A Simulation-Based Validation Study of Incremental Predictive Value for Alzheimer's Disease Prediction Models

Hanshuo Pang [2, *], Yuxuan Pang [1]

[1] College of Basic Medicine and Forensic Medicine, Henan University of Science and Technology, Luoyang, Henan Province, China

[2] Aulin College, Northeast Forestry University, Harbin, Heilongjiang Province, China

* Corresponding Author Email: 221518010411@stu. haust. edu. cn

Abstract. Alzheimer's disease (AD) is the leading cause of dementia worldwide, and its pathological changes often occur several years before the onset of clinical symptoms. Therefore, early identification of individuals with high risk of progression has important clinical significance for subsequent timely intervention and disease management. However, the cumulative predictive value of a single biomarker and the combined use of multiple indicators under different modeling conditions and risk scenarios is still insufficiently quantified. In this study, a Monte Carlo simulation framework was constructed based on the ADEMP (Aims,Data-generating mechanism, Estimand, Methods, Performance measures) structure to systematically evaluate the incremental contribution of four plasma biomarkers (GFAP, pTau181, pTau217 and Aβ42 / 40) to AD risk prediction. We generated survival data covering a variety of scenarios, including demographic variables, cognitive assessment scores, and biomarker measurements. The Cox proportional hazards model was used in the analysis. The basic model was compared with the extended enhanced model with biomarkers, and the predictive performance was evaluated by C index, ΔC index, ΔR2, IDI, risk stratification analysis and sensitivity test. The results showed that GFAP and pTau217 were the most prominent in independent prediction, and the improvement of prediction effect was the most significant when the four markers were combined. This incremental effect remains stable under different sample sizes and censoring rates, indicating that the conclusion has strong robustness. The above findings provide strong support for integrating plasma biomarkers into traditional AD prediction models, and also build a reference methodological framework for future preclinical evaluation of the value of new biomarkers. There are some limitations in this study. It uses simulated data rather than real clinical cases, which cannot fully reflect the heterogeneity and complexity of patient data in the real world. Future research should conduct external validation of the above results, and further explore more flexible models such as adaptive ensemble estimation, in order to ultimately achieve truly personalized AD risk assessment.

Keywords: Alzheimer’s disease, Cox proportional hazards model, Monte Carlo.

1. Introduction

Alzheimer's disease (AD) accounts for 60%–70% of dementia cases [1]. Its neuropathological hallmarks—Aβ plaques and tau tangles—accumulate 10–20 years before symptoms [2]. In current practice, however, diagnosis relies on symptomatic presentation or ancillary tests such as CSF and PET, which become abnormal only after substantial pathology has accrued. Mechanistically, the field has moved to an integrated Aβ-tau model wherein Aβ aggregation primes GSK-3β signaling, promotes tau phosphorylation and microtubule destabilization, leading to neuronal injury; this cascade is modulated by kinase-phosphatase interplay and accelerated by feed-forward loops [3]. GFAP (astrogliosis) and NfL (axonal damage) are now incorporated into the AT(N) framework as neurodegeneration biomarkers [4].

Plasma biomarkers have attracted increasing attention in clinical translation due to their low invasiveness and strong scalability. A multi-index combination including Aβ42 / 40, p-tau181 / 217, GFAP, NfL and ApoEε4 has an AUC of more than 0. 90 for detecting Aβ positive [5]. A number of independent cohort studies have shown that the accuracy of plasma p-tau217 in identifying Aβpathology in cognitively normal people is about 81 %, and the positive predictive value can exceed

90 % after combined PET or cerebrospinal fluid detection [6]. The value of p-tau217 is prominent in the preclinical stage of AD, while GFAP and NfL are more focused on reflecting disease progression [6,7]. In addition, elevated levels of both can predict the risk of dementia 15 years in advance, and the AUC is between 0. 80 and 0. 89 [7].

