

Development, Validation, and Algorithmic Fairness of Traditional Statistical and Machine Learning Models for Dementia Risk

Lingling Zheng, Feng Sha

1 Faculty of Computer Science and Engineering, Shenzhen University of Advanced Technology, Shenzhen, China;

2 Dept. of Computer Information Science, University of Macau, Macau, China; 3 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

1 Background

Machine learning is increasingly used for dementia risk prediction, yet its clinical utility and algorithmic fairness remain uncertain. We developed and validated traditional statistical and machine learning models for dementia risk in the general population, and evaluated their fairness across key subgroups.

2 Methods

Population-based cohort: **483,824 UK Biobank participants** without dementia, followed up to 13 years. Five all-cause dementia models from **56 evidence-based predictors**: Cox proportional hazards (Cox-PH), competing risk (CR), logistic regression, XGBoost, LightGBM. An age-only model served as the comparator. Performance assessed at risk-stratified thresholds; algorithmic fairness evaluated by subgroup analysis.

3 Results

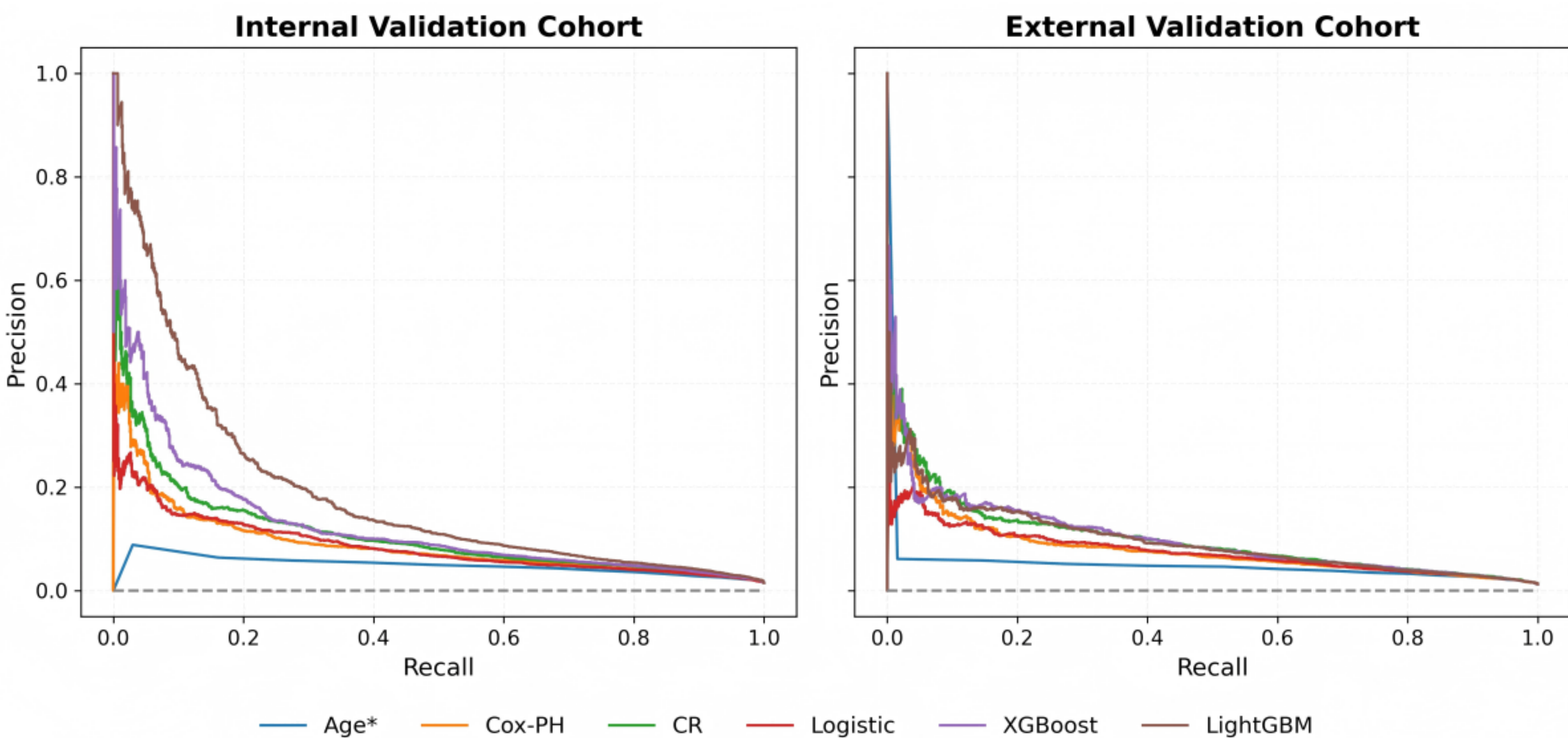
- **7,087 (1.46%)** participants developed dementia during follow-up.
- Similar discrimination across all models: **C-index 0.85–0.90**.
- Traditional models trailed gradient boosting internally (0.85–0.86 vs 0.87–0.90) but matched it externally (0.86–0.87 vs 0.87).
- At a 20% high-risk threshold: **74–76% detection, PPV 5%**; age-only model 68% / PPV 4% externally.
- Performance fell in older adults: **C-index 0.59–0.70 at age >70**.

