Longitudinal Dense-to-Sparse Forecasting: Individual Variability Predicts Conversion Better than Mean Change

Georgi Hrusanov [1 2] **Ivan Stoyanov** [3] **Duy-Thanh Vu** [1 2] **Giovanni d'Ario** [4] **Carsten Magnus** [4 *] **Oliver Y. Chén** [1 2 *]

Abstract

Clinical forecasters trained on dense imaging trajectories are often repurposed to analyse sparse downstream events. What survives this hand-off under cohort shift is rarely investigated. Here, using a cortical-thickness forecaster trained on one Alzheimer's cohort, we compare two label-free risk signals: predicted mean structural change, and predicted. Our results suggest that in-domain variability ranks eventual converters about as well as a fully-supervised clinical baseline. Under transfer, variability remains predictive while mean change degrades. The split is not the usual aleatoric-vs-epistemic: variability from input-residual structure transfers; variability from posterior weight perturbations does not. A controlled covariate-shift experiment reproduces the pattern. Taken together, our work suggests that forecasting error alone does not predict which uncertainty channel transfers.

1. Introduction

Models trained at one clinical site are routinely deployed at others. Scanner, demographic, and visit-schedule shifts can silently degrade risk predictions. For longitudinal neurodegeneration, where conversion events are sparse, expensive to label, and protocol-dependent, it is unclear which forecaster signals survive deployment. We study this through a heteroscedastic longitudinal forecaster trained on the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005) and transferred without retraining to the Australian Imaging, Biomarkers and Lifestyle study (AIBL) (Ellis et al., 2009) (primary external) and the Open Access Series of Imaging Studies release 3 (OASIS-3) (LaMontagne et al., 2019) (underpowered second probe). The forecaster is trained on dense imaging trajectories without conversion supervision ("label-free"). From it we derive two candidate risk signals: a mean structural-change score s_μ and an aleatoric s_σ from a Gaussian residual head, with four sampling-based epistemic approximations, namely Monte Carlo (MC) dropout, deep ensembles, last-layer Laplace (Daxberger et al., 2021), and a spectral-normalised Gaussian process (SNGP) (Liu et al., 2020), and an evidential normal-inverse-gamma (NIG) head (Amini et al., 2020) as controls. Our central finding is a sharp dissociation: the mean forecast, strong in-domain, degrades substantially under cohort shift, while s_σ remains a reliable risk-stratification signal on AIBL and matches a fully-supervised gradient-boosted clinical model. **Contributions.** (1) Predicted individual variability s_σ is able to rank eventual converters under cohort shift while predicted mean change s_μ is not: on AIBL, the dissociation is unambiguous at 54m and directionally consistent at 24m; OASIS-3 replicates the forecasting transfer, although it is underpowered for risk stratification. (2) Within variability channels, the load-bearing distinction is not aleatoric-vs-epistemic: s_σ derived from residual-magnitude likelihoods (heteroscedastic Gaussian, evidential NIG) transfers, while s_σ from posterior weight perturbations (MC dropout, deep ensembles, last-layer Laplace, SNGP) does not. (3) Two heads with near-identical mean absolute error (MAE) produce s_σ scores whose conversion-ranking AUC differs sharply from that of the mean score s_μ; MAE-based model selection is therefore insufficient for transfer. (4) A controlled covariate-shift simulation corroborates the mechanism.

2. Related Work

Longitudinal Alzheimer's progression has been modelled with event-based, probabilistic spatiotemporal, and direct conversion classifiers (Fonteijn et al., 2012; Lorenzi et al., 2019; Cuingnet et al., 2011) that learn from the clinical event itself. Here we instead train a forecaster on the dense imaging trajectory without conversion labels, then

[*]Equal contribution [1]University of Lausanne (UNIL), Lausanne, Switzerland [2]Lausanne University Hospital (CHUV), Lausanne, Switzerland [3]Technical University Munich, Munich, Germany [4]F. Hoffmann-La Roche Ltd, Roche Information Solutions, Basel, Switzerland. Correspondence to: Georgi Hrusanov <georgi.hrusanov@chuv.ch>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

ask whether its outputs can rank subjects by eventual conversion a dense-to-sparse transfer in the spirit of (Raghu et al., 2019; Li et al., 2025). Our backbone builds on graph neural operators (Li et al., 2020; Raissi et al., 2019) and network-propagation models of neurodegeneration (Raj et al., 2012; Iturria-Medina et al., 2014). We adopt the aleatoric/epistemic distinction of Kendall & Gal (2017). A separate line of work calibrates time-series prediction intervals (Lin et al., 2022; Auer et al., 2023; Cini et al., 2025), but it targets forecasting coverage rather than ranking downstream events, which is our focus.

3. Method and Setup

Forecaster. We represent the cortex as a graph over 200 regions using the Schaefer atlas (Schaefer et al., 2018), with k-nearest-neighbour edges (k=10) on parcel centroids and a graph-attention layer that learns a sparse latent adjacency. For subject i with thickness vectors $x_{i,t} \in \mathbb{R}^{200}$ and baseline covariates c_i (diagnosis, age, sex, apolipoprotein E $\varepsilon 4$ allele (APOE4), global thickness), $(\hat{x}_{i,t+\tau}, \sigma_i) = f_\theta(x_{i,t-h}, x_{i,t}, c_i, h, \tau)$ predicts a future structural state and per-region predictive uncertainty from a heteroscedastic Gaussian head. We evaluate five residual heads (no residual, multilayer perceptron (MLP), Gaussian, flow, diffusion) on the shared backbone (Appendix B); deep ensembles, MC dropout, last-layer Laplace (Daxberger et al., 2021), SNGP (Liu et al., 2020), and an evidential NIG head (Amini et al., 2020) serve as uncertainty-quantification (UQ) controls on the same split.

Risk scores. From the predicted future state we define $s_{\mu_i} = \sqrt{(1/N)\|\hat{x}_{i,t+\tau} - x_{i,t}\|_2^2}$ and $s_{\sigma i} = (1/N)\sum_{j=1}^{N} \sigma_{ij}$. We evaluate three downstream uses: *raw* scores (zero labels), *linear* adapters (logistic regression on $[s_\mu, s_\sigma]$), and *quadratic* adapters (degree-two polynomial features). For external transfer, the forecaster and any adapters are trained on ADNI data without target-cohort tuning.

Cohorts and protocol. We consider ADNI data as the main source (204 subjects, 40 converters at 24m). AIBL is the primary external cohort (139 subjects; 14, 15, 29 converters at 24/36/54m) and OASIS-3 data is a community-based second probe (183 subjects; 7–14 converters at 36/54/72m). The forecaster is trained on 12-month thickness prediction; per-subject s_μ and s_σ are fixed features for downstream ranking. We report ROC AUC with 2,000-resample bootstrap confidence intervals (CIs) and mean AUC across repeated subject-level splits. Baselines: diagnosis-only logistic regression, full-clinical-covariate logistic regression and gradient boosting, and an end-to-end supervised imaging classifier.

