Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Accurate Is Not Safe: Uncertainty-Aware Evaluation of Plasma p-tau217 Triage for Deferring Amyloid PET

Tahsien Al-Quraishi
Melbourne Institute of Technology https://orcid.org/0000-0003-3814-0431
Omar Al-Boridi
oboridi@mit.edu.au

Melbourne Institute of Technology
Raza Nowrozy
Melbourne Institute of Technology
Lamyaa Al-Omairi
Melbourne Institute of Technology
Naseer Ali Hussien
Al-Ayen Iraqi University
Enas R. Alwaily
Al-Ayen Iraqi University

Research Article

Keywords: Alzheimer's disease, blood-based biomarkers, p-tau217, amyloid PET, false-omission rate, conformal prediction, selective prediction, ADNI

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11848018/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Accurate Is Not Safe: Uncertainty-Aware Evaluation of Plasma p-tau217 Triage for Deferring Amyloid PET

Tahsien Al-Quraishi[1], Omar Al-Boridi[2,3*], Raza Nowrozy[2,4], Lamyaa Al-Omairi[2], Naseer Ali Hussien[5], Enas R. Alwaily[6], for the Alzheimer's Disease Neuroimaging Initiative[7]

[1]School of Business, Melbourne Institute of Technology, Melbourne, Victoria, Australia.
[2*]School of IT and Engineering (SITE), Melbourne Institute of Technology, 288 La Trobe Street, Melbourne, 3000, Victoria, Australia.
[3]School of Computing Technologies, RMIT University, Melbourne, Victoria, Australia.
[4]College of Art, Business, Law, Education & IT (CoABLEIT), Victoria University, Melbourne, Victoria, Australia.
[5]Department of Computer Engineering Technology, College of Technical Engineering, Al-Ayen Iraqi University (AUIQ), Nasiriyah, Iraq.
[6]College of Science, Al-Ayen Iraqi University (AUIQ), Nasiriyah, Iraq.
[7]Alzheimer's Disease Neuroimaging Initiative.

*Corresponding author(s). E-mail(s): oboridi@mit.edu.au;
Contributing authors: tquraishi@mit.edu.au; rnowrozy@mit.edu.au; lalomairi@academic.mit.edu.au; naseerali@alayen.edu.iq; enas.kazem@alayen.edu.iq;

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract

Plasma phosphorylated tau 217 (p-tau217) offers a promising blood-based approach to identifying Alzheimer's disease pathology, but high predictive accuracy does not necessarily justify deferring confirmatory amyloid positron emission tomography (PET). This study evaluates uncertainty-aware blood-first PET triage by integrating plasma p-tau217, clinical predictors, machine learning, conformal prediction and clinical safety assessment. Using retrospective data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), separate Fujirebio and C2N assay cohorts of symptomatic participants were assembled with blood collected on or before PET. Gradient-boosted models were evaluated with participant-disjoint cross-fitting, and conformal-negative triage, conservative exact-bound safety rules and the clinical two-cut-off approach were compared by the false-omission rate among participants selected to skip PET. The models discriminated amyloid PET status well in both cohorts and were reasonably calibrated, with modest gains from clinical predictors that varied across data splits. Every rule that spared participants PET nevertheless spared some amyloid-positive individuals, including the two-cut-off rule that met current performance benchmarks. The safety rules certified no deferral at the stricter limit, and thresholds chosen to meet the more lenient limit exceeded it when applied to new participants, because the calibration samples were smaller than such limits require. These findings show that predictive discrimination and nominal conformal coverage are insufficient evidence for safe exclusion from confirmatory imaging, and that safety-constrained triage studies must be sized for the decision they evaluate. Independent validation in larger cohorts and formally justified risk-control procedures are required before clinical implementation.

Keywords: Alzheimer's disease, blood-based biomarkers, p-tau217, amyloid PET, false-omission rate, conformal prediction, selective prediction, ADNI

1 Introduction

Dementia affected an estimated 57.4 million people worldwide in 2019, a number projected to reach 152.8 million by 2050, and Alzheimer's disease (AD) is its most common cause [1]. AD is now increasingly defined by its underlying amyloid and tau pathology rather than by clinical presentation alone [2, 3]. Amyloid positron emission tomography (PET) and cerebrospinal fluid (CSF) biomarkers can confirm this pathology, but their reliance on specialist facilities, radiotracers or invasive sampling limits their scalability. This limitation has become pressing since anti-amyloid monoclonal antibodies were shown to slow clinical decline in early symptomatic AD [4, 5], because treatment requires confirmation of amyloid pathology before initiation [6]. Blood-based biomarkers may therefore offer a more accessible way to prioritise confirmatory investigations in patients with cognitive symptoms [7].

Plasma tau phosphorylated at threonine 217 (p-tau217) is strongly associated with AD pathology [8, 9], although its diagnostic performance varies by assay, population and reference standard [10, 11]. Recent evidence supports its use in blood-first pathways. Moghekar et al. [12] validated prespecified dual thresholds for

a p-tau217/amyloid-beta ($A\beta_{1-42}$) ratio in symptomatic cohorts, explicitly retaining an indeterminate group. Bao et al. [13] found that combining plasma biomarkers and clinical covariates provided only modest additional discrimination for amyloid PET positivity in cognitively unimpaired adults. Other studies examined biomarker associations with cognitive decline and disease progression rather than PET-deferral safety [14, 15]. Franco [16] further distinguished diagnostic triage from baseline stratification and longitudinal monitoring. Together, these findings support blood-based assessment while underscoring the need to evaluate each strategy against its intended clinical decision.

A key unresolved question is whether accurate prediction can justify withholding confirmatory PET. Area under the receiver operating characteristic curve (AUROC) measures discrimination across thresholds but does not quantify the risk among patients selected for PET deferral. In particular, a model with high AUROC may still classify amyloid-positive patients as sufficiently low risk to avoid imaging. The relevant decision-level quantity is the *false-omission rate* (FOR): the proportion of amyloid-positive patients among those assigned to the negative, potentially deferred group. Screening efficiency must therefore be considered alongside the number of missed PET-positive cases and uncertainty in the estimated omission risk.

Probability calibration and conformal prediction provide complementary tools for assessing uncertainty. Calibration concerns agreement between predicted and observed risks, whereas conformal prediction constructs prediction sets under specified assumptions [17, 18]. Neither satisfactory average calibration nor nominal marginal conformal coverage guarantees an acceptable FOR among selected negatives. Consequently, the ability to classify participants as low risk must be distinguished from evidence that such classifications meet a prespecified safety criterion.

We address this distinction through a retrospective evaluation of uncertainty-aware blood-first amyloid PET triage in symptomatic participants from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Fujirebio and C2N plasma p-tau217 assays are analysed separately, using blood samples collected on or before the PET scan within a prespecified interval. We examine internally cross-fitted prediction, probability calibration, conformal-negative decisions and conservative calibration-based criteria for negative-only PET deferral, and compare them with the clinical two-cut-off approach. High-risk and uncertain cases are retained for confirmatory testing. The analysis assesses a proposed decision strategy, not a clinically validated replacement for PET.

The study makes three contributions:

1. **Assay-specific prediction assessment.** We evaluate plasma p-tau217 and clinical predictors in timing-restricted Fujirebio and C2N cohorts, reporting discrimination, calibration and the added value of clinical predictors under internal cross-fitting, together with their stability across data splits.
2. **Decision-level evaluation of PET deferral.** We compare conformal-negative triage with the clinical two-cut-off approach by potential deferral, missed PET-positive cases and false-omission risk, explicitly distinguishing prediction-set coverage and population-level benchmarks from safety among selected negatives.
3. **Conservative safety-constrained assessment and its sample-size requirement.** We assess whether exact-bound decision rules meet prespecified FOR limits

when applied to new participants, and apply the exact binomial zero-event bound to show how many calibration participants such limits require, which gives a simple planning check for future triage studies.

The central contribution is thus an evaluation of the gap between predictive discrimination and evidence for safe clinical action, rather than a claim that high AUROC establishes PET replaceability or anti-amyloid treatment eligibility. The remainder of the paper reviews related work (Section 2), details the methodology (Section 3), presents the results (Section 4), discusses the implications and limitations (Section 5), and concludes in Section 6.

