

Supplementary Materials

Table of Contents

<i>Supplementary Methods</i>	2
1. Phenotype Classification via Multinomial Logistic Regression and Risk Scoring	2
2. External validation of the phenotype-based risk score	4
3. Cluster number selection and stability assessment	5
<i>Supplementary Results</i>	6
1. Baseline characteristics of the study population.....	6
2. Kaplan–Meier curves for individual cardiovascular outcomes across clusters	8
3. Predictive Performance of the Multinomial Classifier and Phenotype-Based Risk Score.....	9
3.1 Variable contributions to phenotype discrimination	9
3.2 Overall classification performance	10
3.3 Misclassification patterns	11
3.4 Discrimination ability.....	12
3.5 Clinical risk score distribution	13
4. External validation of the phenotype-based classifier and risk score.....	14
4.1 Baseline characteristics by phenotype in the external validation cohort	14
4.2 Association of phenotypes and risk strata with cardiovascular events	17
4.3 Observed event rates across phenotypes and risk strata	18
4.4 Discrimination of the phenotype-based risk score versus AHI	19
4.5 Calibration of the phenotype-based risk score	20
5. Quantitative metrics for cluster number selection	21

Supplementary Methods

Supplementary Methods provide full details of the multinomial logistic regression classifier and the derivation of the composite phenotype-based risk score, complementing the brief description in the main Methods section.

1. Phenotype Classification via Multinomial Logistic Regression and Risk Scoring

To translate the phenotypic clusters derived from unsupervised analysis into a clinically applicable classification tool, we developed a supervised multinomial logistic regression model to predict phenotype membership based on routinely collected baseline variables. Three phenotypes were pre-defined from the unsupervised clustering results: Original Cluster 1 (Aging–Hypertensive OSA Phenotype), Original Cluster 2 (Healthy Sleep–CVD Profile [reference phenotype]), and Original Cluster 3 (Metabolic–Hypoxic OSA Phenotype).

Fifteen baseline predictor variables were prespecified a priori because they jointly represent clinically complementary dimensions of SDB-related cardiovascular vulnerability, including demographic susceptibility, adiposity and cardiometabolic burden, blood pressure/comorbidity profile, sleep architecture, and nocturnal respiratory-hypoxemic stress. Specifically, age and sex were included to reflect host susceptibility; BMI and waist circumference to capture overall and central adiposity; SBP, DBP, HTN, HL, and DM to reflect cardiovascular and metabolic burden; TST and REMp to represent sleep architecture; and AHI, AI, Min SpO₂, and T90 to quantify respiratory event burden and nocturnal physiological stress. These domains were selected based on prior phenotype-based OSA studies and expert discussions emphasizing that cardiovascular vulnerability in OSA reflects not only respiratory event frequency but also host susceptibility, adiposity, cardiometabolic substrate, sleep fragmentation, and nocturnal hypoxemic burden. To support interpretability, reproducibility, and downstream clinical translation, the predictor set was intentionally restricted to routinely available baseline clinical and polysomnographic measures rather than more specialized endotypic variables. Continuous variables were standardized to Z-scores prior to modeling. From the source cohort, we created a complete case

dataset of N = 4,951 participants (Original Cluster 1: n = 2,084; Original Cluster 2: n = 2,375; Original Cluster 3: n = 492). The dataset was randomly split into a training set (80%, n = 3,960) and an independent test set (20%, n = 991) using stratified sampling to preserve phenotype distribution (Original Cluster 1: n = 417; Original Cluster 2: n = 475; Original Cluster 3: n = 99). A multinomial logistic regression model (LBFGS solver, maximum 1,000 iterations) was fitted to the training set with all 15 predictors included.

Model performance in the test set was evaluated using per-class precision, recall, and F1-score, overall accuracy, a confusion matrix, and multiclass AUROC calculated using a one-vs-rest strategy with macro-averaging. Standardized regression coefficients (β) were extracted and visualized to facilitate interpretation of the contribution of each predictor to phenotype discrimination.

To derive a clinically interpretable risk score, we summed the linear predictors (logit values) for Original Cluster 1 and Original Cluster 3, the two phenotypes that consistently demonstrated higher cardiovascular risk in the derivation cohort:

$$\textbf{Risk Score} = \text{Score}_{\text{Cluster 1}} + \text{Score}_{\text{Cluster 3}}$$

with Original Cluster 2 (Healthy Sleep–CVD Profile [reference phenotype]) contributing zero. The resulting score was intended to represent graded similarity to the higher-risk phenotype space, rather than to serve as a de novo outcome prediction model. Higher values therefore indicate greater phenotypic proximity to the ageing–hypertensive and metabolic–hypoxic profiles.

Risk strata were defined using tertiles of the training-set distribution:

- Low risk: < -7.445
- Intermediate risk: -7.445 to -2.364
- High risk: > -2.364

This score was linked to long-term cardiovascular event probability via the logistic regression–derived equation:

$$P(event) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 \times Risk\ Score)]}$$

where β_0 and β_1 were estimated from the training data cardiovascular outcomes model. This framework allows clinicians to assign phenotype membership and cardiovascular risk probability at the point of care using either the provided web calculator or Excel-based scoring tool.

2. External validation of the phenotype-based risk score

To assess external generalizability, we applied the multinomial logistic regression classifier and phenotype-based risk score to an independent external cohort derived from Fuwai Hospital, Beijing, China. The study was approved by the Ethics Committee of Fuwai Hospital (approval number: 2025-2708). Consecutive adults who underwent clinically indicated polysomnography and cardiovascular assessment as part of routine care between February 2020 and December 2024 were screened. Participants were eligible if they had the baseline clinical and polysomnographic variables required for phenotype assignment. Individuals with missing key predictors or unavailable follow-up information for the composite cardiovascular endpoint were excluded from the complete-case validation analyses. Because this cohort was hospital-based, whereas the SHHS derivation cohort was community-based, the two cohorts differed in referral pattern, disease spectrum, and baseline cardiovascular risk. Therefore, the external validation was intended to assess transportability of the classifier and risk score in an independent clinical setting rather than direct population equivalence with SHHS.

Baseline demographic, clinical, and polysomnographic variables were collected using routine standardized clinical procedures. Sleep studies were scored according to standardized practice, and variables required for phenotype assignment were harmonized to the derivation framework as closely as feasible. Follow-up information for the composite cardiovascular endpoint was obtained through clinical record review and predefined follow-up ascertainment procedures.

The same 15 baseline predictors (age, sex, BMI, waist circumference, SBP, DBP, HTN, HL, DM, TST, REMp, AHI, AI, MinSpO₂, and T90) were standardized using the means and standard deviations estimated in the SHHS training cohort. We then used the previously derived multinomial coefficients to compute phenotype membership probabilities and predicted cluster

labels, as well as the composite phenotype risk score defined as the sum of the linear predictors for the two phenotypes judged to be at higher cardiovascular risk. Training-set tertile cut-offs (−7.445 and −2.364) were used unchanged to categorize individuals into low-, intermediate-, and high-risk groups.

Among these 527 individuals, 419 had complete data on all predictors and adjudicated follow-up for the composite cardiovascular endpoint and were included in the external validation analyses. We examined crude event rates across externally assigned phenotypes and risk groups, fitted logistic regression models to estimate odds ratios (ORs) and 95% confidence intervals (CIs) for the composite endpoint, and compared discrimination of the phenotype-based risk score against AHI using receiver operating characteristic (ROC) curves and the area under the curve (AUC). Calibration of the predicted event probabilities derived from the logistic link was evaluated by plotting observed versus predicted event rates across deciles of predicted risk.

