Proceedings of ICMMGH 2026 Symposium: Biomedical Imaging and AI Applications in Neurorehabilitation
DOI: 10.54254/2753-8818/2026.AU30169

Blood and Digital Biomarkers for Alzheimer's Disease in 2025: A Two-Tier ML Triage Workflow from Screening to Confirmation

Keyi Xiong

Department of Mathematics, South China University of Technology, Guangzhou, China
emilieXKY@outlook.com

Abstract. Imaging-based confirmation such as PET/CSF has long anchored Alzheimer's disease (AD) diagnosis, yet cost, invasiveness, and limited access constrain scale. In 2025, regulated plasma assays, most notably p-tau217, have reached clinic-ready performance, reframing blood tests as a reliable Level-2 "anchor," while passive digital biomarkers (DBMs; speech, wearables, smartphones) offer low-burden reach at Level-1. We synthesize a two-tier machine-learning triage that links DBM pre-screening to p-tau217-centered confirmation, and we formalize evaluation beyond AUC to a calibration net benefit resource framework with thresholds tied to pre-test probability. Across memory clinics and primary care, p-tau217 approaches PET/CSF (AUC ≈0.94–0.96); dual cutoffs yield clear rule-in/rule-out for about 83–88% with only 12–17% indeterminate and enable 57–85% reductions in CSF/amyloid-PET, while sequential blood then tau-PET cuts trial-screening imaging by approximately 94%. Evidence across hundreds of DBM studies shows promising discrimination but sparse external validation and probability calibration, supporting DBMs primarily as pre-screeners that require domain-robust training and transparent reporting. Finally, we codify fairness-first deployment: subgroup calibration and threshold adjustment for device, language, age ≥80, and kidney dysfunction (eGFR<45), aligned with TRIPOD-AI/STARD-AI/PROBAST-AI/DECIDE-AI. In summary, an anchor-at-scale paradigm, digital-as-screen, blood-as-anchor, offers a practical, regulated, and economically defensible path from research prototypes to real-world AD services.

Keywords: Alzheime, digital, blood-based

1. Introduction

Alzheimer's diagnosis remains anchored to PET/CSF, accurate yet invasive, costly, and capacity-limited. Two complementary advances are now mature: passive digital biomarkers (DBMs) that scale with minimal burden and blood-based biomarkers (BBMs), especially plasma p-tau217 on automated platforms, whose discrimination approaches PET/CSF and can be interpreted with dual cut-offs. However, DBMs still lack broad external validation and probability calibration across devices/languages; BBMs need subgroup-aware interpretation and clearer governance of indeterminate results. Heterogeneous endpoints, sparse prospective deployments, and uneven

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.AU30169

privacy safeguards further limit transportability. This review asks, what evidence supports p-tau217 as a regulated "anchor," and how should dual thresholds bound indeterminacy? Under what training/validation standards can passive DBMs serve scalable, fair prescreening? How can a two-tier ML workflow integrate DBM prescreening with blood-anchored decisions so that analytic performance maps to calibrated, resource-aware, and equitable actions?

We synthesize multi-cohort evidence (2019–2025) spanning memory clinics and primary care, prioritize automated p-tau217 and cross-platform comparisons, and review external validations of passive DBMs. Beyond ROC/PR-AUC, we extract calibration metrics (reliability, Brier/ECE, E/O), decision-curve net benefit, and two-stage budget analyses. Reporting quality and bias are appraised against TRIPOD-AI, STARD-AI, PROBAST-AI, and DECIDE-AI. We operationalize a standardized workflow with explicit operating points (dual cut-offs), gray-zone escalation rules, and subgroup recalibration. We provide deployable guidance for a tiered, continuous service, digital as screen, and blood as anchor, including thresholds tied to pre-test risk and care context, fairness-by-design auditing, privacy-preserving deployment, and budget justifications, to enable adoption in everyday clinics.