4. Results

4.1. Similar forecast error, different downstream utility

Table 1 shows the first core result. The load-bearing comparison is s_σ vs. s_μ, not MLP-mean vs. Gaussian-mean: the latter are tied on 12m MAE (0.0818 vs. 0.0819) with heavily overlapping AUC intervals. The raw s_σ achieves mean AUC 0.775 [0.762, 0.789], higher than either mean-derived score, and wins on 8 of 10 splits (p<0.0001). Posterior-predictive controls (lower block) confirm the dissociation: deep ensemble (0.712), MC dropout (0.573), last-layer Laplace (Daxberger et al., 2021) (0.602), and SNGP (Liu et al., 2020) (0.608) all sit substantially below the aleatoric Gaussian. The evidential NIG head (Amini et al., 2020) sharpens the diagnosis: *both* its "aleatoric" ($\sigma_a^2 = \beta/(\alpha-1)$) *and* "epistemic" ($\sigma_e^2 = \beta/(\nu(\alpha-1))$) channels reach 0.766/0.765, near-Gaussian, because the NIG epistemic head is algebraically a re-scaling of the same residual likelihood, not a Monte-Carlo summary over a posterior on weights. The load-bearing axis is therefore not the aleatoric/epistemic label but whether σ is estimated from residual magnitudes (transfers) or posterior weight perturbations (does not trasnfer).

4.2. In-domain utility and label efficiency

The aleatoric s_σ is informative on ADNI data: AUC 0.777 [0.693, 0.856] overall and achieves 0.761 in mild cognitive impairment (MCI). Both raw forecast channels require *zero downstream labels* yet outperform diagnosis-alone logistic regression (0.687); a fully-supervised gradient-boosted clinical model reaches only 0.792 – within 0.015 of raw s_σ. End-to-end supervised classifier never exceeds 0.708 at any label budget.

4.3. External transfer under cohort shift

Table 2 provides the central transfer evidence. The dissociation is unambiguous at 54m (n=29 converters): s_σ 0.769 [0.656, 0.870] vs. mean 0.403 [0.287, 0.523] – non-overlapping CIs. At 24m (n=14 converters), s_σ achieves a higher AUC than the mean score (0.695 vs. 0.534), but their separate 95% CIs overlap, so the gap is not conclusive on its own motivating the paired bootstrap below. A within-method paired bootstrap on AIBL 36m (n=126, 15 converters) controls for small-n noise; we report 36m because it preserves the train-target setup while raising converter count over 24m. To separate s_μ- and s_σ-strength we report two readings (Appendix I): the absolute transfer $\mathrm{AUC}(\hat{\sigma})$ separates the residual-likelihood cluster (Gaussian, NIG aleatoric, NIG epistemic all 0.65–0.67) from the sampling-based cluster (ensemble, Laplace, MC dropout, SNGP all 0.55–0.60); the within-method $\Delta_{\sigma-\mu} := \mathrm{AUC}(\hat{\sigma}) - \mathrm{AUC}(s_\mu)$ is positive only for Gaus-

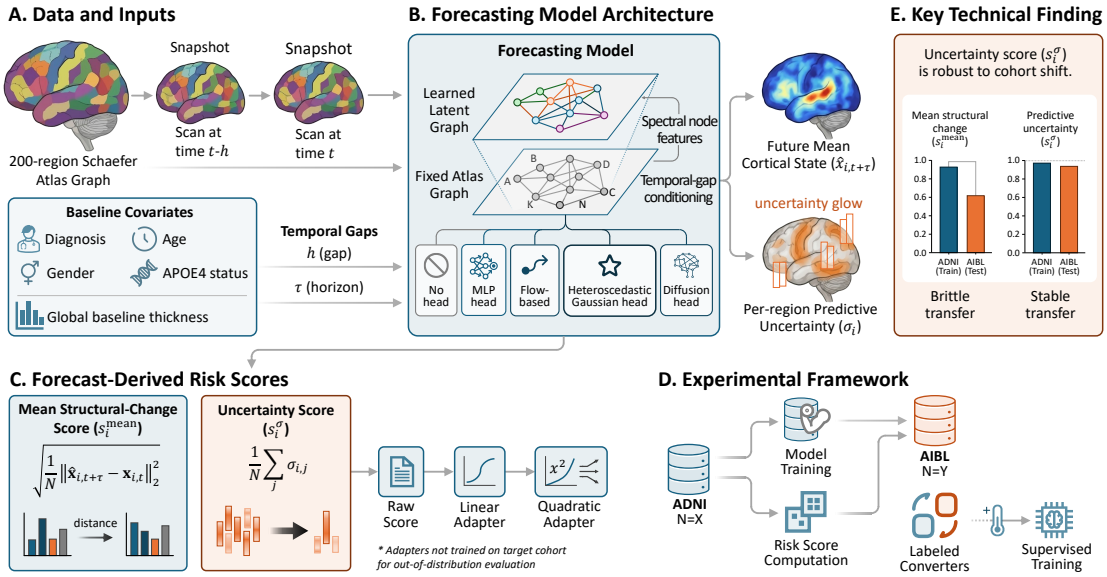

Figure 1. **Method overview. (A)** Inputs: two cortical-thickness snapshots and baseline covariates on a 200-region Schaefer atlas. **(B)** A physics-inspired graph forecaster with five interchangeable heads emits a predicted future state and per-region uncertainty σ_i. **(C)** Two risk scores, mean structural change s_μ and uncertainty s_σ, are optionally calibrated via adapters not trained on the target cohort. **(D)** Model trained on ADNI data and transferred to AIBL data; the same forecaster is additionally probed on OASIS-3 data. **(E)** s_σ transfers stably; s_μ degrades under cohort shift.

sian ($+0.187$) and negative elsewhere, dominated by per-method s_μ strength. A small ADNI-trained quadratic adapter (no AIBL labels) reaches 0.770 on AIBL, exceeding the AIBL-supervised clinical-covariates model. The second probe, on OASIS-3 data, replicates the forecasting transfer (MAE $0.0867 \rightarrow 0.2384$); risk AUCs are directionally consistent at 36m (s_σ 0.726 vs. 0.606) but converge at longer horizons, so results on OASIS-3 data are reported as exploratory (Appendix F). Across the broader UQ panel the same transfer pattern holds on AIBL data at 36m (Appendix H): sampling-based epistemic channels (ensemble, MC dropout, Laplace, SNGP) sit at AUC $[0.55, 0.60]$, while NIG aleatoric and epistemic transfer at 0.65/0.67 – consistent with both being functions of the same residual likelihood.

4.4. Calibration decomposes from ranking

Ranking AUC and probabilistic calibration are distinct properties of $\hat{\sigma}$; the load-bearing-axis claim should hold on both. We calibrate a 90% split-conformal threshold on the validation ADNI data and apply it unchanged to AIBL data at 36m, alongside variance expected calibration error (ECE; the calibration error of $\hat{\sigma}^2$ against squared residual). Residual-likelihood channels are well calibrated *and* discriminative (Gaussian: ECE 0.006, conformal coverage 0.881, s_σ-AUC 0.660; NIG aleatoric 0.067 / 0.754 / 0.650±0.021; NIG epistemic 0.016 / 0.816 / 0.667±0.006). Within this cluster, NIG aleatoric is the weakest on ECE (≈0.07 vs. ≈0.01), although still better than other sampling-based method (0.09–0.11); its s_σ-AUC is unaffected. The four sampling-based epistemic channels combine variance ECE 0.09–0.11 with systematic over-coverage at 0.96–1.00 (Laplace and SNGP at exactly 1.00, ensemble and MC dropout at 0.96) and s_σ-AUC in $[0.55, 0.60]$. Coverage at 1.00 means the conformal threshold absorbs essentially every residual – $\hat{\sigma}$ is inflated and non-discriminative, not merely noisier. The dissociation is therefore visible as a calibration failure as well as a ranking failure, and both align with the residual-likelihood vs. posterior-weight axis.

