Citation: *Korni A, Zandi E. Pairwise External Validation of Plasma Biomarker–Based Machine Learning Models for Amyloid PET Prediction: Implications for Calibration and Clinical Utility. J Exp Neurol. 2026;7(1):70–89.*

Supplementary Table and Figures

Supplementary Table S1. Means and variances for primary biomarkers

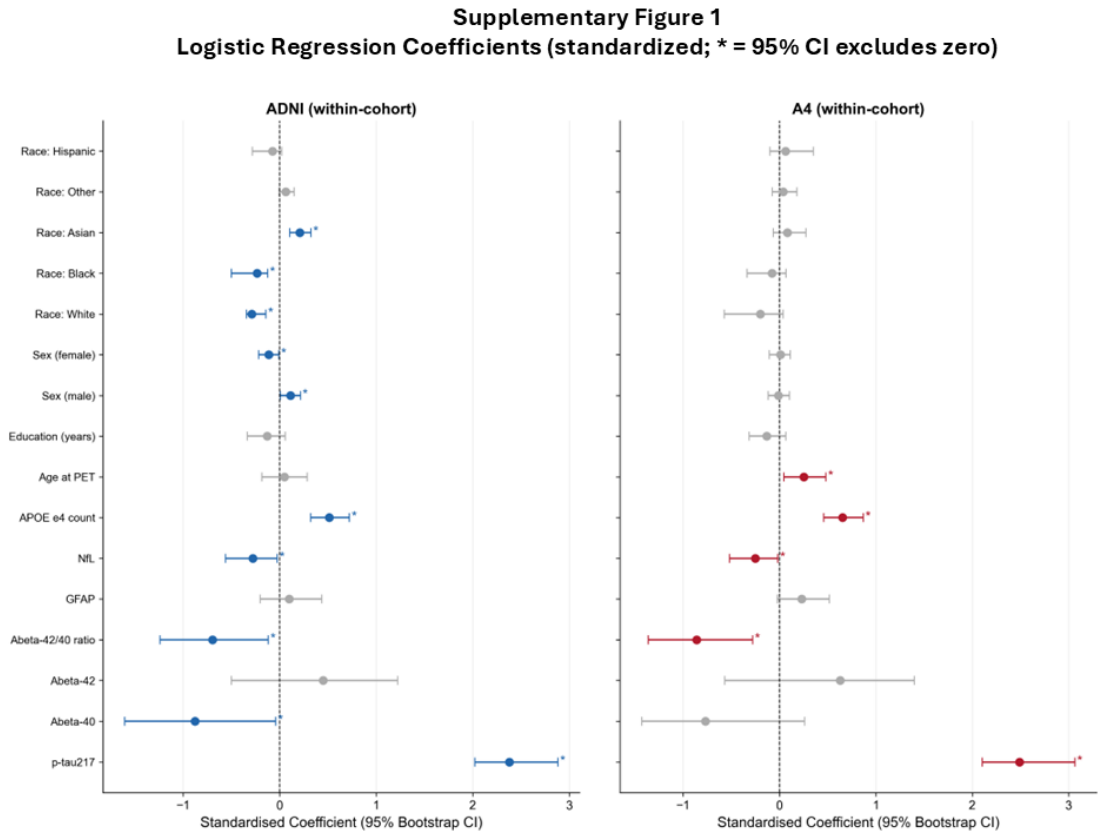
Table: Biomarker Distribution by Cohort (ADNI vs A4)

Values shown as mean ± SD for non-missing observations.

Biomarker	ADNI (mean ± SD)	ADNI N	A4 (mean ± SD)	A4 N
Raw biomarker values				
p-tau217 (AlzPath)	0.382 ± 0.240	885	0.193 ± 0.104	841
Aβ42/40 ratio	0.058 ± 0.011	885	0.093 ± 0.018	828
Aβ40	4.613 ± 0.309	885	5.274 ± 0.343	829
Aβ42	1.924 ± 0.280	885	2.944 ± 0.465	840
p-tau217 / Aβ42 (AlzPath)	0.089 ± 0.092	885	0.016 ± 0.030	840
GFAP	5.035 ± 0.528	885	4.614 ± 0.463	839
NfL	3.103 ± 0.457	885	1.409 ± 0.290	839
Within-cohort z-scored (log scale)				
p-tau217 (z-score)	0.036 ± 1.024	885	0.056 ± 1.028	841
Aβ42/40 ratio (z-score)	-0.016 ± 1.020	885	-0.022 ± 1.000	828
Aβ40 (z-score)	0.011 ± 1.016	885	0.005 ± 0.994	829
Aβ42 (z-score)	0.000 ± 1.027	885	-0.008 ± 1.000	840
p-tau217 / Aβ42 (z-score)	0.034 ± 1.045	885	0.022 ± 1.041	840
GFAP (z-score)	0.043 ± 0.999	885	0.021 ± 1.008	839
NfL (z-score)	0.044 ± 1.004	885	0.022 ± 0.996	839