2 Related Work

This section reviews the evidence on which a blood-first amyloid PET pathway would rest. It first considers the diagnostic accuracy of plasma p-tau217 and the two-cut-off rules proposed for clinical use, then the influence of population, setting and biological factors, then studies that address prognosis rather than confirmatory testing, and finally the methods available for uncertainty-aware abstention and risk control. Table 1 summarises the most relevant studies.

Blood-based biomarkers are increasingly used to inform the diagnostic evaluation of AD, but their clinical role depends on both analytical validity and the decision to be made. Plasma p-tau217 has shown strong discrimination of amyloid pathology, although accuracy varies with assay platform, patient population and reference standard. In ADNI plasma samples, Schindler et al. [19] compared commercial tests including the C2N and Fujirebio assays analysed here and found that p-tau217 measures classified amyloid PET status most accurately, whereas plasma Aβ42/40 performed less well. Barthélemy et al. [20] reported that the mass-spectrometry-based C2N test performed similarly to or better than clinical CSF tests. Ashton et al. [21] evaluated a commercial p-tau217 immunoassay across three observational cohorts and supported a three-range interpretation that preserves an intermediate group for further assessment. Moghekar et al. [12] subsequently studied a plasma p-tau217/Aβ_{1-42} ratio in 499 symptomatic adults, reporting diagnostic accuracy and the proportion of indeterminate results under prespecified thresholds. Explicit intermediate zones have also been proposed for workflow design. Brum et al. [22] combined p-tau217, age and apolipoprotein E (APOE) ε4 status to stratify patients with mild cognitive impairment (MCI) into low, intermediate and high risk, referring only the intermediate group for confirmatory CSF testing; the workflow preserved accuracy while substantially reducing confirmatory tests. Figdore et al. [23] developed assay-specific two-cutpoint rules and showed that the proportion of indeterminate results differed between immunoassays, and Benina et al. [24] found that a p-tau217/Aβ_{1-42} ratio reduced the indeterminate fraction. Yim et al. [25] reversed the order, using magnetic resonance imaging (MRI) atrophy and APOE status to stratify risk and reserving p-tau217 for intermediate-risk patients in Korean and ADNI cohorts. These studies support explicit diagnostic decision zones, but assay-specific cut-offs cannot be assumed to transfer to other measurement platforms or to p-tau217 alone. Moreover, their thresholds are chosen to reach target sensitivity and specificity in development

samples, which does not by itself bound the FOR among patients assigned to the negative zone.

Evidence from intended-use populations is equally important. Palmqvist et al. [26] evaluated blood-based AD testing in primary and secondary care among 1,213 patients with cognitive symptoms, using predefined cut-offs and prospectively assessed samples. Bao et al. [13] examined prediction of amyloid PET positivity in 1,407 cognitively unimpaired adults from the A4 and LEARN studies; adding biomarkers and clinical covariates yielded only modest additional discrimination. In a community-based sample, comorbidities such as chronic kidney disease were associated with higher plasma p-tau levels independently of amyloid [27], and plasma p-tau217 can also identify amyloid pathology before symptoms emerge [28]. Biological factors also shift p-tau217 independently of amyloid: Yun et al. [29] showed that chronic kidney disease, body mass index and anaemia altered optimal cut-offs, and that double cut-offs left 12–39% of results in an intermediate range. These investigations demonstrate the relevance of patient selection and clinical setting: prescreening asymptomatic individuals, supporting diagnostic work-up in memory care and selecting symptomatic patients to forgo PET are related but distinct decisions. The Global CEO Initiative on Alzheimer's Disease proposed minimum accuracy standards for triage and confirmatory blood tests [30], and the Alzheimer's Association clinical practice guideline [31] recommends minimum sensitivity of 90% and specificity of 75% for blood-based tests used for triage, and at least 90% for both measures when replacing PET or CSF testing in specialised care. These are test-performance recommendations for a defined clinical context, not direct bounds on the FOR among a model-selected PET-deferral subgroup.

Other work addresses disease progression rather than immediate confirmatory testing. Liu et al. [14] investigated plasma p-tau217/Aβ_{42} in relation to AD severity, cognition and structural changes in ADNI. Khorsand et al. [15] combined p-tau217, subjective cognitive concerns and digital testing to study longitudinal cognitive outcomes in A4/LEARN. Franco [16] distinguished diagnostic triage from baseline stratification and longitudinal monitoring, cautioning against applying a biomarker threshold indiscriminately across these purposes. Although these studies clarify biological and prognostic utility, they do not establish the error incurred when a PET-positive patient is assigned to omit confirmatory imaging.

Machine-learning models can integrate plasma measurements with genetic, demographic and cognitive predictors, but a high AUROC does not establish safety at a chosen referral threshold. Conformal prediction provides prediction-set coverage under exchangeability and appropriate calibration assumptions [17, 18]; marginal coverage, however, is not equivalent to control of missed amyloid pathology among participants selected for PET deferral. Label-conditional (Mondrian) conformal prediction provides coverage separately within each true class [32]; for the negative label this bounds the probability that an amyloid-positive patient is assigned to the negative set, a sensitivity-type guarantee, but still not the FOR, whose denominator is the selected negatives. In clinical imaging, conformal prediction has been used to flag unreliable predictions, sharply reducing errors among accepted predictions at the cost of referring a minority of cases [33]. Selective classification explicitly studies the trade-off between the number of accepted predictions and their errors [34], while conformal

risk control extends calibration-based methodology to certain prespecified loss functions [35]. These contributions establish that uncertainty-aware abstention and risk control are existing methodological areas. Their guarantees cannot be attributed to a different threshold-search heuristic without showing that its assumptions and selection procedure meet the relevant requirements.

The remaining question is not whether p-tau217 can discriminate amyloid pathology or whether selective prediction exists, but whether an assay-specific, retrospectively developed blood-first model supports any negative-only PET deferrals under explicitly stated omission-risk criteria. We address this question in timing-restricted symptomatic ADNI cohorts by reporting discrimination, conformal-negative and two-cut-off decisions, and conservative calibration-based safety-rule outcomes separately. The study is an exploratory internal evaluation, not independent validation or a formal prospective risk-control guarantee.

Table 1 Representative evidence relevant to blood-first amyloid PET triage and uncertainty-aware decision-making.

Study	Population or focus	Approach	Relevance and limitation for PET deferral
Ashton et al. [21]	Three observational cohorts; AD pathology	Commercial p-tau217 immunoassay; three-range interpretation	Assay-specific thresholds and diagnostic classification; not a PET-deferral safety guarantee
Moghekar et al. [12]	499 symptomatic adults	Prespecified p-tau217/Aβ_{1-42} thresholds	Indeterminate zone and predictive values; ratio-specific cut-offs
Schindler et al. [19]	ADNI plasma; commercial assays	Head-to-head comparison including C2N and Fujirebio	Same data source and assays; accuracy rather than deferral safety
Brum et al. [22]	MCI; two memory-clinic cohorts	p-tau217, age and APOE risk strata; intermediate group to CSF	Two-step workflow; empirical thresholds without an omission-risk bound
Figdore et al. [23]	MCI or mild dementia	Assay-specific two-cutpoint rules	Indeterminate fraction differs by assay; not a selected-subgroup risk bound
Yun et al. [29]	Multicentre memory clinic and community cohorts	Single, subgroup-specific and double cut-offs	Biological confounders and confirmatory-imaging costs; empirical thresholds
Palmqvist et al. [26]	1,213 symptomatic primary- and secondary-care patients	Predefined blood-test thresholds, prospective evaluation	Clinical diagnostic workflow; not the present model-selected deferral policy
Palmqvist et al. [31]	Specialised memory care; practice guideline	Performance-based triage and substitution criteria	Clinical performance benchmarks; not a direct bound on selected-subgroup omission risk
Bao et al. [13]	1,407 cognitively unimpaired adults	Biomarker and combined covariate models	PET endpoint, but different case mix and screening purpose
Liu et al. [14]	ADNI; disease continuum	Plasma biomarker and longitudinal associations	Severity and progression rather than PET-deferral error
Khorsand et al. [15]	A4/LEARN; longitudinal cognition	Plasma, subjective and digital assessments	Prognostic task rather than confirmatory PET deferral
Franco [16]	Biomarker-use perspective	Triage versus stratification versus monitoring	Clarifies intended use and threshold portability
Geifman and El-Yaniv [34]	General machine-learning classification	Selective classification with rejection	Risk–coverage trade-off; not an AD biomarker evaluation
Angelopoulos et al. [35]	General prediction methodology	Conformal risk control	Risk-control framework under specified assumptions; not automatically applicable to our heuristic
This study	**Symptomatic ADNI; Fujirebio and C2N**	**Cross-fitting; conformal, two-cut-off and safety-rule comparison**	**Observed missed positives and permitted deferrals; exploratory internal evaluation only**

3 Materials and Methods

This section describes the data source and cohort construction, the predictors and their preprocessing, the prediction models and cross-fitting scheme, the negative-only triage rules and their clinical comparators, and the measures used to evaluate discrimination, calibration and deferral safety. Figure 1 shows how these components fit together.