2. Advances in blood-based biomarkers for Alzheimer Disease (AD)

2.1. Biological basis and key biomarkers

Within the A/T/N(+I) framework, peripheral readouts reflect central pathology that becomes measurable in plasma. Among candidates, plasma p-tau217 has emerged as the T-class diagnostic anchor with clinic-level performance; GFAP tracks Aβ-linked astroglial reactivity and enriches very-early risk; NfL indexes neuroaxonal injury and is more informative for progression monitoring than disease specificity. These markers are biologically complementary and support staged diagnosis and trial enrichment [1, 2].

2.2. Principal blood biomarkers (2024 - 2025)

Across head-to-head and multi-cohort evaluations, plasma p-tau217 has emerged as the peripheral T-class anchor. On a fully automated platform, p-tau217 detects AD pathology with an AUC of around 0.93 and 0.96; a two-cutoff scheme (<0.22 / >0.34 pg·mL^{-1}) elevates accuracies to 92–94% while explicitly triaging 12–17% to an intermediate zone. Adding the p-tau2Aβ42 ratio does not raise overall accuracy but reduces intermediates to ≤10% and increases NPV in primary care, useful for "blood-first to confirmatory" pathways.

In real-world and prospective clinics beyond high-income settings, p-tau217 alone classifies CSF-defined status with AUC about 0.94, which improves to around 0.98 when combined with Aβ42, providing external validation and supporting site-level robustness [3]. Larger memory-clinic cohorts show AUC is approximately 0.95 and demonstrate that choosing cutoffs at 95% or 97.5% can spare about 57–85% of LP/amyloid-PET with overall accuracy of 95–97%. In cognitively unimpaired populations, p-tau217 ≈ tau-PET for predicting near-term decline; using p-tau217 prescreen to tau-PET confirmation preserves prognostic power while cutting trial imaging loads by around 76–94%. On standardization and regulation, the Lumipulse G p-tau217/Aβ1-42 plasma ratio received FDA 510(k) clearance on May 16, 2025, with positive/negative agreement of 91.7%/97.3% versus PET/CSF and an intended-use population of symptomatic adults ≥50–55 years in specialized care [4]. Independent MS-based panels (e.g., PrecivityAD2, combining %p-tau217 with Aβ42/40)

DOI: 10.54254/2753-8818/2026.AU30169

report AUC is about 0.94 and around 88% amyloid-PET agreement, with 2025 external validations confirming platform-level performance [5].

Finally, the Alzheimer's Association Clinical Practice Guideline (2025) provides performance-based, brand-agnostic recommendations: BBMs may serve as triage tests and, when meeting high sensitivity or specificity thresholds, may substitute for PET or CSF in defined contexts; the guideline endorses two-cutoff interpretation and pretest-probability–aware reporting. According to the materials, 2024–2025 evidence and regulatory milestones position p-tau217-anchored testing and selective use of Aβ ratios as a scalable, auditable entry point for staged diagnosis and trial enrichment, with clear pathways for cost and capacity gains.

Table 1. Key blood-based biomarker studies (2024-2025)

Study	Cohorts	AUC	Cut-points (standardized)	External validation
Palmqvist et al. (2025)	4 memory clinics and 1 primary care; n=1,767	0.93 – 0.96	Single 0.27; Two-cutoff <0.22 / >0.34; Ratio <0.007 / >0.009	Multicenter across care levels; sensitivity analyses by age/eGFR; recommends pre-test probability calibration.
Santos et al. (2025)	Memory-clinic cohort; symptomatic; CSF-referenced	0.94 – 0.98	Uses standardized benchmarks for cross-study comparison	Local external validation in emerging-market setting; ratio improves top-end performance.
Meyer et al. (2024)	Multicenter; symptomatic; n=583	0.94 – 0.94	Algorithmic banding (IFU); mapped to standardized benchmarks for comparison	Independent 2025 reports reproduced high amyloid-PET concordance.
Martínez-Dubarbie et al. (2025)	Real-world memory clinic	0.95 – 0.95	Uses standardized benchmarks for cross-study comparison	Shows 95–97% accuracy at prespecified targets and 57–85% LP/PET sparing in clinic workflow.
Feizpour et al. (2024)	Multicenter; mixed stages	0.92 – 0.95	Uses standardized benchmarks for cross-study comparison	Cross-lab/platform transferability; complements automated IVD ecosystem.

