npj | dementia

Article

https://doi.org/10.1038/s44400-026-00071-1

Machine learning models for dementia risk prediction: evidence from the Sydney Memory and Ageing Study

Rebecca A. Chalmers[1] ✉, Matti Cervin[2], Carol Choo[3], Katya Numbers[4], Karen A. Mather[4], Henry Brodaty[4], Nicole A. Kochan[4], Perminder S. Sachdev[4,5] & Oleg N. Medvedev[1]

Early dementia risk stratification remains challenging despite advances in biomarker research. We evaluated machine learning approaches for predicting 10-year incident dementia among older adults using routinely collected demographic, cardiometabolic, and cognitive measures from the Sydney Memory and Ageing Study. From 1037 community-dwelling Australians aged ≥70 years at baseline, 432 participants were alive and assessed at approximately 10-year follow-up, of whom 119 developed dementia. We compared logistic regression, LASSO regression, random forest, and XGBoost models using baseline predictors. Models were trained on 70% of participants and evaluated on a 30% held-out test set. LASSO regression demonstrated the highest discrimination (AUC = 0.752), outperforming logistic regression (AUC = 0.707), random forest (AUC = 0.657), and XGBoost (AUC = 0.589), retaining four predictors: age, global cognition, fasting glucose, and cardiovascular disease risk score. At the Youden-optimal threshold, sensitivity was 0.698 and specificity was 0.736. Decision-curve analysis indicated greater net clinical benefit across a range of plausible risk thresholds. These findings indicate that a parsimonious model using four accessible variables can support dementia risk stratification up to a decade before diagnosis among older adults who survive to follow-up. The model is intended for use in research and specialist clinical settings where structured cognitive assessment is available; external validation is required prior to broader clinical deployment.

Machine learning (ML) methods are becoming more useful for identifying individuals at heightened dementia risk, potentially enabling earlier interventions in neurodegenerative processes[1]. Predictive models have been developed that incorporate diverse biomarkers including cardiovascular, metabolic, immunological, renal, hepatic, and genetic factors[2]. These biomarkers may signal hidden pathological processes while simultaneously reflecting cumulative risk exposure. Mental health variables uniquely function as both direct risk contributors and potential markers of underlying neurodegeneration. Depression exemplifies this duality, it elevates dementia risk independently while potentially signalling early neurodegenerative changes[3,4]. Despite mounting evidence linking psychological factors with cognitive decline, few studies integrate mental health measures alongside physiological markers within unified predictive frameworks[5].

Recent ML studies have predicted incident dementia using cognitive tests, demographics such as sex and education, electronic health records, neuroimaging, and blood biomarkers, typically reporting test-set AUCs in the ~0.65–0.85 range depending on modality and cohort. Systematic reviews note that cognition and age consistently emerge as the strongest features, with added gains from imaging or select plasma markers in some settings[6]. Notably, plasma phosphorylated tau (p-tau217) combined with age and brief cognitive measures has recently achieved high discrimination (AUC ≈ 0.79–0.96) for future dementia (specifically Alzheimer's Disease; AD) in community cohorts, underscoring the translational promise of blood-based biomarkers[7]. By contrast, "pace-of-aging" and mortality-linked biomarkers such as epigenetic clocks (e.g., GrimAge, DunedinPACE) and related composite aging measures, show robust associations with cognitive decline and dementia risk, but their incremental value in ML prediction remains uncertain and largely experimental, with limited clinical adoption to date[8]. Together, the literature suggests that parsimonious models built from routinely available cognitive and cardiometabolic data perform competitively, while next-generation aging biomarkers may offer future improvement once validated across diverse populations and workflows.

[1]University of Waikato, Hamilton, New Zealand. [2]Lund University, Lund, Sweden. [3]College of Healthcare Sciences, Division of Tropical Health and Medicine, James Cook University, Townsville, QLD, Australia. [4]Centre for Healthy Brain Ageing, Discipline of Psychiatry & Mental Health, School of Clinical Medicine, UNSW, Sydney, NSW, Australia. [5]Neuropsychiatric Institute, Prince of Wales Hospital, Randwick, NSW, Australia. ✉e-mail: rc139@students.waikato.ac.nz

Table 1 | Summary of ML algorithm performance for predicting incident dementia at 10-year follow-up among participants alive and assessed at Wave 6.

Model	AUC	Sensitivity	Specificity	PPV	NPV	Youden Threshold	HL Test
Logistic Regression	0.707	0.558	0.874	0.686	0.800	0.346	28.47**
LASSO L1	0.752	0.698	0.736	0.566	0.831	0.277	NA
Random Forest	0.657	0.535	0.793	0.561	0.775	0.340	NA
XGBoost	0.589	0.651	0.563	0.424	0.766	0.027	NA

PPV = TP/(TP + FP); NPV = TN/(TN + FN); AUC is threshold-independent; PPV/NPV vary with outcome prevalence. All performance metrics reported in this table were computed on the 30% held-out test set (n = 130; 43 dementia cases).

AUC area under the receiver operating characteristic curve, *PPV* positive predictive value, *NPV* negative predictive value.

**$p < 0.001$.

Various biomarkers have demonstrated associations with both cognitive decline and dementia onset including high-density lipoprotein (HDL) cholesterol, which exhibits complex U-shaped relationships with cognitive outcomes, challenging linear assumptions about protective effects[9,10]. Similarly, low-density lipoprotein (LDL) cholesterol has shown non-linear associations requiring careful interpretation[11]. Also, elevated triglycerides have been associated with increased dementia risk through vascular pathways[12]. Studies have shown that anthropometric measures including higher waist-hip ratio (WHR) predict cognitive decline more accurately than higher body mass index (BMI) in elderly populations[13], though higher BMI may paradoxically protect frail individuals[14]. Hyperglycaemia has been associated with accelerated cognitive deterioration in non-diabetic populations[15]. Uric acid demonstrates dose-dependent relationships with neurodegeneration[16]. Creatinine clearance reflects kidney function, linking renal health to cognitive outcomes[17]. Cardiovascular disease (CVD) risk scores integrate multiple factors predicting both vascular dementia and Alzheimer's disease[18]. Together, these findings highlight the complex, often non-linear relationships between metabolic, vascular, and renal biomarkers and cognitive decline, underscoring the need for multifactorial interpretation of dementia risk.

While much work has focused on biological and metabolic risks, psychological and mental health factors may add important predictive value. For example, depression and cognitive performance (i.e., the ability to learn, remember, and reason) have shown potential as predictors. Large-scale studies have demonstrated the association of depression with accelerated cognitive decline, potentially operating through inflammatory pathways, dysregulated stress responses, vascular compromise, genetic vulnerabilities, and reduced engagement in protective behaviours[19–24]. Lower cognitive performance itself predicts future dementia, with declining scores potentially reflecting preclinical neurodegeneration, chronic disease burden, or compromised health literacy affecting self-care[25–27]. Inflammatory markers including C-reactive protein (CRP), interleukin-6 (IL6), and interleukin-8 (IL8) independently predict cognitive decline and dementia onset[28–30]. Genetic factors, especially *APOE* ε4 allele status, substantially modifies dementia risk but are often unavailable during routine clinical screening, limiting their utility for initial risk stratification[31]. Overall, current prediction models remain dominated by biological and metabolic markers, yet growing evidence suggests that incorporating psychological and cognitive factors could enhance predictive accuracy, highlighting a key limitation in integrating these domains into comprehensive dementia risk models.