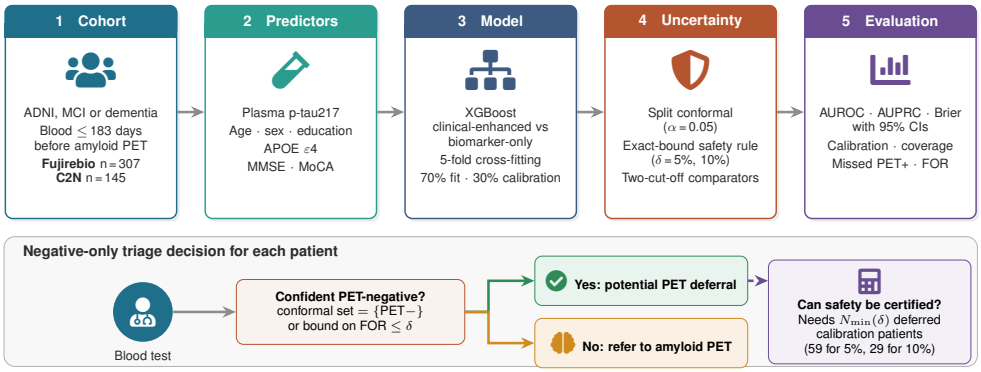

Fig. 1 Study framework. Retrospective evaluation; the workflow is not a validated clinical pathway.

3.1 Data source, study design and participants

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD) [36].

This was a retrospective diagnostic modelling study with a decision-level evaluation, reported following the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis for artificial intelligence (TRIPOD+AI) guideline [37] (checklist in Supplementary Material). The intended population was symptomatic individuals with MCI or dementia. Data were taken from the ADNIMERGE2 R package (version 0.1.1): plasma biomarkers from the University of Pennsylvania Fujirebio table and the C2N PrecivityAD2 table, amyloid PET from the UC Berkeley processed amyloid table, and diagnosis, demographics, apolipoprotein E (APOE) genotype, Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA) from the corresponding ADNI tables. The Fujirebio assay was the primary analysis and the C2N assay a secondary, assay-specific analysis.

Plasma measurements were linked to amyloid PET scans by participant identifier and date. For each participant and assay, all plasma–PET pairs within 183 days were formed and the pair with the shortest absolute interval was kept, so that each participant contributed one record per assay. The diagnosis at the plasma visit was taken from the diagnostic summary at the same visit or, if absent, from the nearest assessment within 90 days, and only participants with MCI or dementia were retained. The primary analysis further required blood to be collected on or before the PET scan:

$$\Delta_i = r_i - b_i, \qquad 0 \le \Delta_i \le 183 \text{ days}, \tag{1}$$

where b_i and r_i are the blood and PET dates, because the intended decision uses the blood result to decide whether PET is needed. A sensitivity analysis used a symmetric window, $|\Delta_i| \leq 183$ days.

Amyloid PET status was the ADNI binary amyloid status from the UC Berkeley processed florbetapir and florbetaben data, which classifies scans using tracer-specific cortical summary standardised uptake value ratio (SUVR) thresholds referenced to the whole cerebellum (positive at SUVR $\geq$1.11 for florbetapir and $\geq$1.08 for florbetaben) [38]. Scans that failed visual quality control were excluded. The study analysed existing ADNI data; no new assays, PET reads or recruitment were performed. The sample size was therefore fixed by data availability rather than calculated in advance; its consequence for the safety rule is examined in Section 4.4.

3.2 Predictors and preprocessing

Plasma p-tau217 was the Fujirebio measurement or the C2N p-tau217 concentration, both in pg/mL and each used on its own scale. Other predictors were age at blood draw, sex, years of education, diagnosis (MCI or dementia), APOE ε4 carrier status, and the most recent MMSE and MoCA scores within 12 months on or before the blood draw. Missing continuous values were replaced by the median of the fitting partition $\mathcal{F}_k$,

$$\widetilde{x}_{ij}^{(k)} = \begin{cases} x_{ij}, & x_{ij} \text{ observed}, \\ \text{median}_{r \in \mathcal{F}_k :\, x_{rj} \text{ observed}}\, x_{rj}, & x_{ij} \text{ missing}, \end{cases} \tag{2}$$

so that no imputation parameter was estimated from calibration or evaluation participants. Missing-data frequencies are reported in Supplementary Table S1.

3.3 Prediction models and cross-fitting

We used extreme gradient boosting (XGBoost) with fixed settings that were not tuned in this analysis: 150 trees, maximum depth 2, learning rate 0.05, row and column subsampling 0.9, and random seed 20261027. Shallow trees and a low learning rate were chosen to limit overfitting in samples of this size. For a predictor vector $\boldsymbol{x}_i$ the model returns

$$f(\boldsymbol{x}_i) = \sum_{m=1}^{150} \eta h_m(\boldsymbol{x}_i), \qquad \widehat{p}_i = \frac{1}{1 + \exp[-f(\boldsymbol{x}_i)]}, \tag{3}$$

with h_m a fitted tree and $\eta = 0.05$. A clinical-enhanced model used all predictors and a biomarker-only model used p-tau217 alone.

Performance was estimated by five-fold, participant-disjoint cross-fitting stratified by PET status. Within each outer development set, 30% of participants (stratified) formed a calibration set and 70% a fitting set; models were fitted on the fitting set only. Each participant therefore received one out-of-fold prediction from a model that

never saw that participant:

$$\widehat{p}_i = \widehat{p}_i^{(-k)} \quad \text{for } i \in \mathcal{I}_k, \qquad \bigcup_{k=1}^{5} \mathcal{I}_k = \{1, \ldots, n\}. \tag{4}$$

3.4 Negative-only triage rules

All rules decide only whether a participant may skip PET (negative-only deferral); everyone else is referred to PET. *Potential deferral* denotes a retrospective classification, not a clinical recommendation.

3.4.1 Split conformal prediction

With nonconformity score $s(\boldsymbol{x}, y) = 1 - \widehat{p}_y(\boldsymbol{x})$, where $\widehat{p}_1 = \widehat{p}$ and $\widehat{p}_0 = 1 - \widehat{p}$, the conformal p-value of label y for a new participant is

$$\pi_y(\boldsymbol{x}) = \frac{1 + |\{j \in \mathcal{C} : s(\boldsymbol{x}_j, y_j) \geq s(\boldsymbol{x}, y)\}|}{|\mathcal{C}| + 1}, \tag{5}$$

computed over the calibration set $\mathcal{C}$, and the prediction set is $\Gamma_\alpha(\boldsymbol{x}) = \{y : \pi_y(\boldsymbol{x}) > \alpha\}$ with $\alpha = 0.05$ [17, 18]. A participant was deferred if $\Gamma_\alpha = \{0\}$. This marginal construction guarantees average coverage of at least 95% under exchangeability, but not a bound on the FOR among deferred participants. As a secondary analysis, label-conditional (Mondrian) p-values used only calibration participants with the same label [32].

3.4.2 Exact-bound safety rule

For a candidate threshold t, let N_t calibration participants have $\widehat{p} \leq t$, of whom M_t are PET positive. The one-sided 95% Clopper–Pearson upper bound on the FOR is

$$U_{0.95}(m, N) = \begin{cases} 1, & m = N, \\ \text{Beta}^{-1}(0.95; m + 1, N - m), & 0 \leq m < N. \end{cases} \tag{6}$$

The rule selects the largest t with $N_t > 0$ and $U_{0.95}(M_t, N_t) \leq \delta$, for $\delta = 0.05$ (primary) and 0.10, and defers evaluation participants with $\widehat{p} \leq t$; if no t qualifies, everyone is referred. Because $U_{0.95}(0, N) = 1 - 0.05^{1/N}$ is the smallest attainable bound, a limit δ can only be met when at least

$$N_{\min}(\delta) = \left\lceil \frac{\ln 0.05}{\ln(1 - \delta)} \right\rceil \tag{7}$$

calibration participants are selected, none of them PET positive: $N_{\min} = 59$ for $\delta = 0.05$ and 29 for $\delta = 0.10$. This is the exact binomial bound for zero observed events, which underlies the familiar rule of three [39]; we use it here as a feasibility check on the triage design. The rule was applied in two ways. In the *per-fold* version, thresholds

were chosen on each fold's calibration set. In the *pooled* version, the threshold for outer fold k was chosen on the out-of-fold predictions of the other four folds and applied to fold k, which uses roughly four times as many participants for threshold selection. Selecting the best of many thresholds on the same data weakens the nominal confidence statement, so both versions are calibration-based heuristics rather than formal guarantees.