4 Table 1 · Model performance — 13-year dementia risk (internal & external cohorts)

Model	Internal validation cohort			External validation cohort		
	C-index	AUROC	AUPRC	C-index	AUC	AUPRC
Age*	0.80 [0.79, 0.81]	0.80 [0.79, 0.81]	0.05 [0.04, 0.05]	0.82 [0.81, 0.83]	0.82 [0.80, 0.83]	0.04 [0.04, 0.05]
Cox-PH	0.85 [0.84, 0.86]	0.84 [0.83, 0.85]	0.10 [0.08, 0.11]	0.86 [0.84, 0.87]	0.85 [0.84, 0.86]	0.08 [0.07, 0.10]
CR	0.86 [0.85, 0.87]	0.85 [0.84, 0.86]	0.11 [0.10, 0.12]	0.87 [0.86, 0.88]	0.86 [0.85, 0.88]	0.10 [0.09, 0.12]
Logistic	0.85 [0.84, 0.86]	0.84 [0.83, 0.85]	0.09 [0.08, 0.11]	0.86 [0.85, 0.87]	0.85 [0.84, 0.87]	0.08 [0.07, 0.10]
XGBoost	0.87 [0.86, 0.88]	0.87 [0.86, 0.87]	0.14 [0.12, 0.16]	0.87 [0.86, 0.88]	0.86 [0.85, 0.88]	0.10 [0.09, 0.12]
LightGBM	0.90 [0.89, 0.91]	0.89 [0.89, 0.90]	0.26 [0.23, 0.28]	0.87 [0.85, 0.88]	0.86 [0.85, 0.87]	0.09 [0.08, 0.11]