This study aimed to develop and validate ML models to predict 10-year incident dementia in older adults who survived to follow-up using baseline data from the Sydney Memory and Ageing Study[32] (MAS). We applied logistic regression, LASSO-penalized regression (Least Absolute Shrinkage and Selection Operator; LASSO), Random Forest (RF), and eXtreme gradient boosting of decision trees (XGB) algorithms using physiological biomarkers, depression scores, cognitive assessments, and inflammatory markers. Following initial model development, we evaluated whether incorporating *APOE* ε4 status would enhance predictive accuracy, recognizing that genetic testing typically follows rather than precedes clinical risk assessment. We hypothesized that ML approaches would provide sufficient accuracy to estimate dementia risk up to ten years before onset. Accordingly, the present study focuses on internally validating a parsimonious risk model in an older cohort, with the aim of informing dementia risk stratification in enriched clinical and research settings rather than population-wide screening.

Results

Using data from the MAS, we trained four supervised models to predict incident dementia up to 10 years before clinical diagnosis. Predictor variables included were measures of age, depression, anxiety, cognitive performance, WHR, education, CVD risk, and blood markers including HDL, LDL, triglycerides, creatinine clearance, uric acid, and glucose. APOE ε4 allele status was added post hoc. All analyses estimate dementia risk conditional on survival and dementia ascertainment at approximately 10-year follow-up.

Models were fit on a randomly selected 70% training subset and evaluated on the remaining 30% held-out test set. We compared standard logistic regression, LASSO, RF, and XGB, using relevant baseline predictors. Discrimination, calibration, threshold-based classification performance, and clinical utility were assessed and are summarised in Table 1, with supporting visualisations in Figs. 1–4. LASSO achieved the highest area under the receiver operating characteristic (ROC) curve with area under the ROC curve (AUC) of 0.752, consistent with acceptable discrimination for clinical risk models. AUC refers to the probability that the model ranks a randomly chosen case with dementia higher than a randomly chosen non-case; threshold-independent measure of discrimination (0.5 no better than chance; ~0.7 acceptable; ~0.8 good; ≥0.9 excellent). Logistic regression also showed acceptable discrimination (AUC = 0.707). RF and XGB showed more modest discrimination (AUC = 0.657 and 0.586, respectively). ROC curves are displayed in Fig. 1; the LASSO curve shows the largest bow away from the diagonal, reflecting the numerically highest AUC.

Operating points (Youden thresholds)

Using the Youden-optimal cut-point for each model, LASSO offered the most balanced sensitivity/specificity (0.698/0.736 at threshold 0.277), exceeding the often-used "fair" benchmark around 0.70 for both indices. Logistic regression prioritised rule-in performance with specificity 0.874 at sensitivity 0.558 (threshold 0.346), resulting in the highest positive predicted value (PPV; 0.686) among models which is useful when false positives (FP) carry meaningful costs. PPV refers to the proportion of predicted positives who truly have dementia, i.e., PPV = TP/(TP + FP); where TP is true positive and depends on prevalence. RF provided intermediate balance (0.535/0.793, threshold 0.340). XGB favoured sensitivity (0.651) at the expense of specificity (0.563) due to a very low threshold (0.027), aligning with a screening-oriented behaviour. These values are reported in Table 1 and visualised by the model in Fig. 1,

Calibration and goodness-of-fit

Calibration for logistic regression is shown in Fig. 2. Decile-based points track the identity line reasonably well at lower/mid probabilities, with some

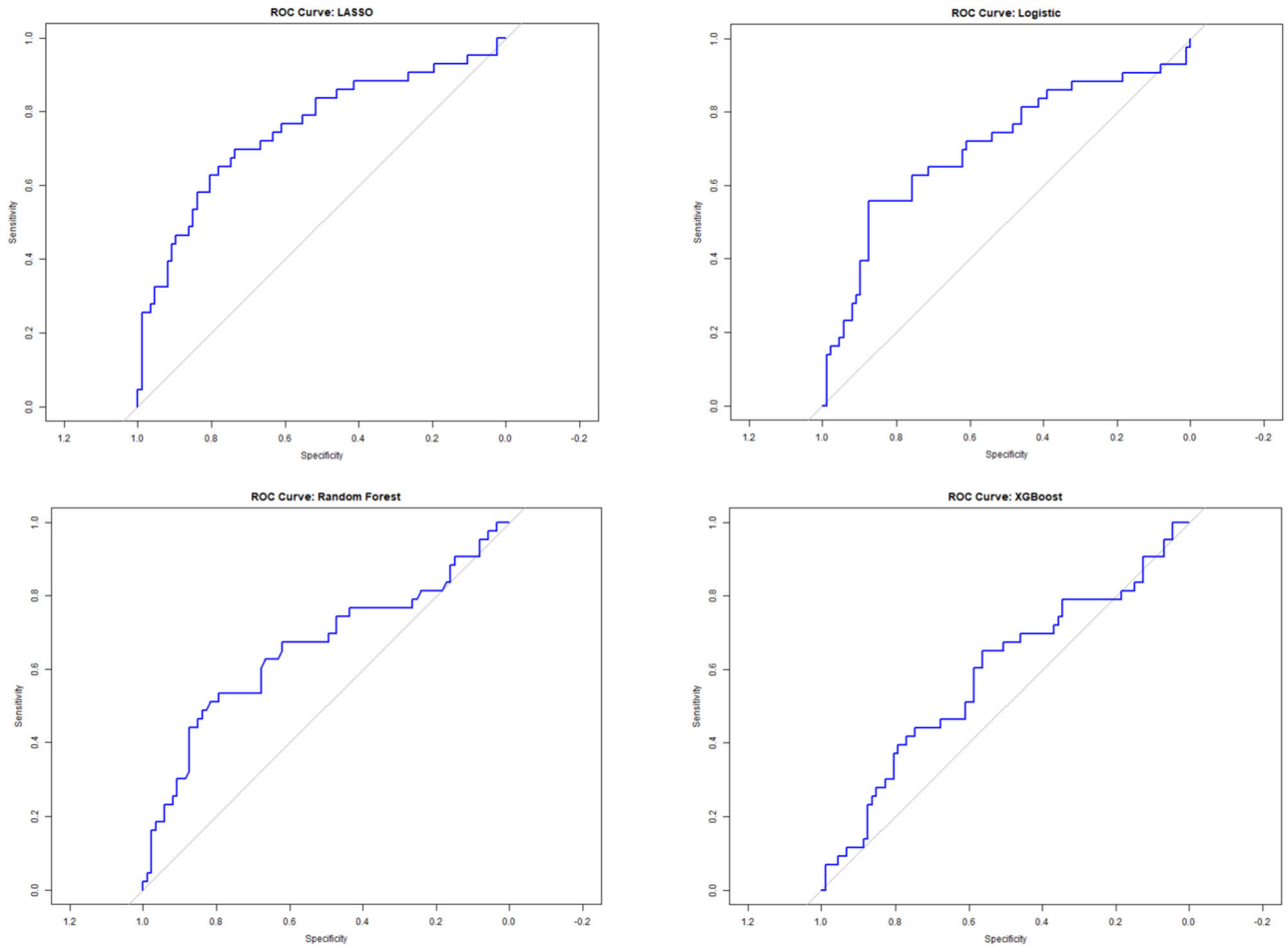

Fig. 1 | ROC curves for LASSO, logistic regression, random forest, and XG boost models predicting incident dementia at approximately 10-year follow-up among participants alive and assessed at Wave 6. LASSO demonstrated the highest discrimination (AUC = 0.752), followed by logistic regression (AUC = 0.707), random forest (AUC = 0.657), and XGBoost (AUC = 0.589). The diagonal line indicates no-discrimination (AUC = 0.50).

deviation at higher predicted risks. The Hosmer–Lemeshow (HL) statistic for logistic regression was $X^2 = 28.47$, $p < 0.001$. Given the HL test's known sensitivity to sample size and grouping, we interpreted it alongside the calibration plot and the threshold-based metrics in Table 1.