3.4.3 Clinical comparators

Two-cut-off rules of the kind used clinically [22, 23] were derived within each outer development set: a lower cut-off keeping sensitivity at least 95% and an upper cut-off keeping specificity at least 95%. Evaluation participants below the lower cut-off were deferred, those at or above the upper cut-off were called positive, and the rest (indeterminate) were referred. The rule was applied to p-tau217 and to the ratio of p-tau217 to Aβ_{42} measured on the same platform [24].

3.5 Evaluation

Discrimination was summarised by the AUROC, estimated as

$$\widehat{\mathrm{AUROC}} = \frac{1}{|P||N|} \sum_{i \in P} \sum_{j \in N} \left[\mathbf{1}(\widehat{p}_i > \widehat{p}_j) + \tfrac{1}{2} \mathbf{1}(\widehat{p}_i = \widehat{p}_j) \right], \tag{8}$$

together with the area under the precision–recall curve (AUPRC) and the Brier score, $n^{-1} \sum_i (\widehat{p}_i - Y_i)^2$, all computed on the pooled out-of-fold predictions. 95% confidence intervals (CIs) used 2,000 bootstrap resamples stratified by PET status. The clinical-enhanced and biomarker-only AUROCs were compared with the DeLong test for correlated curves [40]. Calibration was assessed with calibration plots, the calibration slope and the calibration-in-the-large intercept [41], and conformal validity by empirical coverage, overall and by PET status.

For each triage rule we report the number deferred, missed PET-positive participants, and the FOR among deferred participants,

$$\mathrm{FOR}(D) = \frac{\sum_{i \in D} Y_i}{|D|}, \qquad |D| > 0, \tag{9}$$

with its exact one-sided 95% upper bound (Equation 6); FOR is undefined when no one is deferred. Treating referral to PET as a positive test, we also report the sensitivity and specificity of each rule. Because FOR depends on prevalence, we re-weighted the class-specific deferral rates to PET-positive prevalences of 50% and 30%: $\mathrm{FOR}(\pi) = \pi a / \{\pi a + (1-\pi) b\}$, where a and b are the proportions of PET-positive and PET-negative participants deferred; this assumes that these rates carry over to settings with a different prevalence. To assess how much the results depended on a single random partition, the whole analysis was repeated with 20 different random seeds, each changing the fold assignment, the calibration split and the model randomness. Two further sensitivity analyses used the symmetric blood–PET window and, for

C2N, the percentage of phosphorylated p-tau217 (%p-tau217) instead of the p-tau217 concentration. Cohorts were built in R (version 4.3.3), and analyses used Python (version 3.13.17) with XGBoost (version 3.4.1) and scikit-learn (version 1.9.1).

4 Results

The results follow the order of the analysis: the assay cohorts are described first, followed by the discrimination and calibration of the prediction models, the outcomes of conformal and two-cut-off triage, and finally the outcomes of the exact-bound safety rule, the stability of all results across data splits and the sensitivity analyses.

4.1 Cohorts

Of 418 symptomatic Fujirebio records linked to PET within 183 days, 111 (26.6%) had blood collected after PET, leaving 307 participants, of whom 189 (61.6%) were PET positive (Table 2). The C2N cohort comprised 145 of 178 linked participants, of whom 91 (62.8%) were PET positive; 139 of the 145 C2N participants were also in the Fujirebio cohort. PET-positive participants were older, more often APOE $\varepsilon 4$ carriers and more often had dementia, and they had lower cognitive scores and higher p-tau217 (Table 3, Figure 2). The median blood-to-PET interval was 8 days for Fujirebio and 14 days for C2N. In the C2N cohort, 45 participants, 44 of them PET negative, had the lowest reported p-tau217 value (0.65 pg/mL).

Table 2 Cohort derivation

Step	Fujirebio	C2N
Symptomatic participants with plasma–PET link	418	178
Blood collected after PET, excluded	111 (26.6%)	33 (18.5%)
Analysis cohort	307 (73.4%)	145 (81.5%)
Florbetapir / florbetaben PET	160 / 147	77 / 68

Plasma–PET link: shortest interval within 183 days in either direction; MCI or dementia at the plasma visit. Percentages are of linked participants. 139 participants appear in both cohorts.

4.2 Discrimination and calibration

Cross-fitted AUROC of the clinical-enhanced model was 0.942 (95% CI 0.914–0.967) for Fujirebio and 0.961 (0.926–0.989) for C2N (Table 4). The biomarker-only models reached 0.922 (0.883–0.955) and 0.925 (0.870–0.968), so adding clinical predictors raised AUROC by 0.020 (0.003 to 0.037; DeLong $p = 0.02$) and 0.036 (0.007 to 0.066; $p = 0.02$), respectively (Figure 3). Calibration slopes were 0.81 and 0.84, with intercepts -0.14 and -0.03, indicating predictions that were somewhat too extreme but on average well located (Figure 4).

Table 3 Participant characteristics

Characteristic	All	PET+	PET−
Fujirebio cohort			
n	307	189	118
Age, years	75.0 (8.5)	76.1 (8.2)	73.2 (8.7)
Female, n (%)	147 (47.9)	94 (49.7)	53 (44.9)
Education, years	16.0 (2.6)	16.1 (2.8)	15.8 (2.4)
Dementia, n (%)	83 (27.0)	71 (37.6)	12 (10.2)
APOE ε4 carrier, n (%)	143 (48.3)	123 (67.2)	20 (17.7)
MMSE	26.0 (3.7)	25.2 (4.0)	27.3 (2.7)
MoCA	21.3 (4.6)	20.1 (5.0)	23.0 (3.3)
Plasma p-tau217, pg/mL	0.29 (0.13–0.60)	0.48 (0.32–0.74)	0.11 (0.08–0.15)
Blood-to-PET interval, days	8 (1–34)	8 (1–28)	9 (2–40)
C2N cohort			
n	145	91	54
Age, years	75.3 (8.4)	76.8 (7.8)	72.7 (9.0)
Female, n (%)	79 (54.5)	53 (58.2)	26 (48.1)
Education, years	16.1 (2.7)	16.2 (3.0)	15.9 (2.3)
Dementia, n (%)	33 (22.8)	28 (30.8)	5 (9.3)
APOE ε4 carrier, n (%)	64 (47.8)	55 (64.7)	9 (18.4)
MMSE	26.3 (3.1)	25.8 (3.4)	27.0 (2.4)
MoCA	21.6 (5.0)	20.4 (5.5)	23.6 (3.1)
Plasma p-tau217, pg/mL	2.71 (0.65–5.38)	4.49 (2.84–6.68)	0.65 (0.65–0.65)
Blood-to-PET interval, days	14 (0–41)	12 (1–35)	18 (0–41)

Mean (standard deviation [SD]), n (%) or, for p-tau217 and interval, median (interquartile range). PET+ and PET−: amyloid PET positive and negative. Percentages of APOE carriers are of participants with known genotype. Assay scales are not comparable. Missing data: Supplementary Table S1.