Clinically, MMSE is still one of the most commonly used cognitive assessment tools, and Cox model is still the standard method for time-event risk prediction [2]. When evaluating new markers, we used C index, R 2 and IDI [2]. However, the introduction of multiple markers may not necessarily significantly improve the prediction effect, because its gain depends on the prevalence of the disease and the performance of the baseline model, and the robustness of the system needs to be tested. Monte Carlo simulation can compare the sensitivity under different effect sizes and related structures[3]. To this end, we generated survival data containing demographic and plasma markers under the ADEMP framework, and used the Cox model to evaluate the incremental value and stability of each marker, providing a robust methodological basis for early AD prediction.

2. Methods

2.1. Study Dsign and Sulation Famework

This study employed a Monte Carlo simulation-based design to validate the incremental value of novel biomarkers in AD prediction models. The simulation strictly adhered to the ADEMP framework [7] for Aims, Data-generating mechanisms, Estimands, Methods, and Performance measures,with 500 repetitions per scenario.

2.2. Simulated Dta Gneration

The simulated dataset included demographic variables (age drawn from a normal distribution, sex from a Bernoulli distribution), MMSE (Mini-Mental State Examination) scores, and four novel plasma biomarkers (GFAP, pTau181, pTau217, and Aβ42/40). We set the correlation between biomarkers by preset covariance matrix. The event time was defined as the interval from MCI diagnosis to AD transformation, and the censoring rates were set to 30 %, 50 % and 70 % to simulate different scenarios. The event time was generated based on the Cox proportional hazards model and the Weibull model, respectively. The distribution parameters (mean, variance, and hazard ratio) were calibrated based on empirical estimates from published studies [8]. All simulations and statistical analyses were performed in R with packages including simstudy, MASS, survival, risk Regression and ggplot2.

2.3. Prediction Mdel Cnstruction

In this study, two nested Cox proportional hazard models were established [8]: the baseline model included age, gender and MMSE score; four plasma biomarkers (GFAP, pTau181, pTau217 and Aβ42 / 40) were added to the expanded model. After the model was fitted, the linear predictor (LP) of each individual was calculated for subsequent risk stratification.

2.4. Incremental Pediction Mtrics

This study used the following metrics to evaluate incremental predictive value [9]: the C-index for discrimination performance; ΔC-index for net improvement in discrimination; ΔR^2 for the increase in explained variation; the integrated discrimination improvement (IDI) for risk reclassification; and the Royston D statistic as a supplementary measure. All metrics were computed for each Monte Carlo repetition.

2.5. Visualization and Risk Stratification

This study used Kaplan–Meier curves with log rank tests to compare survival across risk groups defined by the median and tertiles of the linear predictor [10]. Scatter plots of biomarker Z scores against the baseline LP were generated with LOESS smoothing [11]. This study also stratified the

population by baseline LP quintiles and calculated, within each stratum, the difference in mean biomarker values between progressors and stable individuals—termed the separation—to visualize how the incremental effect varies across baseline risk levels.

2.6. Sensitivity Analysis

This study evaluated the effects of sample size (250,500,1000), GFAP risk ratio (1. 42,1. 89,2. 36) and censoring rate (30 %, 50 %, 70 %) on ΔC index [12]. Repeated simulations were performed for each parameter combination, and the results were visualized through a heat map to evaluate their robustness and detect possible saturation effects.

3. Results

3.1. Distribution of ΔC Index and Monte Carlo Stability

Figure 1 displays the kernel density distributions of ΔC index across the five models from 500 Monte Carlo repetitions. The ΔC index of the complete model shifted significantly to the right, concentrated between 0. 14-0. 20 (mean about 0. 17). GFAP and pTau217 were concentrated in the vicinity of 0. 06-0. 10 (mean about 0. 08), pTau181 was in the middle (about 0. 04-0. 08), and Aβ42 / 40 was the lowest (about 0. 02-0. 04). The density curve of the complete model is completely separated from any single model, which intuitively reflects the synergistic effect of the combination of the four markers. All the distributions are unimodal symmetry without abnormal dispersion, suggesting that the simulation design is reasonable and the results are robust