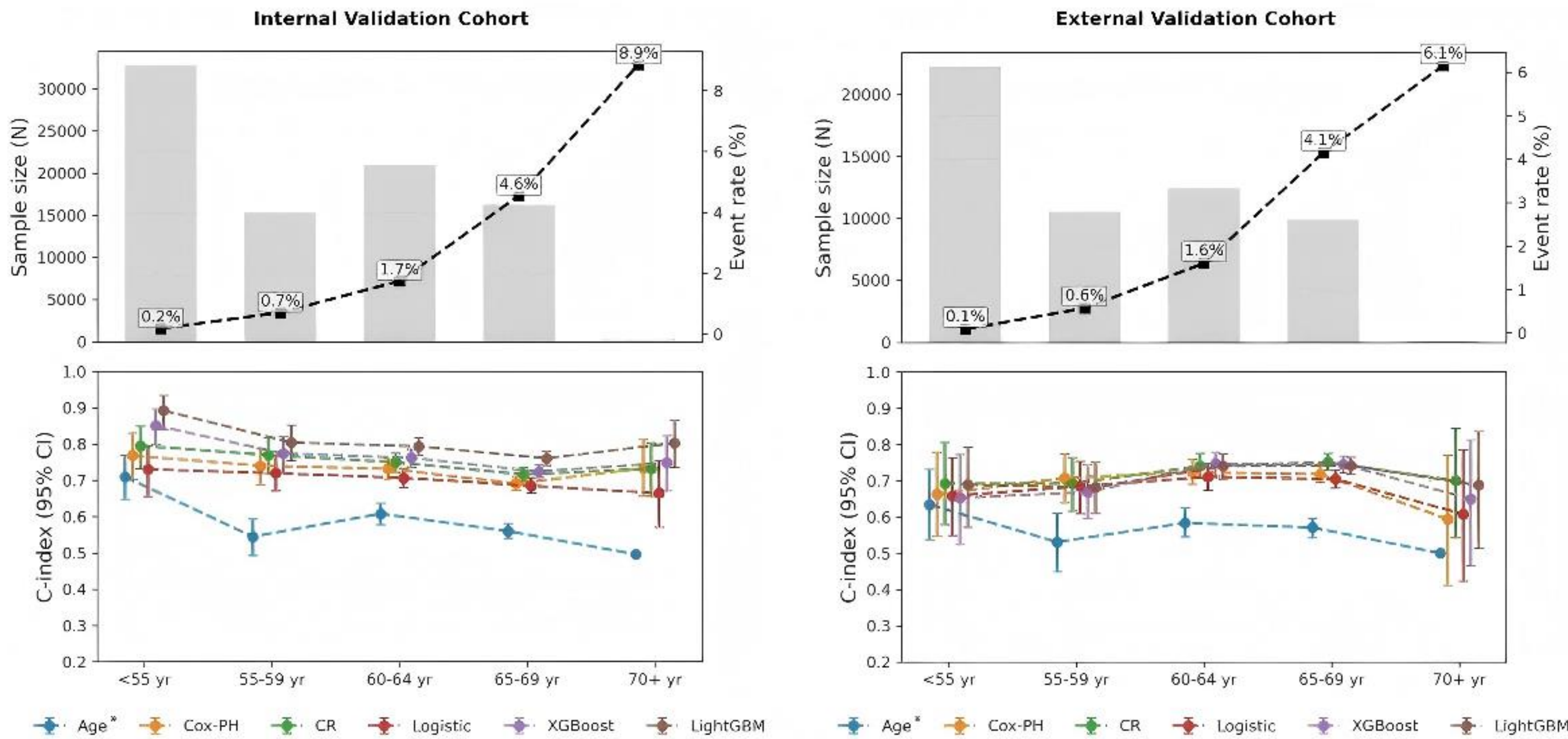
Age* = age-only logistic regression. Values are point estimates (95% CI from bootstrap resampling). AUROC, area under ROC curve; AUPRC, area under precision–recall curve; C-index, Harrell's concordance index.

Figure 1 · Precision–Recall Curves



Precision–recall curves for all six models in the internal (left) and external (right) validation cohorts.

Figure 2 · Clinical Utility Across Age Groups



Sample size, event rate, and C-index (95% CI) by age group for all six models in the internal and external validation cohorts. Discrimination weakens markedly in those aged 70+.

CONCLUSION

Traditional statistical and machine learning models performed comparably for dementia risk prediction in the general population. However, all models showed limited ability to identify high-risk individuals, especially among older adults — indicating limited clinical utility.

KEYWORDS

Dementia prediction · Machine learning · Model performance · Clinical utility · Algorithmic fairness
Corresponding author Feng Sha, feng.sha@siat.ac.cn