2.3. Technical reliability and implementation considerations

Two mature assay routes support deployment are: mass spectrometry and fully automated immunoassays. In memory-clinic settings they show comparable discrimination for AD pathology; in primary care, %p-tau217 is less age-sensitive under a single cut-off, while a two-cutoff scheme, <0.22 / >0.34 pg·mL^{-1} with a conventional single cut-off around 0.27 pg·mL^{-1}, harmonizes platform performance by routing a manageable gray zone to confirmatory PET/CSF. In sequential blood-first to confirmatory workflows, this triage delivers material efficiency gains without loss of prognostic yield, enabling scalable prescreening and routine clinic use. Operationally, p-tau217 is robust to common pre-analytical variation, whereas Aβ42/40 is handling-sensitive, so ratio-based readouts require tighter SOPs. Subgroup-aware calibration is recommended: older age (≥ 80 years) and reduced eGFR/CKD can shift absolute concentrations and bias a single threshold; %p-tau217 or a p-tau217:Aβ42 adjunct mitigates this impact and supports two-cutoff interpretation with explicit gray-zone governance. Together, these practices make the assays platform-agnostic, auditable, and ML-ready: fixed operating points (0.27; <0.22/>0.34 pg·mL^{-1}), clear escalation rules, and predefined recalibration strata translate analytic signal into reproducible clinical decisions.

DOI: 10.54254/2753-8818/2026.AU30169

3. DBMs and ML for AD: evidence, methods, and current state-of-the-art

3.1. Modality-level evidence and representative studies

DBMs for AD have moved from feasibility to early clinical utility. RADAR-AD demonstrated at-home, multimodal monitoring with smartphones and wearables, linking remote measures to clinical scales across disease stages in real-world settings [6]. Intuition (Nature Medicine, n=23,004, all 50 US states) established a blueprint for remote phenotyping and MCI detection at scale using smartphone and smartwatch telemetry, with governance for sampling bias and adherence. At the population level, a Nature Human Behaviour meta-analysis (57 studies; 411,430 adults) associated digital engagement with lower risk of cognitive decline, while an npj Digital Medicine scoping review (431 studies; 86 AI models) reported mean AUCs of 0.887 (AD) and 0.821 (MCI) but highlighted scarce external validation and calibration [7].

3.2. From signals to variables: feature engineering and modeling

Voice pipelines typically combine self-supervised speech embeddings (e.g., wav2vec 2.0, HuBERT) with acoustic–prosodic features, modeled via gradient-boosted trees or compact Transformers to handle language and device variability. Wearable approaches derive gait and cardiorespiratory features (IMU/PPG), often using 1D-CNNs or temporal Transformers; smartphone pipelines fuse passive mobility or interaction rhythms with brief tasks using sequential models (e.g., GBDT/GRU). Across modalities, domain-robust training and device and language normalization are essential for transportability.

3.3. Evaluation and validation: aligning with methodological standards

External validation is the primary evidentiary standard; internal resampling alone is insufficient. Studies should follow TRIPOD+AI for prediction models, STARD-AI for AI diagnostic accuracy, and DECIDE-AI for early live evaluations, with PROBAST+AI to appraise risk-of-bias/applicability. Beyond AUC, calibration and decision-curve analysis should translate outputs into thresholds and actions appropriate to pre-test probability and resource constraints [8, 9].

3.4. Strengths and limitations

Strengths: (1) low-burden, longitudinal capture via speech, wearables, or phones; (2) near-clinical discrimination in pragmatic cohorts (AUCs ≥0.8); demonstrated scalability and bias governance at population scale. Limitations: (1) limited external validation and calibration in many studies; (2) domain shift in language, device, and context; (3) under-reporting of pre-registration and threshold rationale in parts of the literature.

3.5. Fusion strategies: complementary roles of DBMs and BBMs

DBMs (behavioral signal, scale) and BBMs such as p-tau217 (pathology anchor) are methodologically complementary. Three pragmatic fusion strategies are used: early (feature-level), mid-level (shared encoders), and late (stacking/ensembles). Regardless of strategy, integration studies should include external validation, calibration, and decision-curve analysis and report net benefit vs. resource burden; Intuition (remote operations) and RADAR-AD (device/covariate consistency) offer practical templates.

