretable ridge nomogram is positioned as the bedside tool, with the tree ensembles retained as performance benchmarks; the S_lifestyle and S_ses arms further document reverse-causation and suppression effects near the outcome.

Together, these elements offer a transferable, reproducible template—screening separated from diagnosis and circularity quantified rather than assumed—for oral-frailty prediction in hospitalised older adults. The study is reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement[32] and the TRIPOD+AI statement[24].

Methods

Study design, setting and ethics

This retrospective, single-centre observational study was conducted at Guangxi Medical University Cancer Hospital, a tertiary teaching hospital in Nanning, China. Consecutive inpatients admitted between September 2024 and April 2025 were screened. The study was approved by the hospital's Institutional Ethics Committee (approval no. KY2024892, issued 20

December 2024); because the analysis used de-identified, routinely collected clinical data, the requirement for written informed consent was waived. The committee approved retrospective use of such data for patients admitted since September 2024; all data extraction and analyses followed this approval. The study is reported in accordance with STROBE [32]and TRIPOD+AI [24].

Study population

Patients were eligible if they (i) were aged ≥60 years, (ii) had histologically confirmed primary lung cancer, (iii) had complete clinical and oral assessment data and (iv) were able to complete a standardised interview. They were excluded if they had head-and-neck tumours or prior head-and-neck radiotherapy, severe cognitive impairment or psychiatric disease, an expected survival of less than one month, or pre-existing oral conditions that would confound the oral assessment. No duplicate records were identified, and 477 patients were retained for analysis (Supplementary Figure S1).

Data collection and outcome definition

Two trained nurses independently double-entered all data, with discrepancies resolved by consensus under the supervision of the study coordinator. Variables covered socio-demographic characteristics (age, sex, occupation, education, marital status, residence, living arrangement, insurance and monthly income), lifestyle (smoking and drinking duration, dietary structure and taste, weekly exercise), disease and treatment details (histological subtype, stage, metastasis, number of chemotherapy cycles, receipt of chemotherapy, surgery and radiotherapy), comorbidity and polypharmacy, laboratory measurements (albumin, globulin and prealbumin)

and functional status (Barthel index, mMRC dyspnoea grade and physical frailty score). A complete data dictionary is provided as Supplementary material.

The outcome, oral frailty, was assessed with the validated Mandarin Chinese version of the OFI-8 [33], an eight-item weighted instrument covering, among other domains, chewing, swallowing, dry mouth and dentition, with a total score ranging from 0 to 11. A score ≥4 defined oral frailty, consistent with the original OFI-8 derivation and subsequent validations [9]. To avoid circular reasoning, six oral-health variables that overlap directly with OFI-8 items—number of remaining teeth, number of removable prosthetic teeth, denture use, self-reported xerostomia, subjective chewing difficulty and poor self-rated oral health—were excluded from the primary predictor set and retained only for the S_oral sensitivity analysis.

All predictors were obtained from the routine admission record completed within 24 h of presentation (medical and treatment history, physical examination, admission laboratory tests and functional scores), and the OFI-8 was likewise administered within 24 h of admission. Chemotherapy and radiotherapy denote treatment received before, or ongoing at, admission rather than treatment initiated later during the index stay; every predictor therefore precedes the oral-frailty assessment and is available before any dedicated oral examination.

Preprocessing and predictor coding

After exclusion of the six oral variables and the outcome, predictors were coded as follows. Eight nominal variables (occupation, dietary structure, dietary taste, marital status, residence, living arrangement, insurance type and pathological subtype) were one-hot encoded, with the smallest or largest subgroup used as the reference category. Sparse categories were combined beforehand (non-manual occupational groups; unmarried, divorced and widowed as non-married; retired-cadre and self-pay insurance as "other"). BMI, albumin, globulin, prealbumin and the

Barthel index were entered as continuous only; their prespecified categorical forms were discarded and the Barthel category column dropped, leaving no paired continuous–categorical terms. Ordinal variables (monthly income, weekly exercise, mMRC dyspnoea grade and tumour stage) were entered as integer ranks. Continuous variables (age, BMI, smoking and drinking duration, cancer history, chemotherapy cycles, albumin, globulin, prealbumin and Barthel index) were z-standardised using training-set means and standard deviations alone, and binary variables were coded 0/1. Education was not entered into the primary design matrix and appeared only in the S_ses analysis and the Table 1 description. Because no values were missing, complete-case analysis was used.

After the six oral variables were excluded, the full candidate set comprised 52 design terms. Two nested prespecified pools were then defined for the primary analysis. Under Plan B (P0), the five dietary-taste dummies (six taste levels, bland as reference) and the drinking variables (a 0/1 drinking-history flag and a continuous lifetime-years duration) were excluded a priori under a prespecified variable-credibility rule. This left 45 design terms (52−7; the P0 pool). For the prespecified primary analysis (P1), education was additionally excluded a priori, yielding a final pool of 44 design terms (45−1). These rules were fixed before model fitting, not post-hoc choices made to improve performance.

Three concerns motivated the taste and drinking exclusions. First, dietary taste is susceptible to reverse causation in a cross-sectional design: the taste recorded at admission may already reflect post-outcome adaptation. Patients with established oral frailty or treatment-related oral toxicity adopt bland, soft diets, so that a non-bland category can appear "protective" purely through dietary adaptation. Second, the drinking variables are susceptible to sick-quitter bias and collinear suppression: patients who develop cancer or oral symptoms stop drinking, so that ever-

drinking appears "protective" relative to illness-driven abstinence. Entering the binary flag together with the continuous duration produced opposite-sign (suppression) coefficients rather than a coherent association. Third, the taste categories were severely sparse (sweet n=1, greasy n=7, spicy n=20), rendering their estimates unstable.

Education was excluded a priori under the same prespecified credibility rule, because its direction was expected to be sensitive to the socioeconomic adjustment set and its independent discrimination to be near zero. The prespecified S_ses analysis (see below) was designed to verify this expectation and corroborates rather than determines the decision. In that analysis its adjusted odds changed sign after socioeconomic adjustment—from OR 0.90 (unadjusted) to OR 1.71 (fully adjusted, P=0.01), the signature of a suppression/Simpson effect—whereas adding education to the 11-term model changed test AUC by only −0.003. A variable whose direction depends on the covariate set and that contributes near-zero independent discrimination is unsuitable as a prespecified predictor.

Smoking duration, by contrast, was retained because it has a clear temporal ordering (cumulative lifetime years that predate admission). LASSO selected the continuous smoke_yrs term and discarded the binary smoking-history flag, so no time-ordered lifestyle predictor was removed. The prespecified S_lifestyle analysis reported below tested these exclusions; these variables appear only there and in Table 1.

Data cleaning and logic-consistency checks

Before analysis, three logic-consistency checks were applied to the binary flags and their continuous counterparts (Supplementary Table S3). First, one record (a 67-year-old man) had smoke_history=0 but a smoking duration of 30 years; because this is internally inconsistent with abstinence, smoke_history was corrected to 1 (i.e. ever smoked, including former smokers).

Second, one record (a 62-year-old man) had drink_history=1 but a drinking duration of 0 years; this was retained as recorded and logged only, without modification. Third, one record (a 69-year-old woman with stage III disease) had chemotherapy=0 but one recorded chemotherapy cycle; chemotherapy was corrected to 1 (i.e. had received at least one cycle). Throughout, "smoke history" and "drink history" denote current or ever use (including former users who had quit), and the corresponding duration variables denote cumulative lifetime years. After cleaning, the chemotherapy flag was 1 in 350 patients and 0 in 127, and the smoking-history flag was 1 in 334 and 0 in 143. No out-of-range binary coding errors were encountered.

Train/test split and variable selection

The cohort was split 7:3 into a training set and a held-out test set, stratified by oral-frailty status (random_state=42), yielding 333 training patients (236 events, 97 non-events) and 144 test patients (102 events, 42 non-events). On the training set, five-fold cross-validated LASSO-penalised logistic regression[34] was used for variable selection over the 44-term P1 candidate pool, with the penalty strength selected to maximise cross-validated AUC. Both λ~min~ (1.677) and the more parsimonious λ chosen one standard error beyond the minimum (λ~1se~=7.543) were examined. The λ~1se~, prespecified to favour a sparser, more stable set and reduce overfitting, retained 11 design terms. Selection was performed at the level of dummy design terms. Because the selection-stage EPV (236 events across 44 candidates) was only ≈5.4, the optimism and stability of LASSO selection were additionally quantified by 120 bootstrap resamples, reporting the frequency with which each term was retained.

Primary model: ridge-penalised logistic nomogram

With 236 events and 11 LASSO-retained design terms, the events-per-variable ratio was 21.5 (236/11) [35], an acceptable level. Multicollinearity among the retained terms was assessed by variance inflation factors (VIFs). Because the laboratory and functional variables were entered as continuous only, no paired continuous–categorical terms remained, and the maximum VIF was 1.236 (weekly exercise), with all VIFs below 10, indicating no severe multicollinearity. The primary model was therefore a ridge (L2)-penalised multivariable logistic regression [36] fitted on the retained design terms, with the penalty tuned by five-fold cross-validation (best inverse-regularisation strength C=6.866; λ=0.146). Ridge shrinkage stabilises coefficients—particularly for near-separated sparse dummy terms—and further mitigates any residual coefficient instability; the resulting probabilities are displayed as a nomogram [30]. Odds ratios (ORs) and 95% confidence intervals (CIs) were obtained from the penalised fit with penalised standard errors. Because the L2 penalty shrinks coefficients toward zero, the penalised SEs, CIs and P values are descriptive quantities rather than unbiased maximum-likelihood inferences. They should therefore be interpreted cautiously, and the nomogram is intended primarily for risk stratification rather than as a source of formal causal or effect-size claims (see also the Table 3 footnote and Discussion). An unpenalised multivariable logistic fit on the same 11 terms is reported in Supplementary Table S7 as a sensitivity check on the penalised estimates.

Machine-learning benchmarks

All seven classifiers were fitted on the same eleven LASSO-selected design terms (not the 44-term candidate pool), without class weights and with prespecified, conventionally default hyperparameters (no grid search, Bayesian optimisation or cross-validated tuning), so as to keep the comparison transparent and avoid overfitting on the small test set. They comprised the ridge-penalised logistic model (as the transparent reference), a single CART decision tree with

maximum depth selected by five-fold cross-validation (depth=5), a random forest (100 trees, Gini split, √max_features), XGBoost (100 estimators, max_depth=6, learning rate 0.3) [37], a radial-basis-function support-vector machine (RBF-SVM; C=1.0, gamma=scale, with Platt probability estimates), k-nearest neighbours (k-NN; k=5, Euclidean distance) and a Gaussian naïve Bayes classifier. The six non-logistic algorithms served as ML benchmarks against the transparent nomogram. Class weights were deliberately omitted because they distort calibrated probability estimates, compromise DCA interpretation and are natively unsupported by k-NN and naïve Bayes; equal conditions therefore render the head-to-head comparison fair.

Evaluation and statistical inference

Discrimination was quantified by the AUC. At a prespecified 0.5 decision threshold, accuracy, sensitivity, specificity, F1, positive predictive value (PPV) and negative predictive value (NPV) were computed (Table 2), and the corresponding metrics at the Youden-optimal threshold were additionally reported. Threshold sensitivity and cost-weighted threshold selection were deferred to DCA. Test-set AUC 95% CIs were obtained by 2000 bootstrap resamples[38]. Pairwise model comparisons used paired bootstrap of the AUC difference (2000 replicates). Three primary comparisons were prespecified before analysis—XGBoost versus nomogram, random forest versus nomogram and XGBoost versus random forest. Holm correction was applied to these three prespecified comparisons only. All other pairwise comparisons are exploratory and uncorrected. For the three prespecified comparisons, both unadjusted (raw) P values and Holm-corrected P values are reported, whereas the bootstrap CIs were not multiplicity-adjusted.