4.5. Mechanism: why residual-likelihood transfers

Residual-magnitude likelihoods estimate $\sigma(x)$ from the empirical relationship between inputs and residuals at training time; under covariate shift, the same residual-difficulty mapping is recomputed at each *target* input, so $\hat{\sigma}_i$ remains tied to the local geometry of the input distribution rather than to where ADNI training subjects happened to lie. Sampling-based epistemic estimators do the opposite: $\hat{\sigma}_i^{\text{epi}} = \text{Var}_k(\hat{\mu}_i^{(k)})$ is large where ADNI training data are sparse in weight-space and small where they are dense, but those gaps are properties of *source* coverage, with no systematic alignment to target-cohort risk. A controlled covariate-shift simulation (Appendix C, Figure 2) reproduces the dissociation: the mean-score AUC falls from 0.829 to 0.501 under increasing input shift while s_σ-only AUC stays near 0.747, and Appendix C states an idealised proposition (independence assumption explicitly violated in ADNI→AIBL) that captures the qualitative shape.

Table 1. Residual heads (upper) and epistemic/evidential controls (lower) on ADNI. Repeated AUC: mean [95% CI] across 30 runs (10×3 seeds); *5-seed mean $\pm$ std for the epistemic / evidential s_σ block; single-split AUC otherwise. Aleatoric s_σ is label-free and dominates all mean-derived scores. The three sampling-based epistemic channels (ensemble, MC-dropout, Laplace) and SNGP are substantially weaker; the evidential NIG head closes the gap, with both its aleatoric and epistemic channels matching Gaussian aleatoric in-domain. Flow and diffusion heads (Appendix B) are dominated on both MAE and AUC.

Score	Val. MAE ↓	AUC all	AUC MCI	Repeated AUC
MLP (mean)	0.0818	0.755	0.702	0.765 [0.749, 0.779]
Gaussian (mean)	0.0819	0.773	0.710	0.750 [0.736, 0.763]
Gaussian (s_σ, aleatoric)	–	**0.777**	**0.761**	**0.775 [0.762, 0.789]**
Deep ensemble (mean)	0.0875	0.772	0.741	0.772 [0.730, 0.808]
Deep ensemble (s_σ, epistemic)	–	0.647	0.637	$0.712 \pm 0.039^*$
MC dropout (mean)	0.0877	0.772	0.744	0.772 [0.736, 0.803]
MC dropout (s_σ, epistemic)	–	0.574	0.566	$0.573 \pm 0.043^*$
Last-layer Laplace (s_σ, epistemic)	–	–	–	$0.602 \pm 0.019^*$
SNGP (Liu et al., 2020) (s_σ, epistemic)	–	–	–	$0.608 \pm 0.008^*$
Evidential NIG (Amini et al., 2020) (s_σ, aleatoric)	–	–	–	$0.766 \pm 0.007^*$
Evidential NIG (s_σ, epistemic)	–	–	–	$0.765 \pm 0.007^*$

Table 2. Transfer learning between ADNI and AIBL data. 95% bootstrap CIs; forecast-derived scores are label-free. The mean score collapses on the AIBL data while s_σ stays predictive, and a small ADNI-trained adapter recovers strong external performance without using AIBL labels. ADNI data fixed at 24m. † Diagnosis is constant within MCI, so AUC = 0.500 by construction. Full multi-horizon and linear-adapter results are in Appendix G.

Model	ADNI all [95% CI]	ADNI MCI	AIBL all [95% CI]
Zero-label baselines, 24m (ADNI: 204 subj. / 40 conv.; AIBL: 139 subj. / 14 conv.)			
Score only	0.773 [0.689, 0.849]	0.710	0.534 [0.373, 0.693]
s_σ only	0.777 [0.693, 0.856]	0.761	0.695 [0.530, 0.844]
Multi-horizon AIBL transfer (ADNI evaluated at 24m only)			
Score only (54m, n=29 conv.)	–	–	0.403 [0.287, 0.523]
s_σ only (54m, n=29 conv.)	–	–	0.769 [0.656, 0.870]
Supervised baselines			
Diagnosis only†	0.687 [0.616, 0.748]	0.500†	0.622 [0.495, 0.760]
Clinical covariates	0.791 [0.704, 0.867]	0.753	0.701 [0.538, 0.851]
Gradient boosting (clin.)	0.792 [0.710, 0.866]	0.740	–
Adapted forecast scores (ADNI-trained adapter, no AIBL labels)			
Score $+s_\sigma$ (quadratic)	0.789 [0.706, 0.864]	0.755	0.770 [0.655, 0.877]

4.6. Variability as compressed trajectory atypicality

One reading of s_σ is that it measures *trajectory atypicality* meaning how unstable or unusual a subject's structural trajectory is relative to the cohort. This reading makes a testable prediction: s_σ should track each subject's directly observed trajectory variability.Predicted s_σ correlates with the observed within-subject trajectory standard deviation at ρ=0.287 (p<0.0001) yet *outperforms* that direct measurement on conversion (AUC 0.777 vs. 0.656): the forecaster has learned a compressed, portable representation of prediction difficulty. Because ρ=0.287 explains <10% of the variance in s_σ, raw trajectory variability is only a partial account – what additional structure s_σ encodes remains open. The top-s_σ quartile converts at over an order of magnitude above the bottom (median odds ratio (OR) 8.8, range 5.7–10.3 across 5 seeds), with quartile thresholds derived from predicted uncertainty alone – no conversion labels are used. Per-region σ_i correlates positively with conversion across all 200 cortical parcels controlled by (Benjamini–Hochberg false discovery rate (BH-FDR) at q=0.05; ρ=0.26–0.42), peaking in default-mode and temporal-parietal cortex (Palmqvist et al., 2017; Jack et al., 2010) (Appendix E).

5. Discussion and Conclusion

MAE-based model selection is unreliable for transfer. Two heads with indistinguishable forecasting error produced risk AUCs with no consistent winner across splits and because MAE scores only the mean forecast, it carries no information about the uncertainty channel that actually drives the downstream signal. A benchmark setup that selects only on prediction error will silently mis-rank methods on the downstream signal that actually transfers. *Label efficiency is the main operational benefit.* Raw s_σ requires zero downstream labels yet matches fully-supervised clinical models, including under cohort shift to AIBL; even

a tiny ADNI-trained quadratic adapter on $[s_\mu, s_\sigma]$ exceeds an AIBL-supervised clinical-covariates model. *The load-bearing axis is residual-likelihood vs. posterior-weight, not aleatoric vs. epistemic.* The evidential NIG head, where both the "aleatoric" and "epistemic" channels are algebraic functions of the same residual likelihood, behave like the heteroscedastic Gaussian on every metric we measure (in-domain ranking, transfer ranking, variance ECE, conformal coverage), separating from the four sampling-based epistemic estimators that summarise variance across a posterior on weights. The aleatoric/epistemic taxonomy by itself does not predict transfer; what predicts it is the quantity σ is computed from the residual-magnitude likelihood (transfers) or variance across a weight posterior (does not).. Finally, we want to emphasise that our evidence is limited to one disease and one imaging modality; broader validation, prospective evaluation, and fairness analysis across demographic groups remain open before any clinical deployment (Appendix J).

References

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. Deep evidential regression. In *Advances in Neural Information Processing Systems*, 2020.