Citation: *Korni A, Zandi E. Pairwise External Validation of Plasma Biomarker–Based Machine Learning Models for Amyloid PET Prediction: Implications for Calibration and Clinical Utility. J Exp Neurol. 2026;7(1):70–89.*

Supplementary Figure S1



Supplementary Figure S1. Standardized logistic regression coefficients for within-cohort amyloid PET classification models.

Forest plots display standardized regression coefficients (with 95% bootstrap confidence intervals) from logistic regression models predicting amyloid PET positivity within ADNI (left panel) and A4 (right panel). All predictors were z-scored within cohort prior to model fitting. Points represent standardized log-odds coefficients; horizontal lines indicate 95% bootstrap confidence intervals. The vertical dashed line indicates a coefficient of zero. Asterisks (*) denote coefficients for which the 95% confidence interval does not cross zero.

Predictors include demographic variables (age, sex, education, race), genetic risk (APOE ε4 allele count), and plasma biomarkers (p-tau217, Aβ40, Aβ42, Aβ42/40 ratio, GFAP, NfL). Separate within-cohort models were estimated for ADNI and A4.

Standardized coefficients represent:

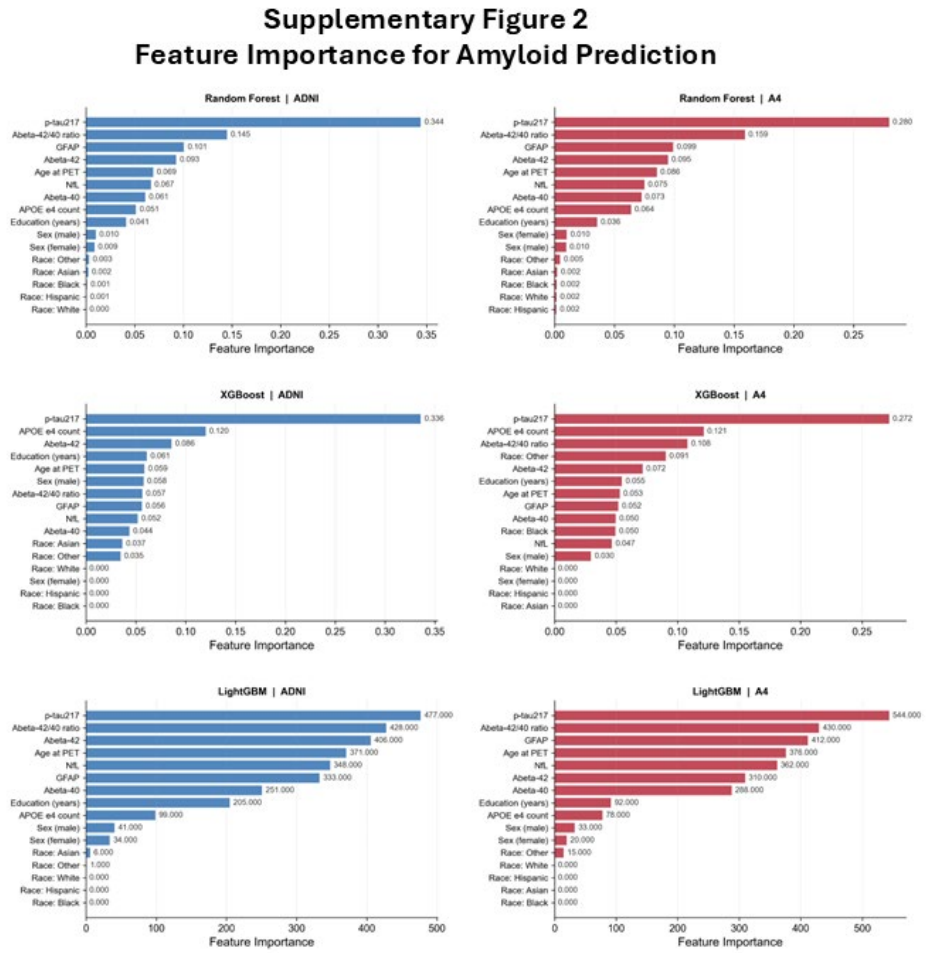
$$\beta_{std} = \beta \times \frac{\sigma_x}{\sigma_y}$$

Because predictors were z-scored, coefficient magnitude directly reflects relative log-odds contribution per standard deviation change in predictor.