3. Cluster number selection and stability assessment

Candidate values of k were evaluated using quantitative measures of cluster compactness/separation and resampling-based stability. Average silhouette width was calculated for each candidate k to summarize within-cluster compactness and between-cluster separation. To assess stability, the k -means procedure was repeated after repeated 80% subsampling of the analytic dataset, and agreement between the full-sample and resampled clustering solutions was quantified using the adjusted Rand index. These quantitative metrics were reviewed together with PCA-based visual separation and clinical interpretability. The three-cluster solution showed the highest average silhouette width and retained high resampling stability. Although the two-cluster solution showed slightly higher resampling stability, it yielded a relatively coarse partition that did not adequately separate the ageing–hypertensive and healthy sleep–CVD profiles. Higher-order solutions showed lower average silhouette widths and/or lower stability and less interpretable partitioning. The quantitative results are summarized in Supplementary Table 5.

Supplementary Results

1. Baseline characteristics of the study population

Baseline characteristics of the full baseline cohort are provided in **Supplementary Table 1** (including crude outcome counts for context).

Supplementary Table 1. Baseline characteristics of the study population

Variable	Total (n=5,682)	Cluster 1 (n=2,335)	Cluster 2 (n=2,789)	Cluster 3 (n=558)	<i>P</i>
Age, years	63.2 ± 11.2	69.1 ± 9.4	57.7 ± 10.0	65.4 ± 10.4	<0.001
Female sex, n (%)	2,992 (52.7)	899 (38.5)	1,909 (68.4)	184 (33.0)	<0.001
<i>Ethnicity</i>					<0.001
Hispanic, n (%)	4,793 (84.4)	2,017 (86.4)	2,294 (82.3)	482 (86.4)	
Latin, n (%)	515 (9.1)	243 (10.4)	216 (7.7)	56 (10.0)	
Unknown, n (%)	374 (6.6)	75 (3.2)	279 (10.0)	20 (3.6)	
BMI, kg/m ²	28.2 ± 5.1	28.1 ± 4.6	27.4 ± 4.8	32.4 ± 6.1	<0.001
Weight, kg	78.1 ± 16.1	78.9 ± 15.0	74.8 ± 15.1	91.6 ± 17.7	<0.001
Height, cm	167.4 ± 9.6	168.4 ± 9.8	166.2 ± 9.2	169.6 ± 9.3	<0.001
Waist circumference, cm	97.0 ± 13.7	98.7 ± 12.2	93.2 ± 13.3	108.1 ± 13.8	<0.001
SBP, mmHg	127.3 (19.3)	132.9 (20.0)	121.8 (17.3)	130.7 (18.4)	<0.001
DBP, mmHg	73.6 (11.6)	73.4 (12.4)	73.4 (10.4)	75.4 (13.4)	<0.001
Smoker, n (%)	2,985 (52.9)	1,436 (61.9)	1,230 (44.4)	319 (57.6)	<0.001
Alcohol intake, n (%)	295 (44.2)	190 (60.7)	82 (28.4)	23 (35.4)	<0.001
Hypertension, n (%)	2,018 (37.1)	1,353 (60.2)	404 (15.2)	261 (49.2)	<0.001
DM, n (%)	391 (7.2)	317 (14.0)	25 (1.0)	49 (9.2)	<0.001
Hyperlipidemia, n (%)	675 (11.9)	482 (20.7)	120 (4.3)	73 (13.2)	<0.001
COPD, n (%)	102 (1.8)	72 (3.1)	17 (0.6)	13 (2.4)	<0.001
Asthma, n (%)	536 (9.5)	236 (10.2)	254 (9.2)	46 (8.3)	0.281
<i>Sleep parameters</i>					

TST, min	360.1 ± 64.3	332.9 ± 64.6	385.8 ± 52.2	345.7 ± 66.1	<0.001
REMP, %	19.8 ± 6.3	17.9 ± 6.3	21.9 ± 5.5	17.2 ± 6.3	<0.001
AHI, events/h	14.6 ± 15.6	14.6 ± 11.1	8.3 ± 8.0	45.9 ± 21.8	<0.001
cAHI, events/h	10.6 ± 10.7	10.9 ± 8.3	6.6 ± 6.4	29.2 ± 15.7	<0.001
oAHI, events/h	14.1 ± 15.0	14.1 ± 10.5	8.1 ± 7.8	44.1 ± 21.1	<0.001
ODI, events/h	12.9 ± 13.1	12.5 ± 9.4	8.2 ± 7.0	38.1 ± 20.0	<0.001
AI, events/h	19.1 ± 10.6	20.4 ± 9.3	15.2 ± 6.6	33.9 ± 16.6	<0.001
Min SpO ₂ , %	85.6 ± 6.2	86.0 ± 4.4	87.6 ± 4.3	74.3 ± 8.8	<0.001
T90, %	3.4 ± 10.2	1.9 ± 3.9	0.9 ± 3.0	22.8 ± 23.1	<0.001
CV Outcome*					
Composite CV					
endpoint, n (%)	1,039 (21.2%)	651 (32.9%)	243 (9.9%)	145 (30.0%)	<0.001
CVD mortality, n (%)	342 (7.0%)	242 (12.2%)	52 (2.1%)	48 (9.9%)	<0.001
CHF, n (%)	608 (12.4%)	400 (20.2%)	108 (4.4%)	100 (20.7%)	<0.001
Incident MI, n (%)	351 (7.2%)	212 (10.7%)	97 (4.0%)	42 (8.7%)	<0.001
Stroke, n (%)	281 (5.7%)	179 (9.0%)	72 (2.9%)	30 (6.2%)	<0.001

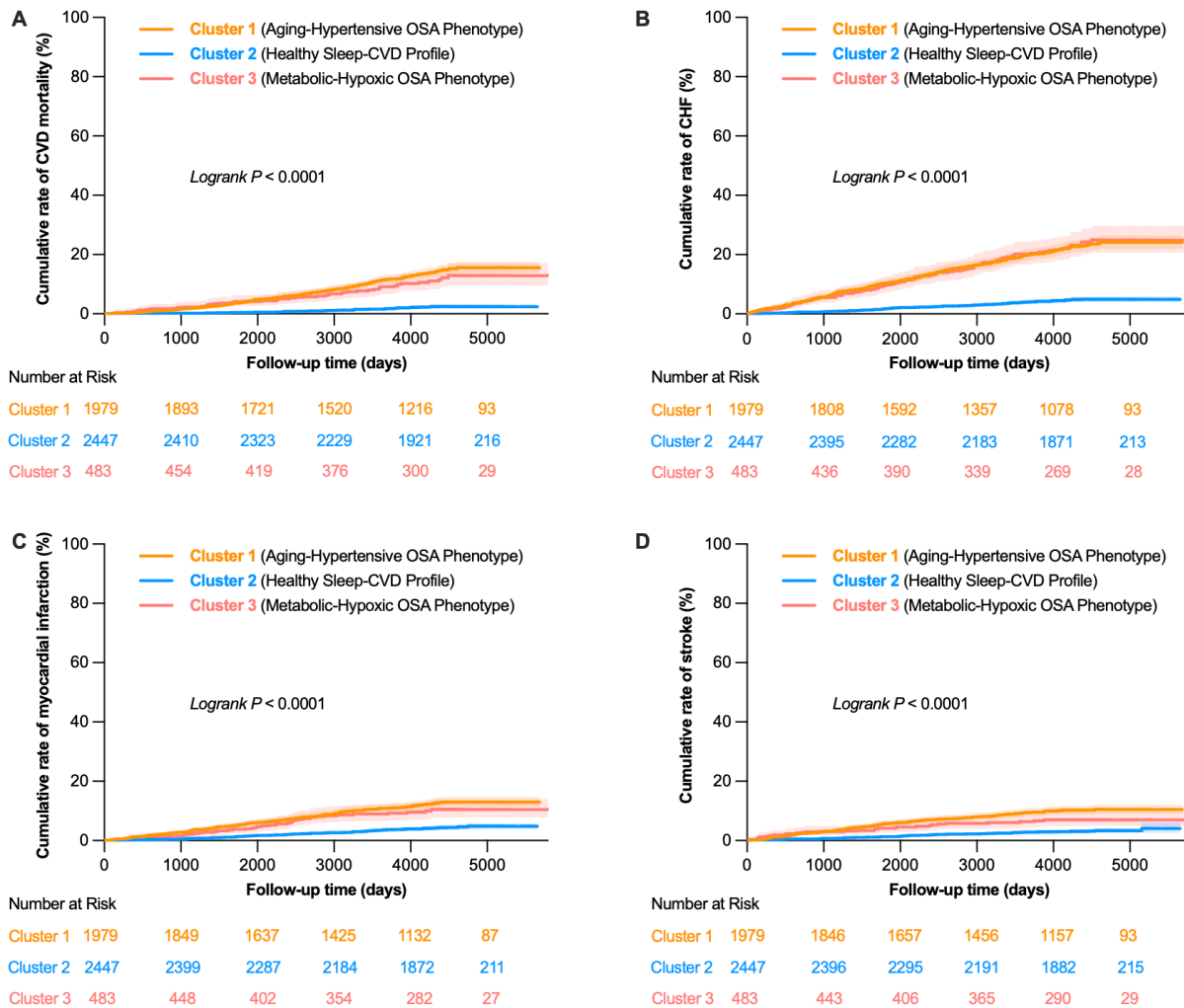
Values are mean (SD) for continuous variables and n (%) for categorical variables.