Auer, A., Gauch, M., Klotz, D., and Hochreiter, S. HopCPT: Conformal prediction for time series with modern Hopfield networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Cini, A., Jenkins, A., Mandic, D., Alippi, C., and Bianchi, F. M. Relational conformal prediction for correlated time series. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=wwYDQ1vXcZ.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., and Colliot, O. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, 56(2):766–781, 2011. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.06.013.

Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., and Hennig, P. Laplace redux: Effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems*, 2021.

Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N. T., Lenzo, N., Martins, R. N., Maruff, P., Masters, C., Milner, A., Pike, K., Rowe, C., Savage, G., Szoeke, C., Taddei, K., Villemagne, V., Woodward, M., Ames, D., and AIBL Research Group. The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, 21 (4):672–687, 2009. doi: 10.1017/S1041610209009405.

Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., Scahill, R. I., Tabrizi, S. J., Ourselin, S., Fox, N. C., and Alexander, D. C. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *NeuroImage*, 60(3):1880–1889, 2012. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.01.062.

Iturria-Medina, Y., Sotero, R. C., Toussaint, P. J., Mateos-Pérez, J. M., Evans, A. C., Alzheimer's Disease Neuroimaging Initiative, et al. Epidemic spreading model to characterize misfolded proteins propagation in aging and associated neurodegenerative disorders. *PLoS Computational Biology*, 10(11):e1003956, 2014.

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., and Trojanowski, J. Q. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010. doi: 10.1016/S1474-4422(09)70299-6.

Kendall, A. and Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

LaMontagne, P. J., Benzinger, T. L., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A., Raichle, M. E., Cruchaga, C., and Marcus, D. OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medRxiv*, 2019. doi: 10.1101/2019.12.13.19014902.

Li, H., Deng, B., Xu, C., Feng, Z., Schlegel, V., Huang, Y.-H., Sun, Y., Sun, J., Yang, K., Yu, Y., and Bian, J. MIRA: Medical time series foundation model for real-world health data, 2025. URL https://arxiv.org/abs/2506.07584.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*, 2020.

Lin, Z., Trivedi, S., and Sun, J. Conformal prediction with temporal quantile adjustments. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax-Weiss, T., and Lakshminarayanan, B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Lorenzi, M., Filippone, M., Frisoni, G. B., Alexander, D. C., Ourselin, S., and Alzheimer's Disease Neuroimaging Initiative. Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease. *NeuroImage*, 190: 56–68, 2019. doi: 10.1016/j.neuroimage.2017.08.059.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). *Alzheimer's & Dementia*, 1(1): 55–66, 2005. doi: 10.1016/j.jalz.2005.06.003.

Palmqvist, S., Schöll, M., Strandberg, O., Mattsson, N., Stomrud, E., Zetterberg, H., Blennow, K., Landau,

S., Jagust, W., and Hansson, O. Earliest accumulation of β-amyloid occurs within the default-mode network and concurrently affects brain connectivity. *Nature Communications*, 8(1):1214, 2017. doi: 10.1038/s41467-017-01150-x.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. Transfusion: Understanding transfer learning for medical imaging. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 0021-9991. doi: 10.1016/j.jcp.2018.10.045.

Raj, A., Kuceyeski, A., and Weiner, M. A network diffusion model of disease progression in dementia. *Neuron*, 73(6): 1204–1215, 2012. doi: 10.1016/j.neuron.2011.12.040.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., and Yeo, B. T. T. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018. doi: 10.1093/cercor/bhx179.

Supplementary Material

A. Dataset Details

Cohorts

Table 3 summarises the three cohorts. ADNI is the source; AIBL and OASIS-3 are fully held-out external cohorts on which the ADNI-trained forecaster is evaluated without further training.

Table 3. Evaluation cohorts. ADNI is the source cohort; AIBL and OASIS-3 are fully held-out external cohorts. OASIS-3 uses multi-horizon windows because sparser visit schedules reduce converter counts at narrow windows.

Cohort	Group	Target scan	Subjects	Converters
ADNI	All	24m exact	204	40
ADNI	Baseline CN	24m exact	92	6
ADNI	Baseline MCI	24m exact	112	34
AIBL	All	24m $\pm$ 6m	139	14
AIBL	Baseline CN	24m $\pm$ 6m	120	9
AIBL	Baseline MCI	24m $\pm$ 6m	19	5
OASIS-3	All	36m	162	7
OASIS-3	All	54m	148	13
OASIS-3	All	72m	131	14

Conversion labels

A subject is labeled positive if any follow-up diagnosis within the conversion window is strictly worse (cognitively normal (CN) $\rightarrow$MCI, CN$\rightarrow$ Alzheimer's disease (AD), or MCI$\rightarrow$AD) than the baseline diagnosis. For ADNI the target structural scan is required at exactly 24 months. For AIBL a ± 6 month tolerance is used because AIBL visits commonly occur at months 18, 36, 54, and 72. For OASIS-3, conversion labels are derived from longitudinal diagnosis codes across the multi-horizon windows listed above; evaluability requires diagnostic follow-up extending to the window end.

Evaluation protocol

Risk models are evaluated by ROC AUC with 2,000 nonparametric bootstrap resamples. Resamples containing only one class are skipped. The high-risk conversion-rate threshold is the 75th percentile of the training score, evaluated on the validation split without using validation labels for the cutoff.

B. Architecture and Training Details

Forecasting backbone

The backbone is a reaction–diffusion graph network over the 200-node Schaefer atlas. The diffusion term uses the graph Laplacian; the reaction term is a graph attention network with spectral positional features and FiLM/AdaIN-style covariate conditioning. A learned sparse latent graph is added as a small residual adjacency component. The residual head is swapped while keeping the same backbone: no residual, deterministic MLP, heteroscedastic Gaussian, normalizing-flow residual, or diffusion residual. Flow and diffusion heads are dominated on both MAE (0.0869, 0.0883) and AUC (0.652, 0.518) and are not the focus of the main results.

For the Gaussian head, the backbone produces a physics-prior forecast μ_{phys}; a lightweight MLP residual head then outputs $(\hat{x}_{t+\tau}, \log \sigma^2)$. The combined loss is

$$\lambda_{\text{nll}} \frac{1}{2} \left[\log \sigma^2 + \frac{(x_{t+\tau} - \hat{x}_{t+\tau})^2}{\sigma^2} \right] + (1 - \lambda_{\text{nll}}) \|\hat{x}_{t+\tau} - x_{t+\tau}\|_2^2,$$

with ROI weighting, plus a backbone forecasting loss, a correlation loss on predicted vs. observed change, and a margin-ranking loss on regional change magnitudes.

Table 4. Main hyperparameters for the Gaussian biomarker experiment.

Category	Setting
Atlas	200-node Schaefer; ADNI source; AIBL/OASIS-3 transfer
Forecast	24-month cortical-thickness prediction
Conv. label	Diagnosis worsening within window
History	Baseline + previous scan; fallback gap 12 m
Split	Stratified, train 0.8, seed 42
Backbone	3 layers, hidden 128, 4 heads, dropout 0.1
Spectral	16 eigenvectors, dim 16; latent dim 32, top-k=10
Optimizer	AdamW, lr 2×10^{-4}, wd 10^{-4}
Training	80 epochs, patience 12, batch 32, noise 0.02
Loss	λ_{phys}=0.5, λ_{corr}=0.5, λ_{rank}=0.2, λ_{nll}=0.7
Parameters	753,557

Hyperparameters

Downstream adapters

Linear adapters are balanced logistic regressions on standardised $[s_\mu, s_\sigma]$ features. The quadratic adapter applies degree-two polynomial features to $[s_\mu, s_\sigma]$ and fits a balanced logistic regression with C=0.1. Clinical covariates are baseline diagnosis, age, sex, APOE4, and baseline mean thickness. All adapters are trained on ADNI and evaluated on the target cohort without any target-cohort tuning.