Bootstrap confidence intervals were used rather than asymptotic Wald intervals to provide more robust uncertainty estimates.

Citation: *Korni A, Zandi E. Pairwise External Validation of Plasma Biomarker–Based Machine Learning Models for Amyloid PET Prediction: Implications for Calibration and Clinical Utility. J Exp Neurol. 2026;7(1):70–89.*

Supplementary Figure S2



Supplementary Figure S2. Feature importance for amyloid PET classification across machine learning algorithms and cohorts.

Bar plots display feature importance rankings for amyloid PET classification models trained within ADNI (left panels, blue) and A4 (right panels, red). Feature importance is shown for three algorithms: Random Forest (top row), XGBoost (middle row), and LightGBM (bottom row).

For tree-based models (Random Forest and XGBoost), feature importance values represent normalized mean decrease in impurity. For LightGBM, importance values reflect total gain contribution across trees.

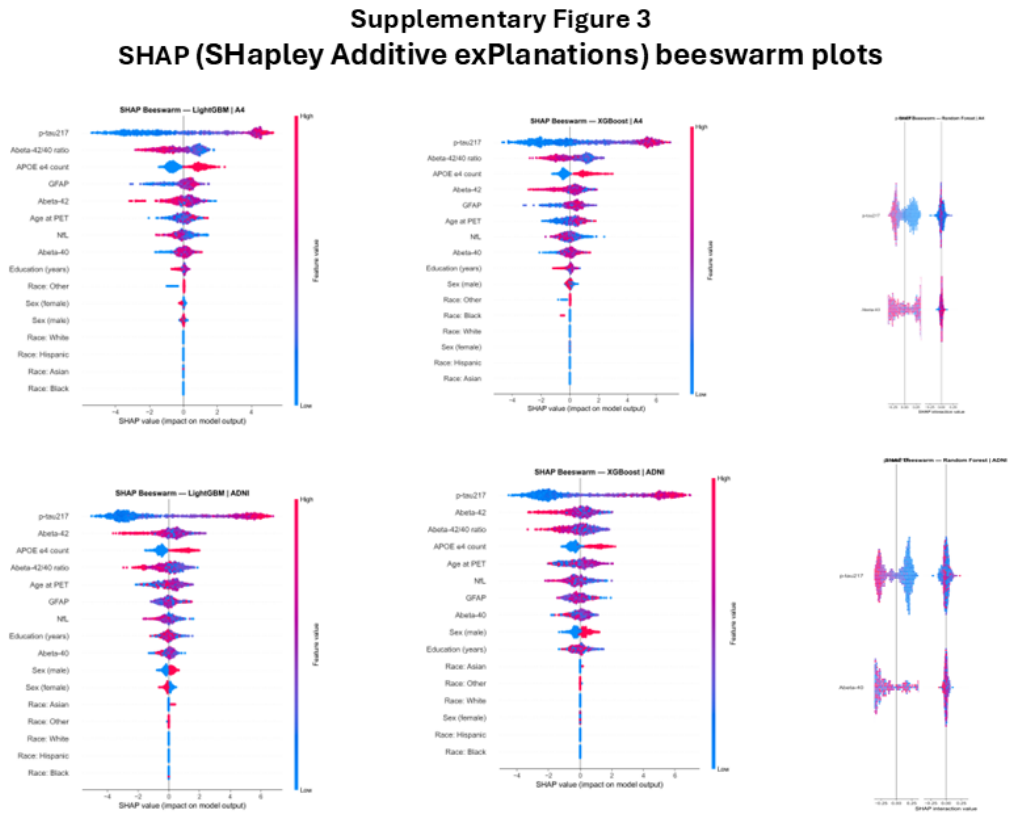
Predictors include plasma biomarkers (p-tau217, Aβ42, Aβ40, Aβ42/40 ratio, GFAP, NfL), demographic variables (age at PET, sex, education, race), and genetic risk (APOE ε4 allele count). Bars are ordered by decreasing importance within each model.