*Cardiovascular outcomes are reported for the survival analysis cohort (n = 4,909; cluster 1, n = 1,979; cluster 2, n = 2,447; cluster 3, n = 483) after excluding participants without adjudicated follow-up data. Comparisons were performed using one-way analysis of variance or Kruskal-Wallis tests for continuous variables, as appropriate, and χ^2 tests for categorical variables. *P* values are two-sided; no adjustment for multiple comparisons was applied. Values smaller than 0.001 are reported as *P* < 0.001.

Abbreviations: AHI, apnea–hypopnea index; AI, arousal index; BMI, body mass index; CHF, congestive heart failure; COPD, chronic obstructive pulmonary disease; CV, cardiovascular; CVD, cardiovascular disease; DBP, diastolic blood pressure; DM, diabetes mellitus; MI, myocardial infarction; ODI, oxygen desaturation index; REMp, rapid eye movement sleep percentage; SBP, systolic blood pressure; SpO₂, oxygen saturation; T90, percentage of total sleep time with oxygen saturation < 90%; TST, total sleep time.

2. Kaplan–Meier curves for individual cardiovascular outcomes across clusters

Kaplan–Meier curves for individual endpoints corroborate the rank ordering across phenotypes and are provided in *Supplementary Figure 1*.



Supplementary Figure 1. Kaplan–Meier curves for individual cardiovascular outcomes across clusters. Cumulative incidence of (A) cardiovascular disease (CVD) mortality, (B) congestive heart failure (CHF), (C) myocardial infarction, and (D) stroke among the three phenotypic clusters: Cluster 1 (Aging–Hypertensive OSA Phenotype), Cluster 2 (Healthy Sleep–CVD Profile), and Cluster 3 (Metabolic–Hypoxic OSA Phenotype). The solid curves represent Kaplan–Meier estimates of cumulative event incidence, and shaded areas indicate the

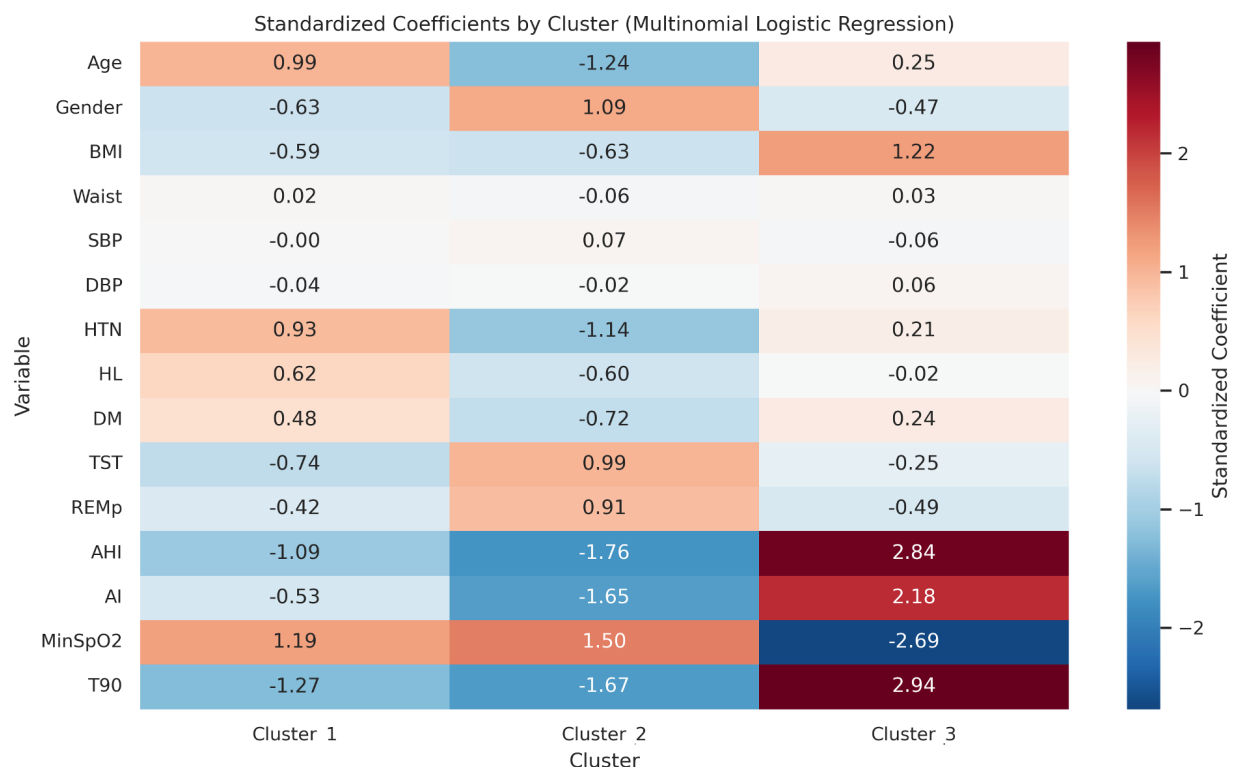
corresponding 95% confidence intervals. The number of participants at risk at each time point is shown below each panel. Comparisons among clusters were performed using the two-sided log-rank test.

Abbreviations: CHF, congestive heart failure; CVD, cardiovascular disease; OSA, obstructive sleep apnea.

3. Predictive Performance of the Multinomial Classifier and Phenotype-Based Risk Score

3.1 Variable contributions to phenotype discrimination

The standardized regression coefficient heatmap (*Supplementary Figure 2*) highlighted distinct variable patterns across phenotypes. Higher AHI, AI, and T90 with lower MinSpO₂ were strongly associated with Original Cluster 3 (Metabolic–Hypoxic OSA Phenotype). Older age, higher SBP/DBP, and shorter TST characterized Original Cluster 1 (Aging–Hypertensive OSA Phenotype). Healthier sleep architecture and lower comorbidity burden defined Original Cluster 2 (Healthy Sleep–CVD Profile [reference phenotype]).