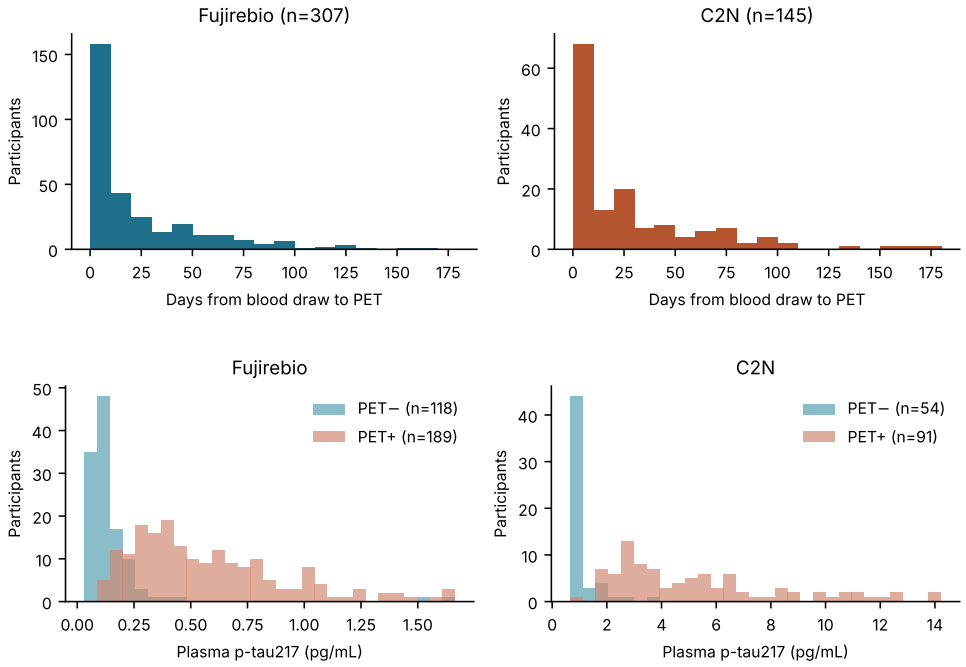

Fig. 2 Blood-to-PET intervals (top) and p-tau217 by PET status (bottom)

Table 4 Discrimination and calibration

Measure	Fujirebio	C2N
AUROC	0.942 (0.914–0.967)	0.961 (0.926–0.989)
AUPRC	0.958 (0.933–0.978)	0.977 (0.955–0.994)
Brier score	0.084 (0.061–0.110)	0.058 (0.028–0.094)
Calibration slope	0.81	0.84
Calibration intercept	−0.14	−0.03
AUROC, biomarker-only	0.922 (0.883–0.955)	0.925 (0.870–0.968)
ΔAUROC	0.020 (0.003 to 0.037)	0.036 (0.007 to 0.066)
DeLong p	0.02	0.02

Clinical-enhanced model unless stated; out-of-fold predictions with 95% stratified bootstrap CIs. Intercept: calibration-in-the-large. ΔAUROC: clinical-enhanced minus biomarker-only.

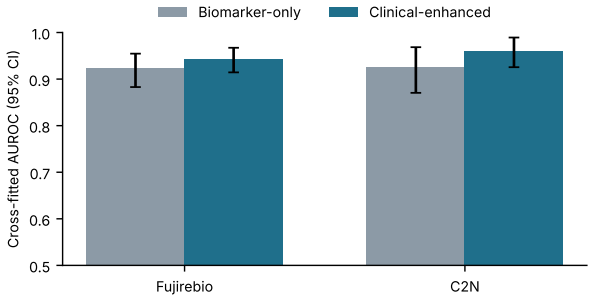

Fig. 3 AUROC of biomarker-only and clinical-enhanced models

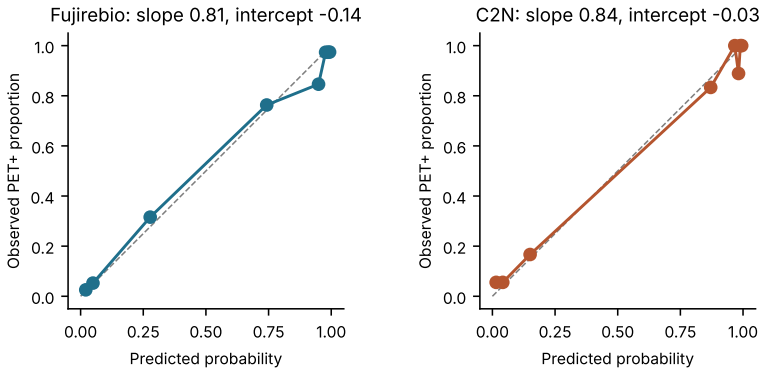

Fig. 4 Calibration of the clinical-enhanced models

4.3 Conformal and two-cut-off triage

Empirical conformal coverage was 97.4% for Fujirebio and 95.2% for C2N against the nominal 95% (Supplementary Table S2). Conformal-negative triage deferred 51 Fujirebio participants (16.6%), of whom 2 were PET positive (FOR 3.9%; upper bound 11.8%), and 31 C2N participants (21.4%), of whom 5 were PET positive (FOR 16.1%;

upper bound 31.0%) (Table 5, Figure 5). The label-conditional version deferred more participants (64 and 48) but also missed more PET-positive cases (7 in each cohort). The two-cut-off p-tau217 rule deferred 103 Fujirebio participants (33.6%) with 10 missed (FOR 9.7%; upper bound 15.9%) and 54 C2N participants (37.2%) with 6 missed (FOR 11.1%; upper bound 20.8%); 17.9% and 2.8% of results were indeterminate. The p-tau217/Aβ_{42} ratio gave similar results. Treating referral as a positive test, the two-cut-off rule had a sensitivity of 94.7% and a specificity of 78.8% for Fujirebio, and 93.4% and 88.9% for C2N, meeting the benchmarks of 90% and 75% recommended for triage tests [31]; the conformal rule had a higher sensitivity (98.9% and 94.5%) but a lower specificity (41.5% and 48.1%). Re-weighted to a PET-positive prevalence of 50%, the point estimate of the conformal FOR would be 2.5% (Fujirebio) and 10.2% (C2N), and that of the two-cut-off rule 6.3% and 6.9% (Supplementary Table S3).

Table 5 Negative-only triage outcomes

Rule	Deferred, n (%)	Missed	FOR (%)	Bound (%)	Indet. (%)
Fujirebio					
Conformal-negative	51 (16.6)	2	3.9	11.8	–
Conformal, label-conditional	64 (20.8)	7	10.9	19.6	–
Two cut-offs, p-tau217	103 (33.6)	10	9.7	15.9	17.9
Two cut-offs, ratio	118 (38.4)	10	8.5	14.0	4.6
Safety rule 5%, per fold	0	0	Undefined	–	–
Safety rule 10%, per fold	0	0	Undefined	–	–
Safety rule 5%, pooled	0	0	Undefined	–	–
Safety rule 10%, pooled	35 (11.4)	4	11.4	24.3	–
C2N					
Conformal-negative	31 (21.4)	5	16.1	31.0	–
Conformal, label-conditional	48 (33.1)	7	14.6	25.7	–
Two cut-offs, p-tau217	54 (37.2)	6	11.1	20.8	2.8
Two cut-offs, ratio	55 (37.9)	6	10.9	20.4	2.1
Safety rule 5%, per fold	0	0	Undefined	–	–
Safety rule 10%, per fold	0	0	Undefined	–	–
Safety rule 5%, pooled	0	0	Undefined	–	–
Safety rule 10%, pooled	16 (11.0)	5	31.2	54.8	–

Out-of-fold decisions in each full cohort at the prespecified seed. Missed: PET-positive participants deferred. Bound: exact one-sided 95% upper bound on the FOR. Indet.: indeterminate, referred by the two-cut-off rule. Ratio: p-tau217/Aβ_{42}. FOR is undefined when no one is deferred. Per fold: thresholds chosen on each calibration set; pooled: thresholds chosen on the other folds' out-of-fold predictions.

4.4 Safety-constrained deferral

With per-fold thresholds, neither the 5% nor the 10% rule deferred any participant. Each calibration set held 74 Fujirebio participants, of whom 28–29 were PET negative, and 35 C2N participants, of whom 13 were PET negative. These numbers are below the 59 selected participants needed for the 5% limit, and for C2N below the 29 needed for the 10% limit (Equation 7); any PET-positive case among those selected raises

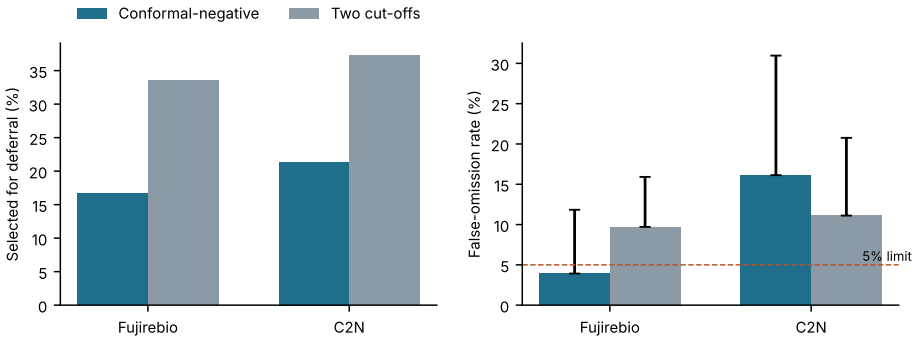

Fig. 5 Deferral and FOR by triage rule

the requirement further. Figure 6 shows how many deferred calibration participants are needed to certify each limit as the true FOR rises. With pooled thresholds, the 5% rule still deferred no one in either cohort. The 10% rule deferred 35 Fujirebio participants (11.4%), of whom 4 were PET positive (FOR 11.4%; upper bound 24.3%), and 16 C2N participants (11.0%), of whom 5 were PET positive (FOR 31.2%; upper bound 54.8%). In both cohorts, the observed FOR among newly deferred participants therefore exceeded the 10% limit that the thresholds were chosen to meet.