Tree-based feature importance measures: Random Forest: Mean decrease in impurity (Gini importance), XGBoost/LightGBM: Gain contribution across boosting iterations

Because importance is algorithm-specific, absolute values are not directly comparable across models; ranking consistency is the key interpretive metric.

Citation: *Korni A, Zandi E. Pairwise External Validation of Plasma Biomarker–Based Machine Learning Models for Amyloid PET Prediction: Implications for Calibration and Clinical Utility. J Exp Neurol. 2026;7(1):70–89.*

Supplementary Figure S3



Supplementary Figure S3. SHAP beeswarm plots illustrating feature contributions to amyloid PET classification across cohorts and algorithms.

SHAP (SHapley Additive exPlanations) beeswarm plots display feature-level contributions to model predictions for amyloid PET status in A4 (top panels) and ADNI (bottom panels). Models include LightGBM, Random Forest, and XGBoost.

Each point represents one participant. The x-axis indicates the SHAP value (impact on model output), with positive values increasing predicted probability of amyloid positivity and negative values decreasing it. Features are ordered by mean absolute SHAP value. Point color reflects feature value (blue = low, red = high).

Vertical gray lines denote zero contribution. Wider horizontal dispersion reflects greater impact on model prediction.

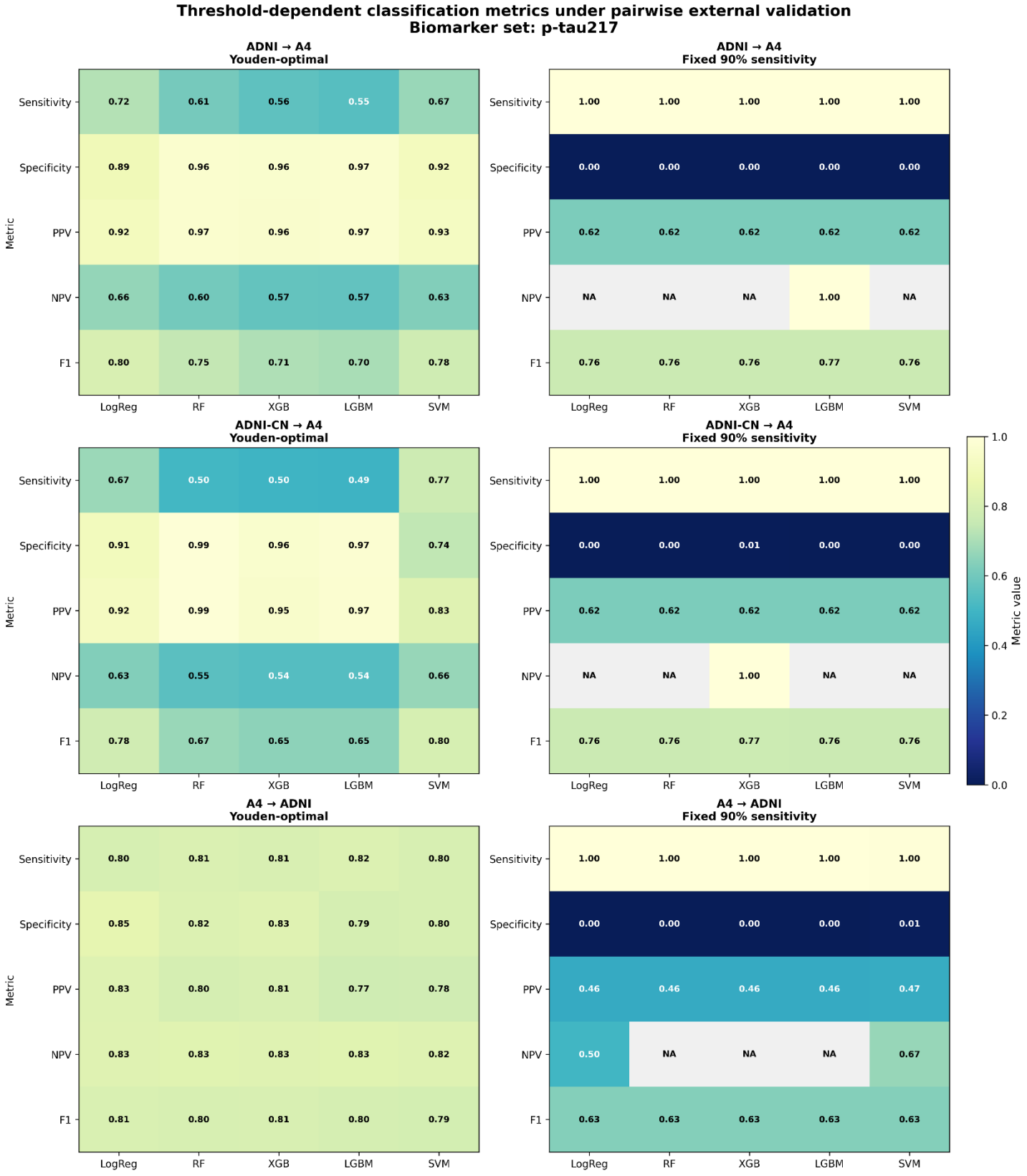
SHAP values are based on Shapley values from cooperative game theory and provide:Local explanation (per individual prediction), additive feature attribution, model-agnostic interpretation