Supplementary Figure 2. Standardized coefficient heatmap of the multinomial logistic regression model. The plot visualizes Z-score standardized β -coefficients of 15 input variables across the three cluster outcomes. Positive coefficients (red) indicate higher values associated with cluster membership, while negative coefficients (blue) suggest inverse associations. For instance, higher T90 and AHI, together with lower MinSpO₂, strongly and positively contribute to prediction of the Metabolic–Hypoxic OSA Phenotype (Original Cluster 3).

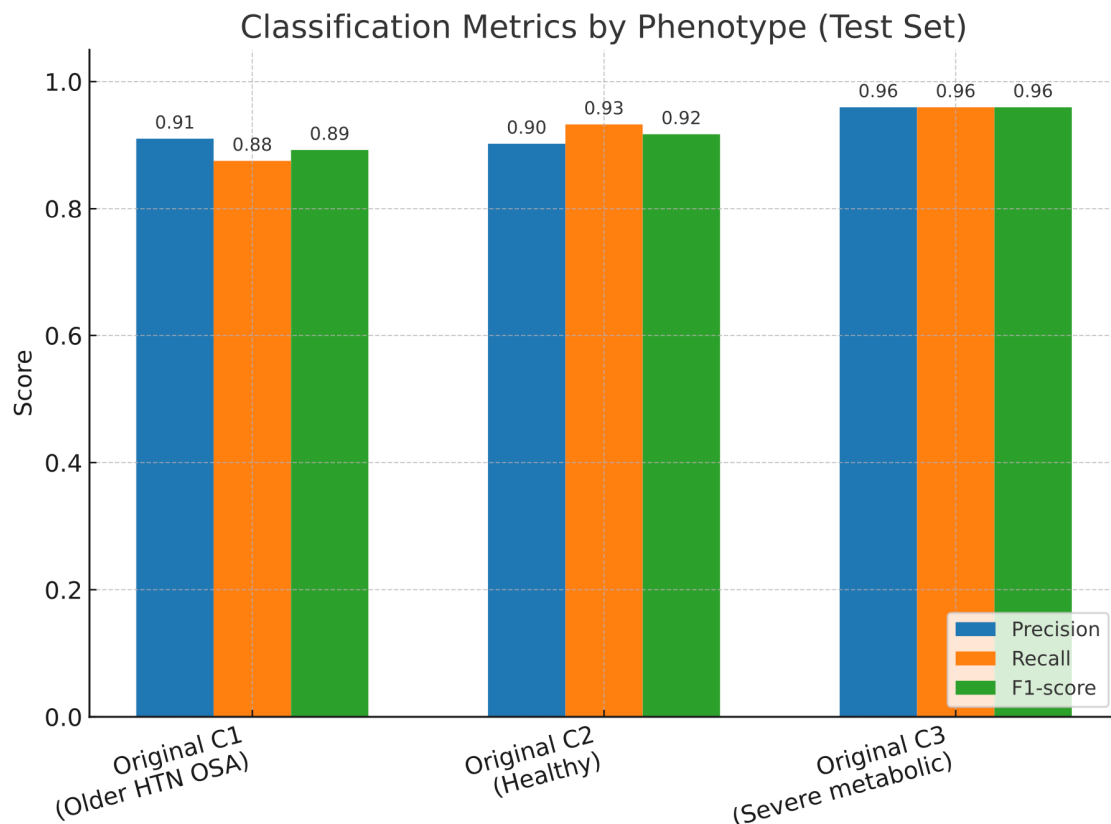
3.2 Overall classification performance

Performance metrics for the independent test set are summarized in *Supplementary Table 2*. The model achieved an overall accuracy of 0.911, with macro-averaged precision, recall, and F1-score of 0.924, 0.923, and 0.923, respectively. Class-specific performance was high across all phenotypes.

Supplementary Table 2. Classification performance metrics for multinomial logistic regression model in the independent test set. Values are reported as precision, recall, F1-score, and support for each phenotype, along with macro-averaged and weighted-averaged scores.

Metric	Original Cluster 1	Original Cluster 2	Original Cluster 3	Accuracy (overall)	Macro- AUC
Precision	0.910	0.902	0.96	0.911	0.988
Recall	0.875	0.933	0.96		
F1-score	0.892	0.917	0.96		

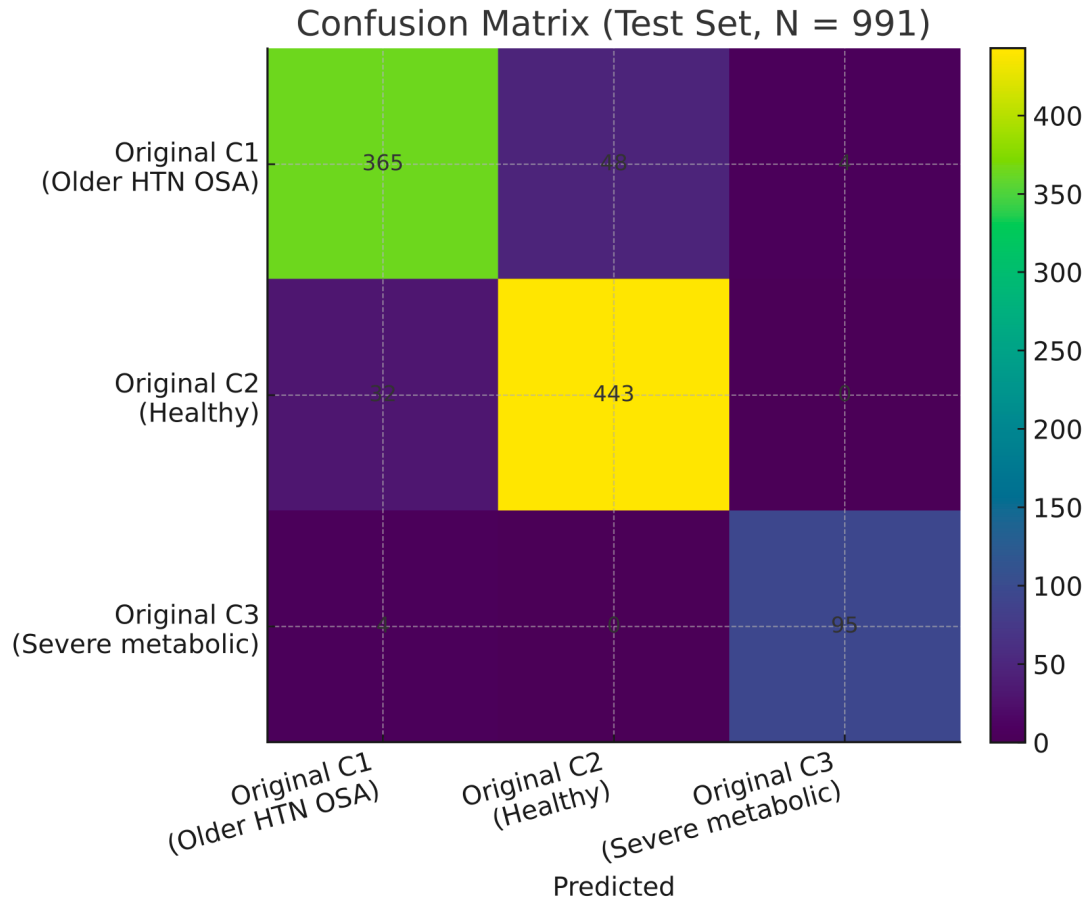
The bar chart in *Supplementary Figure 3* visualizes these metrics, showing that **Original Cluster 3** had the highest precision and recall (both 0.96), while the other two phenotypes also demonstrated balanced and robust performance.