DOI: 10.54254/2753-8818/2026.AU30169

4. Stratified diagnosis and optimized decision-making: a two-tier ML workflow anchored by passive digital biomarkers and blood tests

4.1. Rationale and overall concept

Converging evidence supports a two-tier diagnostic strategy that couples passive, at-home DBMs with a regulated blood anchor. DBMs, speech, gait, smartphone, and wearable signals provide low-burden, population-scale reach but still require broad external validation and probability calibration. By contrast, plasma p-tau217 on fully automated platforms achieves near-PET/CSF discrimination (AUC ≈ 0.93–0.96), performs well under dual cut-offs that confine an indeterminate fraction, and now has an FDA-cleared p-tau217/Aβ1-42 ratio for symptomatic adults. We operationalize these strands into a machine-learning triage: Level-1 screens and follows at home; Level-2 anchors clinic decisions to blood tests with explicit rules for escalation and reporting tied to pre-test probability, decision-curve net benefit, and budget scenarios. The innovation is a regulated pathway in which passive DBMs enter only after robustness/calibration gates, and Level-2 uses common operating points plus subgroup recalibration to deliver actionable, equitable decisions across devices, languages, ages, kidney function, and ancestries.

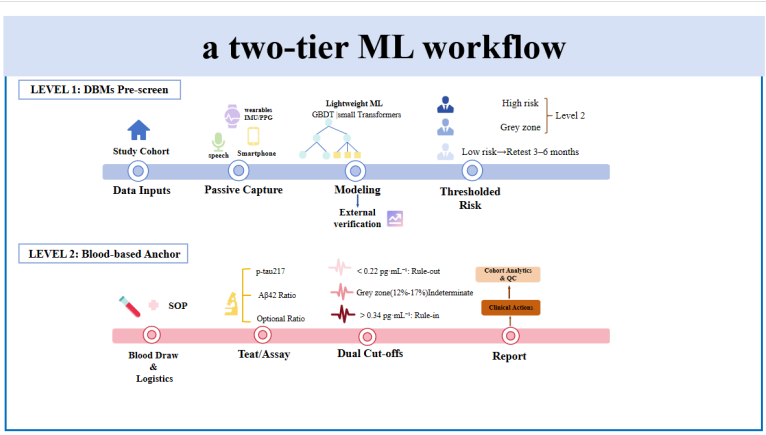

Figure 1. The two-tier ML workflow

4.2. Level 1: at-home pre-screening (DBMs × Lightweight ML)

Level-1 is designed for risk flagging rather than diagnosis, using truly passive, at-home digital signals, speech, gait, smartphone and wearable traces to expand reach with minimal participant burden. To approximate real deployment, we harmonize sampling and coordinates across devices and train lightweight learners (GBDT or compact temporal models such as 1D-CNN/TCN/small Transformers) under grouped cross-validation by site/device/language. Because domain shift is the dominant failure mode, we pair invariance-promoting training (IRM/GroupDRO) with modest domain adaptation, standardized feature pipelines, and post-hoc probability calibration (temperature or isotonic) referenced to the intended pre-test risk of the target population. Evidence must clear true external validation across independent centers/devices/languages and report not only ROC/PR-AUC but also reliability curves, Brier/ECE, E/O, and PPV/NPV at fixed operating quantiles. Outputs are calibrated risk quantiles plus a pre-specified re-test interval (typically 3–6 months), with missingness/adherence handling defined a priori. Fairness is enforced as a first-class constraint: we mandate stratified performance and calibration by device placement, language, age, sex, and ancestry, and we monitor for post-deployment drift with automatic recalibration triggers. In practice,

DOI: 10.54254/2753-8818/2026.AU30169

Level-1 serves as a scalable, auditable gate that enriches yield for Level-2 while avoiding overdiagnosis in the general population.