For tree-based models, TreeSHAP provides exact SHAP values efficiently. SHAP magnitude reflects contribution to log-odds prediction.

Citation: *Korni A, Zandi E. Pairwise External Validation of Plasma Biomarker–Based Machine Learning Models for Amyloid PET Prediction: Implications for Calibration and Clinical Utility. J Exp Neurol. 2026;7(1):70–89.*

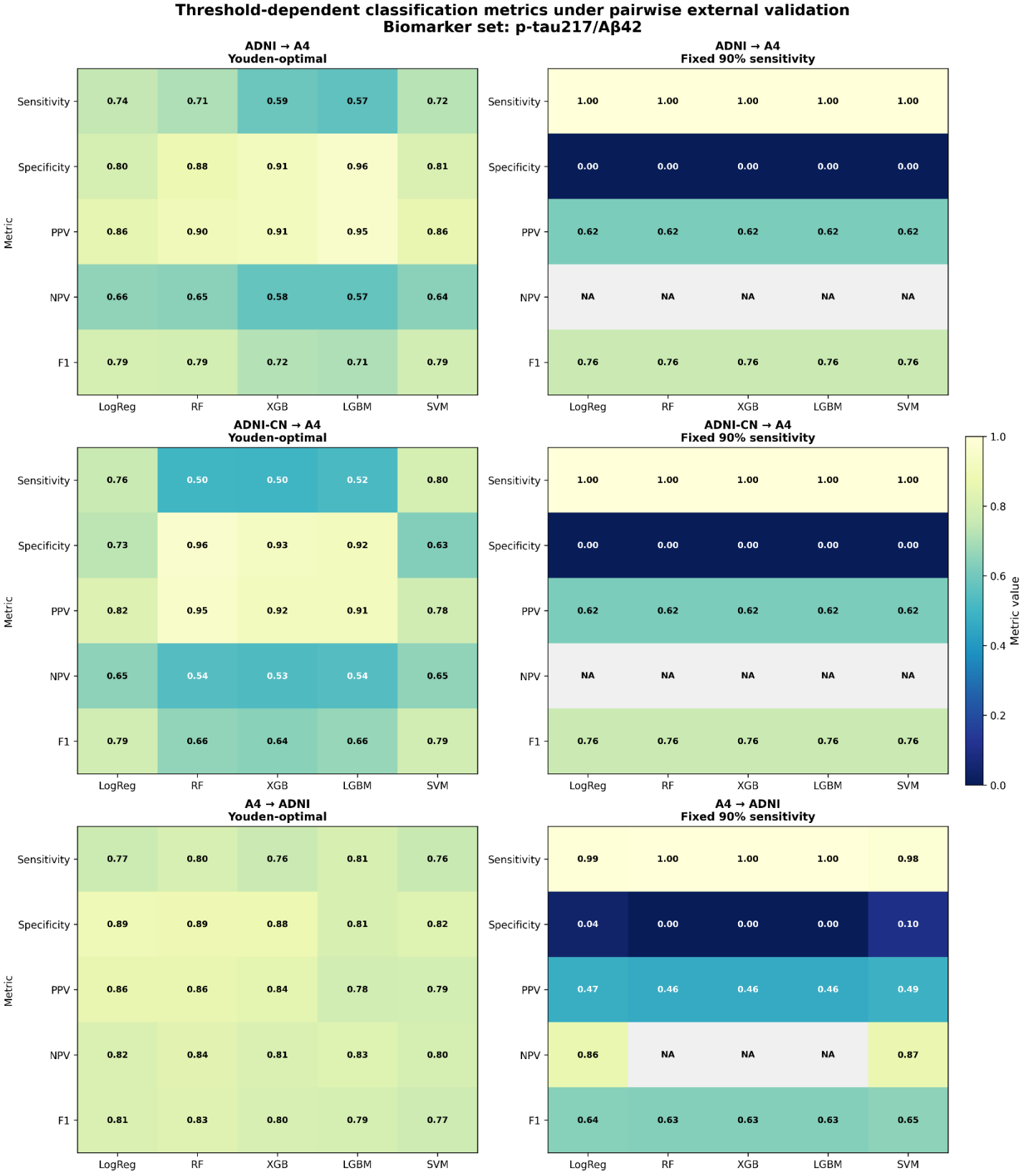
Supplementary Figure S4

(I)



Citation: *Korni A, Zandi E. Pairwise External Validation of Plasma Biomarker–Based Machine Learning Models for Amyloid PET Prediction: Implications for Calibration and Clinical Utility. J Exp Neurol. 2026;7(1):70–89.*

(II)