Supplementary Figure 3. Classification performance metrics for multinomial logistic regression model in the independent test set. Bar chart illustrating precision (blue), recall (orange), and F1-score (green) for each phenotype after label remapping to the original clusters: Original Cluster 1 (Aging–Hypertensive OSA Phenotype), Original Cluster 2 (Healthy Sleep–CVD Profile [reference phenotype]), and Original Cluster 3 (Metabolic–Hypoxic OSA Phenotype). Performance was uniformly high across all phenotypes, with the severe metabolic phenotype achieving the highest precision and recall (both 0.96). Values above each bar represent the exact metric values.

3.3 Misclassification patterns

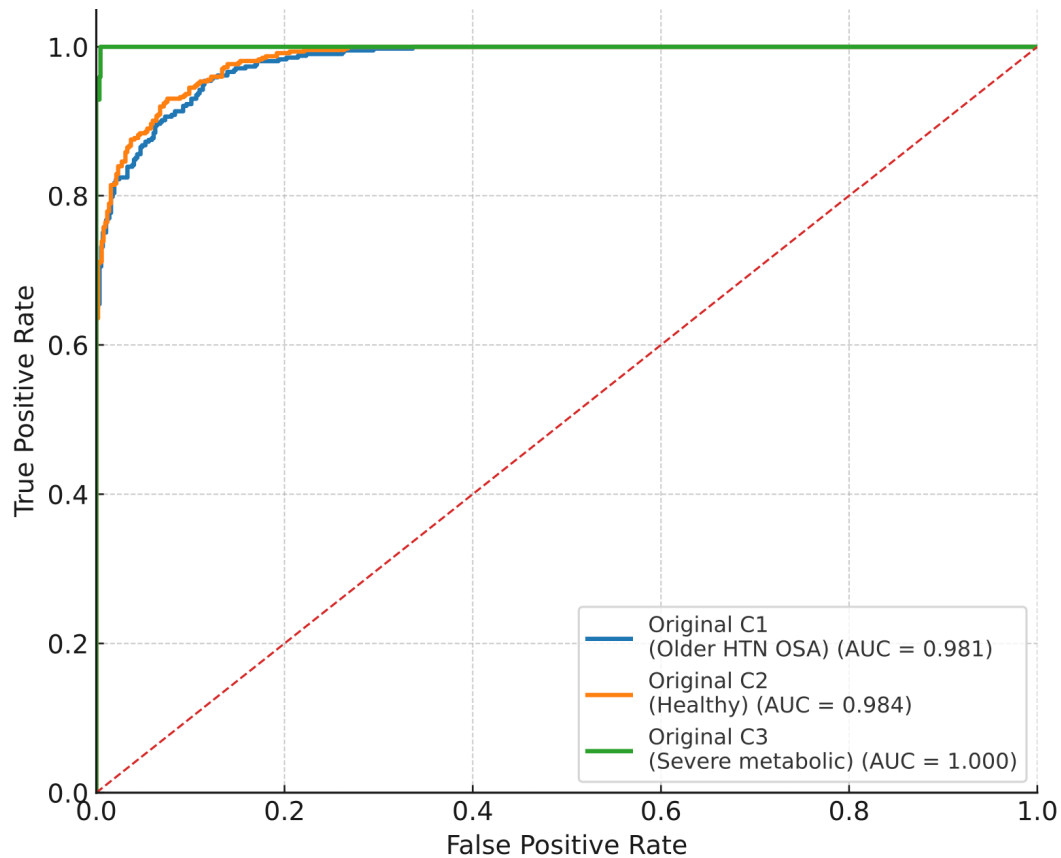
The confusion matrix heatmap (*Supplementary Figure 4*) showed that most errors occurred between Original Cluster 1 and Original Cluster 2, consistent with partial overlap in age, blood pressure, and moderate OSA severity. Original Cluster 3 was rarely misclassified, reflecting its distinctive severe hypoxemia and metabolic burden.



Supplementary Figure 4. Confusion matrix heatmap for phenotype prediction in the independent test set, after label remapping to original phenotypes.

3.4 Discrimination ability

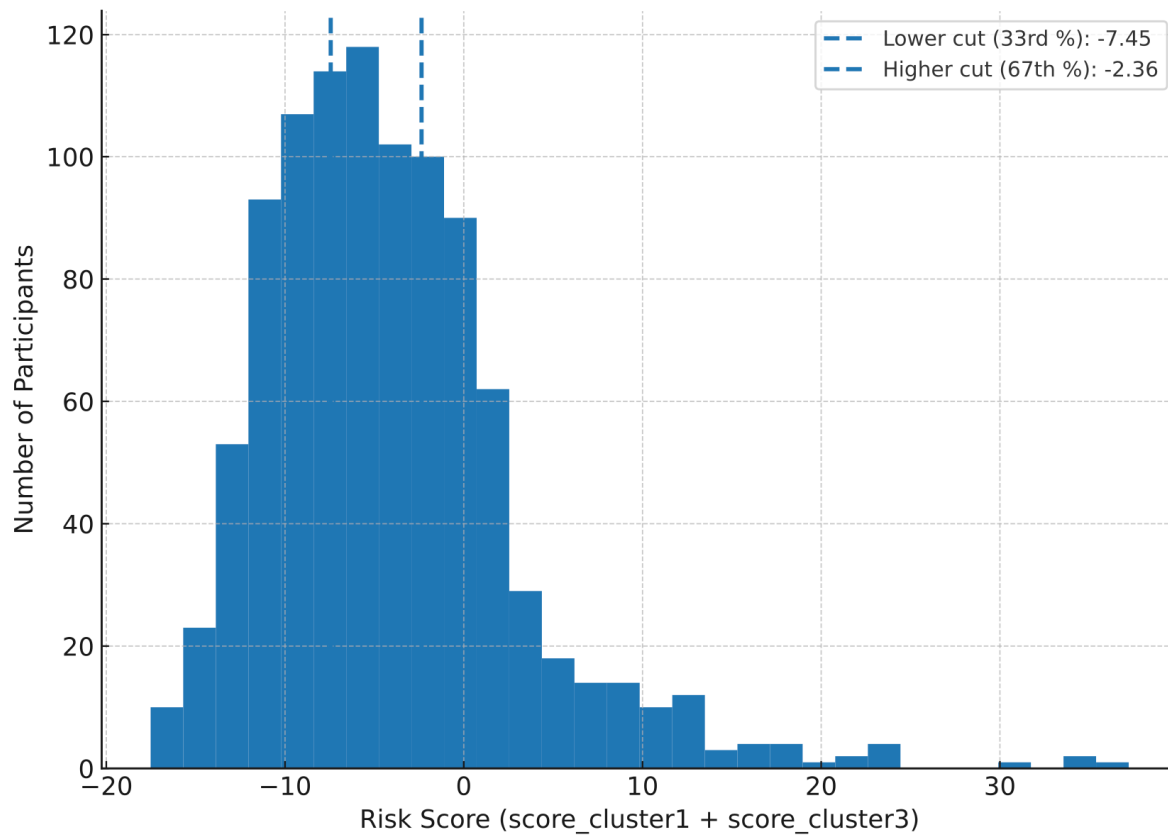
The multiclass one-vs-rest ROC curves (**Supplementary Figure 5**) demonstrated AUC values > 0.95 for all phenotypes, with a macro-AUC of 0.988, confirming excellent discriminative performance across the full clinical spectrum.