Model parsimony and variable importance

LASSO selected four non-zero predictors: Age (+ 0.065), CVDRisk (+ 0.033), Glucose (+ 0.010), and Cognition (− 0.442), indicating higher age, CVD risk, and glucose increase estimated risk while better cognition decreases it. For the non-parametric RF, Fig. 3 shows variable importance by mean decrease in accuracy and Gini; Cognition and Age rank most influential on both scales, broadly concordant with the LASSO selection.

Clinical utility (decision-curve analysis)

Figure 4 compares standardised net benefit across risk thresholds (0.01–0.99) for all models versus "treat-all" and "treat-none". LASSO provides the highest net benefit across a broad, clinically plausible range (≈0.10–0.65), with logistic regression generally second. RF offers modest benefit, and XGB crosses to or below "treat-none" over mid-to-higher thresholds, reflecting its lower specificity at its operating point. Taken together with AUC and calibration, these curves support LASSO as the primary model for risk stratification in this dataset, with logistic regression as a strong, simpler alternative when higher specificity is desired. The magnitude and consistency of net benefit across plausible thresholds suggest that LASSO could provide meaningful support for individual risk-based decisions, though further validation in external cohorts and clinical settings is needed before use in practice.

To assess whether genetic information could further enhance predictive accuracy, we conducted a post-hoc analysis including *APOE* ε4 carrier status in the LASSO model. This was motivated by extensive evidence linking *APOE* ε4 to increased dementia risk, but also recognising that genetic testing is not typically available at initial clinical screening. Adding *APOE* ε4 to the best performing model (Age, Cognition, Glucose, and CVDRisk) did not improve discrimination, calibration, or classification performance. In fact, test-set AUC decreased slightly (from 0.752 to 0.704), with sensitivity declining (0.698 → 0.472) while specificity increased modestly (0.736 → 0.871). Positive and negative predictive values also showed no net benefit. Given this lack of improvement, and in line with our aim to prioritise clinically available measures, we retained the original LASSO model without *APOE* ε4 as the primary predictor set.

Lasso model of the best fit

The final LASSO model retained four predictors: age, cognition, glucose, and CVD risk score, yielding the following regression equation:

$\text{logit}(p) = -6.937 + 0.065 \times \text{Age} - 0.442 \times \text{Cognition} + 0.010 \times \text{Glucose} + 0.033 \times \text{CVDRisk}$ with dementia risk calculated as $p = 1/(1 + \exp(-\text{logit}(p)))$.

This formula can be implemented directly in Excel so that individual risk can be estimated from routine clinical variables. For example, a 75-year-old with a cognition score of −0.40, glucose of 5.8 mmol/L, and a CVD risk score of 20 would yield $\text{logit}(p) = -1.167$ and an estimated 10-year dementia

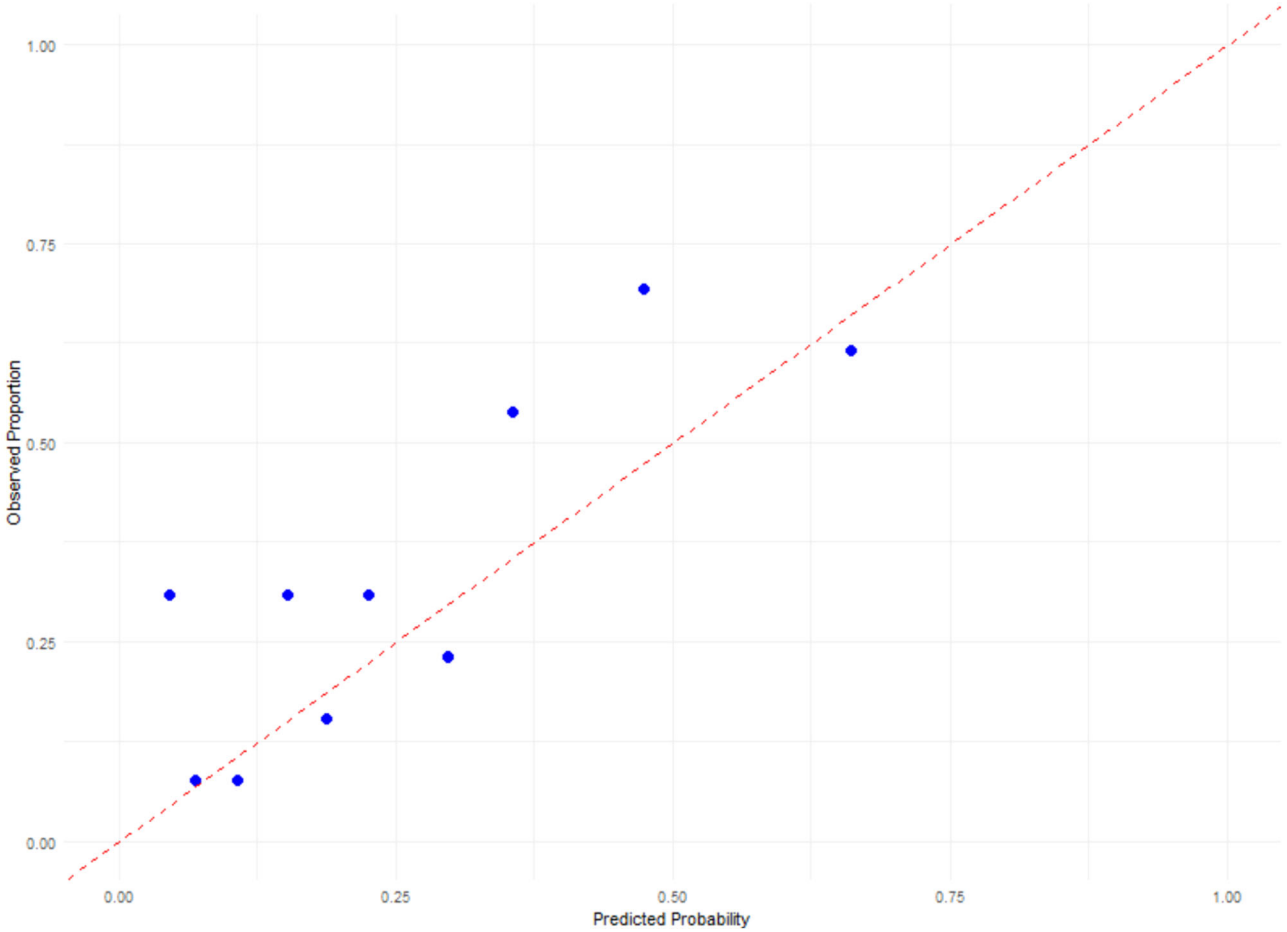

Fig. 2 | Calibration plot for the logistic regression model predicting incident dementia at approximately 10-year follow-up among participants alive and assessed at Wave 6. Points represent observed versus predicted risk across deciles of predicted probability, with the 45° line indicating perfect calibration.

risk of 0.237 (23.7%). Because this value is below the 0.277 cut-off, the individual would be classified as lower risk. The cut-off serves as a classification rule and does not rescale probabilities. The equivalent Excel formula is:

=1/(1 + EXP(-(-6.937 + 0.065*Age - 0.442*Cognition + 0.010*Glucose + 0.033*CVDRisk))) where Age, Cognition, Glucose, and CVDRisk are replaced by the relevant cell references. Probabilities derived from this formula represent the estimated absolute risk of dementia within ten years in the study population, conditional on survival and follow-up.