Calibration was assessed by regressing the observed outcome on predicted log-odds, reporting the intercept (ideal 0) and slope (ideal 1), together with the Brier score. For the nomogram, apparent training-set calibration was compared with 1000-bootstrap bias-corrected calibration.

DCA was performed to estimate net benefit across threshold probabilities[39], with net benefit at key thresholds (0.30, 0.50, 0.70) tabulated in Supplementary Table S6. The XGBoost model was interpreted using SHapley Additive exPlanations (SHAP) [40].

Analyses were performed in Python 3.14.7 (scikit-learn 1.9.1, NumPy 2.5.3, pandas 3.0.6, XGBoost 3.4.1, SHAP 0.52.0, statsmodels 0.15.0, matplotlib 3.11.2) and R 4.5.3 (rms 8.1.1); random_state=42 throughout. All tests were two-sided, and $P<0.05$ was considered significant.

Sensitivity analyses

Three prespecified sensitivity analyses were performed.

S_oral (circular-reasoning inflation). To quantify the bias introduced by including oral-frailty items as predictors, the six excluded oral variables were reintroduced and three prespecified arms were run. Arm (i) refitted XGBoost on the 50-term pool (44 P1 design terms + six oral variables). Arm (ii) was a conditional ridge sensitivity in which the original 11 LASSO-retained terms were forced in and the six oral terms added, with LASSO re-run over the enlarged 17-term design (11+6 terms). Arm (iii) was a symmetric ridge arm that re-ran the identical main-model pipeline—five-fold LASSO λ~1se~ selection followed by five-fold ridge cross-validation—over the full 50-term pool (44 P1 design terms + six oral variables), exactly mirroring the XGBoost 44-to-50 comparison. All arms were fixed before fitting. The resulting AUCs were compared with their respective non-oral benchmarks, and the absolute and relative changes in AUC were reported. This design mirrors the comparison that would have arisen had diagnostic oral items been included naively, thereby providing an empirical, model-specific estimate of circular-reasoning bias rather than a theoretical one.

S_lifestyle (taste and drinking variables). A second sensitivity analysis reintroduced the five dietary-taste dummies (six taste levels, bland as reference), drinking history and drinking duration—seven design terms in total—to the P1 pool, yielding 51 design terms, and refit LASSO, ridge and XGBoost under identical settings. This tested whether these lifestyle variables, excluded a priori from the main model, add discrimination and whether their coefficients are clinically plausible.

S_ses (education suppression). A third sensitivity analysis reintroduced education to the 11-term baseline. Education was modelled as a per-SD ordinal term and added in a stepwise sequence (unadjusted; adjusted for age and sex; adjusted for the full socioeconomic block of income, occupation, residence and marital status; fully adjusted unpenalised; and a descriptive ridge fit) to trace how its association changes across covariate sets. LASSO was also re-run over the 45-term pool (which retained education), its 120-bootstrap selection frequency was reported, and the change in test AUC when education was added to the 11-term model was recorded.

Results

Participant characteristics

After screening, 477 patients were analysed. The median age was 66.0 years (IQR 62.0–70.0), and 364 (76.3%) were men. Oral frailty was present in 338 patients (70.9%). The stratified split allocated 333 patients to the training set (236 events [70.9%], 97 non-events) and 144 to the test set (102 events [70.8%], 42 non-events), preserving an event prevalence close to the cohort average. Between-group comparisons of baseline characteristics are summarised in Table 1. Patients with oral frailty were older (median 67 vs 65 years), had lower BMI (median 21.6 vs 23.0 kg/m^2), longer smoking duration (median 38 vs 20 years), lower serum globulin (median

29.0 vs 29.6 g/L) and prealbumin (median 230 vs 267 mg/L), and lower Barthel scores. They were also more likely to be men, manual workers and rural residents, to report no weekly exercise, mMRC dyspnoea grade ≥2, metastatic disease, chemotherapy and low monthly income. Drinking duration, serum albumin, immunotherapy, marital status and pathological subtype did not differ significantly between groups. In the oral-frailty group, surgery (8.9% vs 15.8%) and targeted therapy (21.6% vs 30.9%) were less frequent, whereas radiotherapy was more frequent (36.7% vs 19.4%; all P<0.05).

Table 1. Baseline characteristics of the study population by oral frailty status

Characteristic	Overall (n=477)	OFI=0 (n=139)	OFI=1 (n=338)	P value
N	477	139	338	—
Age (years)	66.0 (62.0-70.0)	65.0 (62.0-68.0)	67.0 (62.0-71.0)	<0.001
BMI (kg/m²)	21.9 (19.8-24.0)	23.0 (21.0-24.5)	21.6 (19.4-23.7)	<0.001
Smoking years	36.0 (0.0-47.0)	20.0 (0.0-40.0)	38.0 (25.0-48.0)	<0.001
Drinking years	0.0 (0.0-0.0)	0.0 (0.0-0.0)	0.0 (0.0-0.0)	0.457
Cancer detection months	5.0 (2.0-18.0)	9.0 (2.0-24.0)	5.0 (2.0-17.0)	0.078
Chemotherapy cycles	2.0 (0.0-5.0)	2.0 (0.0-5.0)	2.0 (1.0-5.0)	0.017
Albumin (g/L)	38.1 (34.6-40.2)	38.2 (35.4-40.0)	38.1 (34.3-40.3)	0.670
Globulin (g/L)	29.1 (26.1-32.5)	29.6 (26.8-34.1)	29.0 (25.7-31.3)	0.024
Prealbumin (mg/L)	238.0 (185.5-303.0)	266.5 (201.4-323.4)	229.7 (175.5-294.6)	<0.001
Barthel index	95.0 (90.0-95.0)	95.0 (95.0-95.0)	95.0 (85.0-95.0)	<0.001
Number of teeth	23.0 (15.0-26.0)	26.0 (24.0-27.0)	21.0 (12.0-25.0)	<0.001
Number of dentures	1.0 (0.0-8.0)	0.0 (0.0-0.0)	3.0 (0.0-14.0)	<0.001
Sex				0.042

Male	364 (76.3%)	97 (69.8%)	267 (79.0%)	—
Female	113 (23.7%)	42 (30.2%)	71 (21.0%)	—
Occupation				<0.001
Farmer	231 (48.4%)	58 (41.7%)	173 (51.2%)	—
Worker	90 (18.9%)	16 (11.5%)	74 (21.9%)	—
Civil servant	68 (14.3%)	31 (22.3%)	37 (10.9%)	—
Technician	79 (16.6%)	34 (24.5%)	45 (13.3%)	—
Other	9 (1.9%)	0 (0.0%)	9 (2.7%)	—
Smoking history				<0.001
No	143 (30.0%)	66 (47.5%)	77 (22.8%)	—
Yes	334 (70.0%)	73 (52.5%)	261 (77.2%)	—
Drinking history				0.865
No	363 (76.1%)	107 (77.0%)	256 (75.7%)	—
Yes	114 (23.9%)	32 (23.0%)	82 (24.3%)	—
Diet structure				<0.001
Balanced	277 (58.1%)	94 (67.6%)	183 (54.1%)	—
Meat-based	104 (21.8%)	13 (9.4%)	91 (26.9%)	—
Vegetarian	96 (20.1%)	32 (23.0%)	64 (18.9%)	—
Diet taste (descriptive only)				0.004
Bland	208 (43.6%)	68 (48.9%)	140 (41.4%)	—
Moderate	126 (26.4%)	41 (29.5%)	85 (25.1%)	—
Salty	115 (24.1%)	18 (12.9%)	97 (28.7%)	—

Spicy	20 (4.2%)	9 (6.5%)	11 (3.3%)	—
Sweet	1 (0.2%)	1 (0.7%)	0 (0.0%)	—
Greasy	7 (1.5%)	2 (1.4%)	5 (1.5%)	—
Education (descriptive only)				0.055
Primary or below	101 (21.2%)	25 (18.0%)	76 (22.5%)	—
Junior high	213 (44.7%)	54 (38.8%)	159 (47.0%)	—
Senior/tech	117 (24.5%)	45 (32.4%)	72 (21.3%)	—
College+	46 (9.6%)	15 (10.8%)	31 (9.2%)	—
Marital status				0.110
Married	447 (93.7%)	127 (91.4%)	320 (94.7%)	—
Divorced	6 (1.3%)	4 (2.9%)	2 (0.6%)	—
Widowed	24 (5.0%)	8 (5.8%)	16 (4.7%)	—
Monthly income				<0.001
≤1000 RMB	218 (45.7%)	45 (32.4%)	173 (51.2%)	—
1001-3000	107 (22.4%)	31 (22.3%)	76 (22.5%)	—
3001-5000	101 (21.2%)	36 (25.9%)	65 (19.2%)	—
>5000	51 (10.7%)	27 (19.4%)	24 (7.1%)	—
Residence				0.026
Urban	155 (32.5%)	57 (41.0%)	98 (29.0%)	—
Town	93 (19.5%)	27 (19.4%)	66 (19.5%)	—
Rural	229 (48.0%)	55 (39.6%)	174 (51.5%)	—
Living arrangement				0.611

Alone	44 (9.2%)	15 (10.8%)	29 (8.6%)	—
With spouse	168 (35.2%)	45 (32.4%)	123 (36.4%)	—
With spouse & children	224 (47.0%)	66 (47.5%)	158 (46.7%)	—
With children	38 (8.0%)	13 (9.4%)	25 (7.4%)	—
Other	3 (0.6%)	0 (0.0%)	3 (0.9%)	—
Insurance				0.004
Urban employee	194 (40.7%)	72 (51.8%)	122 (36.1%)	—
Urban-rural resident	281 (58.9%)	66 (47.5%)	215 (63.6%)	—
Retired cadre	1 (0.2%)	1 (0.7%)	0 (0.0%)	—
Self-paid	1 (0.2%)	0 (0.0%)	1 (0.3%)	—
Pathology				0.067
Small cell	73 (15.3%)	17 (12.2%)	56 (16.6%)	—
Squamous	94 (19.7%)	21 (15.1%)	73 (21.6%)	—
Adenocarcinoma	295 (61.8%)	99 (71.2%)	196 (58.0%)	—
Adenosquamous	6 (1.3%)	0 (0.0%)	6 (1.8%)	—
Other	9 (1.9%)	2 (1.4%)	7 (2.1%)	—
Tumor stage				<0.001
Stage I	22 (4.6%)	16 (11.5%)	6 (1.8%)	—
Stage II	27 (5.7%)	11 (7.9%)	16 (4.7%)	—
Stage III	126 (26.4%)	28 (20.1%)	98 (29.0%)	—
Stage IV	302 (63.3%)	84 (60.4%)	218 (64.5%)	—
Metastasis				<0.001
No	22 (4.6%)	16 (11.5%)	6 (1.8%)	—

