

Predicting Blood-Based Biomarker Positivity for Pre-Clinical Alzheimer’s Trial Recruitment Using RWD and Predictive Modeling

Sean Black, MS (SiteRx, Machine Learning Engineer); Jackson Dooley (SiteRx, Senior Analyst, Strategy & Data Services); Conner Phalen (SiteRx, Senior Analyst, Strategy & Data Services); Alex Li (SiteRx, Director, Strategy & Data Services); Silvia Metelli, PhD (F. Hoffmann–La Roche AG, Basel, Switzerland)

An interpretable model built on routinely captured EHR data triages patients for blood-based biomarker (BBM) testing (AUC-ROC 0.73) to reduce the high screen-failure rates and cost of pre-clinical Alzheimer’s (PCAD) trial recruitment

INTRODUCTIONMETHODSRESULTSCONCLUSIONSREFERENCES

Background. PCAD research is plagued by high screen-failure rates (88–98%^{1,2}) and in the recent TRAILBLAZER-ALZ 3 trial, 81.2%² of screen failures were due to Plasma P-tau217 assay testing. Blood-based biomarker testing (BBM) is becoming integrated into routine clinical practice and is increasingly being used in upcoming PCAD research.

Objective. To improve patient identification, we developed a data-driven pre-screening model designed to prioritize patients for blood-based biomarker (BBM) testing, reducing patient burden and operational costs for future PCAD clinical research.

Table 1. Baseline participant characteristics by Alzheimer's disease biomarker status

Variable	Overall N = 1,283	Biomarker + n = 653	Biomarker – n = 630
Demographics			
Age, years, mean (SD)	70.4 (5.9)	71.4 (5.6)	69.4 (6.1)
Age category, years, n (%)			
<70	536 (41.8)	224 (34.3)	312 (49.5)
70–74	378 (29.5)	210 (32.2)	168 (26.7)
≥75	369 (28.8)	219 (33.5)	150 (23.8)
Sex, n (%)			
Female	722 (56.3)	366 (56.0)	356 (56.5)
Male	561 (43.7)	287 (44.0)	274 (43.5)
Race/ethnicity, n (%)			
White / Caucasian	410 (32.0)	164 (25.1)	246 (39.0)
Black / African American	44 (3.4)	13 (2.0)	31 (4.9)
Asian	11 (0.9)	7 (1.1)	4 (0.6)
American Indian / Alaska Native	2 (0.2)	1 (0.2)	1 (0.2)
Hispanic or Latinx	47 (3.7)	15 (2.3)	32 (5.1)
Other	27 (2.1)	12 (1.8)	15 (2.4)
Unknown / not recorded	742 (57.8)	441 (67.5)	301 (47.8)
Genetic characteristics			
APOE ε4 genotype, n (%)			
≥1 ε4 allele (carrier)	225 (17.5)	185 (28.3)	40 (6.3)
No ε4 allele	113 (8.8)	59 (9.0)	54 (8.6)
Unknown	945 (73.7)	409 (62.6)	536 (85.1)
Clinical and behavioral characteristics, n (%)			
Subjective cognitive impairment	973 (75.8)	528 (80.9)	445 (70.6)
Hypertension	686 (53.5)	337 (51.6)	349 (55.4)
Hyperlipidemia	671 (52.3)	310 (47.5)	361 (57.3)
Diabetes, type 1 or 2	211 (16.4)	89 (13.6)	122 (19.4)
Pre-diabetes	71 (5.5)	25 (3.8)	46 (7.3)
Sleep disorder (insomnia/hypersomnia)	161 (12.5)	74 (11.3)	87 (13.8)
Recent onset of mood change	78 (6.1)	56 (8.6)	22 (3.5)
Sensory loss	76 (5.9)	44 (6.7)	32 (5.1)
Ever smoker	305 (23.8)	153 (23.4)	152 (24.1)
Current smoker	53 (4.1)	26 (4.0)	27 (4.3)
Family history, n (%)			
Any first-degree relative	531 (41.4)	275 (42.1)	256 (40.6)
Maternal first-degree	365 (28.4)	187 (28.6)	178 (28.3)
Non-maternal first-degree	166 (12.9)	88 (13.5)	78 (12.4)