For clinical interpretation, probabilities should be judged against the Youden-optimal cut-off of 0.277 identified in our analysis. Values above this threshold indicate that an individual is classified as "at elevated risk," whereas values below suggest lower risk, albeit with the trade-offs between sensitivity and specificity described earlier. Because the model was trained in a Sydney cohort of older white Australians, its absolute risk estimates may require recalibration when applied to other populations with different dementia prevalence. This can be achieved by adjusting the model intercept while leaving the predictor coefficients unchanged, using the standard calibration-in-the-large approach based on the difference in log-odds: Intercept_new = Intercept_base + [logit(Prev_target) − logit(Prev_study)], where logit(p) = ln[p/(1−p)]. For example, if target population prevalence is 35% while the study prevalence is 28%, the adjustment is logit(0.35) − logit(0.28) = − 0.619 − (− 0.944) = 0.325, and the new intercept becomes −6.937 + 0.325 = −6.612. This approach, which assumes the predictor effects (slope) remain valid in the new population, mirrors established recalibration procedures for CVD risk models and ensures that risk probabilities remain interpretable and clinically relevant across settings. Note that the simpler approximation ln(Prev_target/Prev_study) is accurate only when outcome prevalence is low (rare disease assumption) and may introduce bias when prevalence exceeds approximately 10%. Given the dementia prevalence in this study (28%), the log-odds formula should be used rather than the ratio approximation.

Adjusting prediction for age

Our model was developed in a cohort aged ≥70 years. When applying it to younger groups (e.g., 40–49, 50–59, 60–69), absolute risks require recalibration to the age-specific dementia prevalence of the target population. We therefore recommend adjusting the model intercept by age band using calibration-in-the-large based on the difference in log-odds: Intercept_new = Intercept_base + [logit(Prev_target,band) − logit(Prev_study,band)], where logit(p) = ln[p/(1−p)], Intercept_new is the age-adjusted intercept, Intercept_base is the intercept from the trained model (− 6.937), Prev_target,band is the dementia prevalence in the target population for the specific age band, and Prev_study,band is the prevalence in the corresponding age stratum of the training data.

This preserves discrimination while aligning predicted probabilities with age-specific disease frequency. The dementia risk calculator included in Supplementary Data 1 is accessible to use and adjustable for age and population prevalence with instructions in Supplementary File 1.

Discussion

This study examined whether machine learning approaches applied to routinely collected baseline data could predict incident dementia up to ten years before diagnosis in the Sydney Memory and Ageing Study. Among the algorithms evaluated, a parsimonious LASSO regression model, incorporating age, global cognition, fasting glucose, and cardiovascular disease risk

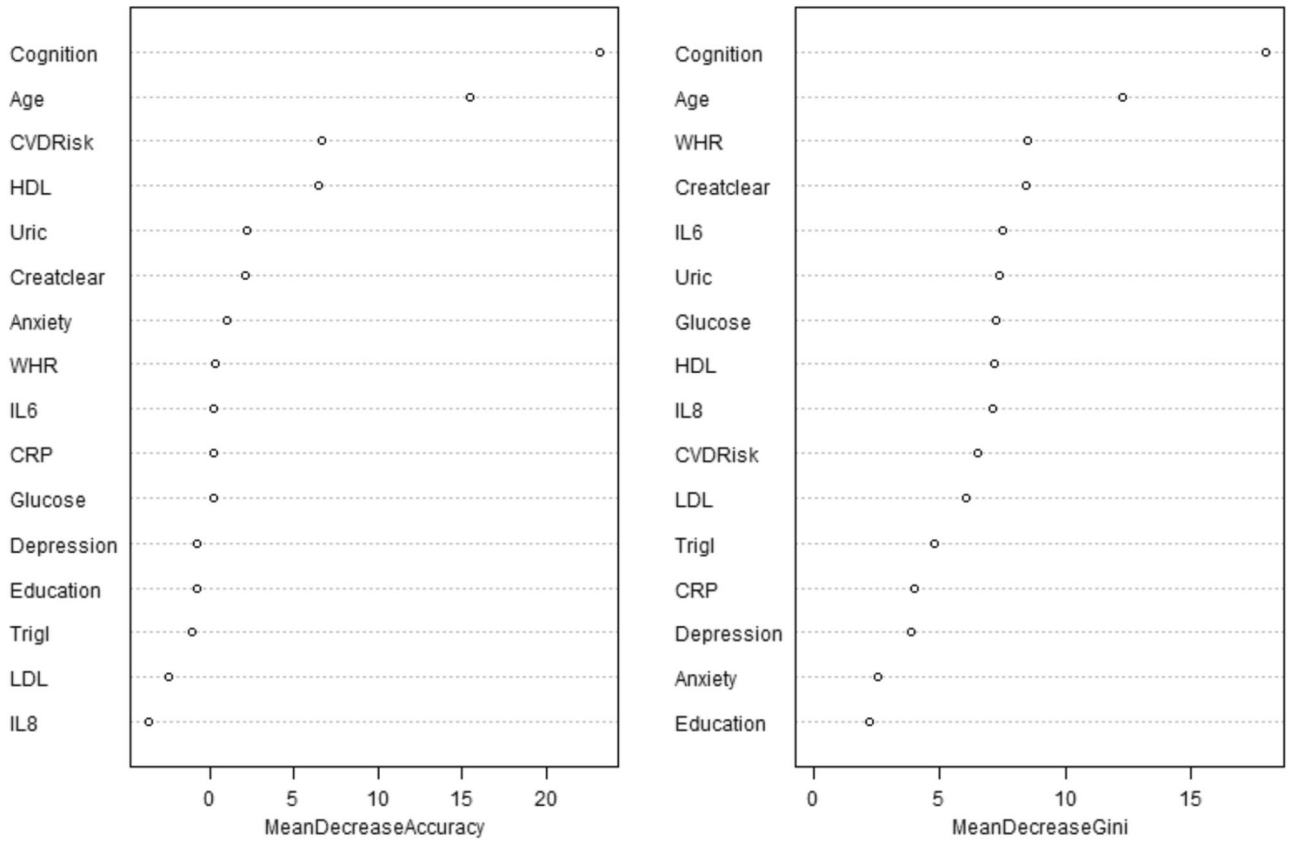

Fig. 3 | Variable importance for the random forest model showing the relative contribution of predictors to classification accuracy. Importance is displayed as mean decrease in accuracy and mean decrease in Gini impurity.

demonstrated the strongest discrimination (AUC = 0.752) and clinical utility. Importantly, these findings should be interpreted as evidence of internal predictive validity among older adults who survived and were assessed at follow-up, rather than as readiness for immediate clinical deployment. Within this context, the results suggest that meaningful dementia risk stratification may be achievable using a small number of accessible predictors without reliance on specialised biomarkers or genetic testing.

Our findings align with prior ML studies demonstrating that age and cognitive performance are among the most robust predictors of future dementia. Reported discrimination in comparable cohort-based studies typically ranges from moderate to good, depending on the predictors and follow-up period. While emerging blood-based biomarkers and neuroimaging measures have shown promise in improving prediction in some settings, these approaches remain costly and are not yet widely available for routine use. In contrast, the present study demonstrates that a parsimonious model based on readily obtainable cognitive and cardiometabolic measures can achieve discrimination comparable to more complex approaches, supporting its potential utility in settings where advanced biomarkers are unavailable.

Although age and cognitive performance are well-established risk factors for dementia, the value of the present work lies not in identifying novel predictors, but in quantifying their joint predictive contribution within a transparent and reproducible modelling framework. Clinical judgement alone does not provide calibrated estimates of absolute risk or formal evaluation of trade-offs between sensitivity and specificity. By translating known risk factors into a validated probabilistic model with demonstrated net benefit, this study contributes a reproducible tool for risk stratification rather than a restatement of clinical intuition.

The addition of APOE ε4 carrier status did not improve predictive performance in this cohort. This finding should not be interpreted as diminishing the biological importance of APOE ε4, but rather reflects its limited incremental value for long-term dementia prediction when age, cognition, and cardiometabolic risk are already accounted for. These results support the prioritisation of non-genetic predictors for initial risk stratification, particularly in settings where genetic testing is not routinely available.