Supplementary Figure 5. One-vs-Rest ROC curves for multinomial classification. Receiver operating characteristic (ROC) curves for each phenotype against the other two combined, with AUC values all exceeding 0.95 and a macro-AUC of 0.988.

3.5 Clinical risk score distribution

The histogram in *Supplementary Figure 6* depicts the distribution of composite phenotype risk scores in the test set, with vertical dashed lines marking the training-set-derived tertile cut-offs (−7.445 and −2.364). These thresholds stratified participants into low-, intermediate-, and high-risk groups, which corresponded to increasing similarity to high-risk phenotypes (Original Cluster 1 and Original Cluster 3). High-risk participants demonstrated the greatest predicted probability of cardiovascular events, while the low-risk group had the lowest estimated risk.



Supplementary Figure 6. Distribution of composite phenotype risk scores in the independent test set. Histogram illustrating the distribution of risk scores (defined as *score_cluster1* + *score_cluster3*) among test set participants (N = 991). Vertical dashed lines represent the training set-derived tertile cut-offs: lower cut-off (33rd percentile) at -7.45 and higher cut-off (67th percentile) at -2.36 . These cut-offs were used to classify participants into low (< -7.45), intermediate (-7.45 to -2.36), and high (> -2.36) risk strata, corresponding to increasing similarity to high-risk phenotypes (Original Cluster 1 and Original Cluster 3).

4. External validation of the phenotype-based classifier and risk score

4.1 Baseline characteristics by phenotype in the external validation cohort

Among the 527 individuals in the external dataset, 419 had complete information on all 15 predictors and adjudicated follow-up for the composite cardiovascular endpoint and were therefore included in the external validation analyses. Application of the multinomial logistic regression classifier yielded 40 participants (9.5%) assigned to the Aging-Hypertensive OSA

phenotype (Cluster 1), 249 (59.4%) to the Healthy Sleep–CVD Profile (Cluster 2), and 130 (31.0%) to the Metabolic–Hypoxic OSA phenotype (Cluster 3) (*Supplementary Table 3*).

Baseline profiles in the external cohort were broadly consistent with those observed in the SHHS derivation cohort. Cluster 1 comprised the oldest patients (mean age 72.4 ± 7.1 years), with intermediate adiposity (BMI 27.4 ± 4.5 kg/m², waist circumference 94.6 ± 14.1 cm) and the highest systolic blood pressure (134.4 ± 20.6 mmHg). Cluster 2 represented a comparatively younger and metabolically favorable group (59.8 ± 9.9 years), with the lowest waist circumference (93.7 ± 12.9 cm) and SBP (119.4 ± 14.9 mmHg), and relatively mild sleep-disordered breathing (AHI 7.4 ± 6.2 events/h, AI 15.2 ± 6.1 events/h). Cluster 3 was characterized by more severe obesity (BMI 30.5 ± 5.7 kg/m², waist circumference 103.1 ± 14.3 cm) and marked OSA-related abnormalities, including substantially higher AHI (33.1 ± 19.6 events/h), higher arousal index (28.4 ± 13.7 events/h), lower Min SpO₂ ($83.9 \pm 4.1\%$), and longer T90 ($7.9 \pm 14.0\%$ of sleep time spent with SpO₂ < 90%). Sleep duration and architecture also differed: Cluster 2 showed the longest TST (383.0 ± 52.0 min) and highest REM ($21.9 \pm 5.2\%$), whereas Cluster 1 had shorter sleep (292.5 ± 59.2 min) and reduced REM ($16.9 \pm 6.9\%$), and Cluster 3 had intermediate TST (342.1 ± 64.6 min) with similarly reduced REM ($17.2 \pm 6.6\%$).

Gender and comorbidities followed patterns that again mirrored the derivation cohort. Cluster 1 had the highest prevalence of hypertension (67.5%) and diabetes (35.0%), with a predominance of women (72.5%). Cluster 2 was enriched for women (59.0%) but had the lowest burden of hypertension (22.5%), diabetes (2.8%), and hyperlipidemia (8.8%). Cluster 3 contained more men (54.6%) and showed intermediate-to-high prevalence of cardiometabolic comorbidities (hypertension 52.3%, diabetes 9.2%, hyperlipidemia 18.5%). Omnibus comparisons across phenotypes were statistically significant for most clinical and polysomnographic variables (all $P < 0.001$, except DBP; *Supplementary Table 3*), supporting that the externally assigned phenotypes retain distinct and clinically coherent profiles. Because the external validation cohort was hospital-based, whereas the SHHS derivation cohort was community-based, differences in referral pattern and baseline risk profile should be considered when interpreting transportability and generalizability, despite the broadly consistent phenotype structure observed across cohorts.

Supplementary Table 3. Baseline characteristics of the external validation cohort by externally assigned phenotype.

Variable	Cluster 1 (n = 40)	Cluster 2 (n = 249)	Cluster 3 (n = 130)	P
Age, years	72.38 (7.12)	59.82 (9.90)	67.96 (10.44)	<0.001
BMI, kg/m ²	27.38 (4.46)	27.11 (4.55)	30.49 (5.65)	<0.001
Waist circumference, cm	94.64 (14.05)	93.65 (12.93)	103.08 (14.27)	<0.001
SBP, mmHg	134.40 (20.56)	119.37 (14.90)	130.42 (16.97)	<0.001
DBP, mmHg	70.85 (15.63)	71.61 (9.53)	72.39 (10.46)	0.665
AHI, events/h	7.48 (5.94)	7.35 (6.20)	33.12 (19.62)	<0.001
AI, events/h	19.12 (8.43)	15.19 (6.09)	28.43 (13.70)	<0.001
Min SpO ₂ , %	90.45 (2.36)	89.20 (2.61)	83.85 (4.12)	<0.001
TST, min	292.45 (59.18)	382.95 (51.99)	342.14 (64.61)	<0.001
REMP, %	16.90 (6.94)	21.86 (5.22)	17.20 (6.64)	<0.001
T90, %	0.49 (0.97)	0.50 (1.81)	7.88 (13.99)	<0.001
Female, n (%)	29 (72.5)	147 (59.0)	59 (45.4)	0.004
Hypertension, n (%)	27 (67.5)	56 (22.5)	68 (52.3)	<0.001
DM, n (%)	14 (35.0)	7 (2.8)	12 (9.2)	<0.001
Hyperlipidemia, n (%)	14 (35.0)	22 (8.8)	24 (18.5)	<0.001

Values are mean (SD) for continuous variables and n (%) for categorical variables unless otherwise indicated. Phenotypes were assigned using the multinomial logistic regression classifier trained in the SHHS derivation cohort and applied to the external validation dataset after Z-score standardization of continuous predictors based on SHHS means and standard deviations. Statistical comparisons were performed using one-way analysis of variance or Kruskal-Wallis tests for continuous variables, as appropriate, and χ^2 tests for categorical variables. P values are two-sided; no adjustment for multiple comparisons was applied.

4.2 Association of phenotypes and risk strata with cardiovascular events

During follow-up, 94 composite cardiovascular events occurred among the 419 externally validated participants (22.4%). When modeled using logistic regression with the Healthy Sleep–CVD Profile (Cluster 2) as the reference, both high-risk phenotypes showed materially higher odds of the composite endpoint (*Supplementary Table 4*). The Aging–Hypertensive OSA phenotype (Cluster 1) was associated with an odds ratio (OR) of 2.59 (95% CI, 1.20–5.39; $P = 0.012$) for the composite endpoint, and the Metabolic–Hypoxic OSA phenotype (Cluster 3) with an OR of 2.57 (95% CI, 1.56–4.23; $P < 0.001$), relative to Cluster 2.