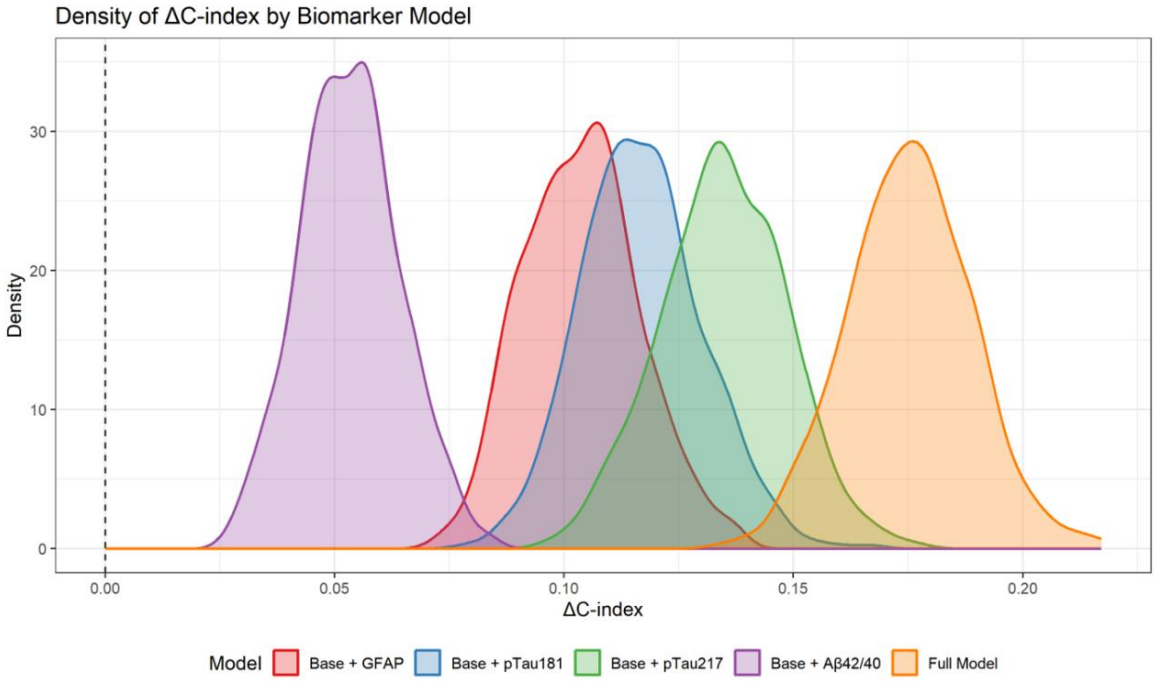


Fig. 1 The kernel density map of the ΔC index of each model under 500 Monte Carlo repetitions. (Photo/Picture credit: Original)

The horizontal axis is the ΔC index and the vertical axis is the density. The complete model curve shifted significantly to the right, concentrated in 0. 14-0. 20 ; GFAP and pTau217 were concentrated near 0. 06-0. 10, pTau181 was in the middle, and Aβ42 / 40 was the lowest. The complete model curve almost does not overlap with any single marker curve.

3.2. Sensitivity Analysis

When the risk ratio of GFAP was fixed at 1. 89, the ΔC index of the complete model was stable between 0. 173 and 0. 178 under all nine combinations of sample size and censoring rate (Figure 2),

with a range of no more than 0. 005. This stability is maintained even under the conditions of small sample (n = 250) and high censoring rate (70 %), suggesting that the incremental effect is highly robust.