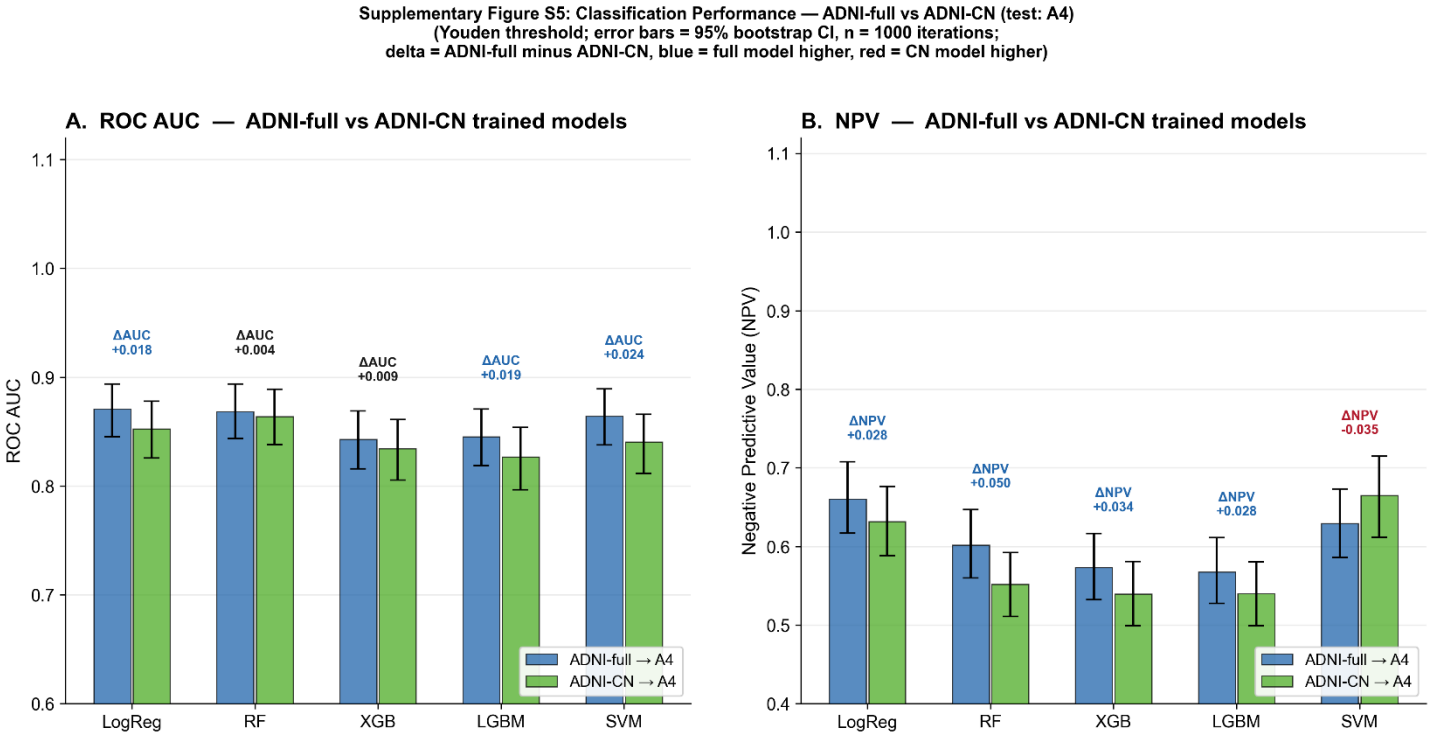


Supplementary Figure S4. Threshold-dependent classification metrics under pairwise external validation across biomarker configurations.

Heatmap representation of threshold-dependent classification performance metrics across pairwise external validation settings, machine learning models, and thresholding strategies. Rows correspond to transfer directions (ADNI→A4, ADNI-CN→A4, and A4→ADNI), and columns correspond to machine learning models. Panels are shown separately for Youden-optimal thresholds and fixed 90% sensitivity thresholds for both plasma p-tau217-only (I) and combined p-tau217/Aβ42 (II) biomarker configurations. Heatmap values represent sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score. Under fixed high-sensitivity thresholding, specificity frequently collapsed toward zero across several transfer directions and models, resulting in unstable predictive values despite relatively preserved discrimination performance.

Citation: *Korni A, Zandi E. Pairwise External Validation of Plasma Biomarker–Based Machine Learning Models for Amyloid PET Prediction: Implications for Calibration and Clinical Utility. J Exp Neurol. 2026;7(1):70–89.*

Supplementary Figure S5



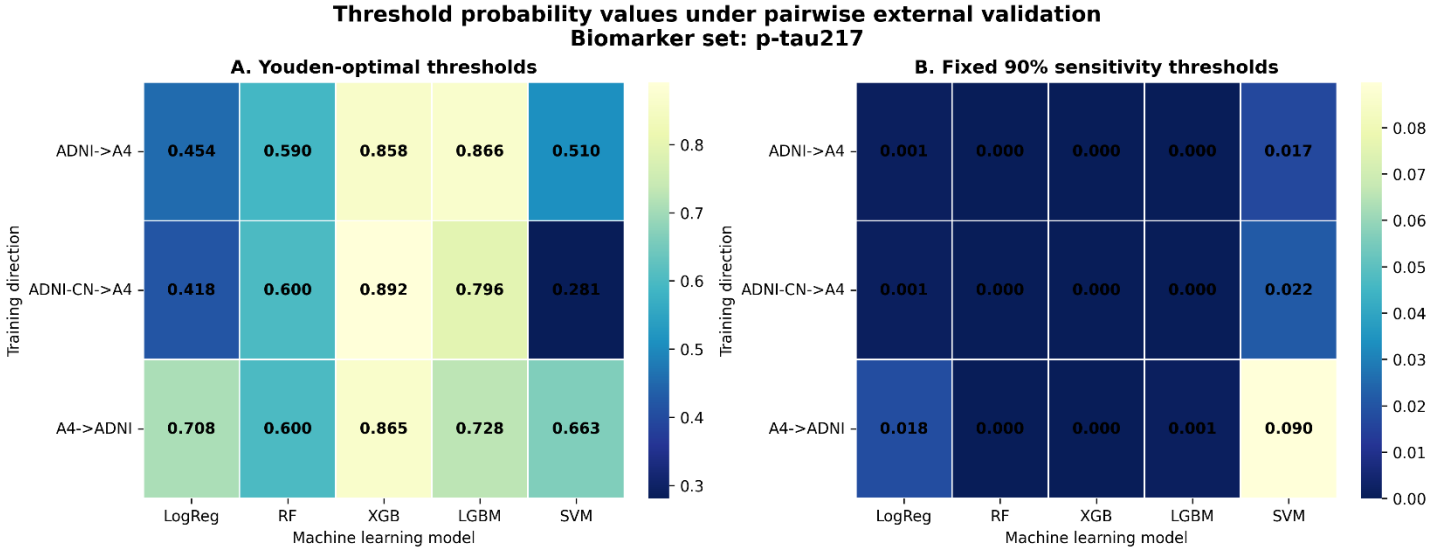
Supplementary Figure S5. Effect of restricting ADNI to cognitively normal participants on pairwise external validation performance.