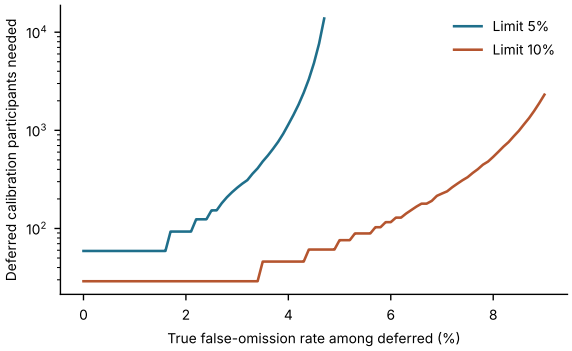

Fig. 6 Calibration participants needed to certify a FOR limit

4.5 Stability and sensitivity analyses

Across 20 random seeds (Table 6), AUROC varied little, but the other results varied considerably. The gain from clinical predictors ranged from −0.006 to 0.020 for Fujirebio, so the prespecified split lay at the upper end of this range, and from 0.013 to 0.036 for C2N. The number of conformal-negative deferrals ranged from 51 to 97 (Fujirebio) and 26 to 46 (C2N), and the conformal FOR ranged from 3.8% to 10.1% and from 0.0% to 16.1%. The two-cut-off rule was more stable, with a FOR between

7.5% and 13.8%. The pooled 5% rule deferred participants in only one of the 40 analyses, in which 3 of the 19 deferred Fujirebio participants were PET positive (FOR 15.8%). The pooled 10% rule deferred participants in most splits, and in 26 of the 36 splits with any deferral the observed FOR exceeded 10%.

In the symmetric-window analysis (418 Fujirebio and 178 C2N participants), AUROC was 0.942 and 0.968, conformal triage deferred 120 and 40 participants with 13 and 2 missed, and the pooled 5% rule deferred no one (Supplementary Table S4). Using C2N %p-tau217 raised AUROC to 0.984 (0.965–0.998; biomarker-only 0.981); conformal triage then deferred 41 participants with 1 missed (FOR 2.4%; upper bound 11.1%), but the pooled 5% rule again deferred no one, and the pooled 10% rule deferred 13 participants with 3 missed (FOR 23.1%).

Table 6 Stability across 20 random splits

Measure	**Fujirebio**	**C2N**
AUROC	0.941 (0.931–0.954)	0.972 (0.961–0.985)
ΔAUROC	0.007 (−0.006 to 0.020)	0.025 (0.013 to 0.036)
Conformal-negative: deferred, n	76 (51–97)	38 (26–46)
Conformal-negative: FOR (%)	5.8 (3.8–10.1)	5.1 (0.0–16.1)
Two cut-offs, p-tau217: deferred, n	103 (101–106)	54 (52–58)
Two cut-offs, p-tau217: FOR (%)	9.8 (7.8–11.3)	9.3 (7.5–13.8)
Pooled 5%: splits with any deferral	1/20	0/20
Pooled 10%: splits with any deferral	19/20	17/20
Pooled 10%: splits with FOR >10%	13/19	13/17

Median (range) across 20 repetitions of the whole analysis, each with a different random seed for fold assignment, calibration split and model fitting. ΔAUROC: clinical-enhanced minus biomarker-only. Pooled: exact-bound safety rule with pooled thresholds.

5 Discussion

This study asked whether accurate blood-based prediction of amyloid PET positivity supports a defensible decision to skip confirmatory PET. This section interprets the three main findings, relates them to previous studies of blood-based amyloid testing and uncertainty-aware prediction, and sets out the limitations and the work needed before clinical use. First, plasma p-tau217 models discriminated PET status well in both cohorts and were reasonably calibrated, although calibration slopes below one indicate somewhat overconfident predictions. Clinical predictors improved discrimination modestly; the gain was consistent for C2N but varied in sign across data splits for Fujirebio, so the significant improvement at the prespecified split should not be overinterpreted. Discrimination, however, describes ranking across all thresholds, not the error among the patients who would actually skip PET.

Second, every negative-only rule deferred some amyloid-positive participants. The two-cut-off rule, which mirrors current practice, met the sensitivity and specificity benchmarks recommended for triage tests [31] in both cohorts, yet about one in ten of the participants it would have spared PET was amyloid positive. Conformal-negative

triage deferred fewer participants and, at the prespecified split, achieved a low FOR for Fujirebio but a high FOR for C2N, with upper bounds well above 5% in both; across splits its FOR ranged widely. Meeting a population-level benchmark or a nominal coverage level therefore does not bound the FOR among the patients who would actually skip PET, and in samples of this size the FOR of any selected subgroup is itself highly uncertain. Third, a rule designed to certify a FOR limit could not do so. With per-fold calibration sets, the 5% limit was unattainable by design, as Equation 7 predicts. Pooling calibration information made deferral possible at the 10% limit, but the observed FOR among newly deferred participants exceeded that limit at the prespecified split and in most repeated splits: a threshold chosen on a few dozen PET-negative participants is optimistic when applied to new patients. The practical lesson is that a safety-constrained evaluation must be sized for its risk limit before it is run; otherwise it either certifies nothing or certifies thresholds that do not hold. For study planning, Equation 7 and Figure 6 give the calibration size a blood-first triage evaluation needs before its safety claim can be tested.

These findings complement the growing clinical evidence for plasma p-tau217. Moghekar et al. [12] and Figdore et al. [23] showed that prespecified lower and upper thresholds leave an indeterminate group and achieve high predictive values. Our analysis asks a narrower question: among patients a rule would let skip PET, is the missed-amyloid burden compatible with a stated limit, and can a study of a given size show it? Bao et al. [13] studied cognitively unimpaired adults, and Liu et al. [14], Khorsand et al. [15] and Franco [16] addressed prognosis rather than confirmatory testing; our endpoint is binary amyloid PET status within a defined interval and should not be extended to those tasks. The uncertainty results also need careful reading. Marginal conformal coverage does not bound the FOR, and label-conditional coverage bounds the probability that a PET-positive patient is deferred, a sensitivity-type quantity, rather than the FOR itself [32]; in our data it deferred more participants and missed more PET-positive cases. Formal selective-risk methods such as conformal risk control [35] target the right quantity, but they face the same sample-size requirement. Clinically, blood testing is clearly useful for prioritising referrals; whether it can safely replace PET for a subset of patients depends on the omission risk among that subset, the acceptable risk limit and a study large enough to measure both.

5.1 Limitations and future work

First, this was a retrospective analysis of ADNI participants selected for having both plasma and PET measurements. The cohorts were small, with high PET-positive proportions (189/307 and 91/145), so omission estimates are imprecise and predictive values will differ in memory clinics with other prevalence; the prevalence re-weighting illustrates this but does not replace validation. ADNI participants are predominantly White and highly educated, and plasma biomarker performance has been reported to differ by race [42]. The two assay cohorts overlap almost completely (139 of the 145 C2N participants are also in the Fujirebio cohort), so the C2N analysis is not independent replication, and earlier exploratory iterations of this work used the same participants, so data-informed choices of model settings cannot be fully excluded. Second, the shortest-interval pairing was done before the direction of the interval

was considered and may have excluded participants with an eligible blood-before-PET pair; PET status inherits the uncertainty of tracer-specific thresholds near the cut-off [38]; and requiring blood before PET orders the measurements but does not show that a blood-first decision could have been made prospectively. Many C2N PET-negative participants had the lowest reported p-tau217 concentration, which limits ranking among them; %p-tau217 discriminated better but did not change the safety conclusions. Single median imputation ignores imputation uncertainty, and reserving 30% of each development set for calibration reduces the data available for model fitting. Third, conformal coverage is not clinical safety, and the threshold search in the safety rule does not preserve a formal guarantee after selection, as the pooled results show.