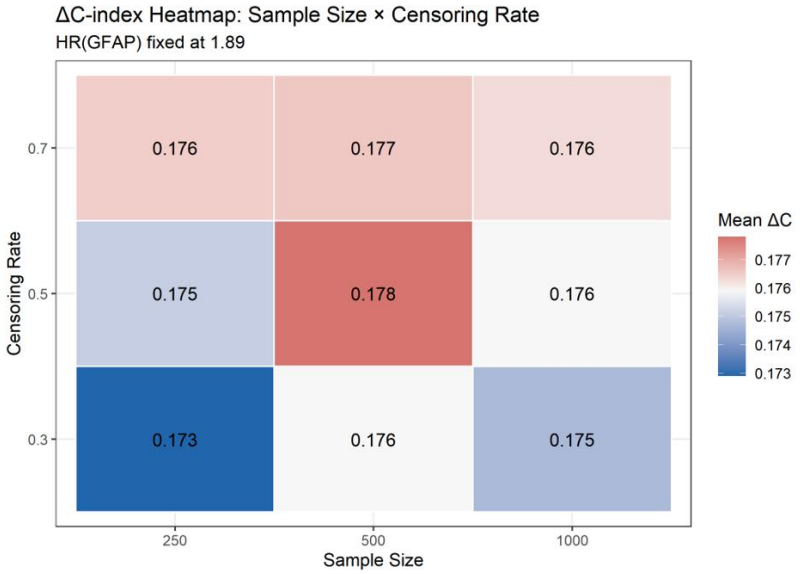


Fig. 2 Sensitivity heat map. (Photo/Picture credit: Original)

The change of ΔC index of the complete model under different sample sizes and censoring rates (GFAP risk ratio = 1. 89). The unit cell value is the average ΔC index under each scenario, and the deeper the color scale, the higher the index. The ΔC index under the nine combinations was stable at 0. 173-0. 178.

3.3. Incremental Performance across Different Baseline Risk Strata

After stratifying the population according to the baseline LP quintile (Figure 3, Table 1), GFAP and pTau217 showed positive separation in each group. The separation degree reached the peak in Q3 group (GFAP 0. 94, pTau217 0. 95), and the lowest in Q2 group (0. 70 and 0. 75, respectively). Even in the Q1 group with the lowest risk, the degree of separation was still close to a standard deviation. In the highest risk group (Q5), the resolution of pTau217 remained 0. 92, higher than that of GFAP (0. 75).

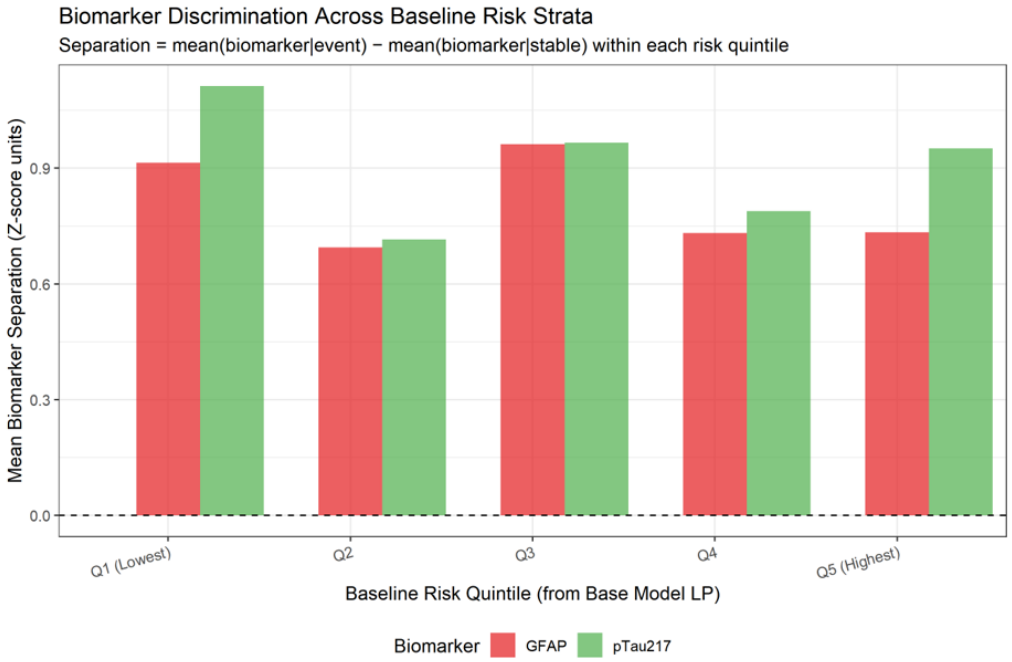


Fig. 3 Trend plot of separation for GFAP and pTau217 across baseline risk quintiles (Q1 to Q5). (Photo/Picture credit: Original)

Separation is defined as the difference in standardized mean biomarker values between progressors and stable individuals.