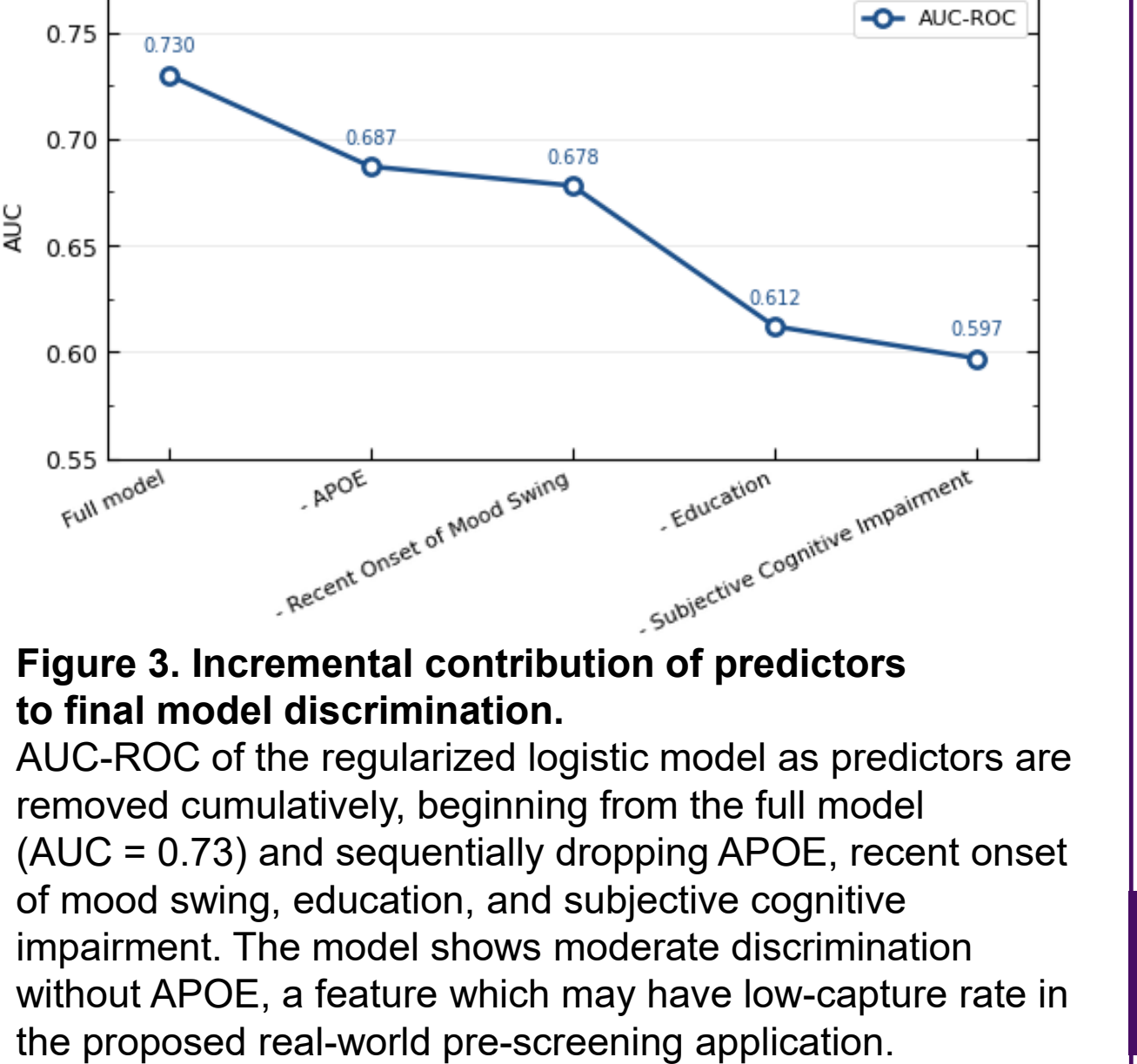
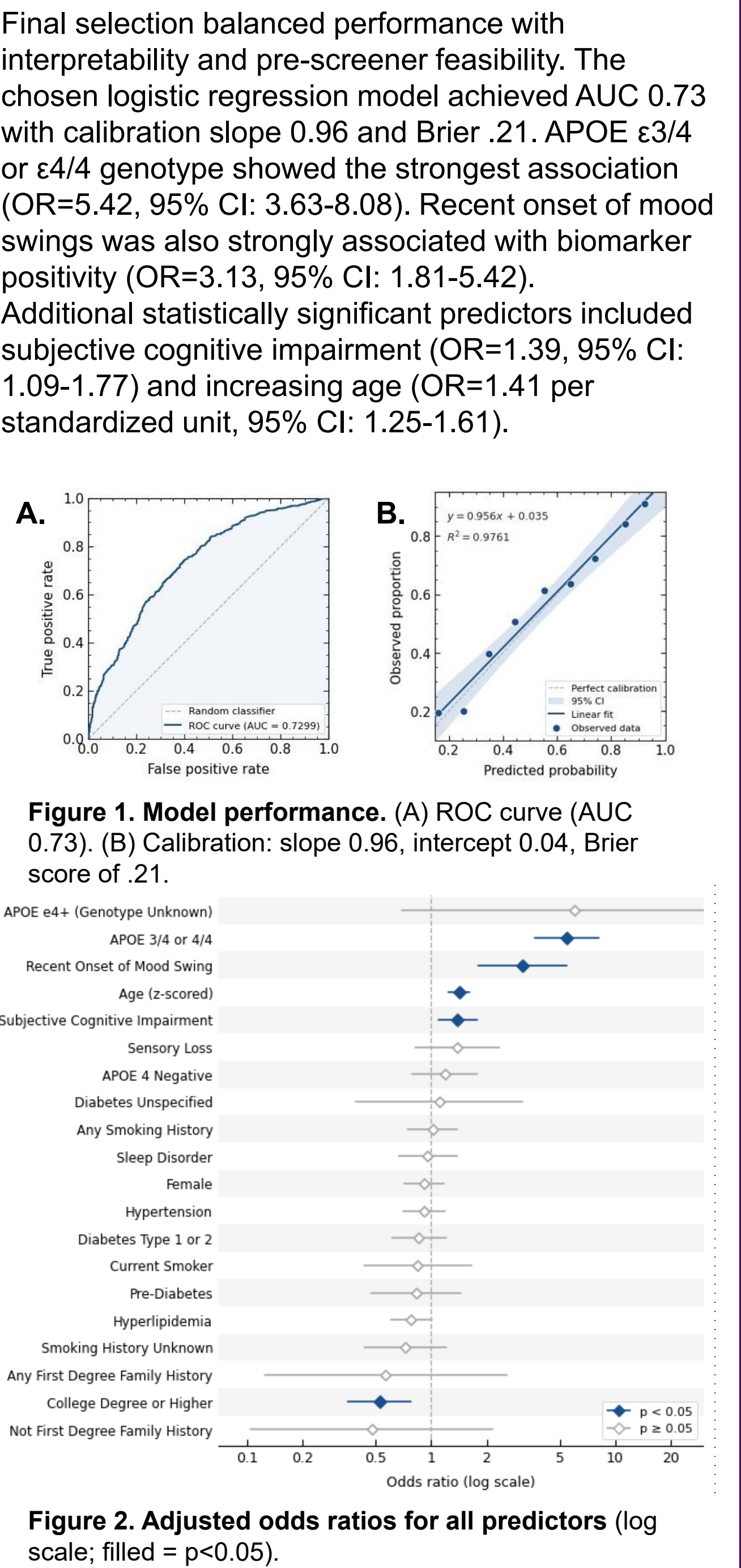
Sample. Using SiteRx’s de-identified real-world EHR data, a dataset of 1,283 patients with confirmed biomarker outcomes aligned to common PCAD trial protocol was assembled. Of the 1,283 patients, 653 (50.9%) were biomarker-positive and 630 (49.1%) were biomarker-negative, with tests conducted during PCAD research screening. The sample was 56.3% female (n = 722), with a mean age of 70.4 years (SD 5.9; range 55-80). Baseline characteristics: Table 1.