Future work should prespecify the risk limit and analysis, size the calibration data using Equation 7, evaluate formal risk-control procedures with calibration sets held out from threshold selection in independent, clinically representative cohorts with locally calibrated assays, add biological covariates such as kidney function [29], and link any claim of PET savings to an economic evaluation that includes the costs of missed and delayed diagnoses [43].

6 Conclusion

This study examined whether accurate blood-based prediction of amyloid PET status is enough to justify skipping confirmatory PET in symptomatic patients. Using timing-restricted ADNI cohorts for the Fujirebio and C2N plasma p-tau217 assays, we combined gradient-boosted prediction under participant-disjoint cross-fitting with negative-only triage rules: split conformal prediction, a conservative exact-bound safety rule and the clinical two-cut-off approach, all evaluated by the number of deferred participants, missed PET-positive cases and the FOR among those deferred. Plasma p-tau217 discriminated amyloid PET status well in both cohorts, and routine clinical predictors added a modest and partly unstable gain. This accuracy, however, did not carry over to the decision that matters for patients. Every rule that spared participants PET also spared some amyloid-positive participants, including the two-cut-off rule that met current performance benchmarks. No rule certified the stricter limit, and thresholds chosen to meet the more lenient limit exceeded it when applied to new participants, because the calibration data contained fewer PET-negative participants than such limits require. The zero-event bound and the sample-size curve shown here give future studies a simple way to plan calibration data before testing a safety claim. These conclusions are drawn from a retrospective, internally cross-fitted analysis of small, overlapping research cohorts with high amyloid prevalence, and they do not establish a clinical pathway. Before blood-first triage is used to withhold PET, studies should prespecify an acceptable omission risk, size their calibration data to test it, use formal risk-control methods with held-out calibration, and validate the chosen rule in independent, clinically representative cohorts, together with an economic evaluation of missed and delayed diagnoses. Table 7 lists the abbreviations used in this manuscript.

Table 7 Abbreviations

Abbreviation	Definition
$A\beta$	Amyloid-beta
AD	Alzheimer's disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
APOE	Apolipoprotein E
AUPRC	Area under the precision–recall curve
AUROC	Area under the receiver operating characteristic curve
CI	Confidence interval
CSF	Cerebrospinal fluid
FOR	False-omission rate
MCI	Mild cognitive impairment
MMSE	Mini-Mental State Examination
MoCA	Montreal Cognitive Assessment
MRI	Magnetic resonance imaging
PET	Positron emission tomography
p-tau217	Tau phosphorylated at threonine 217
SD	Standard deviation
SUVR	Standardised uptake value ratio
TRIPOD+AI	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis for artificial intelligence
XGBoost	Extreme gradient boosting

Supplementary information

Supplementary Material (Online Resource 1) contains the following tables. Table S1: Missing data by variable. Table S2: Conformal coverage and prediction sets. Table S3: False-omission rate re-weighted to other prevalences. Table S4: Sensitivity analysis with a symmetric blood–PET window (±183 days).

Acknowledgements

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19 AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen

Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

Declarations

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval and consent to participate. This study is a secondary analysis of de-identified data from the Alzheimer's Disease Neuroimaging Initiative. ADNI was approved by the institutional review boards of all participating institutions, and written informed consent was obtained from all participants or their authorised representatives at enrollment. Data were accessed under the ADNI Data Use Agreement. Because the study used only de-identified data obtained under the ADNI Data Use Agreement and involved no contact with participants, no additional institutional ethics approval was required.

Consent for publication. Not applicable.

Clinical trial number. Not applicable.

Data availability. The ADNI data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative under approved data access and are subject to the initiative's data use requirements; they cannot be redistributed by the authors but are available to qualified researchers on application at https://adni.loni.usc.edu.

Code availability. The analysis code, which reads data from each user's own approved ADNI download, is available from the corresponding author on reasonable request.

Author contributions (CRediT). **T.A.**: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Data curation, Visualisation, Supervision, Writing – original draft, Writing – review & editing. **O.A.**, **Raza Nowrozy (R.N.)**, **L.A.**, **N.H.**, and **E.A.**: Writing – review & editing. All authors read and approved the final manuscript. The Alzheimer's Disease Neuroimaging Initiative is listed as a group credit as required by the ADNI Data Use Agreement; ADNI investigators did not participate in the analysis or writing of this report.

Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this work, the authors used Claude Opus 5.5 (Anthropic), a large language model assistant, to assist with code development, running the analyses on the authors' approved ADNI data, and drafting, editing, and

checking of the manuscript text. All study design decisions, analyses, and conclusions were reviewed and verified by the authors. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] GBD 2019 Dementia Forecasting Collaborators: Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. The Lancet Public Health **7**(2), 105–125 (2022) https://doi.org/10.1016/S2468-2667(21)00249-8 . PMID: 34998485

[2] Jack Jr., C.R., Bennett, D.A., Blennow, K., Carrillo, M.C., Dunn, B., Haeberlein, S.B., Holtzman, D.M., Jagust, W., Jessen, F., Karlawish, J., *et al.*: NIA-AA research framework: Toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia **14**(4), 535–562 (2018) https://doi.org/10.1016/j.jalz.2018.02.018 . PMID: 29653606

[3] Jack Jr., C.R., Andrews, J.S., Beach, T.G., *et al.*: Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's association workgroup. Alzheimer's & Dementia **20**(8), 5143–5169 (2024) https://doi.org/10.1002/alz.13859 . PMID: 38934362

[4] Dyck, C.H., Swanson, C.J., Aisen, P., *et al.*: Lecanemab in early Alzheimer's disease. New England Journal of Medicine **388**(1), 9–21 (2023) https://doi.org/10.1056/NEJMoa2212948 . PMID: 36449413

[5] Sims, J.R., Zimmer, J.A., Evans, C.D., *et al.*: Donanemab in early symptomatic Alzheimer disease: The TRAILBLAZER-ALZ 2 randomized clinical trial. JAMA **330**(6), 512–527 (2023) https://doi.org/10.1001/jama.2023.13239 . PMID: 37459141

[6] Cummings, J., Apostolova, L., Rabinovici, G.D., *et al.*: Lecanemab: Appropriate use recommendations. Journal of Prevention of Alzheimer's Disease **10**(3), 362–377 (2023) https://doi.org/10.14283/jpad.2023.30 . PMID: 37357276

[7] Teunissen, C.E., Verberk, I.M.W., Thijssen, E.H., Vermunt, L., Hansson, O., Zetterberg, H., Flier, W.M., Mielke, M.M., Del Campo, M.: Blood-based biomarkers for Alzheimer's disease: towards clinical implementation. The Lancet Neurology **21**(1), 66–77 (2022) https://doi.org/10.1016/S1474-4422(21)00361-6 . PMID: 34838239

[8] Therriault, J., Brum, W.S., Trudel, L., *et al.*: Blood phosphorylated tau for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. The Lancet Neurology **24**(9), 740–752 (2025) https://doi.org/10.1016/S1474-4422(25)00227-3 . PMID: 40818474

[9] Pahlke, S., Kahale, L.A., Mahinrad, S., *et al.*: Blood-based biomarkers for detecting Alzheimer's disease pathology in cognitively impaired individuals within specialized care settings: A systematic review and meta-analysis. Alzheimer's & Dementia **21**(11), 70828 (2025) https://doi.org/10.1002/alz.70828 . PMID: 41193403

[10] Janelidze, S., Bali, D., Ashton, N.J., *et al.*: Head-to-head comparison of 10 plasma phospho-tau assays in prodromal Alzheimer's disease. Brain **146**(4), 1592–1601 (2023) https://doi.org/10.1093/brain/awac333 . PMID: 36087307

[11] Warmenhoven, N., Salvadó, G., Janelidze, S., *et al.*: A comprehensive head-to-head comparison of key plasma phosphorylated tau 217 biomarker tests. Brain **148**(2), 416–431 (2025) https://doi.org/10.1093/brain/awae346 . PMID: 39468767

[12] Moghekar, A.R., Rock, J.A., Mohs, R.C., *et al.*: Multicenter validation of plasma p-tau217/amyloid beta 1-42 ratio in symptomatic Alzheimer's disease. Alzheimer's & Dementia **22**(7), 71644 (2026) https://doi.org/10.1002/alz.71644 . PMID: 42418465