Table 1. Separation of GFAP and pTau217 by baseline risk quintile.

Baseline risk quintile	GFAP separation	pTau217 separation
Q1 (lowest risk)	0. 91	0. 95
Q2	0. 70	0. 75
Q3	0. 94	0. 95
Q4	0. 75	0. 82
Q5 (highest risk)	0. 75	0. 92

Note: Separation is defined as the mean biomarker Z score in the progressor group minus that in the stable group. Positive values indicate higher biomarker levels in progressors.

3.4. Kaplan–Meier Survival Curves

The Kaplan–Meier curves for the high and low risk groups defined by the full model LP median separated early during follow up (Figure 4a, log rank χ^2 = 316. 14, P < 0. 001). Stratification by the baseline model was considerably weaker (Figure 4b, χ^2 = 41. 34). Tertile based stratification of the full model LP revealed a clear risk gradient (Figure 4c, χ^2 = 401. 18). Notably, GFAP tertiles alone also achieved substantial risk discrimination (Figure 4d, χ^2 = 203. 14), indicating that GFAP is a strong independent prognostic marker. (a) High versus low risk groups defined by the median LP of the full model (log rank χ^2 = 316. 14, P < 0. 001). (b) High versus low risk groups defined by the median LP of the baseline model (age + sex + MMSE; log rank χ^2 = 41. 34, P < 0. 001). (c) Low, intermediate, and high risk groups defined by the tertiles of the full model LP (log rank χ^2 = 401. 18, P < 0. 001). (d) Risk groups defined by GFAP tertiles alone (log rank χ^2 = 203. 14, P < 0. 001). Shaded areas indicate 95% pointwise confidence bands for each survival curve.