Decision-curve analysis suggested that the LASSO model provided greater net benefit than alternative approaches across a range of plausible risk thresholds. This finding indicates potential value for risk stratification and prioritisation in selected settings, rather than definitive guidance for treatment decisions. As with all prediction models, the appropriate risk threshold will depend on the clinical context, downstream interventions, and tolerance for false-positive and false-negative classifications.

When applied to populations with different dementia prevalence or age structure, recalibration of absolute risk estimates is required. Consistent with established practices in cardiovascular risk prediction, calibration-in-the-large can be achieved by updating the model intercept using population-specific disease frequency, while retaining the original predictor coefficients. This is analogous to widely accepted recalibration practices in CVD risk prediction, such as recalibrating Framingham or SCORE models using local population data[33]. These methods have proven effective: for example, a population-based recalibration of survival neural network models for CVD corrected substantial underestimation of risk (up to 60%) by leveraging summary statistics from the new population without altering the original model structure[34]. Such recalibration preserves discrimination but depends on the availability of reliable external epidemiological data. Without recalibration, predicted probabilities should be interpreted cautiously and primarily in relative terms.

Taken together, these findings support the feasibility of developing parsimonious dementia risk models based on routinely collected data in ageing cohorts. However, further work is required to evaluate stability across

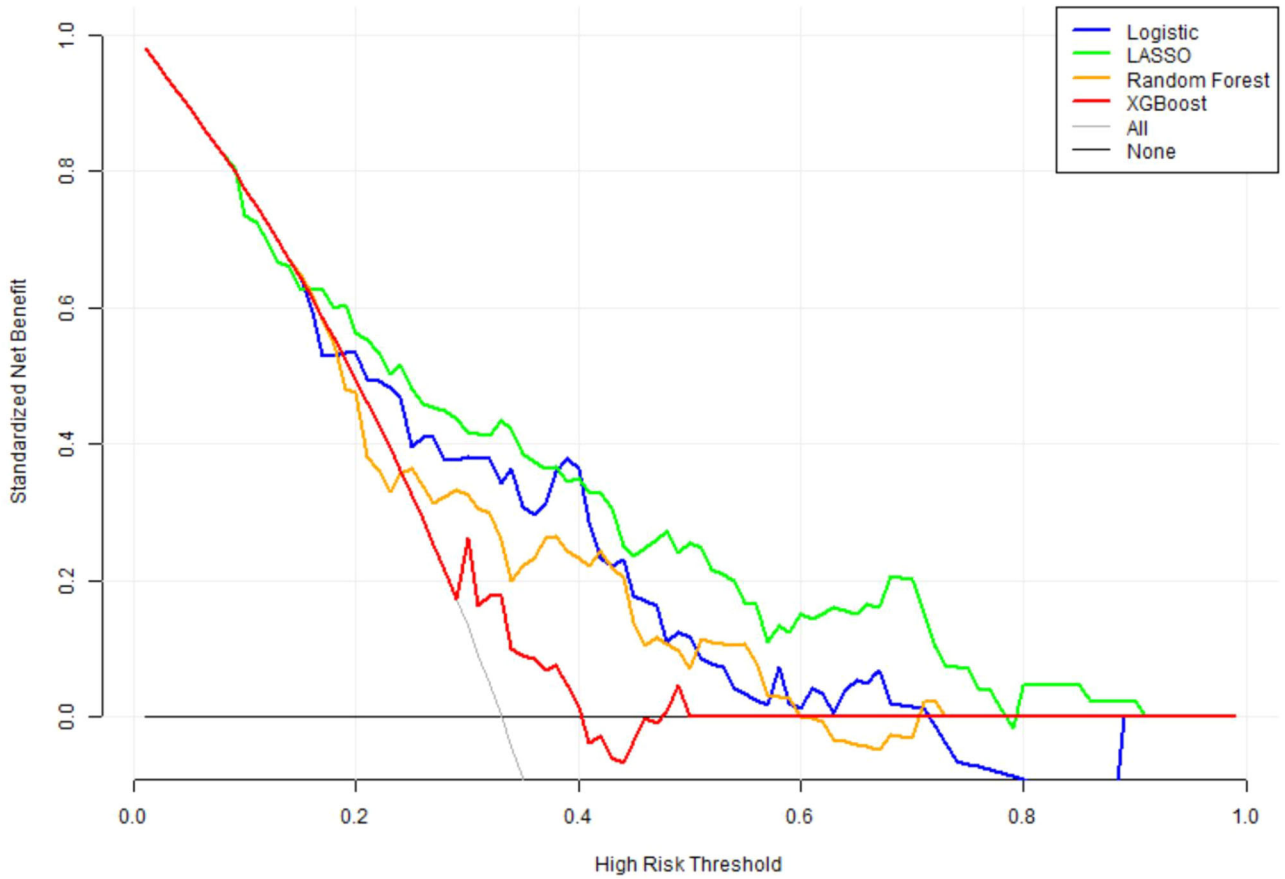

Fig. 4 | Decision-curve analysis comparing standardised net benefit across threshold probabilities for logistic regression, LASSO, random forest, and XGBoost models relative to "treat-all" and "treat-none" strategies.

resampling frameworks, validate performance in independent populations, and assess utility using alternative cognitive measures more commonly employed in primary care. Ultimately, translation into practice will depend not only on predictive performance, but also on clinician acceptability, patient understanding of probabilistic risk information, and evidence that early risk stratification leads to improved outcomes.

Several limitations should be considered when interpreting these findings. First, model development and evaluation were conditioned on survival and dementia ascertainment at approximately 10-year follow-up. Participants who died, withdrew, or were otherwise unavailable for assessment were excluded, introducing potential selection bias and limiting generalisability to the baseline cohort. Consequently, predicted risks reflect dementia probability among individuals who survived and were assessed, rather than unconditional population risk.

Second, internal validation relied on a single stratified train–test split (70/30) with a relatively small effective sample ($n = 432$; 119 events). Although this approach provides an unbiased estimate of out-of-sample performance, with only 43 events in the test set, discrimination and calibration metrics may be sensitive to the particular random split and should be interpreted as point estimates subject to sampling variability. Confidence intervals were not computed for the primary metrics; future work should employ repeated cross-validation, bootstrap internal validation with optimism correction, or external validation in independent cohorts to better characterise model stability and provide interval estimates for key performance metrics.

Third, the global cognition predictor was derived from a comprehensive neuropsychological battery and standardised to the study's normative reference group. While this enhances sensitivity to subtle cognitive differences, it limits immediate applicability to primary care settings where brief screening instruments are more commonly used. Published studies report strong correlations between the Mini Mental State Examination (MMSE) and comprehensive neuropsychological batteries ($r = 0.84$ with WAIS-R Verbal IQ), and Telephone Interview for Cognitive Status-Modified (TICS-M) scores correlate strongly with face-to-face neuropsychological composites ($r \approx 0.81$) and show high concordance with MMSE. Future studies should therefore evaluate whether MMSE or TICS-M scores, transformed to z-scores using population norms, can serve as acceptable proxies for comprehensive cognitive assessment in this prediction model. If validated, such substitution would substantially enhance the model's scalability to routine primary care and telephone-based screening contexts.

Fourth, although we examined the incremental value of APOE ε4 carrier status, other emerging biomarkers (e.g., plasma proteomics, polygenic risk scores, longitudinal cognitive change) were not included and may improve prediction in future models once clinically validated.

Finally, this study did not address downstream implementation considerations, including integration into clinical workflows, user acceptability, or cost-effectiveness. These factors, along with external validation in independent cohorts and evaluation using alternative cognitive measures, represent important directions for future research. Demonstrating clinical utility will ultimately require evidence that early risk stratification leads to improved outcomes through targeted monitoring or preventive interventions.