Yes	455 (95.4%)	123 (88.5%)	332 (98.2%)	—
Chemotherapy				<0.001
No	127 (26.6%)	57 (41.0%)	70 (20.7%)	—
Yes	350 (73.4%)	82 (59.0%)	268 (79.3%)	—
Surgery				0.040
No	425 (89.1%)	117 (84.2%)	308 (91.1%)	—
Yes	52 (10.9%)	22 (15.8%)	30 (8.9%)	—
Radiotherapy				<0.001
No	326 (68.3%)	112 (80.6%)	214 (63.3%)	—
Yes	151 (31.7%)	27 (19.4%)	124 (36.7%)	—
Targeted therapy				0.041
No	361 (75.7%)	96 (69.1%)	265 (78.4%)	—
Yes	116 (24.3%)	43 (30.9%)	73 (21.6%)	—
Immunotherapy				0.513
No	279 (58.5%)	85 (61.2%)	194 (57.4%)	—
Yes	198 (41.5%)	54 (38.8%)	144 (42.6%)	—
Chronic conditions				0.182
1	257 (53.9%)	82 (59.0%)	175 (51.8%)	—
≥2	220 (46.1%)	57 (41.0%)	163 (48.2%)	—
Polypharmacy				<0.001
≤4	268 (56.2%)	95 (68.3%)	173 (51.2%)	—
≥5	209 (43.8%)	44 (31.7%)	165 (48.8%)	—

Denture (descriptive)				<0.001
No	231 (48.4%)	104 (74.8%)	127 (37.6%)	—
Yes	246 (51.6%)	35 (25.2%)	211 (62.4%)	—
Xerostomia (descriptive)				<0.001
No	193 (40.5%)	95 (68.3%)	98 (29.0%)	—
Yes	284 (59.5%)	44 (31.7%)	240 (71.0%)	—
Masticatory difficulty (descriptive)				<0.001
No	264 (55.3%)	124 (89.2%)	140 (41.4%)	—
Yes	213 (44.7%)	15 (10.8%)	198 (58.6%)	—
Barthel category				<0.001
100	26 (5.5%)	14 (10.1%)	12 (3.6%)	—
61-99	418 (87.6%)	125 (89.9%)	293 (86.7%)	—
41-60	30 (6.3%)	0 (0.0%)	30 (8.9%)	—
21-40	1 (0.2%)	0 (0.0%)	1 (0.3%)	—
≤20	2 (0.4%)	0 (0.0%)	2 (0.6%)	—
Exercise frequency				<0.001
0/wk	150 (31.4%)	12 (8.6%)	138 (40.8%)	—
1-3/wk	126 (26.4%)	36 (25.9%)	90 (26.6%)	—
≥4/wk	201 (42.1%)	91 (65.5%)	110 (32.5%)	—
mMRC dyspnea				<0.001
Grade 0	170 (35.6%)	91 (65.5%)	79 (23.4%)	—
Grade 1	193 (40.5%)	34 (24.5%)	159 (47.0%)	—

Grade 2	93 (19.5%)	11 (7.9%)	82 (24.3%)	—
Grade 3	21 (4.4%)	3 (2.2%)	18 (5.3%)	—
Physical frailty				<0.001
No	392 (82.2%)	135 (97.1%)	257 (76.0%)	—
Yes	85 (17.8%)	4 (2.9%)	81 (24.0%)	—
Poor oral health (descriptive)				<0.001
No	211 (44.2%)	109 (78.4%)	102 (30.2%)	—
Yes	266 (55.8%)	30 (21.6%)	236 (69.8%)	—

Note. Values are median (IQR) for continuous variables and n (%) for categorical variables. P values from Mann-Whitney U (continuous) or chi-square (categorical) tests. OFI=0 = no oral frailty; OFI=1 = oral frailty. Dietary taste, drinking history and drinking duration are shown descriptively but were excluded a priori from the primary multivariable model under the prespecified variable-credibility rule (see S_lifestyle). In the primary model, continuous variables were entered as per-SD increments and ordinal variables as per-level increments.

LASSO variable selection

Five-fold cross-validated LASSO retained 11 predictors at λ~1se~=7.543 (Figure 1): age, BMI, smoking duration (smoke_yrs), monthly income, tumour stage, chemotherapy, radiotherapy, globulin, Barthel index, weekly exercise and mMRC dyspnoea grade. The λ~1se~ choice yielded a sparser, more stable set than λ~min~ (1.677); polypharmacy was shrunk to zero at λ~1se~=7.543. Dietary taste dummies, drinking history, drinking duration and education had been removed from the candidate pool a priori (see S_lifestyle and S_ses below), and none of the six excluded oral variables was selected. LASSO selection stability over 120 bootstrap resamples was high: smoking duration, Barthel index and weekly exercise were selected in 100% of resamples, mMRC grade in 95.8%, age in 88.3%, globulin in 85.8%, BMI in 81.7%, stage in

79.2%, radiotherapy in 73.3%, chemotherapy in 55.8% and income in 48.3% (Supplementary Table S4 reports the full bootstrap frequency). Sex, occupation, dietary structure, dietary taste, drinking history, drinking duration, cancer history, chemotherapy cycles, serum albumin, prealbumin, metastasis, surgery, targeted therapy, immunotherapy, chronic-disease count, polypharmacy, physical frailty, marital status, residence, living arrangement, insurance type and pathological subtype were not retained. After adjustment, the systemic, treatment and functional markers carried more independent information on oral frailty than these broader oncological or demographic labels.

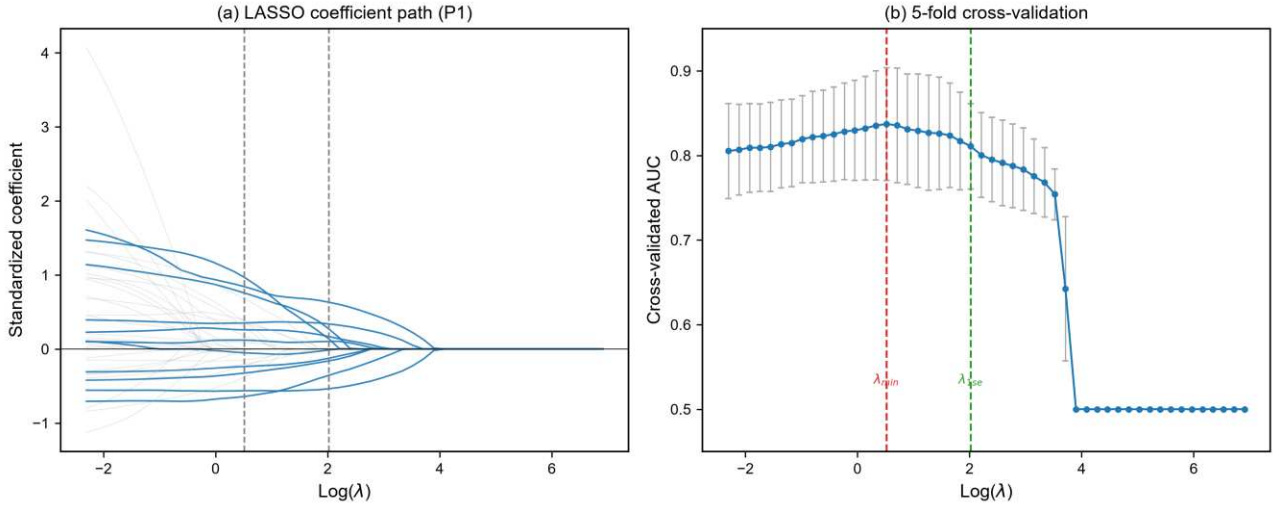


Figure 1. LASSO variable selection on the training set. (a) Coefficient path as a function of log(λ); (b) five-fold cross-validated AUC (mean ± SD across folds) against log(λ). The vertical dashed line indicates λ.1se = 7.543, at which 11 design terms were retained from the prespecified P1 candidate pool of 44 design terms.

Discrimination and machine-learning benchmark comparison

Test-set performance of all seven classifiers is summarised in Table 2 and Figure 2 (the full metric set for all seven models is given in Supplementary Table S2). XGBoost showed the highest AUC (0.959, 95% CI 0.920–0.988; accuracy 0.903, sensitivity 0.902, specificity 0.905,

F1 0.929), followed by random forest (AUC 0.945, 95% CI 0.902–0.978; accuracy 0.903, sensitivity 0.951, specificity 0.786, F1 0.933). The ridge nomogram (0.888, 95% CI 0.827–0.942) was followed by RBF-SVM (0.874, 95% CI 0.797–0.934), naïve Bayes (0.864, 95% CI 0.796–0.927), the decision tree (0.851, 95% CI 0.777–0.915) and k-NN (0.846, 95% CI 0.761–0.913). The three prespecified pairwise comparisons (XGBoost versus nomogram, random forest versus nomogram and XGBoost versus random forest) are reported in the paired-bootstrap section below and in Table 2.

Table 2. Test-set performance of seven classifiers

Model	AUC (95% CI)	Acc	Sens	Spec	F1	PPV	NPV	Brier	Cal slope	Cal int
XGBoost	0.959 (0.920–0.988)	0.903	0.902	0.905	0.929	0.958	0.792	0.071	0.773	0.591
Random forest	0.945 (0.902–0.978)	0.903	0.951	0.786	0.933	0.915	0.868	0.092	1.808	0.029
Ridge logistic (nomogram)	0.888 (0.827–0.942)	0.819	0.863	0.714	0.871	0.880	0.682	0.122	1.065	0.295
RBF-SVM	0.874 (0.797–0.934)	0.840	0.902	0.690	0.889	0.876	0.744	0.120	0.998	0.435
Naive Bayes	0.864 (0.796–0.927)	0.778	0.765	0.810	0.830	0.907	0.586	0.158	0.506	0.814
Decision tree	0.851 (0.777–0.915)	0.785	0.853	0.619	0.849	0.845	0.634	0.145	0.241	0.972
K-NN	0.846 (0.761–0.913)	0.826	0.931	0.571	0.884	0.841	0.774	0.140	1.308	-0.007

Note. All seven classifiers were fitted on the same 11 LASSO-selected admission variables without class weights. Classification metrics (Acc, Sens, Spec, F1, PPV, NPV) are at the prespecified fixed 0.5 decision threshold; AUC 95% CIs are bootstrap-based

(2000 replicates). Test set n = 144 (102 events, 42 non-events). Ridge logistic is the nomogram; the other six classifiers use prespecified default hyperparameters.