4.3. Level 2: clinic decision-making (blood anchor + two-cutoff + limited confirmation)

Level-2 converts flagged risk into accountable clinical decisions using a blood anchor centered on plasma p-tau217, which on automated platforms achieves near-PET/CSF discrimination (AUC around 0.93–0.96) and supports dual thresholds that provide rule-out and rule-in determinations while confining approximately 12–17% to a governed gray zone. Indeterminate cases follow explicit escalation rules—repeat blood testing on a short interval or small-n PET/CSF—and adding a p-tau217:Aβ42 ratio (or %p-tau217 on MS platforms) further reduces indeterminacy and enables age-tailored interpretation in very old adults or patients with reduced kidney function. Reporting binds thresholds to pre-test probability (e.g., SCD vs MCI) and quantifies clinical utility through decision-curve analysis and two-stage budget scenarios (direct PET/CSF vs blood then PET/CSF vs blood-only), typically yielding substantial resource savings without loss of prognostic yield; in trial recruitment, a sequential p-tau217→tau-PET strategy can reduce imaging volume by roughly 75–94%. Generalization is documented via cross-platform validation (immunoassay vs mass spectrometry) and care-tier transfer (primary vs specialty clinics), with subgroup calibration (e.g., ≥80 y, eGFR strata, language/device/ancestry) implemented via stratified temperature scaling or threshold offsets. Importantly, the FDA-cleared p-tau217/Aβ1-42 ratio supplies traceable intended-use language and performance claims, embedding Level-2 within a regulated pathway that maps analytic performance to equitable, auditable decisions [10, 11].

4.4. Workflow evaluation and metrics

We move beyond AUC-only to a decision-tied, three-pillar framework.

Diagnostic performance: Report ROC/PR-AUC; sensitivity, specificity, and PPV/NPV at fixed operating quantiles. Stratify Level-1 by device/language/age/sex/ethnicity; Level-2 by assay platform (immunoassay vs MS) and care tier (primary vs specialty).

Calibration & external validation: Provide reliability plots, Brier/ECE, E/O under true external validation—Level-1 requires cross-device/setting; Level-2 requires cross-platform/tier. Bind thresholds to pre-test probability (e.g., SCD ≈0.20–0.25; MCI ≈0.40–0.60).

Net benefit & budget: Present decision-curve analysis and two-stage budgets (direct PET/CSF vs. Blood, then PET/CSF vs. blood-only), including marginal cost per confirmed case and resources saved.

Fairness & generalization: Mandate stratified performance & calibration (device placement, language, age including ≥80 y, sex, ethnicity, renal/metabolic comorbidity) with pre-specified recalibration (e.g., stratified temperature scaling or threshold offsets). This unified scheme clarifies Level-1 reach and Level-2 decisiveness, aligning evidence with deployable practice.

5. Methods and translational challenges

5.1. Evidence of posture and model behavior

As for the actionable reliability, models must yield calibrated probabilities that are externally credible and explicitly bounded by intended use. Reporting focuses on what is safe to act on (risk quantiles, re-test cadence, rule-in/out logic), with preregistered analysis plans, versioned datasets/model cards, and clear off-label boundaries.

DOI: 10.54254/2753-8818/2026.AU30169

5.2. Governance and compliance by design

Because speech traces, mobility, and multi-modal logs may constitute special-category data, pipelines adopt data-minimizing, privacy-preserving processing with auditable data-lineage diagrams and DPIA/consent language. Change-control and monitoring (drift, failure, and override rates) are specified before deployment, and economic justifications are documented as budget-impact statements rather than post-hoc claims.

5.3. Fairness, portability, and maintenance

Equity is treated as a design constraint; performance and calibration are routinely audited by device/placement, language, age (≥80 years), sex, ancestry, and renal/metabolic comorbidity; subgroup-specific recalibration (simple temperature/threshold offsets) and mismatch costs are reported, with triggers for retraining or threshold revision. Portability is maintained through pre-specified external checks (new sites/devices/languages) and lightweight adaptation, ensuring that probabilities remain stable, auditable, and clinically usable as the system scales.