C. Formal Analysis: When Does Aleatoric Uncertainty Transfer?

We formalise the toy mechanism and give conditions under which the rank of the aleatoric uncertainty channel is preserved under cohort shift while the rank of the mean channel degrades.

Setup. Let $z_i \in \mathbb{R}$ be a scalar latent disease severity, with conversion label $y_i = \mathbf{1}[z_i > \tau_y]$, where τ_y is the conversion threshold (distinct from the forecast horizon τ in the main text). On the *source* cohort

$$x_i = g(z_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2(z_i)),$$

where g is the source mean mapping and $\sigma^2(z_i)$ is the local variance, higher for subjects far from the stable-population centre z_0. A forecaster trained on the source learns a **mean channel** $\hat{\mu}_i \approx g(z_i)$ and a **sigma channel** $\hat{\sigma}_i \approx \sigma(z_i)$. On the *target* cohort an additive offset $\delta_i \sim \mathcal{N}(0, \Delta^2)$, $\delta_i \perp z_i$, shifts observations without affecting labels.

Assumptions. **A (monotone mean):** g is strictly increasing. **B (monotone variance):** $\sigma^2(z)$ is strictly increasing in $|z - z_0|$. Combined with $y_i = \mathbf{1}[z_i > \tau_y]$ and $\tau_y > z_0$, this implies $\mathrm{Corr}(\sigma(z_i), y_i) > 0$. **C (learned sigma):** $\hat{\sigma}_i = \sigma(z_i) + \xi_i$ where ξ_i is small-variance noise independent of δ_i.

Proposition 1 (Idealised controlled-shift mechanism)**.** *Under Assumptions A–C* and *the idealised independence $\delta_i \perp z_i$, let $\rho_\mu(\Delta)$ and $\rho_\sigma(\Delta)$ be the population rank correlations of each channel with y_i on the target cohort as a function of shift magnitude Δ.*

(i) $\rho_\mu(\Delta)$ *is* decreasing *in Δ and $\rho_\mu(\Delta) \to 0$ as $\Delta \to \infty$.*

(ii) $\rho_\sigma(\Delta) = \rho_\sigma(0)$ *for all $\Delta \geq 0$.*

We treat this proposition as a mechanism statement, not a transfer guarantee for real cohorts: the $\delta_i \perp z_i$ assumption is violated in ADNI→AIBL (see Remark below). Empirical support comes from the controlled simulation (Figure 2); transfer claims in the main text are empirical, not derived from this proposition.

Proof sketch. **(i).** On the target cohort $\hat{\mu}_i \approx \hat{g}(g(z_i) + \delta_i + \varepsilon_i)$. Linearising around g gives $\hat{\mu}_i = az_i + b\delta_i + \varepsilon'_i$. Because $\delta_i \perp z_i \perp y_i$:

$$\rho_\mu(\Delta) = \frac{a\,\mathrm{Cov}(z_i, y_i)}{\sqrt{a^2\mathrm{Var}(z) + b^2\Delta^2 + \mathrm{Var}(\varepsilon')}} \xrightarrow{\Delta\to\infty} 0.$$

(ii). By Assumption C, $\hat{\sigma}_i = \sigma(z_i) + \xi_i$ is a function of z_i alone; since an additive shift $\delta_i \perp z_i$ leaves the joint distribution of $(\hat{\sigma}_i, y_i)$ unchanged, $\rho_\sigma(\Delta) = \rho_\sigma(0)$. □

Remark 1 (Assumption violation in ADNI→AIBL)**.** *The proposition requires $\delta_i \perp z_i$. In practice ADNI and AIBL differ in converter base rate (≈20% vs. ≈10%), suggesting $\mathrm{Cov}(\delta_i, z_i) \neq 0$. Under this violation the proposition does not guarantee exact s_σ invariance. The empirical result – s_σ AUC drops -0.082 (ADNI 0.777 → AIBL 0.695) vs. mean AUC drops -0.239 (0.773 → 0.534) – is consistent with the qualitative prediction even when the exact independence condition is relaxed.*

Corollary (sampling-based epistemic uncertainty does not transfer). Sampling-based epistemic signals $\hat{\sigma}_i^{\mathrm{epi}} = \mathrm{Var}_k(\hat{\mu}_i^{(k)})$ – as produced by deep ensembles, MC dropout, last-layer Laplace, and SNGP – track parameter-space coverage gaps in the *source* training distribution. Under cohort shift the set of x values encountered changes, so the epistemic signal is no longer aligned with the source-derived label structure. This explains the empirical observation that Gaussian-head aleatoric s_σ transfers from ADNI (0.777) to AIBL (0.695, drop of 0.082), while the four sampling-based epistemic s_σ channels (ensemble, MC dropout, Laplace, SNGP) span in-domain ADNI-val AUCs 0.712, 0.573, 0.602, 0.608 respectively (range $[0.55, 0.71]$, with the deep ensemble setting the upper bound and the other three clustered in $[0.57, 0.61]$) and converge to a tighter AIBL 36m cluster $[0.55, 0.60]$ under transfer (Section 4.3).

The corollary applies specifically to sampling-based posterior approximations, where epistemic $\hat{\sigma}$ is computed from variance across model samples or from a Laplace fit at the last layer. Evidential approaches whose "epistemic" channel is a closed-form function of the residual likelihood (e.g. NIG with $\sigma_e^2 = \beta/(\nu(\alpha - 1))$ (Amini et al., 2020)) are *not* covered by this argument: their predicted σ is an algebraic re-scaling of the same likelihood that drives the aleatoric channel, and they empirically transfer like aleatoric heads (NIG aleatoric AIBL 0.650, NIG epistemic AIBL 0.667 – close to heteroscedastic Gaussian, far from the sampling-based cluster).

Controlled covariate-shift simulation

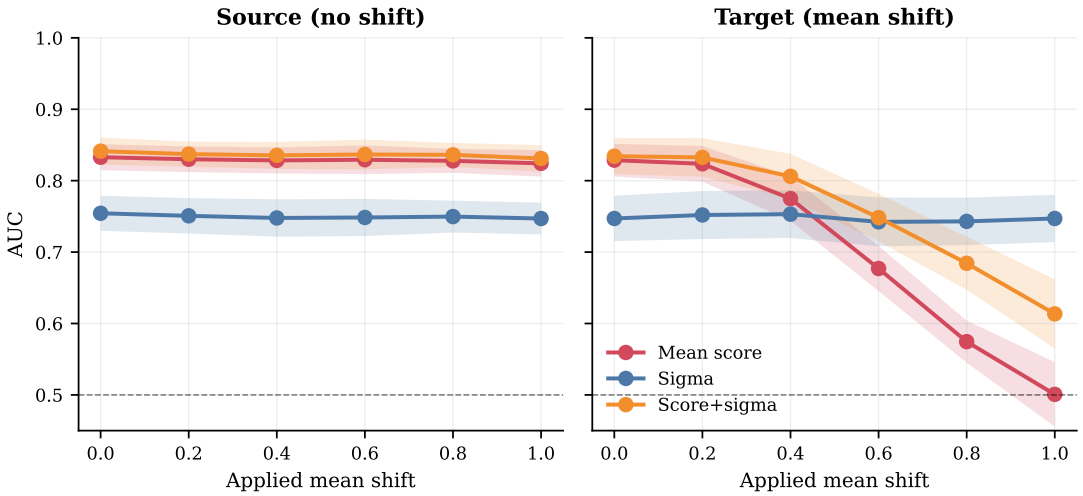

Figure 2. Controlled cohort-shift simulation. x-axis: applied mean shift as a fraction of the source-cohort mean-score standard deviation (0 = no shift, 1 = one source σ). Increasing shift collapses the mean score's target AUC toward chance while s_σ stays near 0.75; the joint channel degrades more gracefully than the mean alone.