Clinical record review and temporal anchoring. Features were recorded only up to each patient’s biomarker-testing date (mean look-back 8.88 months) to prevent leakage. Two members of SiteRx clinical excellence team validated abstracted features via structured chart review (definitions: S3).

Alignment with PCAD trial eligibility criteria. Sample and predictors were aligned to typical PCAD inclusion/exclusion criteria: age 55–80; cognitive status (CDR=0 or equivalent); exclusionary medical conditions (e.g., other CNS, and uncontrolled cardiovascular, renal, and hepatic disorders); and concomitant AD medications.

Feature identification. Candidates were drawn from (i) a targeted literature review of AD risk/protective factors (references 3-30), (ii) Clinical excellence assessment of EHR in accordance with operational definitions, and (iii) review with Roche medical experts for clinical plausibility.

Feature engineering & modeling. APOE was collapsed to ε4-carrier status; education to "high school or less" vs "college degree or higher"; insomnia/hypersomnia and OSA into a sleep-disturbance composite; and diabetes with pre-diabetes into a metabolic-risk construct. Age was z-scored. Missingness was encoded as an explicit "unknown" category; race/ethnicity and smoking pack-years (insufficient or non-informative capture) were excluded; capture for retained features: see S1. Regularized logistic regression, random forest, and decision tree were trained with stratified 5-fold cross-validation and evaluated on discrimination (AUC-ROC) and calibration.



The proposed model provides a practical, interpretable, and tunable framework for improving PCAD trial pre-screening. It uses routinely captured EHR signals from both chronic and recently developed conditions to discriminate biomarker positivity to triage individuals for BBM testing.

Because it is trained on real-world data with incomplete capture, the model handles sub-100% data availability natively, and its calibrated probabilities let sites tune the threshold toward recall or precision based on screening objectives.

Notably, our model did not show comorbidities commonly associated with an increased risk of Alzheimer’s Disease (diabetes, hypertension, sleep disorders, etc.) as statistically significant features. We attribute this to our sample being a research interested patient population who exhibited similar rates across major comorbidities.

Future work will generate larger datasets to study the divergence between research patients and the at-risk community patient population as well as pursue external and prospective validation; an early look in a small ex-US sample is encouraging but requires further confirmation.

1. USC Schaeffer Center. (n.d.). *Key barriers for clinical trials for Alzheimer's disease*. <https://schaeffer.usc.edu/research/key-barriers-for-clinical-trials-for-alzheimers-disease/>

2. Yaari, R., Holdridge, K. C., Williamson, M., Wessels, A. M., Shcherbinin, S., Kotari, V., Reiman, E. M., Tariot, P. N., Alexander, R., Langbaum, J. B., & Sims, J. R. (2025). Donanemab in preclinical Alzheimer's disease: Screening and baseline data from TRAILBLAZER-ALZ 3. *Alzheimer's & Dementia*, 21(9), Article e70662. <https://doi.org/10.1002/alz.70662>

Full reference list (3–30) and supplementary tables S1–S3 are available via the QR code.

CONTACT



Alex Li
ali@siterx.com

Jack Dooley
jdooley@siterx.com

Conner Phalen
cphaeln@siterx.com

SiteRx, New York, NY, USA

Control # 2026-DTA-13448-AAIC