Method

Participants

The sample at baseline comprised 1037 participants without dementia from the Sydney Memory and Ageing Study (MAS) run by the Centre for Healthy Brain Ageing (CHeBA) in New South Wales, Australia. For more details see Sachdev et al.[32]. Dementia status was ascertained at approximately 10-year

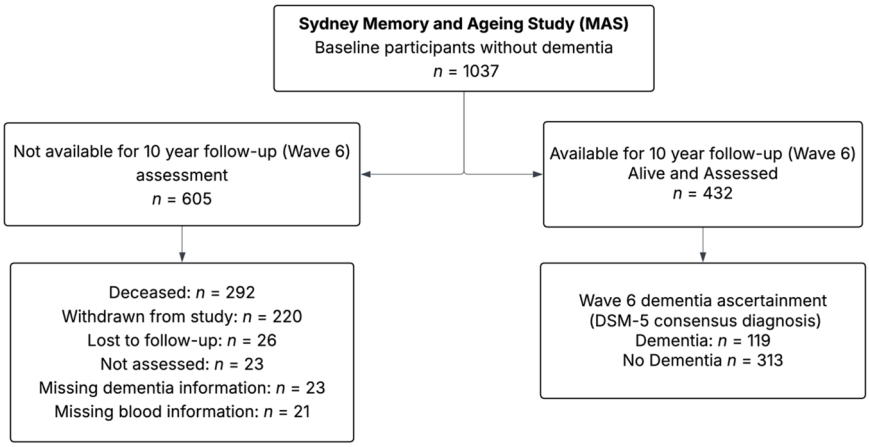

Fig. 5 | Participant flow diagram. Flow of participants from baseline of the Sydney Memory and Ageing Study to the final analytic sample. The analytic cohort comprises participants who survived and were assessed at 10-year follow-up (Wave 6). Attrition reflects death, withdrawal, loss to follow-up, missing dementia ascertainment, and missing baseline blood data.

Table 2 | Baseline characteristics of included vs excluded participants

Variable	Unit	Included in study			Not included in study			sig diff?
		n	Mean	*SD*	*n*	Mean	*SD*	
n		432			605			
Age	years	432	76.61	4.20	605	79.56	4.85	$p < 0.001$
Depression		432	1.84	1.70	600	2.60	2.26	$p < 0.001$
Anxiety		418	1.08	1.82	589	1.15	1.91	
Global cognition	z-score	430	-0.38	1.24	602	-0.98	1.41	$p < 0.001$
IL-8	pg/ml	419	19.65	14.63	497	20.50	12.24	
IL-6	pg/ml	419	6.27	9.76	497	6.81	6.88	
CRP	mg/L	431	2.73	5.35	502	3.29	5.53	
HDL Cholesterol	mmol/L	430	1.44	0.41	501	1.44	0.46	
LDL Cholesterol	mmol/L	428	2.84	0.87	500	2.76	0.88	
Triglyceride	mmol/L	430	1.06	0.54	503	1.07	0.55	
Creatinine Clearance		428	0.70	0.18	493	0.63	0.19	$p < 0.001$
Uric acid	mmol/L	326	0.30	0.09	396	0.31	0.10	
Glucose	mmol/L	431	5.81	1.12	499	5.93	1.25	
Waist-Hip Ratio		412	0.89	0.08	582	0.91	0.08	$p < 0.001$
Education		432	5.92	7.76	605	5.66	7.58	
CVD Risk		418	16.72	3.64	580	17.73	3.37	$p < 0.001$

follow-up (Wave 6). The analytic sample for the present study included participants who were alive and assessed at Wave 6 and had complete data for dementia ascertainment and baseline predictors. Participant inclusion and attrition are summarised in Fig. 5. All participants provided written consent to participate in this study, which was approved by the University of New South Wales Human Ethics Review Committee (HC 05037, 09382, 14327) and conducted in accordance with the Declaration of Helsinki. Clinical trial number: not applicable.

Baseline characteristics of included versus excluded participants are shown in Table 2. Excluded participants were significantly older at baseline, with higher depression scores, lower cognition scores, lower creatinine clearance, higher waist-hip ratio and higher CVD risk.

Measures

Dementia at Wave 6 was diagnosed in accordance with DSM-5 criteria based on the results of the Mini Mental State Exam (MMSE) and Addenbrooke's Cognitive Examination III (ACE-III). Participants were taken for review by the consensus diagnosis panel if they met one or more of the following criteria: MMSE ≤ 24; ACE-3 ≤ 82; drop in MMSE ≥ 3 points; elevated NPI data; previous dementia diagnosis. Consensus diagnoses were made by an expert panel of clinicians including psychiatrists, neuropsychiatrists, clinical neuropsychologists and clinical psychologists specialising in older people, using all available clinical data and MRI where available. Dementia status at Wave 6 was treated as a binary outcome indicating incident dementia during follow-up among participants assessed at that wave.

CVD risk was quantified using the sex-specific multivariable risk algorithms for "general CVD" from the Framingham Heart Study[33]. To maintain data granularity in our older sample, we used the point-based scoring system rather than probability percentages. Points were allocated based on: current smoking status, diabetic status, systolic blood pressure (average of two seated readings), total cholesterol, HDL cholesterol, and use of antihypertensive medication (which calibrates the weighting of the blood pressure score). Cholesterol and HDL values were converted from mmol/L to mg/dL (multiplying by 38.67) for point allocation. When phlebotomy was unavailable, BMI (kg/m^2) was substituted for cholesterol data per protocol. Importantly, participants' age was excluded from the total to prevent ceiling effects and to allow age to remain a distinct covariate. While diabetic status is a component of the CVD risk score, it does not capture glucose variability across the full sample. Therefore, fasting glucose (mg/dL) was included as a separate continuous variable in the model to account for glycaemic variability.

As outlined by Trollor et al.[35], a comprehensive neuropsychological test battery was administered by trained research assistants. Ten tests were conducted representing the diverse array of cognitive functions impaired with aging. To assess the memory domain, several tests were conducted, including Logical Memory Story A[36] (delayed recall), Rey Auditory Verbal Learning Test[37] (total learning, short-term and long-term delayed recall) and Benton Visual Retention Test[38] recognition. Trail Making Test B[39], and a Controlled Oral Word Association Test[40] were part of the executive function domain. Both the 30-item Boston Naming Test[41] and semantic fluency (Animal Naming Task[42]) were utilised to examine the language domain. Trail Making Test A[43] and Digit–Symbol Coding[44] task were used to assess the processing speed domain. The Block Design[44] task assessed the visuo-spatial domain.

Global cognition was calculated as the average of the five cognitive domain composite scores (attention/processing speed; memory; executive function; language; and visuo-spatial ability). Individual neuropsychological test scores were first standardised as z-scores using the mean and standard deviation of a normative "healthy" reference group at baseline. Domain composites were the average of constituent test z-scores, and the global cognition score was the average of these composites and standardised relative to the normative reference group. Scores were only calculated for participants with at least 8 of the 12 individual tests present[32].

Depression was measured by the Geriatric Depression Scale[45] (GDS-15). The 15 item, dichotomous, self-rated scale calculates a depression score by summing up points allocated to a 'yes' or 'no' answer. A score from 0 to 4 indicates absence of depression; 5 to 8 indicates mild depression; 9 to 11 indicates moderate depression; and 12 to 15 indicates severe depression. The scale has been proven reliable and valid in the assessment of depressive symptoms in the elderly.

Anxiety was measured on the Goldberg Anxiety Scale[46] (GAS). The 9-item, dichotomous, self-rated scale calculates an anxiety score by summing up points allocated to a 'yes' or 'no' answer. The first four items act as screening questions; endorsement of two or more prompts administration of the remaining five items. A score of 5 or more on the full scale suggests clinically significant anxiety. The GAS has demonstrated good reliability and validity as a brief measure of anxiety symptoms in community and clinical samples.