Figure 2 confirms the mean-vs-sigma divergence under controlled shift: target AUC for the mean score falls from 0.829 at zero shift to 0.501 at full corruption, while s_σ-only AUC stays near 0.747.

D. Physics and Backbone Ablations

The physics-ablation study retrains the Gaussian pipeline after removing one backbone component at a time. Across 5 random subject splits, only removing the diffusion term (the "Reaction only" row, which keeps the reaction component but ablates the graph-Laplacian diffusion) produces a consistent, large drop; all other single-component ablations fall within the

full model's 95% CI. This supports the statement in Section 4.1 that the s_σ advantage is localised to the heteroscedastic Gaussian head, not to any specific backbone component.

Table 5. Gaussian backbone ablations on the 24-month ADNI split (5 random splits; mean AUC with 95% CI from a Student's t-distribution with 4 degrees of freedom). Only removing the diffusion term ("Reaction only") produces a consistent, large drop.

Configuration	Mean AUC	95% CI
Full model	0.762	[0.707, 0.817]
No spectral features	0.759	[0.705, 0.812]
No fixed atlas graph	0.741	[0.679, 0.802]
No latent graph	0.762	[0.707, 0.816]
No temporal gap	0.754	[0.692, 0.816]
Reaction only	0.689	[0.606, 0.773]
Diffusion only	0.764	[0.692, 0.836]

E. Trajectory Atypicality and Regional Distribution of Predictive Uncertainty

Sigma tracks observed trajectory variability

Figure 3 provides the visual evidence for the trajectory-atypicality interpretation in Section 4.6. Predicted s_σ tracks observed within-subject variability (ρ=0.287), and the top-s_σ quartile concentrates conversion (median OR 8.8 across 5 seeds), with quartile thresholds derived from the forecaster's predicted uncertainty alone – no conversion labels are used.

Regional gradient

Per-region σ_i correlates positively with conversion across all 200 cortical parcels (Benjamini–Hochberg FDR at q=0.05; ρ range 0.26–0.42), indicating that elevated σ_i is a cortex-wide signature of unstable subjects rather than a focal hotspot. Within this diffuse elevation a clear gradient emerges: the strongest correlations concentrate in the inferior parietal lobule / salience network (ρ=0.42) and default-mode prefrontal cortex (ρ=0.42), with a secondary temporal-parietal gradient (ρ=0.38) – precisely the regions implicated in earliest amyloid accumulation and default-mode degeneration in Alzheimer's disease (Palmqvist et al., 2017; Jack et al., 2010). The model therefore assigns highest predictive uncertainty where neuroscience independently expects earliest pathology, without any conversion supervision.

F. OASIS-3 External Transfer

A second external cohort, OASIS-3 (183 subjects, community-based, independent scanner), provides a further transfer probe. Forecasting MAE rises to 0.2384 (vs. 0.0867 on ADNI), consistent with scanner and protocol shift – the forecasting transfer itself is well-powered. Risk stratification is directionally positive only: at 36m s_σ clearly exceeds the mean score (0.726 vs. 0.606); at 54m and 72m the differences are within noise. All OASIS-3 risk AUCs should be read as exploratory (n=7–14 converters).

Table 6. OASIS-3 transfer (community-based, independent scanner). MAE rises to 0.2384 vs. 0.0867 on ADNI, consistent with protocol shift. Risk AUCs are exploratory (7–14 converters): s_σ clearly beats score at 36m (0.726 vs. 0.606); 54m/72m differences are negligible with fully overlapping CIs.

	ADNI	OASIS-3
Forecasting		
MAE (mm)	0.0867	0.2384
Risk AUC (exploratory; n=7–14 conv.)		
Score (36m)	0.784	0.606 [0.299, 0.893]
s_σ (36m)	0.778	0.726 [0.537, 0.883]
Score (54m)	–	0.618 [0.446, 0.784]
s_σ (54m)	–	0.638 [0.501, 0.771]
Score (72m)	–	0.612 [0.441, 0.764]
s_σ (72m)	–	0.618 [0.477, 0.751]