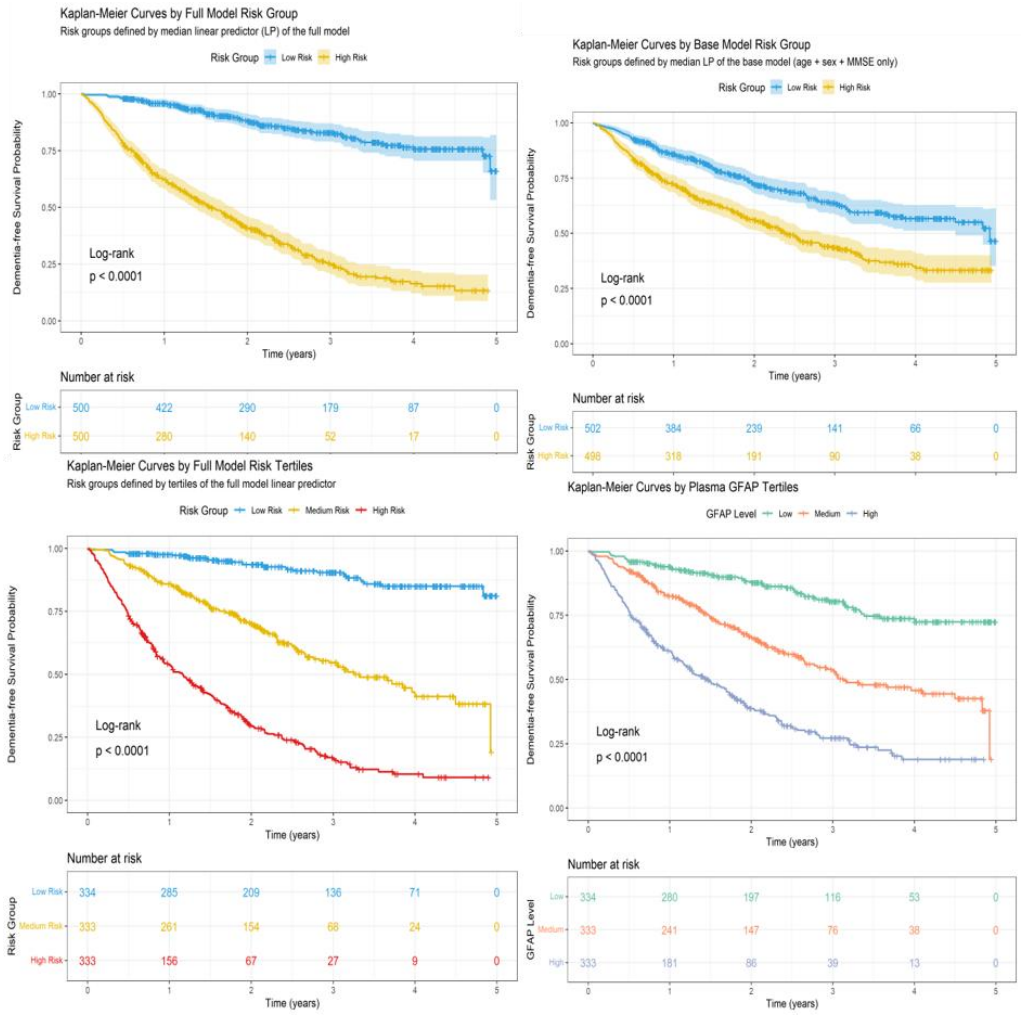


Fig. 4 Kaplan–Meier survival curves. (Photo/Picture credit: Original)

3.5. Individual Level Association between Biomarkers and Baseline Risk

The scatter plots in Figure 5 show that GFAP, pTau181, and pTau217 each increased with the baseline LP (pTau217 showing the steepest slope), whereas Aβ42/40 trended downward, consistent with its role as a protective factor. At any given level of baseline LP, progressors (red) had systematically higher biomarker values than stable individuals (green), and this separation was most pronounced in the high LP region. Each point represents an individual; red points indicate progressors to dementia, green points indicate stable individuals. The black curve is a LOESS smooth fit. The horizontal axis shows the baseline LP; the vertical axis shows the Z score for each biomarker. (a) GFAP; (b) pTau181; (c) pTau217; (d) Aβ42/40.