The phenotype-based risk score, categorized using the tertile cut-offs learned in the SHHS training cohort, further stratified risk in the external dataset. Compared with the low-risk group, individuals in the intermediate-risk stratum had a numerically higher but not statistically significant odds of composite events (OR 1.63; 95% CI, 0.67–4.57; $P = 0.31$). The high-risk group demonstrated a markedly elevated odds of events (OR 3.71; 95% CI, 1.62–10.07; $P = 0.004$), indicating that the risk score retains prognostic information beyond simple dichotomization of OSA severity. Full model estimates are presented in *Supplementary Table 4*.

Supplementary Table 4. Association of externally assigned phenotypes and phenotype-based risk groups with the composite cardiovascular endpoint in the external validation cohort

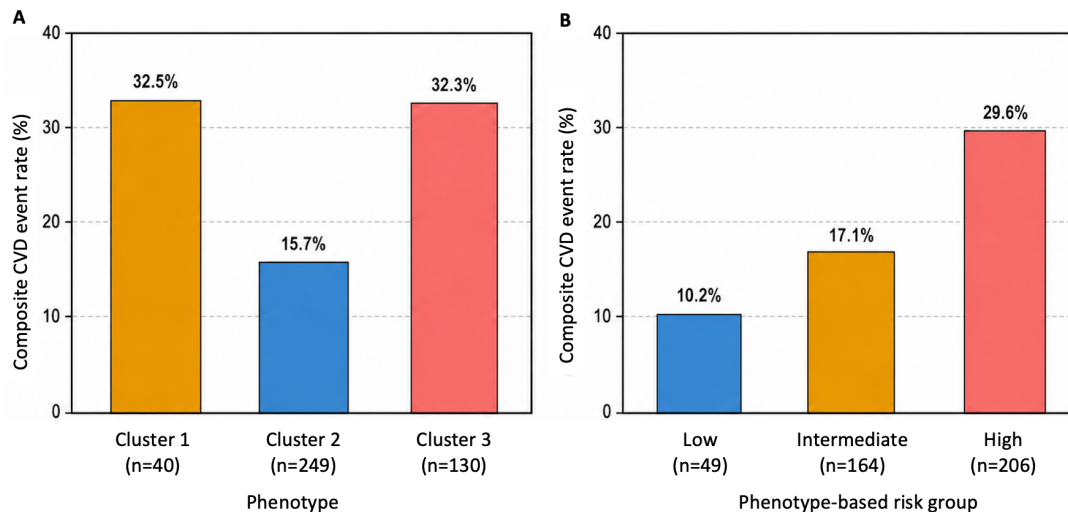
Group		ORs	95% CI	<i>P</i>
Phenotype (ref = Cluster 2)	Cluster1	2.59	1.20-5.39	0.0121
Phenotype (ref = Cluster 2)	Cluster3	2.57	1.56-4.23	0.0002
Risk group (ref = Low)	Intermediate	1.63	0.67-4.57	0.3090
Risk group (ref = Low)	High	3.71	1.62-10.07	0.0042

Odds ratios (ORs) and 95% confidence intervals (CIs) were estimated using logistic regression models with the composite cardiovascular endpoint (cardiovascular death, heart failure, myocardial infarction, or stroke) as the dependent variable. The Healthy Sleep–CVD Profile (Cluster 2) was used as the reference phenotype, and the low-risk group as the reference for risk group analyses. P values are two-sided; no adjustment for multiple comparisons was applied.

4.3 Observed event rates across phenotypes and risk strata

Observed event rates by externally assigned phenotype and by phenotype-based risk group are visualized in **Supplementary Figure 7**. Consistent with the derivation cohort, Cluster 1 and Cluster 3 showed substantially higher incidence of the composite cardiovascular endpoint than Cluster 2. Event rates were 32.5% for Cluster 1, 15.7% for Cluster 2, and 32.3% for Cluster 3 (Supplementary Figure 7A), reinforcing the adverse prognostic profile of the Aging–Hypertensive and Metabolic–Hypoxic phenotypes.

A similar gradient was evident when participants were stratified by the composite phenotype-based risk score. Using the fixed tertile thresholds (−7.445 and −2.364), 49 (11.7%), 164 (39.1%), and 206 (49.2%) individuals were classified as low-, intermediate-, and high-risk, respectively. The corresponding composite event rates were 10.2% in the low-risk group, 17.1% in the intermediate-risk group, and 29.6% in the high-risk group (Supplementary Figure 7B). These gradients demonstrate a clear stepwise increase in observed cardiovascular risk as the phenotype-based risk score increases.

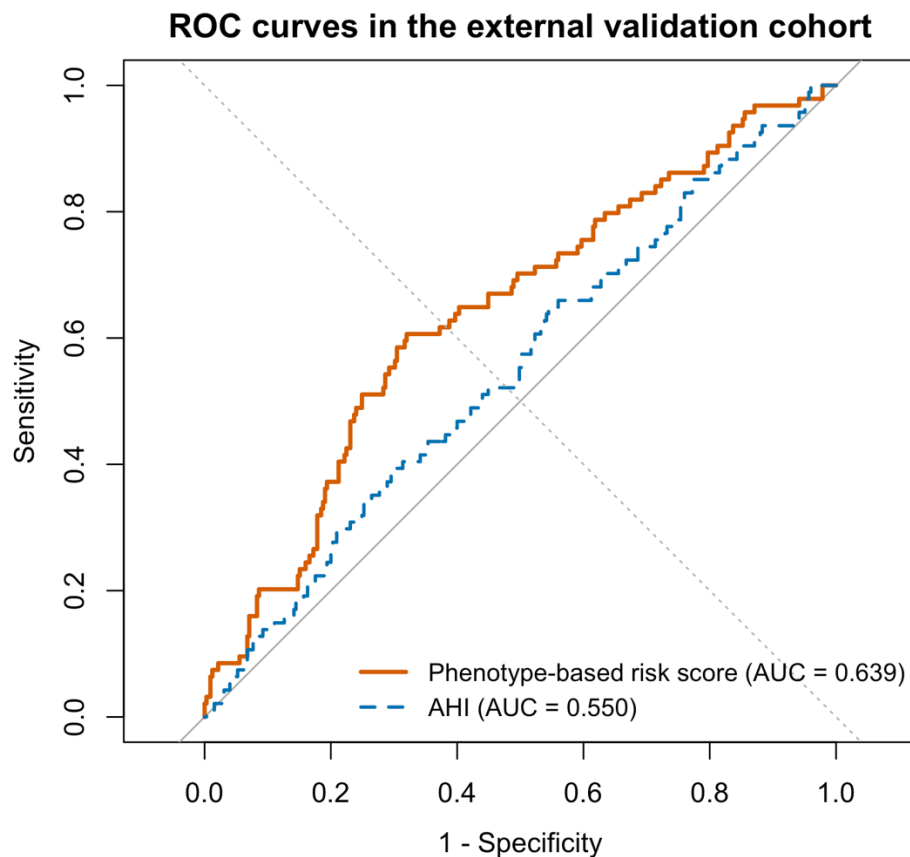


Supplementary Figure 7. External validation: composite cardiovascular event rates by phenotype and by phenotype-based risk group. Panel A shows the observed incidence of the composite cardiovascular endpoint in the external validation cohort according to externally assigned phenotype (n = 419 overall; Cluster 1, n = 40; Cluster 2, n = 249; Cluster 3, n = 130). Event rates were 32.5%, 15.7%, and 32.3% in Clusters 1, 2, and 3, respectively. Panel B displays

the corresponding event rates by phenotype-based risk group using tertile cut-offs learned in the SHHS training cohort (low risk, $n = 49$; intermediate risk, $n = 164$; high risk, $n = 206$). Event rates were 10.2%, 17.1%, and 29.6% across low-, intermediate-, and high-risk strata, demonstrating a stepwise increase in cardiovascular risk with higher risk-score categories.