Comparison of classification performance between models trained on the full ADNI cohort and models trained using cognitively normal ADNI participants only (ADNI-CN), both evaluated in A4 using Youden-optimal thresholds. Panel A shows ROC AUC with 95% bootstrap confidence intervals. Panel B shows negative predictive value (NPV) with 95% bootstrap confidence intervals estimated from 1,000 bootstrap iterations. Delta values represent the difference between ADNI-full and ADNI-CN performance for each model. Blue annotations indicate higher performance for the full ADNI-trained models, whereas red annotations indicate higher performance for the ADNI-CN-trained models. Restricting ADNI to cognitively normal participants produced only modest changes in discrimination and NPV, suggesting that clinical heterogeneity contributes to, but does not fully account for, cross-cohort attenuation.

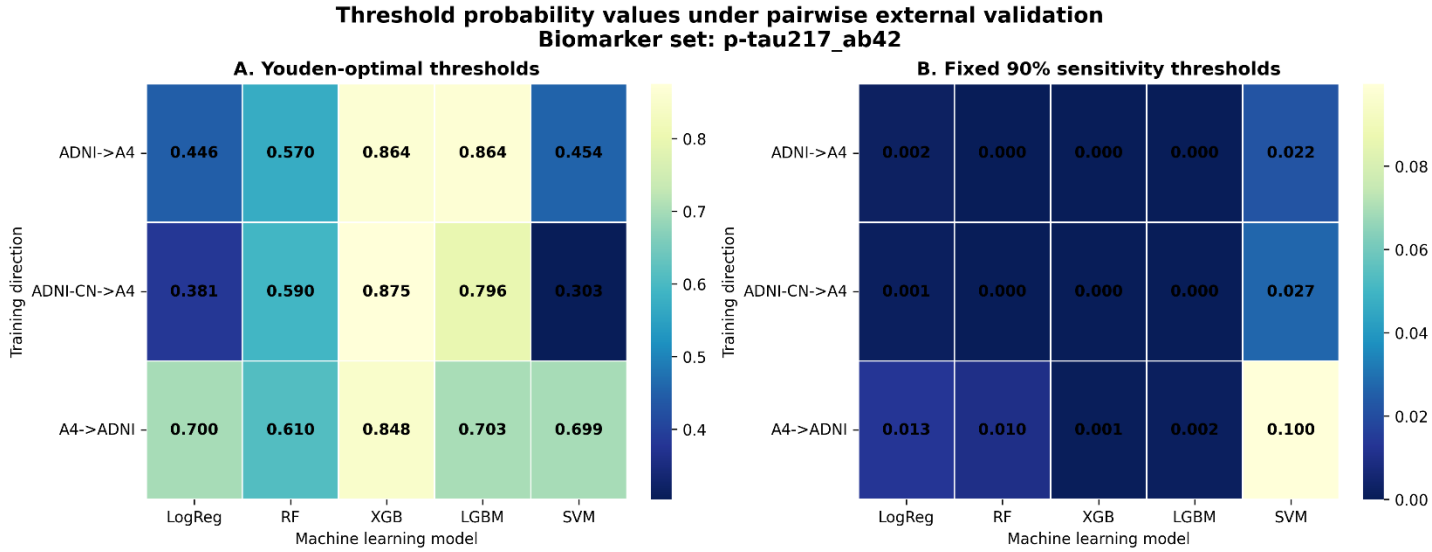
Citation: *Korni A, Zandi E. Pairwise External Validation of Plasma Biomarker–Based Machine Learning Models for Amyloid PET Prediction: Implications for Calibration and Clinical Utility. J Exp Neurol. 2026;7(1):70–89.*

Supplementary Figure S6

(I)



(II)



Supplementary Figure S6. Threshold probability values under pairwise external validation across biomarker configurations.

Heatmap representation of probability thresholds derived under pairwise external validation across transfer directions, machine learning models, and biomarker configurations. Separate heatmaps are shown for plasma p-tau217-only (I) and combined p-tau217+Aβ42 (II) biomarker configurations. Panel A shows Youden-optimal thresholds, whereas Panel B shows thresholds selected to achieve approximately 90% sensitivity in the training cohort. Rows correspond to transfer directions and columns correspond to machine learning models. Under fixed sensitivity thresholding, threshold values frequently collapsed toward zero across several transfer directions and models, reflecting substantial reduction in specificity under high-sensitivity operating conditions.