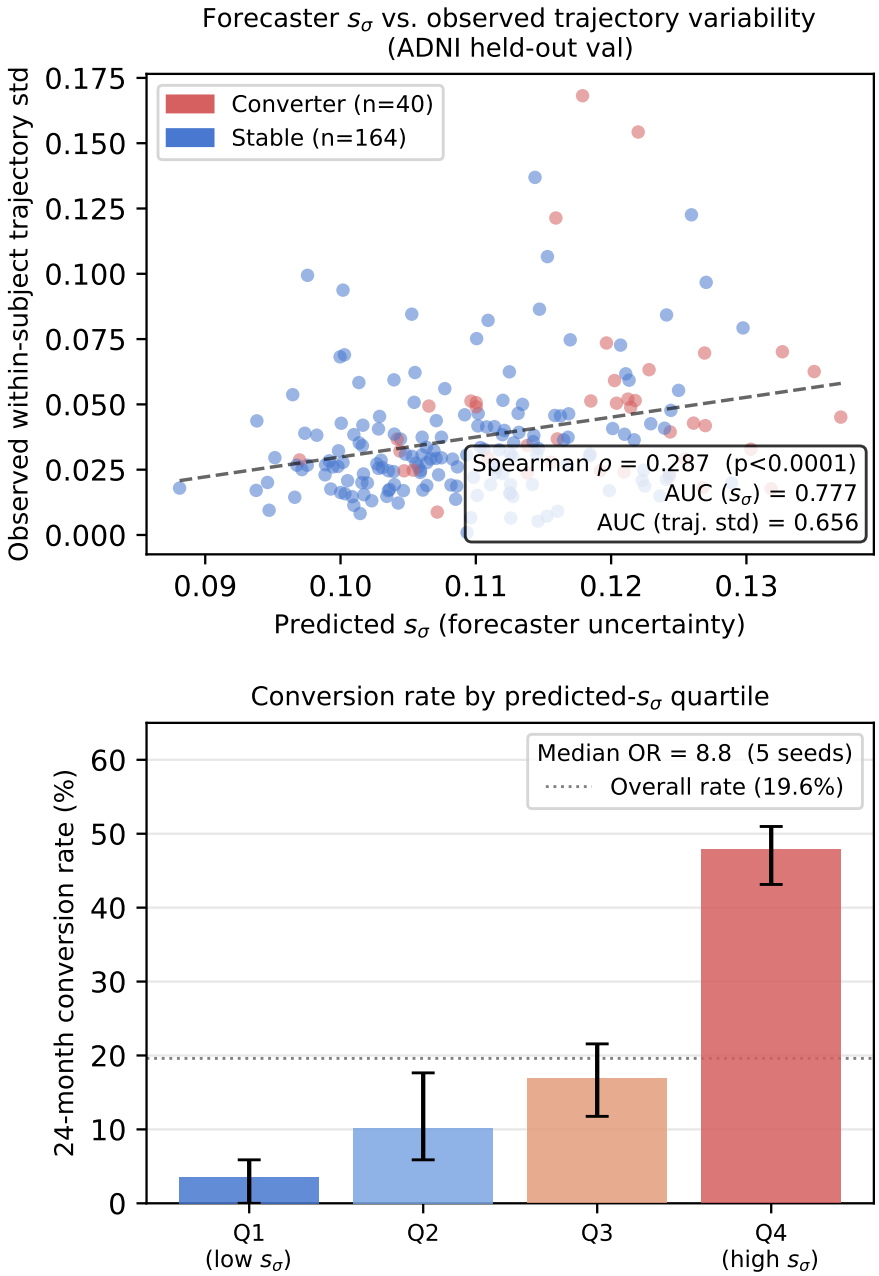

Figure 3. **Sigma tracks trajectory instability** (ADNI held-out val). *Top:* predicted s_σ vs. observed within-subject trajectory standard deviation; Spearman ρ=0.287 (p<0.0001), AUC 0.777 vs. 0.656 for raw std. *Bottom:* conversion rate by s_σ-quartile (5 seeds); Q4 vs. Q1–3 median OR = 8.8 (5.7–10.3).

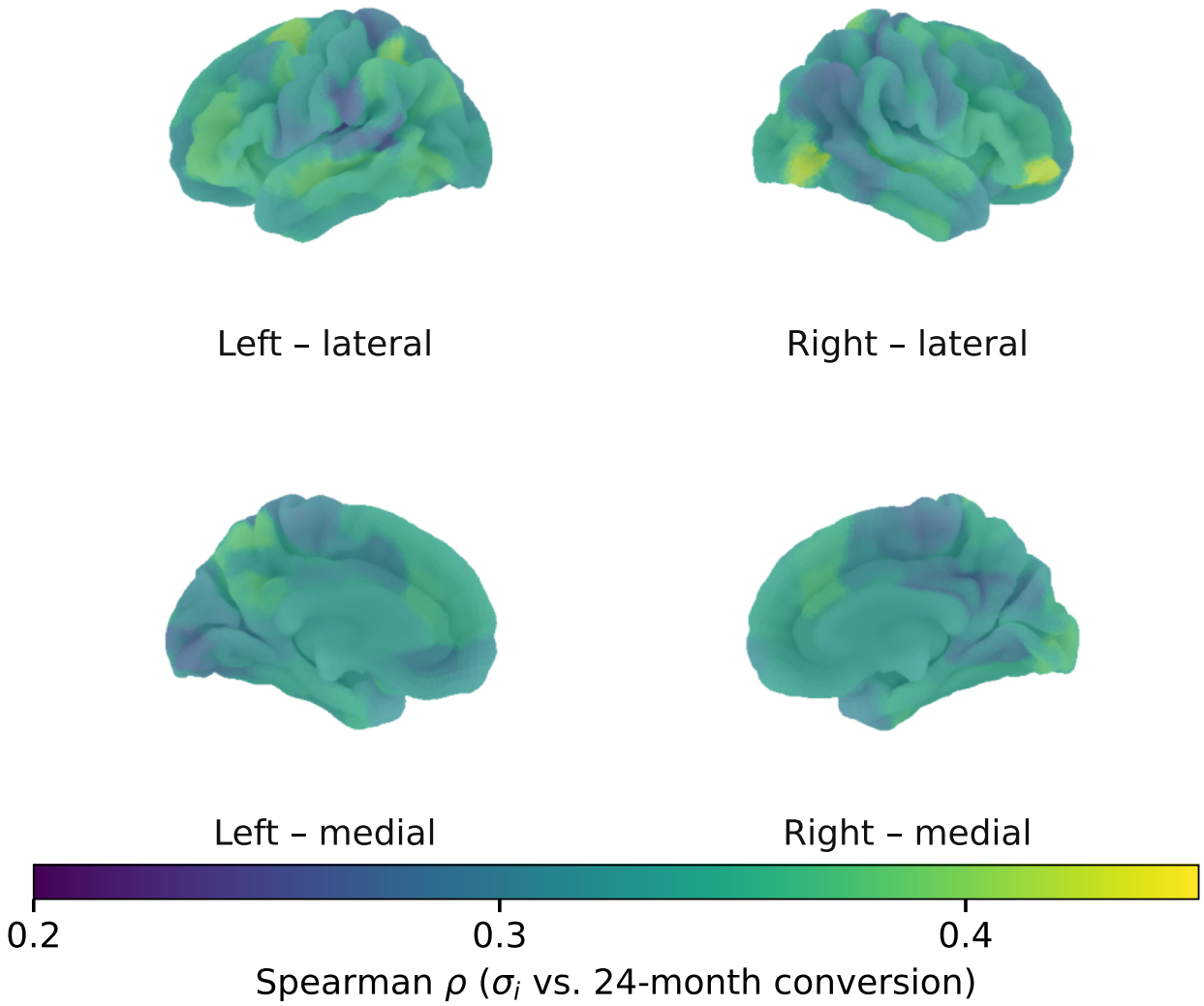

Figure 4. Per-region Spearman ρ (σ_i vs. 24-month conversion; ADNI, n=204; Benjamini–Hochberg FDR at q=0.05). All 200 regions significant (ρ=0.26–0.42); gradient peaks in temporal-parietal cortex and default-mode regions (ρ=0.42), consistent with earliest Alzheimer's pathology.

G. Supervised Baselines and Adapter Variants

Table 7 provides the full ADNI–AIBL transfer results, including the 36m AIBL horizon and the linear-adapter row omitted from Table 2 for space, and reproducing the supervised-baseline rows for one-table reference. Both raw forecast channels are label-free; clinical baselines use full clinical covariates (diagnosis, age, sex, APOE4, baseline mean thickness). Sample sizes shrink at longer AIBL horizons due to follow-up requirements (126/15 at 36m, 119/29 at 54m).

Table 7. Full ADNI–AIBL transfer results. 95% bootstrap CIs; forecast-derived scores are label-free. ADNI fixed at 24m. † Diagnosis is constant within MCI, so AUC = 0.500 by construction.