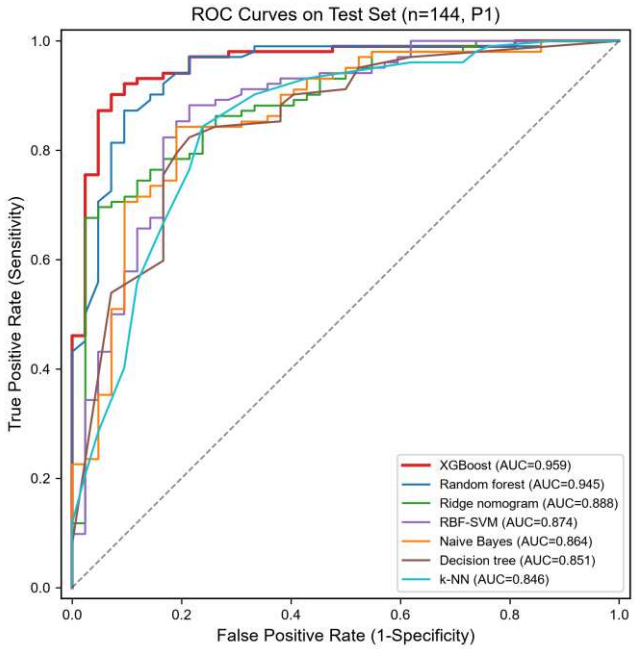
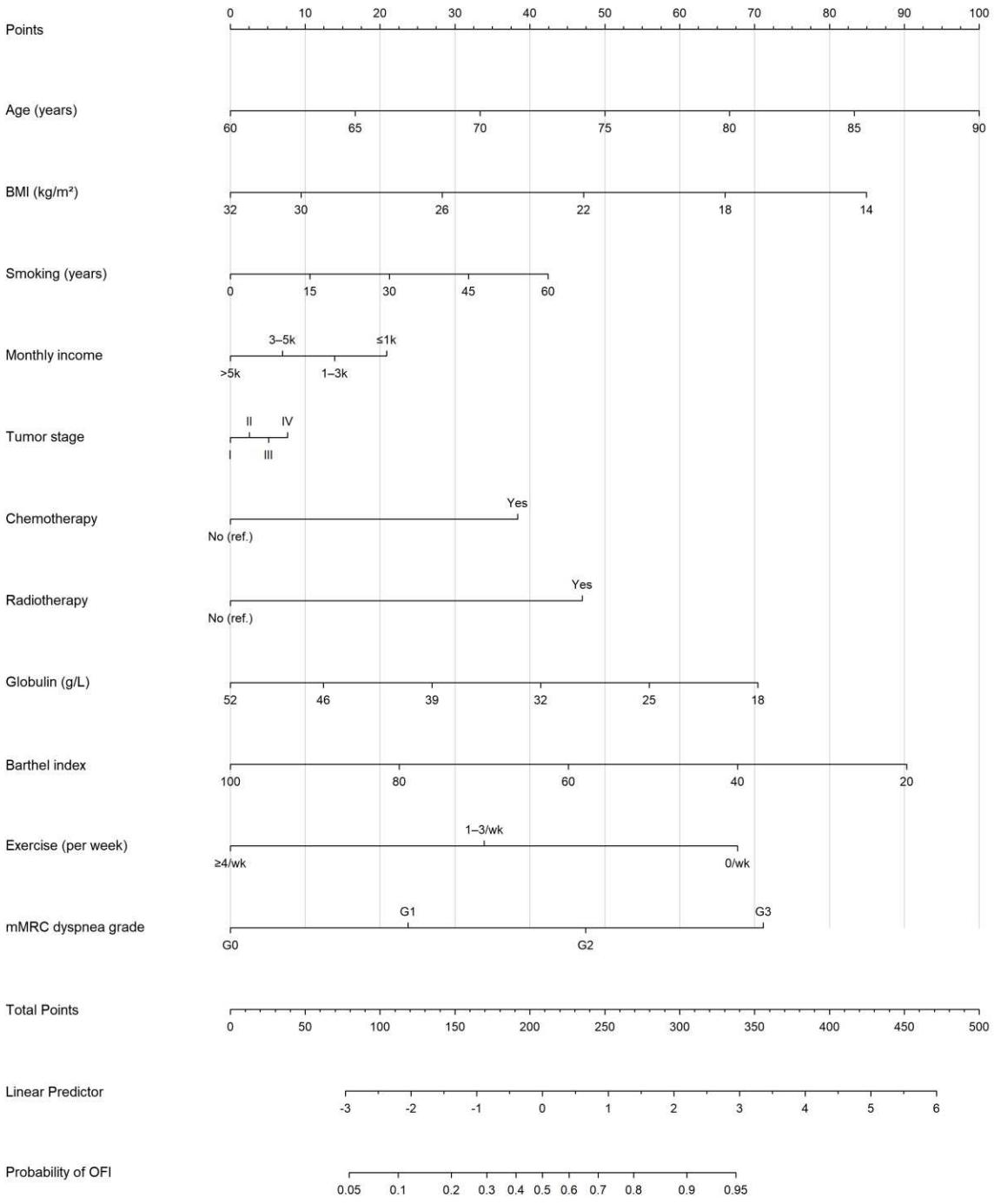


Figure 2. Receiver operating characteristic curves in the test set (n=144) for seven classifiers fitted on the same 11 LASSO-selected admission variables (the ridge nomogram plus six machine-learning algorithms); AUC 95% CIs are bootstrap-based.

Primary ridge-penalised nomogram

The final ridge model retained all 11 LASSO-selected terms, with an events-per-variable ratio of 21.5. Multicollinearity was negligible: because the laboratory and functional variables were entered as continuous only, no paired continuous–categorical terms remained, and the maximum VIF was 1.236 (weekly exercise), all well below 10. Adjusted ORs are reported in Table 3. Variables independently associated with higher odds of oral frailty included radiotherapy (OR 2.96, 95% CI 1.37–6.42; P=0.006), chemotherapy (OR 2.61, 95% CI 1.27–5.36; P=0.009), smoking duration per SD (OR 2.39, 95% CI 1.71–3.33; P<0.001) and mMRC grade per level

(OR 1.60, 95% CI 1.05–2.44; P=0.028). Age per SD (OR 1.45, 95% CI 0.99–2.12; P=0.058) was borderline. Variables independently associated with lower odds included Barthel index per SD (OR 0.45, 95% CI 0.22–0.93; P=0.031), weekly exercise per level (OR 0.52, 95% CI 0.36–0.73; P<0.001), globulin per SD (OR 0.70, 95% CI 0.50–1.00; P=0.050), BMI per SD (OR 0.71, 95% CI 0.50–1.01; P=0.055) and monthly income per level (OR 0.88, 95% CI 0.64–1.20; P=0.412). Tumour stage (OR 1.09, 95% CI 0.80–1.49; P=0.584) was not statistically significant in the multivariable model, and polypharmacy, education and the six oral variables were not retained. The nomogram visualising these contributions is shown in Figure 3; its 11 axes span the full range observed in the development cohort.


Nomogram for Predicting Oral Frailty (P1 model, 11 predictors)
Points
0 10 20 30 40 50 60 70 80 90 100
Age (years)
60 65 70 75 80 85 90
BMI (kg/m²)
32 30 26 22 18 14
Smoking (years)
0 15 30 45 60
Monthly income
3–5k ≤1k
>5k 1–3k
Tumor stage
II IV
I III
Chemotherapy
Yes
No (ref.)
Radiotherapy
Yes
No (ref.)
Globulin (g/L)
52 46 39 32 25 18
Barthel index
100 80 60 40 20
Exercise (per week)
1–3/wk
≥4/wk 0/wk
mMRC dyspnea grade
G1 G3
G0 G2
Total Points
0 50 100 150 200 250 300 350 400 450 500
Linear Predictor
-3 -2 -1 0 1 2 3 4 5 6
Probability of OFI
0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95

384

Figure 3. R rms nomogram of the ridge-penalised logistic regression model (11 predictors) for predicting oral frailty. Continuous predictors are plotted on their native clinical scales—age (years), BMI (kg/m²), cumulative smoking duration (years), globulin (g/L) and Barthel index (points); ordinal predictors (monthly income, tumour stage I–IV, weekly exercise frequency, mMRC grade 0–3) and binary predictors (chemotherapy, radiotherapy) appear as categorical ticks, with reference categories labelled. Summing the points from each axis gives a total score that maps to the predicted probability of oral frailty (0.05–0.95).

Table 3. Ridge-penalised logistic regression coefficients and adjusted odds ratios (OR)

Variable	β	SE	OR (95% CI)	P value
Age (per SD)	0.371	0.195	1.45 (0.99–2.12)	0.058
BMI (per SD)	-0.338	0.177	0.71 (0.50–1.01)	0.055
Smoking duration (per SD)	0.870	0.169	2.39 (1.71–3.33)	<0.001
Monthly income (per level)	-0.131	0.160	0.88 (0.64–1.20)	0.412
Tumour stage (per level)	0.086	0.158	1.09 (0.80–1.49)	0.584
Chemotherapy (yes vs no)	0.958	0.368	2.61 (1.27–5.36)	0.009
Radiotherapy (yes vs no)	1.086	0.394	2.96 (1.37–6.42)	0.006
Globulin (per SD)	-0.350	0.178	0.70 (0.50–1.00)	0.050
Barthel index (per SD)	-0.800	0.371	0.45 (0.22–0.93)	0.031
Weekly exercise (per level)	-0.661	0.179	0.52 (0.36–0.73)	<0.001
mMRC grade (per level)	0.472	0.215	1.60 (1.05–2.44)	0.028
Intercept	0.513	—	—	—

Note. ORs, 95% CIs and P values were derived from the L2-ridge-penalised fit (best inverse-regularisation strength C=6.866; λ=0.146, tuned by five-fold cross-validation) with penalised standard errors. Because the L2 penalty shrinks coefficients toward zero, the penalised SEs, CIs and P values are descriptive quantities rather than unbiased maximum-likelihood inferences and should be interpreted cautiously; the model is intended primarily for risk stratification. Continuous variables are per-SD increments and ordinal variables are per-level increments; the binary chemotherapy and radiotherapy terms are coded yes vs no.

Nomogram discrimination

The primary nomogram achieved a training AUC of 0.867 (95% CI 0.819–0.908) and a held-out test AUC of 0.888 (95% CI 0.827–0.942) (Figure 4). At the prespecified 0.5 threshold on the test set, accuracy was 0.819, sensitivity 0.863, specificity 0.714, F1 0.871, PPV 0.880 and NPV 0.682. At the Youden-optimal threshold (maximising sensitivity plus specificity), accuracy was 0.833, sensitivity 0.873, specificity 0.738, F1 0.881, PPV 0.890 and NPV 0.705.

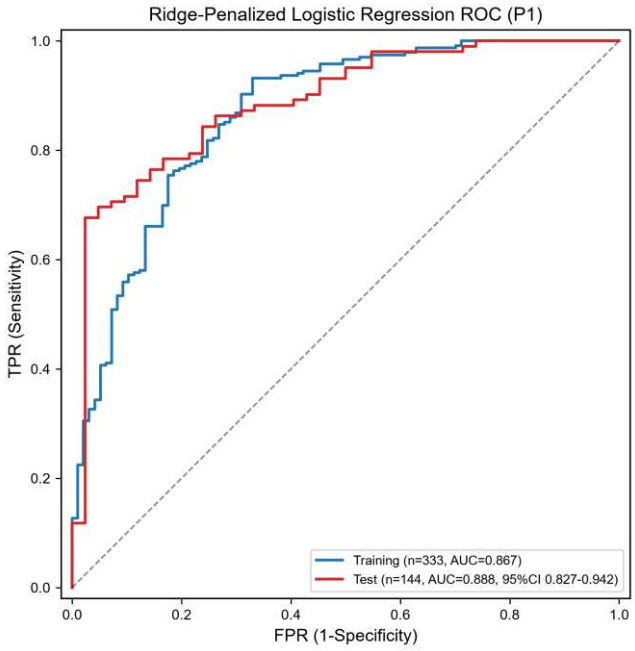


Figure 4. ROC curves of the primary ridge-penalised nomogram (11 LASSO-selected admission variables) in the training set (AUC=0.867) and test set (AUC=0.888, 95% CI 0.827–0.942).

Calibration

Apparent training calibration of the nomogram showed a slope of 0.995, intercept of 0.002 and Brier score of 0.122; 1000-bootstrap bias correction yielded a slope of 0.979 and intercept of 0.001, indicating that optimism in training calibration was minimal. On the held-out test set, the

calibration slope was 1.065, intercept 0.295 and Brier score 0.122 (Figure 5). The test slope was close to unity (1.065), indicating near-ideal calibration. The persistent positive intercept (0.295), however, indicates systematic underestimation of absolute risk in unseen patients and warrants recalibration-in-the-large before use elsewhere. The +0.295 value is specific to this test set, and each new setting requires an intercept estimated from its own case-mix and prevalence rather than a portable correction. The Brier score of 0.122 remains well below the no-information Brier score of approximately 0.21 given the 71% event rate. Slopes below and above 1 denote over-confident and under-confident predictions, respectively. Both ensembles had calibration slopes farther from 1 than the nomogram (1.065; random forest 1.808, XGBoost 0.773). The random-forest intercept (0.029) was closer to zero than the nomogram's (0.295), whereas XGBoost had the largest positive intercept (0.591). The nomogram thus retained the slope closest to the ideal value of 1 but showed a persistent positive intercept drift.