[13] Bao, R., Shi, W., Bao, H., Zhang, T., Li, X., Ding, W.: Plasma p-tau217, p-tau181, and Aβ42 predict amyloid PET positivity in cognitively unimpaired adults. Alzheimer's Research & Therapy **18**(1), 136 (2026) https://doi.org/10.1186/s13195-026-02080-x . PMID: 42163294

[14] Liu, W.-Z., Xue, M., Sheng, Z.-H., *et al.*: Baseline plasma p-tau217/Aβ42 as a sensitive marker for the severity of Alzheimer's disease continuum. Journal of Translational Medicine **24**(1), 305 (2026) https://doi.org/10.1186/s12967-026-07939-z . PMID: 41761301

[15] Khorsand, B., Teichrow, D., Ghanbarian, E., *et al.*: Scalable markers for early cognitive decline: Plasma p-tau217, subjective cognitive concerns, and digital testing: Results from the A4/LEARN studies. Alzheimer's & Dementia **22**(5), 71505 (2026) https://doi.org/10.1002/alz.71505 . PMID: 42204876

[16] Franco, R.: Plasma p-tau217 in Alzheimer's disease is a dynamic gauge of progression heterogeneity, not a baseline stratifier. Exploration of Neuroprotective Therapy **5**, 1004118 (2025) https://doi.org/10.37349/ent.2025.1004118

[17] Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning **16**(4), 494–591 (2023) https://doi.org/10.1561/2200000101

[18] Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005). https://doi.org/10.1007/b106715

[19] Schindler, S.E., Petersen, K.K., Saef, B., *et al.*: Head-to-head comparison of leading blood tests for Alzheimer's disease pathology. Alzheimer's & Dementia **20**(11), 8074–8096 (2024) https://doi.org/10.1002/alz.14315 . PMID: 39394841

[20] Barthélemy, N.R., Salvadó, G., Schindler, S.E., *et al.*: Highly accurate blood test for Alzheimer's disease is similar or superior to clinical cerebrospinal fluid tests. Nature Medicine **30**(4), 1085–1095 (2024) https://doi.org/10.1038/s41591-024-02869-z . PMID: 38382645

[21] Ashton, N.J., Brum, W.S., Di Molfetta, G., *et al.*: Diagnostic accuracy of a plasma phosphorylated tau 217 immunoassay for Alzheimer disease pathology. JAMA Neurology **81**(3), 255–263 (2024) https://doi.org/10.1001/jamaneurol.2023.5319 . PMID: 38252443

[22] Brum, W.S., Cullen, N.C., Janelidze, S., *et al.*: A two-step workflow based on plasma p-tau217 to screen for amyloid β positivity with further confirmatory testing only in uncertain cases. Nature Aging **3**(9), 1079–1090 (2023) https://doi.org/10.1038/s43587-023-00471-5 . PMID: 37653254

[23] Figdore, D.J., Griswold, M., Bornhorst, J.A., *et al.*: Optimizing cutpoints for clinical interpretation of brain amyloid status using plasma p-tau217 immunoassays. Alzheimer's & Dementia **20**(9), 6506–6516 (2024) https://doi.org/10.1002/alz.14140 . PMID: 39030981

[24] Benina, N., Buitrago, L., De Simone, F.I., *et al.*: Plasma pTau 217/β-amyloid 1-42 ratio for enhanced accuracy and reduced uncertainty in detecting amyloid pathology. Brain **149**(4), 1168–1181 (2026) https://doi.org/10.1093/brain/awag001 . PMID: 41562409

[25] Yim, S., Park, S., Lim, K., *et al.*: Integrating MRI volume and plasma p-tau217 for amyloid risk stratification in early-stage Alzheimer disease. Neurology **105**(6), 213954 (2025) https://doi.org/10.1212/WNL.0000000000213954 . PMID: 40829110

[26] Palmqvist, S., Tideman, P., Mattsson-Carlgren, N., *et al.*: Blood biomarkers to detect Alzheimer disease in primary care and secondary care. JAMA **332**(15), 1245–1257 (2024) https://doi.org/10.1001/jama.2024.13855 . PMID: 39068545

[27] Mielke, M.M., Dage, J.L., Frank, R.D., *et al.*: Performance of plasma phosphorylated tau 181 and 217 in the community. Nature Medicine **28**(7), 1398–1405 (2022) https://doi.org/10.1038/s41591-022-01822-2 . PMID: 35618838

[28] Salvadó, G., Janelidze, S., Bali, D., *et al.*: Plasma phosphorylated tau 217 to identify preclinical Alzheimer disease. JAMA Neurology **82**(11), 1122–1134 (2025) https://doi.org/10.1001/jamaneurol.2025.3217 . PMID: 40952756

[29] Yun, J., Lee, J., Shin, D., *et al.*: Plasma phosphorylated tau 217 cutoffs for

amyloid pathology and kidney function, body mass index, and anemia. JAMA Neurology **83**(3), 269–279 (2026) https://doi.org/10.1001/jamaneurol.2025.5530 . PMID: 41627837

[30] Schindler, S.E., Galasko, D., Pereira, A.C., *et al.*: Acceptable performance of blood biomarker tests of amyloid pathology – recommendations from the Global CEO Initiative on Alzheimer's Disease. Nature Reviews Neurology **20**(7), 426–439 (2024) https://doi.org/10.1038/s41582-024-00977-5 . PMID: 38866966

[31] Palmqvist, S., Whitson, H.E., Allen, L.A., *et al.*: Alzheimer's association clinical practice guideline on the use of blood-based biomarkers in the diagnostic workup of suspected Alzheimer's disease within specialized care settings. Alzheimer's & Dementia **21**(7), 70535 (2025) https://doi.org/10.1002/alz.70535 . PMID: 40729527

[32] Vovk, V.: Conditional validity of inductive conformal predictors. In: Proceedings of the Asian Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 25, pp. 475–490 (2012)

[33] Olsson, H., Kartasalo, K., Mulliqi, N., *et al.*: Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. Nature Communications **13**(1), 7761 (2022) https://doi.org/10.1038/s41467-022-34945-8 . PMID: 36522311

[34] Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems, vol. 30, pp. 4878–4887 (2017)

[35] Angelopoulos, A.N., Bates, S., Fisch, A., Lei, L., Schuster, T.: Conformal risk control. In: International Conference on Learning Representations (ICLR) (2024)

[36] Petersen, R.C., Aisen, P.S., Beckett, L.A., *et al.*: Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology **74**(3), 201–209 (2010) https://doi.org/10.1212/WNL.0b013e3181cb3e25 . PMID: 20042704

[37] Collins, G.S., Moons, K.G.M., Dhiman, P., *et al.*: TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ **385**, 078378 (2024) https://doi.org/10.1136/bmj-2023-078378 . PMID: 38626948

[38] Royse, S.K., Minhas, D.S., Lopresti, B.J., *et al.*: Validation of amyloid PET positivity thresholds in centiloids: a multisite PET study approach. Alzheimer's Research & Therapy **13**(1), 99 (2021) https://doi.org/10.1186/s13195-021-00836-1 . PMID: 33971965

[39] Hanley, J.A., Lippman-Hand, A.: If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA **249**(13), 1743–1745 (1983). PMID: 6827763

[40] DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics **44**(3), 837–845 (1988). PMID: 3203132

[41] Van Calster, B., McLernon, D.J., Smeden, M., Wynants, L., Steyerberg, E.W.: Calibration: the Achilles heel of predictive analytics. BMC Medicine **17**(1), 230 (2019) https://doi.org/10.1186/s12916-019-1466-7 . PMID: 31842878

[42] Schindler, S.E., Karikari, T.K., Ashton, N.J., *et al.*: Effect of race on prediction of brain amyloidosis by plasma Aβ42/Aβ40, phosphorylated tau, and neurofilament light. Neurology **99**(3), 245–257 (2022) https://doi.org/10.1212/WNL.0000000000200358 . PMID: 35450967

[43] Mattke, S., Chen, J., Hanson, M., *et al.*: Estimation of the value-based price of a blood test for Alzheimer's disease pathology in primary and specialty care in the U.S. Journal of Prevention of Alzheimer's Disease **12**(7), 100219 (2025) https://doi.org/10.1016/j.tjpad.2025.100219 . PMID: 40494736