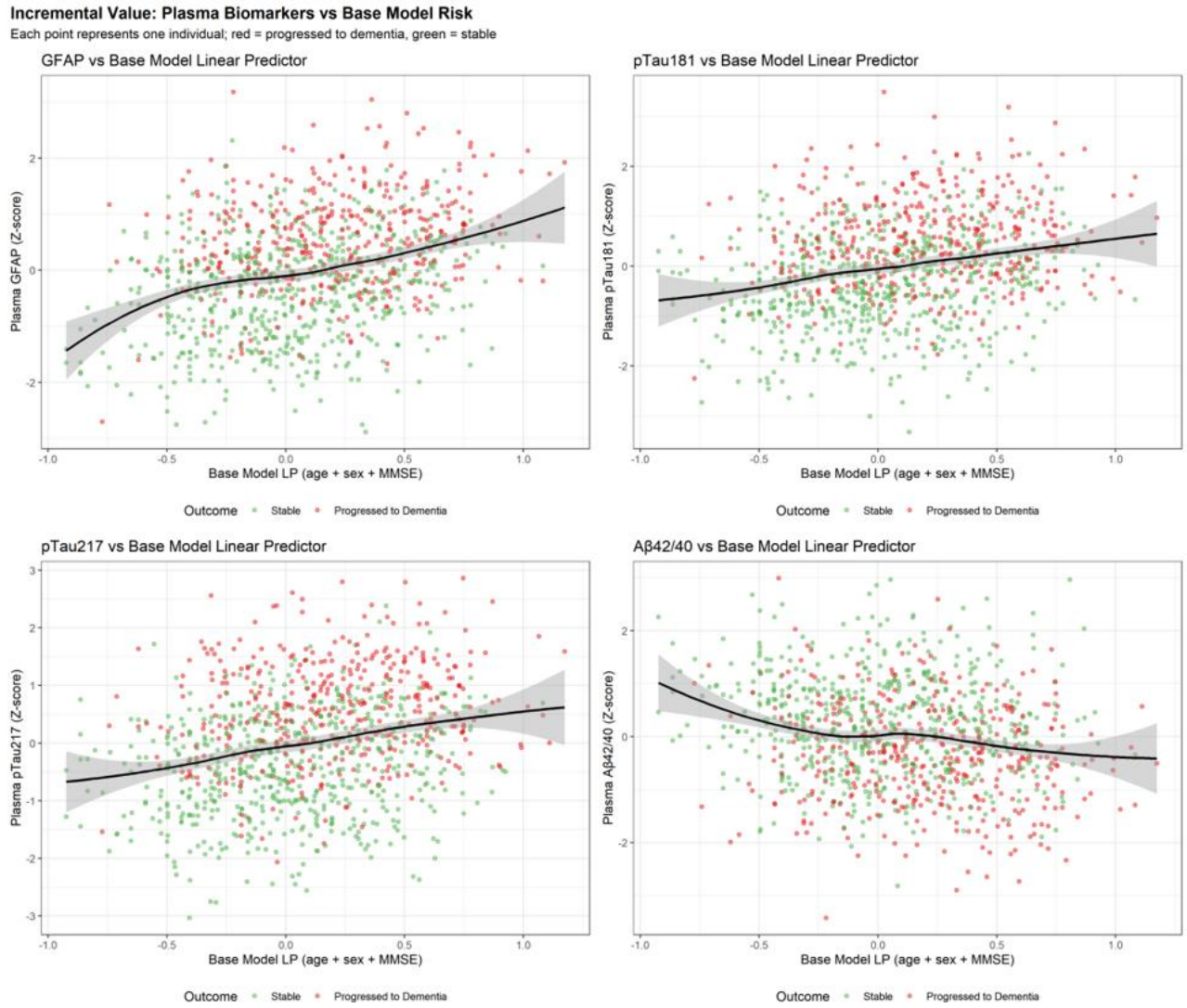


Fig. 5 Scatter plots of the four plasma biomarkers against the baseline linear predictor (LP_base). (Photo/Picture credit: Original)

4. Discussion

The multi-label extension model was significantly better than the baseline in terms of incremental indicators, and GFAP and pTau217 contributed most of the increments, which was consistent with the existing research [6,7]. The logarithmic rank $\chi 2$ of the complete model increased from 41. 34 to 316. 14, an increase of nearly seven times ; only GFAP can achieve 64 % discriminant performance of the complete model, which is consistent with the C index (0. 69) of GFAP in real queues such as Honey [3].

When the sample size increased from 250 to 1000, the ΔC index was stable at 0. 175-0. 180, suggesting that small samples could still be reliably detected. When the risk ratio of GFAP increased

from 1. 42 to 1. 89, the ΔC index increased by about 75 %. However, when further increased to 2. 36, the increase was only 0. 005, showing a saturation effect, which was consistent with the results of Palmqvist et al. (ΔAUC increase ≤ 0. 01) [5]. The C index of the complete model is about 0. 795, which is higher than that of Honey [3] (0. 71). The difference may be due to the fact that the simulated data does not contain measurement errors and non-AD pathology, so it is the upper limit of optimism.

The results showed that in the early AD risk stratification, GFAP and pTau217 were not only additive factors, but also provided predictive information as complementary information carriers. The saturation effect also shows that after the marker effect reaches moderate intensity, the incremental improvement of continuing to optimize a single marker is limited. This study provides a quantitative benchmark for the incremental value of AD prediction model, and has reference significance for model development and screening strategy formulation.