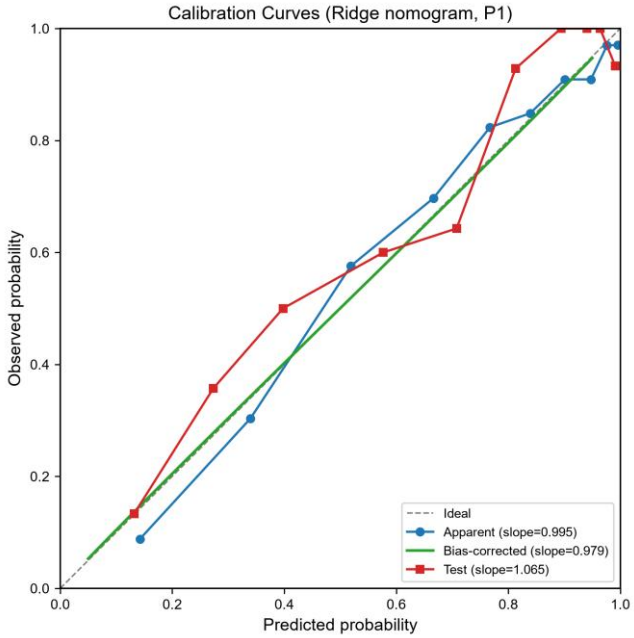


Figure 5. Calibration curves of the nomogram. Apparent (training), 1000-bootstrap bias-corrected, and test-set (loess) calibration are shown alongside the 45° ideal line. Test-set calibration slope=1.065, intercept=0.295, Brier=0.122.

SHAP interpretation of XGBoost

SHAP analysis of XGBoost (Figure 6) ranked the leading predictors by mean absolute SHAP value as smoking duration (1.763), mMRC dyspnoea grade (0.947), weekly exercise (0.898), globulin (0.718) and chemotherapy (0.641). Longer smoking duration, higher mMRC grade, receipt of chemotherapy and radiotherapy, older age and higher stage shifted predicted risk upward, whereas more frequent exercise, higher globulin, higher BMI, higher income and higher Barthel index shifted it downward. The directions of effect for all retained predictors were consistent with the corresponding ridge ORs (chemotherapy: mean SHAP +0.46 in treated vs −1.15 in untreated, consistent with its ridge OR of 2.61). Because SHAP depicts marginal effects within a non-linear ensemble rather than adjusted logistic coefficients, this directional

concordance indicates that the transparent nomogram and the machine-learning model converge on the same underlying risk structure.

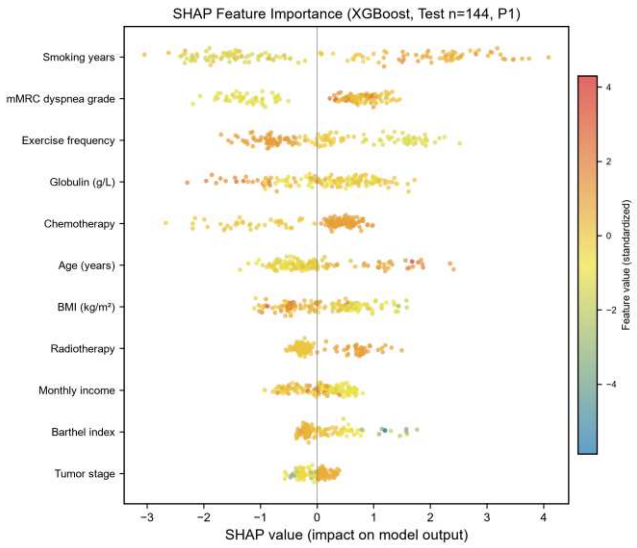


Figure 6. SHAP beeswarm plot of the XGBoost model, computed on the test set (n=144), showing the top features by mean absolute SHAP value (smoking duration, mMRC grade, weekly exercise, globulin, chemotherapy). Feature value is colour-coded (red=high, blue=low).

Clinical utility (decision-curve analysis)

DCA (Figure 7) showed that, across a broad range of threshold probabilities encompassing clinically plausible decision points, the nomogram yielded greater net benefit than either the treat-all or treat-none reference strategy. At the observed prevalence of approximately 71%, the treat-all strategy gave a high baseline net benefit; the nomogram's incremental benefit arose from targeting patients most likely to benefit while sparing low-risk patients from unnecessary procedures. Net benefit at key thresholds is summarised in Supplementary Table S6: at threshold probabilities of 0.30, 0.50 and 0.70, the nomogram's net benefit was 0.605, 0.528 and 0.433, respectively. This exceeded the treat-all reference (0.583, 0.417 and 0.028) at each of these thresholds, whereas XGBoost attained 0.646, 0.611 and 0.569. The nomogram's net benefit was

close to treat-all at the lowest threshold (0.10: 0.679 vs 0.676). It diverged increasingly from treat-all as the threshold rose, indicating that the model's incremental value is greatest when clinicians are willing to withhold intervention from lower-risk patients. A decision threshold of 0.30–0.50 corresponded to the range over which the nomogram showed a consistent net-benefit advantage.

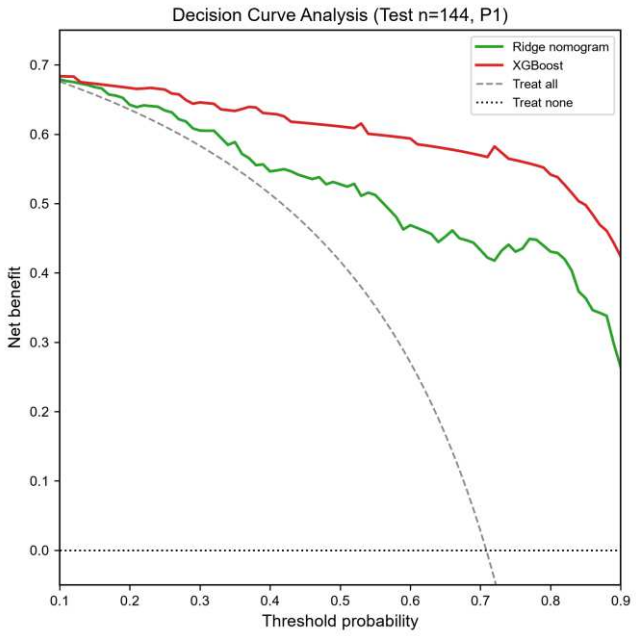


Figure 7. Decision curve analysis in the test set. The four curves are the ridge nomogram, XGBoost, the treat-all strategy (dashed) and the treat-none strategy (dotted); net benefit is plotted against threshold probability.

Head-to-head paired-bootstrap comparison

Paired bootstrap of the AUC difference (2000 replicates) showed that XGBoost outperformed the nomogram by ΔAUC +0.071 (95% CI 0.025–0.124), which was statistically significant after Holm correction (P_raw=0.0065, Holm P=0.020; Figure 8a). Random forest outperformed the nomogram by a ΔAUC of +0.057 (95% CI 0.019–0.101; P_raw=0.010, Holm P=0.020; Figure 8b). XGBoost and random forest, by contrast, did not differ significantly (ΔAUC +0.014, 95%

CI −0.004–0.037; P=0.156; Figure 8c). Bootstrap CIs were not multiplicity-adjusted; the reported Holm P values correct for the three prespecified comparisons.

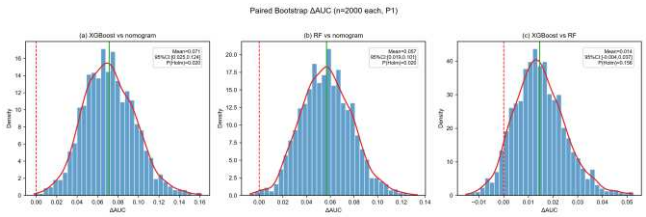


Figure 8. Paired bootstrap distributions (2000 replicates) of ΔAUC for the three prespecified comparisons. (a) XGBoost versus nomogram: ΔAUC=+0.071, 95%CI 0.025–0.124, P_raw=0.0065, Holm P=0.020. (b) Random forest versus nomogram: ΔAUC=+0.057, 95%CI 0.019–0.101, P_raw=0.010, Holm P=0.020. (c) XGBoost versus random forest: ΔAUC=+0.014, 95%CI −0.004–0.037, P=0.156. The red dashed line in each panel indicates no difference (ΔAUC=0).

Sensitivity analyses

S_oral (reintroducing six oral variables). The principal contrast differed by model class. When the six oral variables were added to the P1 pool (50 design terms) and XGBoost was refitted, the test-set AUC changed from 0.959 to 0.950 (95% CI 0.900–0.986), an absolute ΔAUC of approximately −0.009 (paired-bootstrap 95% CI −0.065 to 0.040, containing zero; relative change ≈ −0.9%). Adding the items therefore did not reduce the discrimination of the non-linear model, while avoiding circular reasoning and outcome leakage. For the transparent model, the symmetric arm that re-ran the identical main-model LASSO→ridge pipeline over the full 50-term pool served as the primary test: LASSO at λ~1se~=5.18 retained 18 design terms (five of the six oral items). The ridge test AUC then rose from 0.888 to 0.968 (95% CI 0.940–0.987), an absolute ΔAUC of +0.080 (relative +9.0%), with the retained oral items carrying large positive coefficients (ORs 12.6–26.3 for denture use, xerostomia and masticatory difficulty), as expected for items overlapping the OFI-8 definition. A supporting conditional arm, in which the original

11 terms were forced in and the six oral terms added before re-running LASSO over the 17-term design, retained five oral items and yielded a ridge AUC of 0.969 (Δ=+0.081; +9.1%), with ORs of 4.8–7.1 for denture use, xerostomia and masticatory difficulty. This contrast—no gain for XGBoost but a marked mechanical inflation for the transparent linear model—indicates that the oral items act as near-outcome variables for that model. A per-item mapping of the six variables to the OFI-8 domains, their crude association with the outcome and their symmetric-arm ridge coefficients is provided in Supplementary Table S8.

S_lifestyle (reintroducing taste and drinking variables). When dietary taste dummies, drinking history and drinking duration were reintroduced (51 design terms) and LASSO, ridge and XGBoost were refitted under identical settings, LASSO at λ~1se~=5.179 retained 17 features. The ridge model's test AUC decreased from 0.888 to 0.874 (absolute Δ≈−0.014), and the XGBoost AUC changed from 0.959 to 0.949 (Δ≈−0.010). Several reintroduced coefficients ran counter to biological expectation: the "moderate taste vs bland" coefficient yielded an OR≈0.28 (OR<1), and the drinking-history coefficient yielded an OR≈0.10 (OR<1). The former is consistent with reverse causation—patients with established oral frailty or treatment-related oral toxicity switch to bland, soft diets—whereas the latter is consistent with the sick-quitter effect—patients quit drinking after illness. Collinear suppression between drinking history and drinking duration further destabilised these coefficients. Because these variables contribute near-zero or negative discrimination and exhibit reverse-causation or sick-quitter bias, they were excluded from the main model and are reported only in this sensitivity analysis (Supplementary Table S5). They remain fully described in Table 1.

S_ses (reintroducing education). Education was distributed across four levels (primary or below n=101; junior high n=213; senior high/vocational n=117; college or above n=46), with

oral-frailty prevalences of 75.2%, 74.6%, 61.5% and 67.4%, respectively. Its adjusted odds per SD changed markedly across covariate sets: unadjusted OR 0.90 (P=0.375); adjusted for age and sex OR 0.86; after addition of the socioeconomic block (income, occupation, residence, marital status) OR 1.28; and in the fully adjusted unpenalised model OR 1.71 (95% CI 1.14–2.58; P=0.010), with a descriptive ridge OR of 1.70. In the 45-term LASSO pool (which retained education), education was selected in 86 of 120 bootstrap resamples (71.7%), yet adding it to the 11-term model changed test AUC by only −0.003. The sign reversal after socioeconomic adjustment is the signature of a suppression/Simpson effect, and the near-zero independent discrimination increment supports the a priori exclusion of education from the prespecified primary model (Supplementary Table S4).