6. Conclusion

By 2025, an anchor-at-scale paradigm for Alzheimer's diagnosis will be within reach: digital as screen, blood as anchor. Passive, no-task digital biomarkers, from speech, gait, and smartphone/wearable telemetry, extend low-burden reach for population pre-screening and longitudinal follow-up. In clinics, plasma p-tau217 provides a regulated decision anchor whose discrimination approaches PET/CSF; dual cut-offs (e.g., <0.22 / >0.34 pg·mL^{-1}) bound indeterminacy and enable targeted, small-n confirmation. Taken together, diagnosis becomes a tiered, continuous service rather than an imaging-limited event.

The evidence base now supports resource-aware impact: in memory clinics, p-tau217 strategies deliver high accuracy with substantial reductions in lumbar puncture/amyloid-PET; in cognitively unimpaired cohorts, a sequential blood→tau-PET path preserves prognostic power while markedly reducing scans per participant. DBMs are feasible and inherently scalable when outputs are externally calibrated and coupled to pre-specified retest intervals, feeding Level-2 decisions with auditable risk estimates.

Key gaps remain. DBMs still require broader externalization and probability calibration across devices, languages, and sites; blood tests need age/eGFR-aware thresholds and consistent reporting of gray-zone governance. Heterogeneous datasets, inconsistent endpoints, and sparse prospective deployments limit comparability and transportability. Privacy and governance, especially for speech and geolocation, demand clearer, enforceable safeguards.

Accordingly, priorities must shift from marginal AUC gains to calibrated, externally validated, fair, and privacy-preserving systems with explicit net-benefit and budget justifications. We call for (1) registered, multicenter real-world studies that pre-specify calibration, decision-curve analysis, and two-stage budgets; (2) fairness-by-design auditing with stratified recalibration and mismatch-cost reporting (device, language, age ancestry, and renal/metabolic comorbidity); (3) privacy-preserving deployment (on-device feature extraction; and PETs with documented data flows); and (4) living guidance linking thresholds to pre-test risk and care context, with drift monitoring and scheduled recalibration. Success should be judged by well-calibrated, equitable decisions whose net benefit and costs justify adoption in everyday clinics, not by higher AUC alone.

DOI: 10.54254/2753-8818/2026.AU30169

References

[1] Palmqvist, S., et al. (2025) Plasma phospho-tau217 for Alzheimer's disease diagnosis in primary and secondary care using a fully automated platform. Nature Medicine, 1-8.
[2] Grande, G., et al. (2025) Blood-based biomarkers of Alzheimer's disease and incident dementia in the community. Nature Medicine: 2027-2035.
[3] Santos, L. E., et al. (2025) Performance of plasma biomarkers for diagnosis and prediction of dementia in a Brazilian cohort. Nature Communications, 16(1): 2911.
[4] Hu, S., Yu, H. & Gao, J. (2025) The pTau217/Aβ1-42 plasma ratio: The first FDA-cleared blood biomarker test for diagnosis of Alzheimer's disease. Drug Discoveries & Therapeutics, 19, 208-209.
[5] Coppinger, J., et al. (2025) Independent validation of the PrecivityAD2™ blood test to identify presence or absence of brain amyloid pathology in individuals with cognitive impairment. npj Dementia 1: 23.z
[6] Lentzen, M., et al. (2025) RADAR-AD: assessment of multiple remote monitoring technologies for early detection of Alzheimer's disease. Alzheimer's research & therapy 17(1): 29.
[7] Qi, W.H., et al. (2025) Alzheimer's disease digital biomarkers multidimensional landscape and AI model scoping review. npj Digital Medicine 8(1): 366.
[8] Sounderajah, V., et al. (2025) The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nature Medicine, 1-7.
[9] Collins, G. S. et al. (2024) TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385, e078378.
[10] Verberk, Inge MW, et al. (2025) Evidence-based standardized sample handling protocol for accurate blood-based Alzheimer's disease biomarker measurement: Results and consensus of the Global Biomarker Standardization Consortium. Alzheimer's & Dementia, 21(10): e70752.
[11] Ossenkoppele, R., et al. (2025) Plasma p-tau217 and tau-PET predict future cognitive decline among cognitively unimpaired individuals: implications for clinical trials. Nature Aging, 1-14.