4.4 Discrimination of the phenotype-based risk score versus AHI

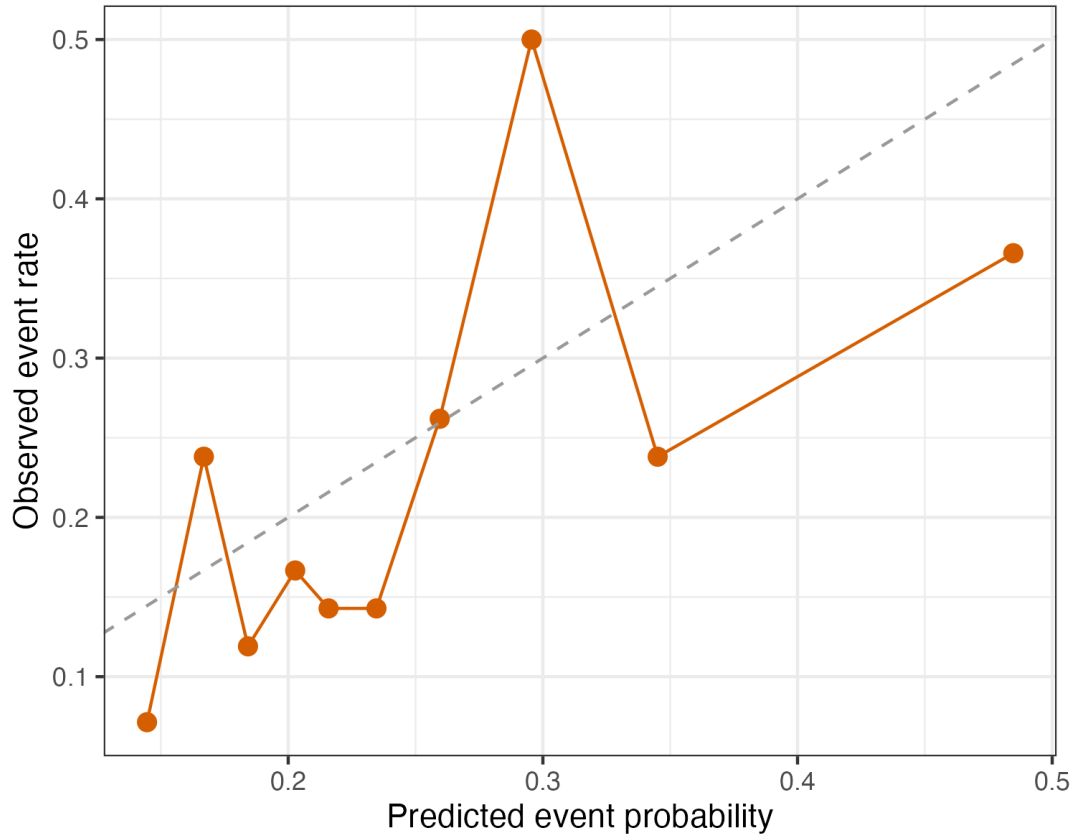
To compare discrimination of the composite phenotype-based risk score with that of conventional OSA severity metrics, ROC curves for the composite endpoint were constructed in the external validation cohort (*Supplementary Figure 8*). The phenotype-based risk score achieved an AUC of 0.64 (95% CI, 0.59–0.71) for predicting the composite cardiovascular endpoint, whereas AHI alone yielded an AUC of 0.55 (95% CI, 0.48–0.62). Thus, in this independent dataset, the phenotype-based score provided better separation of individuals with and without subsequent major cardiovascular events than AHI, despite being derived from the same underlying clinical and polysomnographic predictors.



Supplementary Figure 8. Discrimination of the phenotype-based risk score versus AHI in the external validation cohort. Receiver operating characteristic (ROC) curves for predicting the composite cardiovascular endpoint in the external validation cohort (n = 419). The solid orange line represents the phenotype-based composite risk score, which achieved an AUC of 0.64 (95% CI, 0.59–0.71). The dashed blue line represents AHI alone, with an AUC of 0.55 (95% CI, 0.48–0.62). The diagonal line indicates the line of no discrimination. The phenotype-based risk score provided better discrimination than AHI in the external dataset.

4.5 Calibration of the phenotype-based risk score

Calibration of the logistic mapping from the risk score to absolute event probability was examined by plotting observed versus predicted composite event rates across deciles of predicted risk (*Supplementary Figure 9*). Mean predicted probabilities closely tracked the observed event rates over most of the distribution, with points lying near the 45° line. Only modest underestimation of risk was observed in the highest deciles, suggesting that the risk score retains acceptable calibration when transported to the external cohort. Taken together with the discrimination and gradient analyses, these findings support that the phenotype-based classifier and composite risk score demonstrate clinically meaningful and externally transportable prognostic performance.



Supplementary Figure 9. Calibration of the phenotype-based risk score in the external validation cohort (n = 419). Calibration plot comparing observed versus predicted probabilities of the composite cardiovascular endpoint across deciles of predicted risk in the external validation cohort. Each point represents one decile of predicted risk (x-axis: mean predicted probability; y-axis: observed event rate), with the solid line connecting deciles and the dashed diagonal line indicating perfect calibration. The plot demonstrates overall good agreement between predicted and observed risks, with only modest miscalibration at the highest risk deciles.

5. Quantitative metrics for cluster number selection

Quantitative cluster-selection metrics are summarized in *Supplementary Table 5*. The three-cluster solution showed the highest average silhouette width and retained high resampling stability, while providing the most clinically interpretable phenotype structure. Although the two-cluster solution showed slightly higher resampling stability, it yielded a relatively coarse

partition that did not adequately separate the ageing–hypertensive and healthy sleep–CVD profiles. Solutions with four or more clusters showed lower average silhouette widths and/or lower stability and less clinically coherent partitioning, suggesting over-partitioning. Therefore, the three-cluster solution was selected before outcome analyses and retained for all downstream analyses.

Supplementary Table 5. Quantitative metrics supporting cluster number selection across candidate values of k

Candidate k	Average silhouette width	Resampling stability, adjusted Rand index, mean $\pm$ SD	Interpretation
2	0.161	0.924 ± 0.066	High resampling stability, but produced an overly coarse partition that did not adequately separate the ageing–hypertensive and healthy sleep–CVD profiles.
3	0.179	0.899 ± 0.088	Highest average silhouette width and retained high stability, while providing the most clinically interpretable three-phenotype structure. Selected solution.
4	0.139	0.481 ± 0.025	Lower average silhouette width than the three-cluster solution and substantially reduced stability, suggesting partition fragmentation.
5	0.104	0.737 ± 0.124	No meaningful improvement in compactness and lower stability than the three-cluster solution, suggesting over-partitioning.
6	0.120	0.532 ± 0.012	Lower average silhouette width than the three-cluster solution and markedly lower stability, indicating less robust partitioning.

Values were calculated using the prespecified clustering feature matrix. Average silhouette width was used to assess within-cluster compactness and between-cluster separation. Resampling-based stability was assessed using repeated 80% subsampling, and agreement between the full-sample and resampled clustering solutions was quantified using the adjusted Rand index. The final cluster number was selected by considering quantitative metrics together with PCA-based separation and clinical interpretability, without reference to outcome data.