Discussion

Principal findings and comparison with the literature

In this single-centre cohort of 477 older inpatients with lung cancer, oral frailty was highly prevalent (70.9%). A transparent ridge-penalised nomogram built from 11 routine admission variables, none overlapping the OFI-8 diagnostic items, provided good discrimination (test AUC 0.888) with a near-ideal calibration slope (1.065) but a positive intercept drift (0.295). The two tree ensembles achieved modestly higher AUCs and significantly outperformed the nomogram after Holm correction, yet did not differ from each other and were less well calibrated; the printable, interpretable nomogram is therefore preferred for bedside risk stratification. The central methodological finding is that the circularity claimed in earlier work is here quantified rather than asserted: in a symmetric re-run, reintroducing the diagnostic oral items mechanically inflated the transparent linear model by approximately 9% while leaving the non-linear model

essentially unchanged. Together, these results support the nomogram as a risk-stratification rather than an absolute-risk tool and provide an empirical estimate of the bias that arises when near-outcome items serve as predictors.

The observed prevalence of approximately 71% is consistent with cross-sectional reports in older hospitalised patients with cancer [20] and Chinese population surveys[21], and is higher than the pooled prevalence of 51% (95% CI 47–55%) reported in a recent meta-analysis of 27 studies enrolling 11,575 older inpatients[41]. Even the cancer subgroup in that meta-analysis (62%, 95% CI 57–66%) was lower than that in the present cohort, possibly reflecting the treatment-related oral burden of lung cancer and the composition of the study centre. The test AUC of 0.888 is comparable to the community-dwelling oral-frailty prediction model of Lv et al[42]. and to risk models for cognitive frailty in lung cancer [43] and for oral frailty in oesophageal cancer [44]. The present model benefits from an explicit circularity control that these earlier models did not implement. However, these cross-cohort AUC comparisons are indicative only because populations, predictors and case-mix differ. The leading predictors—long smoking history, chemotherapy, radiotherapy, lower BMI and nutritional markers and reduced physical activity—are biologically plausible and echo associated factors identified in systematic reviews[45, 46] and in oncology nutrition guidance[47], with smoking–periodontal pathways well established [48]. Treatment-related oral toxicity from chemotherapy and radiotherapy, together with immunotherapy adverse effects [4, 5], plausibly concentrates risk in treated patients and explains the large ORs for chemotherapy and radiotherapy. Finally, although both tree ensembles showed higher point-estimate AUCs than the nomogram, the absolute gains were modest (ΔAUC ≈0.014–0.071). The ensembles did not differ from each other, so these gains may not outweigh the practical advantages of a printable, interpretable nomogram at the bedside.

Having established how the model performs relative to prior work, the discussion turns to why these predictors emerged and where the evidence remains limited.

Interpretation of key predictors

Radiotherapy was most strongly associated with higher odds of oral frailty (adjusted OR 2.96), plausibly through salivary-gland dysfunction and fibrosis[49], and was more frequent in the oral-frailty group on unadjusted comparison (36.7% vs 19.4%); the crude and adjusted directions were therefore concordant. Smoking duration (OR 2.39 per SD) is a known driver of periodontal breakdown and tooth loss [50]. Chemotherapy (OR 2.61) likely reflects chemotherapy-induced xerostomia, mucositis and reduced oral intake [6, 51]. Older age per SD (OR 1.45) was borderline, whereas mMRC dyspnoea grade (OR 1.60 per level) was independently associated with higher odds. Conversely, weekly exercise (OR 0.52 per level), higher Barthel index (OR 0.45 per SD) and higher BMI (OR 0.71 per SD) were independently associated with lower odds, consistent with preserved functional reserve and nutritional status. Higher globulin (OR 0.70 per SD) and higher monthly income (OR 0.88 per level) were also independently associated with lower odds. Education was retained only in the S_ses sensitivity analysis, where its association proved unstable across socioeconomic adjustment (see below) and contributed near-zero independent discrimination.

As with education, dietary taste dummies, drinking history and drinking duration were deliberately excluded from the primary model. In the S_lifestyle sensitivity analysis, these variables lowered the ridge AUC by ≈0.014 and the XGBoost AUC by ≈0.010, yet produced clinically unreasonable inverse coefficients (moderate-taste vs bland OR≈0.28; drinking-history OR≈0.10). Three patterns together explain why they were excluded rather than interpreted as causal. First, a crude unadjusted association for a salty taste was in the expected direction. Once

579 smoking, occupation, income and other systemic confounders were adjusted for, the apparent
580 effect dissolved, indicating that the crude taste association was carried by these broader
581 confounders rather than by taste itself. Second, the moderate-vs-bland coefficient pointed
582 opposite to biological expectation (OR≈0.28). This is most plausibly explained by reverse
583 causation: patients with established oral frailty or treatment-related oral toxicity switch to bland,
584 soft diets, so that a non-bland taste appears "protective" purely because of post-outcome dietary
585 adaptation. Third, drinking showed essentially no crude association with oral frailty, yet after the
586 binary drinking flag and the continuous drinking duration were entered together the adjusted
587 drinking-history coefficient changed sign, yielding OR≈0.10. This sign reversal is the signature
588 of collinear suppression compounded by the sick-quitter effect, in which patients quit drinking
589 after developing cancer or oral symptoms, so that ever-drinking appears "protective" relative to
590 illness-driven abstinence. Table 1 continues to describe the distribution of these variables by
591 oral-frailty status, but they are not used as nomogram predictors.

592 Education deserves separate comment. In the S_ses sensitivity analysis, its unadjusted
593 association was slightly protective (OR 0.90, P=0.38), but after adjustment for the full
594 socioeconomic block it reversed (OR 1.28) and became a statistically significant positive
595 association in the fully adjusted model (OR 1.71, P=0.01). This sign reversal across covariate
596 sets is the hallmark of a suppression/Simpson effect in which education acts as a proxy for
597 correlated socioeconomic resources rather than as an independent risk factor. Adding education
598 to the 11-term model changed the AUC by only −0.003. A predictor whose direction depends on
599 the adjustment set and that adds near-zero discrimination is unsuitable for a prespecified bedside
600 tool, and it was therefore excluded a priori, with its distribution reported in Table 1 and its
601 sensitivity analysis in Supplementary Table S4. Because all of these associations are cross-

sectional, the retained predictors are described throughout as "independently associated with" higher or lower odds rather than as risk or protective factors. Only predictors with a clear temporal ordering or mechanistic basis (smoking duration, chemotherapy and radiotherapy, age, and the functional and nutritional markers) are discussed in causal terms.

Machine learning versus the transparent nomogram

Although both tree ensembles achieved higher point-estimate AUCs than the nomogram, three considerations temper this finding. The gains were modest (ΔAUC +0.071 for XGBoost, +0.057 for random forest) and, although both reached Holm-corrected significance (P=0.020), the ensembles did not differ from each other (ΔAUC +0.014, P=0.156), consistent with evidence that ML offers little consistent advantage over logistic regression in small clinical datasets [29]. Their shared benefit reflects flexible non-linear modelling rather than any single algorithm. A high AUC does not imply usable probabilities: the ensemble calibration slopes (random forest 1.808, XGBoost 0.773) were both further from the ideal value of 1 than the nomogram's (1.065). Tree ensembles also require computational infrastructure and are neither printable nor auditable at the bedside, whereas SHAP confirmed that their dominant drivers are the same familiar variables (smoking duration, mMRC, exercise, globulin, chemotherapy). The ridge nomogram is therefore recommended as the deployable risk-stratification tool, with the ensembles serving as benchmarks. This pattern is sensitive to the prespecified variable set. The ML classifiers used untuned default hyperparameters, so tuned ML might perform differently, and although XGBoost showed higher decision-curve net benefit, its poorer calibration renders that estimate optimistic.

Methodological strengths

The principal methodological contribution lies in operationalising, rather than merely stating, the separation of screening from diagnosis. Excluding a priori the six oral-health variables that overlap directly with the outcome-defining oral items constrains the nomogram to admit only information recorded before any dedicated oral examination, so that outcome leakage is removed by construction rather than corrected after the fact. The circularity that this exclusion guards against is quantified rather than asserted: the symmetric S_oral arm re-ran the identical LASSO-to-ridge pipeline on the enlarged predictor pool and, in a head-to-head contrast with the identically refitted non-linear ensemble on the same enlarged pool, showed that the diagnostic items inflate the transparent linear model to a marked degree while leaving the non-linear model essentially unchanged. This makes the cost of circularity measurable and model-specific, and offers a transferable, reproducible template—screening separated from diagnosis, circularity quantified rather than assumed—for oral-frailty and related hospital-based prediction models.

A second contribution is an equal-condition comparison across all fitted classifiers that clarifies the trade-off between discrimination, calibration, interpretability and deployability. Under identical predictors, identical train/test splits and untuned default hyperparameters, the tree ensembles achieved modestly higher point-estimate AUCs but showed calibration slopes farther from unity and required computational infrastructure that prevents printing or bedside audit; the transparent ridge nomogram, although slightly less discriminating, retained the calibration slope closest to the ideal value and remains printable and auditable. This pattern empirically supports the wider observation that a high AUC does not by itself imply usable, calibrated probabilities [27, 29], and it aligns with the TRIPOD+AI emphasis on transparent reporting and on calibration alongside discrimination [24].

Third, the prespecified S_lifestyle and S_ses arms operationalise the management of variable-credibility problems common in observational prediction research but seldom reported: reintroducing dietary-taste and drinking variables yielded counter-intuitive inverse coefficients consistent with reverse causation through post-outcome dietary adaptation and with the sick-quitter pattern, whereas reintroducing education produced a sign reversal after socioeconomic adjustment—the signature of a suppression or Simpson effect—with a negligible increment in independent discrimination. Documenting these arms alongside the primary model makes the a priori exclusions transparent and reproducible. The analysis followed TRIPOD+AI [24], with a final events-per-variable ratio of 21.5, ridge stabilisation [35, 36], seven classifiers compared without class weights and Holm correction of three prespecified comparisons. Entering laboratory and functional variables as continuous only removed continuous–categorical collinearity (maximum VIF 1.24), while calibration, Brier score, decision-curve analysis [39] and SHAP [40] supplemented discrimination within a transparent, printable framework [28].

Limitations

The above estimates and conclusions are qualified by the following constraints, which follow from the single-centre retrospective design, the modest held-out sample and the self-reported oral-frailty outcome.

1. **Single-centre design and absence of external validation.** This was a single-centre retrospective study with internal bootstrap validation only. No independent external or temporal validation was available, and the model cannot be generalised directly to other institutions, regions or populations. Discrimination and calibration must be re-estimated prospectively at each new site, with site-specific recalibration-in-the-large expected given the positive test intercept of

0.295. Multicentre external validation across regions and hospital levels, together with temporal validation, is the highest-priority next step.

2. **Small test set, low selection-stage EPV and optimism.** The test set contained only 144 patients (102 events, 42 non-events), yielding wide AUC confidence intervals. With only 42 non-events, specificity and NPV estimates were unstable (for example, k-NN specificity 0.571). Although the final EPV was 21.5, the selection-stage EPV (236/44≈5.4) was low, and selection followed by refitting can remain optimistic despite the 120-bootstrap stability assessment. Nested cross-validation with bootstrap optimism correction and .632+ estimation would be preferable.