5. Conclusion

Based on Monte Carlo simulation, this study evaluated the incremental value of plasma biomarkers in AD risk prediction. The baseline model (age, gender, MMSE) and the extended model (adding GFAP, pTau181, pTau217, Aβ42 / 40) were compared, and the performance improvement was analyzed by multiple incremental indicators. The results showed that GFAP and pTau217 contributed the most independently, and the improvement was the most significant when the four were combined. The effect was stable under different sample sizes and censoring rates, and the framework was robust.

The pathology of AD is several years earlier than the symptoms. Early identification of high-risk groups is crucial for timely intervention and stratification. Blood markers (especially GFAP and pTau217) can enhance the traditional model and improve risk reclassification, and the multi-index synergistic effect supports the application of the combined panel in accurate screening. Simulation evaluation of new markers is of great value before large-scale clinical implementation.

However, the simulation is based on preset assumptions, which makes it difficult to fully reflect the complexity of real measurement and disease heterogeneity, and no external clinical verification is carried out. Subsequently, it is necessary to verify in the longitudinal queue and explore the nonlinear survival model and dynamic modeling. Integrating multi-omics and imaging markers is expected to improve the accuracy of prediction and promote the development of personalized risk assessment tools.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Bun S, Ito D, Tezuka T, et al. Performance of plasma Aβ42/40, measured using a fully automated immunoassay, across a broad patient population in identifying amyloid status. Alzheimer‘s Research & Therapy. 2023;15(1):149. doi:10. 1186/s13195-023-01296-5

[2] Austin PC, Pencinca MJ, Steyerberg EW. Predictive accuracy of novel risk factors and markers: A simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model. Statistical Methods in Medical Research. 2015;26(3):1053-1077. doi:10. 1177/0962280214567141

[3] Honey MIJ, Van Maurik IS, Van Harten AC, et al. Individualized prediction of clinical progression to dementia using plasma biomarkers in non-demented elderly. Alzheimer‘s Research & Therapy. 2025;18(1):4. doi:10. 1186/s13195-025-01925-1

[4] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019;38(11):2074-2102. doi:10. 1002/sim. 8086

[5] Palmqvist S, Stomrud E, Cullen N, et al. An accurate fully automated panel of plasma biomarkers for Alzheimer’s disease. Alzheimer‘s & Dementia. 2022;19(4):1204-1215. doi:10. 1002/alz. 12751

[6] Salvadó G, Janelidze S, Bali D, et al. Plasma phosphorylated tau 217 to identify preclinical Alzheimer disease. JAMA Neurology. 2025;82(11):1122. doi:10. 1001/jamaneurol. 2025. 3217

[7] Wang X, Shi Z, Qiu Y, Sun D, Zhou H. Peripheral GFAP and NfL as early biomarkers for dementia: longitudinal insights from the UK Biobank. BMC Medicine. 2024;22(1):192. doi:10. 1186/s12916-024-03418-8

[8] Sanusi B, Cai J, Hudgens MG. Nonparametric estimation of marked survival data in the presence of dependent censoring. Statistics in Medicine. 2023;42(12):1995-2008. doi:10. 1002/sim. 9710

[9] An S, Eo W, Lee S. Prognostic significance of modified Shine and Lal index in patients with non-small cell lung cancer undergoing surgical resection. Biomedicines. 2025;13(4):937. doi:10. 3390/biomedicines13040937

[10] Andrade C. Survival analysis, Kaplan-Meier curves, and Cox regression: basic concepts. Indian Journal of Psychological Medicine. 2023;45(4):434-435. doi:10. 1177/02537176231176986

[11] Shi Y, Zhang L, Do KA, Jenq RR, Peterson CB. survivalContour: visualizing predicted survival via colored contour plots. Bioinformatics Advances. 2024;4(1):vbae105. doi:10. 1093/bioadv/vbae105

[12] Muralidar S, Ambi SV, Sekaran S, Thirumalai D, Palaniappan B. Role of tau protein in Alzheimer‘s disease: the prime pathological player. International Journal of Biological Macromolecules. 2020;163:1599-1617. doi:10. 1016/j. ijbiomac. 2020. 07. 327