As outlined in Trollor et al.[35] and Lipnicki et al.[47], fasting blood was collected and various biomarkers were analysed including cholesterol, triglycerides, creatinine, glucose, uric acid, IL6, IL8, and CRP. For the final LASSO model, the four retained predictors are specified as follows: Age is measured in years at baseline; Global Cognition is a standardised z-score derived from the composite of domain scores (mean = 0, SD = 1 in the MAS normative reference group, with higher scores indicating better performance); Glucose is fasting plasma glucose measured in mmol/L; and CVD Risk is the Framingham 10-year cardiovascular risk score expressed as integer points (range approximately 0–30 in this sample), calculated using age, sex, smoking status, diabetic status, systolic blood pressure, total cholesterol, HDL cholesterol, and antihypertensive medication status. No additional scaling or transformation was applied to these predictors in the final model equation.

DNA was extracted from peripheral blood or donated saliva samples using standard procedures. Genotyping was undertaken examining the two SNPs, rs7412 and rs429358, to determine the *APOE* genotype and ε4 carrier status as described in Song et al.[48].

Waist and hip circumference were measured to calculate waist:hip ratio. Education was coded by highest qualification achieved from completion of primary school (1), incomplete high school (2), completed high school (3), incomplete tertiary (4), completed tertiary (5).

Data analyses

We evaluated four supervised learning algorithms: standard logistic regression, LASSO-penalised logistic regression, Random Forest (RF), and XGBoost (XGB) to predict incident dementia up to 10 years before diagnosis using baseline measures from the MAS. Models were trained on a randomly selected 70% subset of participants and evaluated on the remaining 30% held-out test set to provide an unbiased estimate of out-of-sample performance.

The binary outcome was incident dementia status at approximately 10-year follow-up. Candidate predictors were routinely available demographic, cognitive, cardiometabolic, and inflammatory measures, including the composite CVD risk score. Analyses were conducted on the imputed baseline dataset. Unused variables such as time-to-event were removed, and the outcome was coded as a binary factor. A stratified 70/30 train–test split was applied (set.seed = 123) to preserve outcome prevalence across subsets.

Model specifications were as follows. Logistic regression was fit with all candidate predictors entered simultaneously. LASSO regression ($\alpha = 1$) was fit using 10-fold cross-validation on the training set, with the penalty parameter chosen at the value of lambda that minimised cross-validated error (lambda.min). Predictors were internally standardised by glmnet to stabilise penalisation; selected coefficients were later back-transformed to the raw data scale for interpretation and implementation. Random Forest models were fit with 1000 trees and default mtry; XGBoost was trained with a binary logistic objective and AUC as the optimisation metric, with 300 boosting rounds. Note that while LASSO underwent hyperparameter tuning via cross-validation, Random Forest and XGBoost were evaluated using default or minimally tuned settings. This reflects our primary aim of identifying a parsimonious, interpretable model rather than optimising ensemble methods, and means that the reported performance of RF and XGBoost should be interpreted as baseline estimates that might improve with systematic hyperparameter optimisation.

Discrimination was assessed by computing the area under the receiver-operating characteristic curve (AUC) on the test set. Operating thresholds for binary classification were defined using Youden's J statistic (maximising sensitivity + specificity − 1). This criterion was chosen because it provides an objective, prevalence-independent method for identifying the threshold that optimally balances sensitivity and specificity, facilitating fair comparison of model performance across algorithms. At these thresholds, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were estimated.

Calibration was evaluated by plotting observed versus predicted risk across deciles of predicted probability. For logistic regression, calibration was also formally tested using the Hosmer–Lemeshow statistic, recognising its sensitivity to sample size.

Clinical utility was examined using decision-curve analysis (DCA) over threshold probabilities ranging from 0.01 to 0.99, comparing each model to "treat-all" and "treat-none" strategies.

Finally, we conducted a post-hoc sensitivity analysis that added *APOE* ε4 status to the LASSO model to evaluate whether this genetic risk factor improved prediction. *APOE* ε4 was coded as carrier versus non-carrier. Models were re-fit on the training set with *APOE* included, and predictive performance was re-evaluated on the held-out test set.

All analyses were conducted in R[49]. We used the glmnet package for LASSO regression, pROC for ROC/AUC estimation and Youden thresholds, caret for classification metrics, randomForest for RF models, xgboost for gradient boosting, ResourceSelection for the Hosmer–Lemeshow test, rmda for decision-curve analysis, and ggplot2 for data visualisation.

This study was reported in accordance with the TRIPOD recommendations for prediction model development and validation.

Data availability

The terms of consent for research participation stipulate that an individual's data can only be shared outside of the MAS investigators group if the group has reviewed and approved the proposed secondary use of the data. This consent applies regardless of whether data has been de-identified. Access is mediated via a standardised request process managed by the CHeBA Research Bank, who can be contacted at [ChebaData@unsw.edu.au] (mailto:ChebaData@unsw.edu.au).

Code availability

The underlying R code for this study is available in Supplementary File 1.

Received: 16 October 2025; Accepted: 19 February 2026;
Published online: 27 April 2026