3. **Descriptive penalised estimates and sample-size basis.** The ridge-penalised ORs, SEs, CIs and P values are descriptive rather than unbiased maximum-likelihood inferences. An unpenalised multivariable fit is reported in Supplementary Table S7, and bootstrap percentile CIs were not computed. Sample-size adequacy was gauged from the EPV criterion rather than a formal Riley calculation.

4. **Self-reported outcome and absent objective measures.** Oral frailty was ascertained by self-reported OFI-8 rather than objective examination. Comprehension and cognitive limitations may introduce reporting bias, acute treatment-related xerostomia may distort long-term self-report, and the absence of objective measures (plaque index, bite force, stimulated salivary flow, tongue pressure, masticatory efficiency) may leave residual confounding. The database retained only the binary OFI-8≥4 outcome and not the item-level scores of the eight components, precluding item-response analyses.

5. **Retrospective observational design and selection bias.** Data were extracted retrospectively from electronic medical records, entailing information bias and complete-case selection bias; as in any observational study, the design precludes causal inference.

6. **Fixed decision threshold.** Classification metrics were reported principally at a fixed 0.5 threshold, with Youden-optimal values added and no cost-weighted threshold chosen; at the observed prevalence of 70.9%, the PPV and NPV at 0.5 must be interpreted cautiously.

7. **Short recruitment window and untuned ML hyperparameters.** The eight-month window (September 2024–April 2025) may not capture seasonal variation in oral symptoms, and machine-learning hyperparameters were fixed at published defaults rather than tuned.

Clinical and nursing implications

The nomogram is designed for use within 24 hours of admission, before the formal oral examination, and can be embedded in the electronic health record to support nurse-led risk stratification [52-54]. On admission, the responsible ward nurse completes the admission assessment record, into which the eleven nomogram variables—age, BMI, smoking duration, monthly income, tumour stage, receipt of chemotherapy and radiotherapy, serum globulin, Barthel index, weekly exercise and mMRC dyspnoea grade—are entered. All are already part of the routine admission data set, so no additional measurement is required. A built-in calculator then returns the predicted probability. On the basis of the decision-curve thresholds tabulated in Supplementary Table S6, patients are triaged into three strata. Those at low risk (predicted probability <0.30) receive routine oral-health education and self-care guidance [55]. Those at intermediate risk (0.30–0.70) receive structured oral-health education and a repeat assessment at discharge. Those at high risk (≥0.70) are prioritised for dental consultation, nutritional support and targeted oral-care protocols, with reassessment at discharge or sooner if the clinical status changes. The model also supports resource allocation during periods of high bed occupancy. Because the tool uses variables already collected on admission, it adds no documentation burden

and is well suited to the nursing-led, whole-person care model increasingly emphasised in geriatric oncology.

Worked example (illustrative, not a real patient). To illustrate bedside use, consider a 72-year-old patient with stage III lung cancer, BMI 22 kg/m², 30 years of smoking, monthly income in the 1–3 k RMB band, who has received chemotherapy but not radiotherapy, serum globulin 30 g/L, Barthel index 90, weekly exercise 1–3 times and mMRC dyspnoea grade 1. Reading down each axis of Figure 3, the individual points sum to approximately 270 on the Total Points scale. This corresponds to a linear predictor of approximately 1.8 and, on the Probability axis, a predicted probability of approximately 0.86. Direct substitution of the locked ridge coefficients confirms the arithmetic: after z-standardisation of the continuous and ordinal variables, the log-odds is $0.513 + \Sigma(\beta \cdot z) \approx 1.82$, and $1/(1+e^{(-1.82)}) \approx 0.86$, agreeing with the figure reading within rounding. Because this exceeds the 0.70 threshold, the patient falls in the high-risk stratum and would be prioritised for dental consultation, nutritional support and targeted oral care. A low-risk patient (probability <0.30) would instead receive routine oral-health education and self-care guidance. This example is illustrative only and does not correspond to any individual in the cohort.

These bedside properties carry implications for health-service organisation that remain to be demonstrated in practice. As populations age and the lung-cancer burden concentrates in older adults [1], and given the high prevalence of oral frailty observed among hospitalised older patients [41], a nurse-led, admission-day risk stratification that uses variables already recorded on the routine admission record may help direct limited dental, nutritional and oral-care resources toward the patients most likely to benefit, and may support closer collaboration among ward nursing, dietetics and dental services [54, 55]. Because oral frailty is linked to malnutrition

[10], to treatment-related oral toxicity [4-6] and to broader physical and cognitive decline [11-14], earlier identification has the potential to reduce preventable complications during chemotherapy and radiotherapy, although such benefit is yet to be evaluated in an interventional study. The printable, no-additional-measurement design is also relevant to settings with limited electronic-infrastructure capacity, where a non-auditable ensemble would be difficult to deploy; whether this improves equity of access to oral-care screening across hospital levels remains to be examined.

The generalisability of these expectations is explicitly limited. At present the nomogram is derived and internally validated in a single tertiary cancer centre, and its positive calibration intercept means that absolute-risk outputs require site-specific recalibration before use elsewhere. Pending multicentre external and temporal validation and recalibration, the broader framework—the leakage-free predictor set, the symmetric circularity check and the nurse-led triage logic—may be adaptable to other cancer populations in which treatment-related oral toxicity is prominent, to community-dwelling older adults, and to hospitals in other regions. Any such extension should be regarded as a future research direction rather than as a demonstrated benefit of the present model.

Future directions

Future work should adopt a multicentre design spanning community and tertiary settings across regions and hospital levels, with prospective external validation, temporal validation and an interventional evaluation of screening-driven referral. Incorporating objective oral measures, item-level OFI-8 scores, longitudinal follow-up and multiple imputation would further refine the model, and a web-based calculator alongside the printable nomogram would facilitate point-of-care use.

Conclusions

Among older inpatients with lung cancer, a ridge-penalised nomogram derived exclusively from routine admission variables—and deliberately excluding the diagnostic oral-frailty items—provides a transparent, printable and well-calibrated means of stratifying oral-frailty risk at admission, before any dedicated oral examination. Because every predictor is recorded on admission and these diagnostic items are removed a priori, screening is separated from diagnosis by design, avoiding the circular reasoning and outcome leakage that inflate published tools. The calibration slope is close to unity, whereas the tree ensembles, although slightly more discriminating, are less well calibrated and cannot be printed or audited at the bedside; the ensembles are therefore retained as performance benchmarks rather than as deployable tools. The persistent positive calibration intercept indicates that the absolute-risk output is descriptive of this internal cohort and must be recalibrated at each new site before use. Empirically, reintroducing the diagnostic oral items inflated the transparent model by approximately 9% while leaving the non-linear model essentially unchanged, which quantifies rather than presupposes the circularity that the a priori exclusion guards against and offers a transferable safeguard for screening-model development.

The following work, in priority order, is required before bedside deployment. First, multicentre external validation across regions and hospital levels, supplemented by temporal validation on later admission cohorts. Second, site-specific recalibration-in-the-large (intercept adjustment, with slope re-estimation where needed). Third, nested cross-validation with bootstrap optimism correction and .632+ estimation, given the small test set and low selection-stage events-per-variable ratio. Fourth, incorporation of objective oral measures (plaque index, bite force, stimulated salivary flow and tongue pressure) alongside the self-reported oral-frailty assessment.

Fifth, development of a deployable web calculator that embeds the nomogram for nurse-led risk stratification.

Taken together, the study contributes a reproducible, leakage-free template for screening-model development in oral frailty and, more broadly, for transparent prediction in hospitalised older patients; its bedside benefit, and its transferability to other populations, remain contingent on the external validation and recalibration steps outlined above.

Declarations

Abbreviations

AUC, area under the receiver-operating characteristic curve; BMI, body mass index; CI, confidence interval; DCA, decision-curve analysis; DT, decision tree; EPV, events-per-variable; k-NN, k-nearest neighbours; LASSO, least absolute shrinkage and selection operator; ML, machine learning; mMRC, modified Medical Research Council; NB, naïve Bayes; OFI-8, Oral Frailty Index-8; OR, odds ratio; RF, random forest; SHAP, SHapley Additive exPlanations; RBF-SVM, radial-basis-function support-vector machine; TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; VIF, variance inflation factor; XGBoost, Extreme Gradient Boosting.

Availability of data and materials

The datasets generated and analysed during the current study are available from the corresponding author on reasonable request and subject to institutional data-governance approval.

Ethics approval and consent to participate

The study was approved by the Institutional Ethics Committee of Guangxi Medical University Cancer Hospital (approval no. KY2024892, 20 December 2024). The ethics committee approved retrospective use of de-identified, routinely collected data from patients admitted since September 2024. The requirement for written informed consent was waived, and all analyses were conducted after this approval, including the waiver of consent.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

Zhihua Fan: Conceptualization, Investigation, Data curation, Formal analysis, Visualization, Writing – original draft. Rongquan Wei: Conceptualization, Methodology, Supervision, Project administration, Writing – review & editing. Guijie Ren: Investigation, Data curation, Resources. Huiling Wang: Investigation, Data curation, Formal analysis, Validation. Xuemei Chen: Investigation, Data curation, Resources. Xinghong Liu: Investigation, Data curation, Resources. All authors read and approved the final manuscript and agree to be accountable for all aspects of the work.

Acknowledgements

The authors thank the nursing and medical records staff of the participating centre for data collection support.

References

1. Sung H, Filho AM, Laversanne M, Ferlay J, Siegel RL, Soerjomataram I, Jemal A, Bray F: **Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 34 cancers in 186 countries**. *CA Cancer J Clin* 2026, **76**(4):e70090.
2. Liu H, Chen Y, Wang S: **Frailty as a Key Clinical Indicator in Older Patients with Advanced Lung Cancer: A Systematic Review and Meta-Analysis of Prevalence, Risk Factors, and Clinical Implications**. *J Palliat Med* 2026:10966218261439657.
3. McGovern J, Dolan RD, Skipworth RJ, Laird BJ, McMillan DC: **Cancer cachexia: a nutritional or a systemic inflammatory syndrome?** *Br J Cancer* 2022, **127**(3):379–382.
4. Elad S, Yarom N, Zadik Y: **Immunotherapy-Related Oral Adverse Effects: Immediate Sequelae, Chronicity and Secondary Cancer**. *Cancers (Basel)* 2023, **15**(19).
5. Long X, Fang C, Qiu X, Yu F, Chen L, Liu X, Li Y, Chen Y, Wang R, Zhang T: **Immunonutrients for radiation-induced oral mucositis in head and neck cancer: current evidence and future perspectives**. *Front Immunol* 2026, **17**:1885556.
6. Elad S, Cheng KKF, Lalla RV, Yarom N, Hong C, Logan RM, Bowen J, Gibson R, Saunders DP, Zadik Y *et al*: **MASCC/ISOO clinical practice guidelines for the management of mucositis secondary to cancer therapy**. *Cancer* 2020, **126**(19):4423–4431.
7. Tanaka T, Hirano H, Ikebe K, Ueda T, Iwasaki M, Minakuchi S, Arai H, Akishita M, Kozaki K, Iijima K: **Consensus statement on "Oral frailty" from the Japan Geriatrics Society, the Japanese Society of Gerodontology, and the Japanese Association on Sarcopenia and Frailty**. *Geriatr Gerontol Int* 2024, **24**(11):1111–1119.
8. Minakuchi S, Tsuga K, Ikebe K, Ueda T, Tamura F, Nagao K, Furuya J, Matsuo K, Yamamoto K, Kanazawa M *et al*: **Oral hypofunction in the older population: Position paper of the Japanese Society of Gerodontology in 2016**. *Gerodontology* 2018, **35**(4):317–324.
9. Tanaka T, Hirano H, Ohara Y, Nishimoto M, Iijima K: **Oral Frailty Index-8 in the risk assessment of new-onset oral frailty and functional disability among community-dwelling older adults**. *Arch Gerontol Geriatr* 2021, **94**:104340.
10. Iwasaki M, Motokawa K, Watanabe Y, Shirobe M, Inagaki H, Edahiro A, Ohara Y, Hirano H, Shinkai S, Awata S: **Association between Oral Frailty and Nutritional Status among Community-Dwelling Older Adults: the Takashimadaira Study**. *J Nutr Health Aging* 2020, **24**(9):1003–1010.
11. Tanaka T, Takahashi K, Hirano H, Kikutani T, Watanabe Y, Ohara Y, Furuya H, Tetsuo T, Akishita M, Iijima K: **Oral Frailty as a Risk Factor for Physical Frailty and Mortality in Community-Dwelling Elderly**. *J Gerontol A Biol Sci Med Sci* 2018, **73**(12):1661–1667.