Model	ADNI all [95% CI]	ADNI MCI	AIBL all [95% CI]
Zero-label baselines, 24m			
Score only	0.773 [0.689, 0.849]	0.710	0.534 [0.373, 0.693]
s_σ only	0.777 [0.693, 0.856]	0.761	0.695 [0.530, 0.844]
Multi-horizon AIBL transfer			
Score (36m, 15 conv.)	–	–	0.473 [0.306, 0.643]
s_σ (36m, 15 conv.)	–	–	0.660 [0.490, 0.816]
Score (54m, 29 conv.)	–	–	0.403 [0.287, 0.523]
s_σ (54m, 29 conv.)	–	–	0.769 [0.656, 0.870]
Supervised baselines			
Diagnosis only†	0.687 [0.616, 0.748]	0.500†	0.622 [0.495, 0.760]
Clinical covariates	0.791 [0.704, 0.867]	0.753	0.701 [0.538, 0.851]
Gradient boosting (clin.)	0.792 [0.710, 0.866]	0.740	–
Adapted forecast scores (ADNI-trained adapter, no AIBL labels)			
Score $+s_\sigma$ (linear)	0.788 [0.704, 0.862]	0.758	0.741 [0.603, 0.866]
Score $+s_\sigma$ (quadratic)	0.789 [0.706, 0.864]	0.755	0.770 [0.655, 0.877]

H. UQ-Panel Cross-Cohort Transfer

Table 8 reports the AIBL 36-month transfer of the s_σ channel for all six UQ methods evaluated on the shared backbone, sourcing the in-domain ADNI-val numbers in Table 1. The four sampling-based epistemic methods (ensemble, MC dropout, last-layer Laplace, SNGP) cluster in the range $[0.55, 0.60]$, while the heteroscedastic Gaussian aleatoric (0.660) and the two evidential NIG channels (0.650 aleatoric, 0.667 epistemic) sit cleanly above. The split is consistent with the residual-likelihood vs. posterior-weight distinction discussed in §4.1: every channel that is a closed-form function of the residual-magnitude likelihood transfers; every channel estimated from variation across model samples does not. The Gaussian ADNI-val and AIBL numbers are reported on the single calibration seed used for the registry; the other methods' AIBL numbers are 5-seed means.

Table 8. AIBL 36-month transfer of the s_σ channel for the full UQ panel. Mean $\pm$ std across 5 seeds (42–46) where available; Gaussian AIBL is the single calibration-registry seed. ADNI-val column reproduces the s_σ AUC from Table 1 for direct comparison. The four sampling-based epistemic methods cluster low; the heteroscedastic Gaussian and the two evidential NIG channels sit above the cluster.

UQ channel (s_σ AUC)	ADNI val	AIBL 36m
Residual-likelihood-based (transfers)		
Heteroscedastic Gaussian (aleatoric)	0.792 ± 0.022	0.660
Evidential NIG (aleatoric)	0.766 ± 0.007	0.650 ± 0.021
Evidential NIG (epistemic)	0.765 ± 0.007	0.667 ± 0.006
Posterior-weight-based (does not transfer)		
Deep ensemble (epistemic)	0.712 ± 0.039	0.600 ± 0.070
SNGP (epistemic)	0.608 ± 0.008	0.574 ± 0.111
Last-layer Laplace (epistemic)	0.602 ± 0.019	0.579 ± 0.052
MC dropout (epistemic)	0.573 ± 0.043	0.555 ± 0.022

I. Paired Bootstrap on AIBL 36m

Because the marginal 95% bootstrap CIs for s_σ and s_μ overlap at the AIBL 24m horizon (n=14 converters), we run a within-method paired bootstrap on AIBL 36m (n=126, 15 converters), where the larger converter count restores power

without changing the train-target setup. For each seed we resample subjects with replacement and recompute $\Delta_{\sigma-\mu} = \mathrm{AUC}(\hat{\sigma}) - \mathrm{AUC}(s_\mu)$ on the same resample, then aggregate the per-seed point estimates into a paired-by-seed 95% CI. Both $\Delta_{\sigma-\mu}$ and the absolute transfer $\mathrm{AUC}(\hat{\sigma})$ appear in Table 9. Two readings are useful and do not agree: $\Delta_{\sigma-\mu}$ asks "does s_σ beat s_μ within this method?" and is sensitive to how strong s_μ itself is; $\mathrm{AUC}(\hat{\sigma})$ asks "does s_σ rank converters on AIBL?" and is the quantity the main claim turns on.

On $\Delta_{\sigma-\mu}$, only the heteroscedastic Gaussian is positive ($+0.187$, single calibration seed); every other method is negative. This reflects that the NIG and sampling-based heads inherit a stronger learned s_μ predictor on this transfer set ($\mathrm{AUC}(\hat{\mu})\approx$0.65–0.71) than the Gaussian checkpoint we calibrate against ($\mathrm{AUC}(\hat{\mu})$=0.473), which raises the bar for s_σ to clear. NIG aleatoric and epistemic exclude zero on the negative side ($[-0.089, -0.025]$ and $[-0.068, -0.014]$); the deep ensemble, Laplace, and MC dropout exclude zero on the negative side too; SNGP's CI ($[-0.159, +0.011]$) just includes zero.

On $\mathrm{AUC}(\hat{\sigma})$ the residual-likelihood cluster is clearly separated from the sampling-based one. Heteroscedastic Gaussian, NIG aleatoric, and NIG epistemic all sit at 0.650–0.667; deep ensemble, Laplace, MC dropout, and SNGP all sit at 0.555–0.600. A reviewer therefore reads $\Delta_{\sigma-\mu}$ as a within-method robustness check whose sign is dominated by the strength of s_μ, and reads $\mathrm{AUC}(\hat{\sigma})$ as the load-bearing transfer signal.

Table 9. Per-method AIBL 36m paired bootstrap and absolute transfer $\mathrm{AUC}(\hat{\sigma})$. $\Delta_{\sigma-\mu}$ is paired by-seed (mean and 95% CI across 5 seeds, except Gaussian which uses the single calibration-registry seed). $\mathrm{AUC}(\hat{\sigma})$ is the cross-cohort transfer AUC of the per-method σ channel (mean across seeds).

Method	$\Delta_{\sigma-\mu}$	95% CI	$\mathrm{AUC}(\hat{\sigma})$
Heteroscedastic Gaussian	$+0.187$	–	0.660
NIG aleatoric	-0.057	$[-0.089, -0.025]$	0.650
NIG epistemic	-0.041	$[-0.068, -0.014]$	0.667
Deep ensemble	-0.091	$[-0.175, -0.007]$	0.600
Last-layer Laplace	-0.115	$[-0.200, -0.030]$	0.579
MC dropout	-0.116	$[-0.183, -0.049]$	0.555
SNGP	-0.074	$[-0.159, +0.011]$	0.574

J. Limitations and Responsible Use

The evidence supports using aleatoric uncertainty as a transfer-stable risk signal, but does not establish clinical readiness. AIBL has only 14 converters under the 24-month definition (29 at 54m). OASIS-3 has 7–14 converters across multi-horizon windows, yielding wide confidence intervals for risk AUC; forecasting transfer metrics (MAE) are well-powered and show strong generalisation. The uncertainty score is a predictive-dispersion signal from a heteroscedastic likelihood, not a full decomposition of epistemic and aleatoric uncertainty. Downstream labels are diagnosis transitions, which are clinically meaningful but noisy and visit-schedule dependent.

All experiments should be interpreted as a retrospective model evaluation on curated neuroimaging cohorts. The model should not be used for clinical decision-making without additional validation, calibration, and fairness analysis across demographic groups, and prospective assessment under the intended deployment protocol.

Societal impact. The s_σ AUC of 0.695 on AIBL transfer is below any clinically actionable threshold for individual screening decisions. This model is a research tool: its outputs should not be used for patient-level diagnostic or treatment decisions without prospective validation, regulatory approval, and fairness analysis across demographic groups (age, sex, race, APOE4 status). Conversion labels in both ADNI and AIBL are derived from retrospective clinical assessments and do not constitute a ground-truth biomarker of AD pathology.

Data and code. ADNI and AIBL data access requires a data use agreement; see `https://adni.loni.usc.edu` and `https://aibl.com.au` respectively. Code will be released upon acceptance.