References

1. Javeed, A. et al. Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions. *J. Med. Syst.* **47**, 17 (2023).
2. Belsky, D. W. et al. Quantification of biological aging in young adults. *Proc. Natl. Acad. Sci.* **112**, E4104–E4110 (2015).
3. Cummings, J. The Role of Neuropsychiatric Symptoms in Research Diagnostic Criteria for Neurodegenerative Diseases. *Am. J. Geriatr. Psychiatry* **29**, 375–383 (2021).
4. Hayley, S., Hakim, A. M. & Albert, P. R. Depression, dementia and immune dysregulation. *Brain* **144**, 746–760 (2021).
5. Walker, E. R., McGee, R. E. & Druss, B. G. Mortality in Mental Disorders and Global Disease Burden Implications: A Systematic Review and Meta-analysis. *JAMA Psychiatry* **72**, 334–341 (2015).
6. Nori, V. S. et al. Machine learning models to predict onset of dementia: A label learning approach. *Alzheimers Dement. Transl. Res. Clin. Interv.* **5**, 918–925 (2019).
7. Xiao, Z., Zhou, X., Zhao, Q., Cao, Y. & Ding, D. Significance of plasma p-tau217 in predicting long-term dementia risk in older community residents: Insights from machine learning approaches. *Alzheimers Dement. J. Alzheimers Assoc.* **20**, 7037–7047 (2024).
8. Sugden, K. et al. Association of Pace of Aging Measured by Blood-Based DNA Methylation With Age-Related Cognitive Impairment and Dementia. *Neurology* **99**, (2022).
9. Madsen, C. M., Varbo, A. & Nordestgaard, B. G. Extreme high high-density lipoprotein cholesterol is paradoxically associated with high mortality in men and women: two prospective cohort studies. *Eur. Heart J.* **38**, 2478–2486 (2017).
10. Trimarco, V. et al. High HDL (High-Density Lipoprotein) Cholesterol Increases Cardiovascular Risk in Hypertensive Patients. *Hypertension* **79**, 2355–2363 (2022).
11. Johannesen, C. D. L., Langsted, A., Mortensen, M. B. & Nordestgaard, B. G. Association between low density lipoprotein and all cause and cause specific mortality in Denmark: prospective cohort study. *BMJ Br. Med. J. Online* **371**, m4266 (2020).
12. Liu, J. et al. Effects of blood triglycerides on cardiovascular and all-cause mortality: a systematic review and meta-analysis of 61 prospective studies. *Lipids Health Dis* **12**, 159 (2013).
13. Srikanthan, P., Seeman, T. E. & Karlamangla, A. S. Waist-Hip-Ratio as a Predictor of All-Cause Mortality in High-Functioning Older Adults. *Ann. Epidemiol.* **19**, 724–731 (2009).
14. Jayanama, K. et al. Relationship of body mass index with frailty and all-cause mortality among middle-aged and older adults. *BMC Med* **20**, 404 (2022).
15. Kesavadev, J. et al. Blood glucose levels should be considered as a new vital sign indicative of prognosis during hospitalization. *Diabetes Metab. Syndr. Clin. Res. Rev.* **15**, 221–227 (2021).
16. Fini, M. A., Elias, A., Johnson, R. J. & Wright, R. M. Contribution of uric acid to cancer risk, recurrence, and mortality. *Clin. Transl. Med.* **1**, 16 (2012).
17. Oterdoom, L. H. et al. Urinary creatinine excretion, an indirect measure of muscle mass, is an independent predictor of cardiovascular disease and mortality in the general population. *Atherosclerosis* **207**, 534–540 (2009).
18. Yusuf, S. et al. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *The Lancet* **395**, 795–808 (2020).
19. Cuijpers, P. et al. Comprehensive Meta-Analysis of Excess Mortality in Depression in the General Community Versus Patients With Specific Illnesses. *Am. J. Psychiatry* **171**, 453–462 (2014).
20. Meng, R. et al. Association of Depression With All-Cause and Cardiovascular Disease Mortality Among Adults in China. *JAMA Netw. Open* **3**, e1921043 (2020).
21. Miloyan, B. & Fried, E. A reassessment of the relationship between depression and all-cause mortality in 3,604,005 participants from 293 studies. *World Psychiatry* **16**, 219–220 (2017).
22. Rondón Bernard, J. E. Depression: A Review of its Definition. *MOJ Addict. Med. Ther.* **5**, (2018).
23. Triolo, F. et al. Late-life depression and multimorbidity trajectories: the role of symptom complexity and severity. *Age Ageing* **52**, afac315 (2023).
24. Zhang, Z., Jackson, S. L., Gillespie, C., Merritt, R. & Yang, Q. Depressive Symptoms and Mortality Among US Adults. *JAMA Netw. Open* **6**, e2337011 (2023).
25. Connors, M. H. et al. Cognition and mortality in older people: the Sydney Memory and Ageing Study. *Age Ageing* **44**, 1049–1054 (2015).
26. Hayat, S. A. et al. Understanding the relationship between cognition and death: a within cohort examination of cognitive measures and mortality. *Eur. J. Epidemiol.* **33**, 1049–1062 (2018).
27. Shipley, B., Der, G., Taylor, M. D. & Deary, I. J. Cognition and all-cause mortality across the entire adult age range: Health and lifestyle survey. *Psychosom. Med. J. Biobehav. Med.* **68**, 17–24 (2006).
28. Baune, B. T., Rothermundt, M., Ladwig, K. H., Meisinger, C. & Berger, K. Systemic inflammation (Interleukin 6) predicts all-cause mortality in men: results from a 9-year follow-up of the MEMO Study. *AGE* **33**, 209–217 (2011).
29. Moreno et al. Serum IL8 is not associated with cardiovascular events but with all-cause mortality. *BMC Cardiovasc. Disord.* **19**, 34 (2019).
30. Proctor, M. J. et al. Systemic Inflammation Predicts All-Cause Mortality: A Glasgow Inflammation Outcome Study. *PLOS ONE* **10**, e0116206 (2015).
31. Belloy, M. E., Napolioni, V. & Greicius, M. D. A Quarter Century of APOE and Alzheimer's Disease: Progress to Date and the Path Forward. *Neuron* **101**, 820–838 (2019).
32. Sachdev, P. S. et al. The Sydney Memory and Ageing Study (MAS): methodology and baseline medical and neuropsychiatric characteristics of an elderly epidemiological non-demented cohort of Australians aged 70–90 years. *Int. Psychogeriatr.* **22**, 1248–1264 (2010).
33. D'Agostino, R. B. et al. General Cardiovascular Risk Profile for Use in Primary Care. *Circulation* **117**, 743–753 (2008).
34. Liu, X., Zhang, J., Liu, L., Tang, X. & Gao, P. A population-based recalibration method for updating survival neural networks models for cardiovascular risk prediction in United Kingdom and China. *J. Clin. Epidemiol.* **185**, 111895 (2025).
35. Trollor, J. N. et al. The association between systemic inflammation and cognitive performance in the elderly: the Sydney Memory and Ageing Study. *AGE* **34**, 1295–1308 (2012).
36. Wechsler, D. Wechsler Memory Scale-Revised. *Psychol. Corp.* https://cir.nii.ac.jp/crid/1573105974508393472 (1987).
37. Rey, A. Rey Auditory Verbal Learning Test. https://doi.org/10.1037/t27193-000 (2016).
38. Benton Sivan, A. Benton Visual Retention Test - Fifth Edition. https://doi.org/10.1037/t14985-000 (1991)
39. Arbuthnott, K. & Frank, J. Trail Making Test, Part B as a Measure of Executive Control: Validation Using a Set-Switching Paradigm. *J. Clin. Exp. Neuropsychol.* https://www.tandfonline.com/doi/abs/10.1076/1380-3395%28200008%2922%3A4%3B1-0%3BFT518 (2000).
40. Ross, T. P. et al. The reliability and validity of qualitative scores for the Controlled Oral Word Association Test. *Arch. Clin. Neuropsychol.* **22**, 475–488 (2007).
41. Kaplan, E., Goodglass, H. & Weintraub, S. *Boston Naming Test*. (Lea & Febiger, Philadelphia, PA, 1983).

42. Goodglass, H. & Kaplan, E. *The Assessment of Aphasia and Related Disorders*. (Lea & Febiger, Philadelphia, PA).
43. Reitan, R. M. Validity of the Trail Making Test as an Indicator of Organic Brain Damage. *Percept. Mot. Skills* https://doi.org/10.2466/pms.1958.8.3.271 (1958).
44. Wechsler, D. WAIS-R: Manual: Wechsler adult intelligence scale-revised. (1981).
45. Yesavage, J. A. & Sheikh, J. I. 9/Geriatric Depression Scale (GDS). *Clin. Gerontol*. https://doi.org/10.1300/J018v05n01_09 (1986).
46. Goldberg, D., Bridges, K., Duncan-Jones, P. & Grayson, D. Detecting anxiety and depression in general medical settings. *BMJ* **297**, 897–899 (1988).
47. Lipnicki, D. M. et al. Risk Factors for Late-Life Cognitive Decline and Variation with Age and Sex in the Sydney Memory and Ageing Study. *PLoS ONE* **8**, e65841 (2013).
48. Song, F. et al. Plasma Apolipoprotein Levels Are Associated with Cognitive Status and Decline in a Community Cohort of Older Individuals. *PLOS ONE* **7**, e34078 (2012).
49. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing (2025).

Acknowledgements

We thank the participants and their informants for their time and generosity in contributing to this research. We also acknowledge the MAS research team: https://cheba.unsw.edu.au/research-projects/sydney-memory-and-ageing-study. The Sydney Memory and Ageing Study has been funded by the National Health and Medical Research Council (NHMRC) Program Grants [ID350833, ID568969, and APP1093083].

Author contributions

R.A.C. lead and designed the study, analysed data, and wrote the manuscript. M.C. supervised the study and edited the manuscript. C.C. supervised the study and edited the manuscript. K.N. managed data, collaborated with the study design and edited the manuscript. K.A.M. supervised data collection and edited the manuscript. H.B. sourced funding, collected data and edited the manuscript. N.A.K. supervised data collection and edited the manuscript. P.S.S. supervised data collection, the study design and edited the manuscript. O.N.M. supervised the study and data analyses and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44400-026-00071-1.

Correspondence and requests for materials should be addressed to Rebecca A. Chalmers.