12. Kimble R, Papacosta AO, Lennon LT, Whincup PH, Weyant RJ, Mathers JC, Wannamethee SG, Ramsay SE: **The Relationship of Oral Health with Progression of Physical Frailty among Older Adults: A Longitudinal Study Composed of Two Cohorts of Older Adults from the United Kingdom and United States**. *J Am Med Dir Assoc* 2023, **24**(4):468–474.e463.
13. Komiyama T, Gallagher JE, Hattori Y: **Relationship between tooth loss and progression of frailty: Findings from the English longitudinal study of aging**. *Arch Gerontol Geriatr* 2024, **127**:105572.
14. Tanaka T, Lyu W, Hirano H, Shirobe M, Iijima K: **Oral frailty and the trajectories of psychological well-being and cognitive function: findings from the 12-year community-based Kashiwa study**. *J Gerontol A Biol Sci Med Sci* 2026, **81**(3).
15. Puranen T, Hiltunen K, Pitkälä KH, Roitto HM, Mäntylä P, Saarela RKT: **Association of oral frailty with falls in long-term care residents**. *Eur Geriatr Med* 2025, **16**(1):191–195.
16. Matsuo K, Onoda D, Kobayashi K, Mibu K, Tamiya S: **Association of oral frailty with 1-year all-cause mortality in older patients with heart failure: A retrospective cohort study**. *Clin Nutr ESPEN* 2026, **73**:103128.
17. Pearce J, Martin S, Heritage S, Khoury EG, Kucharczak J, Nuamek T, Cairns DA, Velikova G, Richards SH, Clegg A *et al*: **Frailty and outcomes in adults undergoing systemic anticancer treatment: a systematic review and meta-analysis**. *J Natl Cancer Inst* 2025, **117**(7):1316–1339.
18. Lee HJ, Jr., Boscardin J, Walter LC, Smith AK, Cohen HJ, Giri S, Williams GR, Presley CJ, Singhal S, Huang LW *et al*: **Associations of frailty with survival, hospitalization, functional decline, and toxicity among older adults with advanced non-small cell lung cancer**. *Oncologist* 2025, **30**(8).
19. Iwasaki M, Shirobe M, Motokawa K, Tanaka T, Ikebe K, Ueda T, Minakuchi S, Akishita M, Arai H, Iijima K *et al*: **Prevalence of oral frailty and its association with dietary variety, social engagement, and physical frailty: Results from the Oral Frailty 5-Item Checklist**. *Geriatr Gerontol Int* 2024, **24**(4):371–377.
20. Wang X, Lv M, Zhang T, Zhou K, Xia W, Chen Y: **Prevalence and risk factors of oral frailty in elderly cancer inpatients: A cross-sectional study**. *Medicine (Baltimore)* 2025, **104**(50):e46378.
21. Li W, Tang S, Tian L, Zhai J, He J, Li J, Zhao Z, Zhou Z, Xu M: **Oral Frailty and Multidimensional Health Among Community-Dwelling Older Adults in China: A Cross-Sectional Study**. *Nutrients* 2026, **18**(14).
22. Lv K, Yu P, Xue Y, Su J, Ren Y, Tang J: **Construction and validation of an oral frailty risk prediction model for community-dwelling older adults**. *BMC Geriatr* 2025, **25**(1):808.
23. Chen Y, Tang H, Li B, Yuan L: **Risk prediction models for oral frailty in older adults: a scoping review**. *Front Med (Lausanne)* 2026, **13**:1868632.
24. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M *et al*: **TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods**. *Bmj* 2024, **385**:e078378.
25. Obermeyer Z, Emanuel EJ: **Predicting the Future - Big Data, Machine Learning, and Clinical Medicine**. *N Engl J Med* 2016, **375**(13):1216–1219.

26. Shen L, Wu J, Lan J, Chen C, Wang Y, Li Z: **Interpretable machine learning-based prediction of 28-day mortality in ICU patients with sepsis: a multicenter retrospective study**. *Front Cell Infect Microbiol* 2024, **14**:1500326.
27. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K: **The practical implementation of artificial intelligence technologies in medicine**. *Nat Med* 2019, **25**(1):30–36.
28. Balachandran VP, Gonen M, Smith JJ, DeMatteo RP: **Nomograms in oncology: more than meets the eye**. *Lancet Oncol* 2015, **16**(4):e173–180.
29. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B: **A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models**. *J Clin Epidemiol* 2019, **110**:12–22.
30. Zhao L, Leng Y, Hu Y, Xiao J, Li Q, Liu C, Mao Y: **Understanding decision curve analysis in clinical prediction model research**. *Postgrad Med J* 2024, **100**(1185):512–515.
31. Liu H, Hu X, Liu Q, Shao M, Huang X, Gu J, Luo J, Fang S, Tao X, Lin J *et al*: **Development and internal validation of a risk prediction model for oral frailty in hospitalized older adults with chronic diseases**. *Front Public Health* 2026, **14**:1717485.
32. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP: **The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies**. *J Clin Epidemiol* 2008, **61**(4):344–349.
33. Yang CC, Chen HT, Iijima K, Tanaka T, Dai CY, Yu SJ, Chuang HY: **Establishment of a Mandarin Chinese Version of the Oral Frailty Index-8 and Exploration of the Association Between Oral Frailty and Sarcopenia**. *Geriatrics (Basel)* 2025, **10**(2).
34. Robert T: **Regression Shrinkage and Selection via the Lasso**. *Journal of the Royal Statistical Society Series B (Methodological)* 1996, **58**(1):267–288.
35. Riley R, Ensor J, Snell K, Harrell F, Martin G, Reitsma J, Moons K, Collins G, van Smeden M: **Calculating the sample size required for developing a clinical prediction model**. *Bmj-British Medical Journal* 2020, **368**.
36. Hoerl A, Kennard R: **Ridge regression: Biased estimation for nonorthogonal problems**. *Technometrics* 2000, **42**(1):80–86.
37. Ester M, Kriegel H, Xu X: **XGBoost: A scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (vol , pg 785, 2016)**. *Geographical Analysis* 2022.
38. Efron B, Tibshirani R: **Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy**. *Statistical Science* 1986, **1**(1):54–75.
39. Vickers AJ: **Decision Curve Analysis: A Novel Method for Evaluating Prediction Models**. *Medical Decision Making* 2006, **26**(6):565–574.
40. Wu L: **A review of the transition from Shapley values and SHAP values to RGE**. *Statistics* 2025, **59**(5):1161–1183.
41. Gong J, Tan L, Yan J, Qian Y, Li Y: **Systematic review and meta-analysis of oral frailty prevalence among older hospitalized patients**. *Frontiers in Public Health* 2025, **13**.

42. Lv K, Yu P, Xue Y, Su J, Ren Y, Tang J: **Construction and validation of an oral frailty risk prediction model for community-dwelling older adults**. *Bmc Geriatrics* 2025, **25**(1).
43. Li J, Wang Y, Zhai M, Qin M, Zhao D, Xiang Q, Shao Z, Wang P, Lin Y, Dong Y *et al*: **Risk factors and a nomogram for predicting cognitive frailty in Chinese patients with lung cancer receiving drug therapy: A single-center cross-sectional study**. *Thoracic Cancer* 2024, **15**(11):884–894.
44. Lv J, Li J, Wang Y, Liu S, Du Y, Wang L, Wang H, Shi Y: **Construction and validation of a risk prediction model for oral frailty in elderly patients with esophageal cancer**. *Frontiers in Oncology* 2026, **15**.
45. Chen Y, Zhang L, Yan W, Liu F: **Factors associated with oral frailty in older adults: a systematic review and meta-analysis**. *Frontiers in Public Health* 2025, **13**.
46. Cha J, Jung D, Kim H, Lim H: **Factors affecting oral frailty and nutritional status in older adults living alone**. *Frontiers in Nutrition* 2025, **12**.
47. Arends J, Bachmann P, Baracos V, Barthelemy N, Bertz H, Bozzetti F, Fearon K, Hütterer E, Isenring E, Kaasa S *et al*: **ESPEN guidelines on nutrition in cancer patients**. *Clinical Nutrition* 2017, **36**(1):11–48.
48. Johnson G, Guthmiller J: **The impact of cigarette smoking on periodontal disease and treatment**. *Periodontology 2000* 2007, **44**:178–194.
49. Le Guevelou J, Palard-Novello X, Kammerer E, Baty M, Perazzi M, Larnaudie A, De Crevoisier R, Castelli J: **Assessment and Prediction of Salivary Gland Function After Head and Neck Radiotherapy: A Systematic Review**. *Cancer Medicine* 2024, **13**(24).
50. Furuta M, Takeuchi K, Takeshita T, Shibata Y, Suma S, Kageyama S, Asakawa M, Shimazaki Y, Hata J, Ninomiya T *et al*: **Baseline periodontal status and modifiable risk factors are associated with tooth loss over a 10-year period: Estimates of population attributable risk in a Japanese community**. *Journal of Periodontology* 2022, **93**(4):526–536.
51. Mercadante V, Smith D, Abdalla-Aslan R, Andabak-Rogulj A, Brennan M, Jaguar G, Clark H, Fregnani E, Gueiros L, Hovan A *et al*: **A systematic review of salivary gland hypofunction and/or xerostomia induced by non-surgical cancer therapies: prevention strategies**. *Supportive Care in Cancer* 2025, **33**(2).
52. Chae S, Davoudi A, Song J, Evans L, Bowles K, Mcdonald M, Barrón Y, Min S, Oh S, Scharp D *et al*: **Developing a clinical decision support framework for integrating predictive models into routine nursing practices in home health care for patients with heart failure**. *Journal of Nursing Scholarship* 2025, **57**(1):165–177.
53. van Velzen M, de Graaf-Waar H, Ubert T, van der Willigen R, Muilwijk L, Schmitt M, Scheper M, van Meeteren N: **21st century (clinical) decision support in nursing and allied healthcare. Developing a learning health system: a reasoned design of a theoretical framework**. *Bmc Medical Informatics and Decision Making* 2023, **23**(1).
54. Mitchell G, Stark P, Wilson C, Tsakos G, Brocklehurst P, Lappin C, Quinn B, Holland N, Mckenna G: **'Whose role is it anyway?' Experiences of community nurses in the delivery and support of oral health care for older people living at home: a grounded theory study**. *Bmc Nursing* 2023, **22**(1).
55. Kylen J, Björns S, Hägglin C, Grönbeck-Lindén I, Piper L, Wårdh I: **Decisional needs for older adults, home health care nurses and dental hygienists during team-based**

oral health assessments in ordinary home settings - a qualitative study. *Bmc Geriatrics* 2024, **24**(1).

Supplementary Files

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

- SupplementaryMaterial.docx