# AUDITING FORECAST OUTPUTS UNDER COHORT SHIFT: RESIDUAL LEARNING RETAINS CONVERSION-RANKING INFORMATION WHILE PREDICTED CHANGE DOES NOT

**Anonymous authors**
Paper under double-blind review

(a)

**Train forecaster on ADNI**
Inputs: current and previous cortical thickness, demographics, and baseline diagnosis.
Outputs: future cortical thickness and per-region standard deviation.
Future diagnoses are not used in training.

↓ Freeze forecaster; extract predicted change $s_\mu$, residual scale $s_\sigma$, and model variability $s_{\mathrm{PW}}$.

**Evaluate unchanged on AIBL**
No parameters are fitted on target data.

**Raw-score audit (primary)**
Rank subjects by each score; evaluate against future conversion.
No conversion-label training.

**Source-supervised adapter**
Fit logistic regression on ADNI conversion labels; apply unchanged to AIBL.

(b)

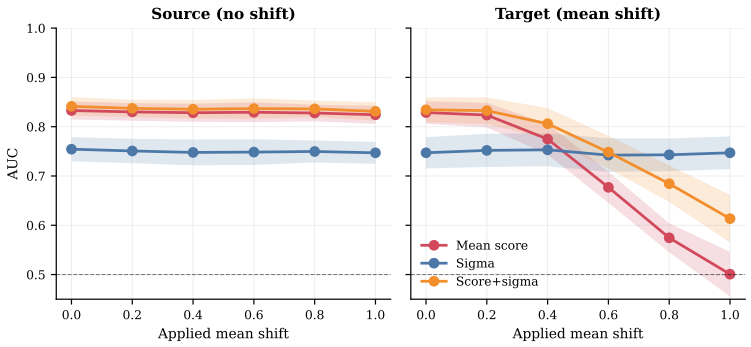



Figure 1: (a) Evaluation protocol. The ADNI-trained forecaster is evaluated unchanged on AIBL; target diagnoses are used only for evaluation. The volumetric analysis follows the same separation within ADNI. (b) Synthetic cohort-shift simulation. Both scores reflect latent disease severity, which is higher in future converters. As shift strength $\rho$ increases from 0 to 1, predicted change loses its severity association in the target (right): its AUC falls from 0.83 to chance (0.5, dashed), while residual-scale AUC stays near 0.75. The source is shown at left. Curves show medians and 5th–95th percentiles over 30 simulations; parameters are not fitted to ADNI or AIBL.

## ABSTRACT

A probabilistic forecaster produces several scores for each subject: its predicted change, its predicted residual spread, and its variability across model fits. These scores need not remain equally informative when the model is applied to a new cohort. We study this question in Alzheimer's disease (AD), where cortical thinning tracks disease progression. Here, we train a probabilistic forecaster on longitudinal Magnetic Resonance Imaging (MRI) data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study to predict future regional brain cortical



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

thickness, without using future-conversion diagnoses. We then freeze it and evaluate, on the Australian Imaging, Biomarker and Lifestyle (AIBL) data, how well each of three scores identifies subjects whose clinical diagnosis will worsen within 24 months (converters): predicted-change magnitude $s_\mu$, learned residual scale $s_\sigma$, and model variability $s_{\mathrm{PW}}$. As an in-domain reference, we evaluate a separate volumetric forecasting analysis within ADNI data, without cohort shift: on 127 held-out subjects (26 converters), its predicted change is informative (ROC-AUC $0.755$), residual scale is weaker ($0.656$), and model variability is near chance ($0.532$). In contrast, when the cortical-thickness forecaster is transferred from ADNI to AIBL 139 subjects; 14 converters), residual scale identifies converters with ROC-AUC $0.695$, predicted change performs near chance ($0.534$), and model-variability scores do not exceed chance ($0.40$–$0.52$). The advantage of residual scale over predicted change on AIBL is consistent across training seeds and leave-one-converter-out checks; its 95% interval includes zero at 24 months but excludes zero when the outcome window is extended to 36 and 54 months. We explain these differences theoretically: predicted change transfers only if its relationship to disease severity is preserved across cohorts; residual scale transfers if its distribution given severity is stable within each outcome class; and model variability reflects parameter sensitivity rather than disease state. These results argue for evaluating each forecast output separately under cohort shift; they do not show that uncertainty is generally superior, nor do they provide a calibrated clinical risk model.

## 1 INTRODUCTION

A forecasting model produces more than a point prediction. Alongside its expected output, it exposes a residual scale; additionally, perturbing its parameters or training configuration yields slightly different answers. These distinct quantities arise from different mechanisms, yet practitioners often treat them as interchangeable summaries of the same model. When deploying a model on a population that differs in scanners, visit schedules, or baseline severity profiles, there is no reason to expect every output to degrade at the same rate. Which forecast-derived outputs retain their usefulness for a downstream task after such a shift? This is the central question we study.

We investigate this problem in longitudinal brain imaging. A model is trained to predict future regional cortical thickness from current and past Magnetic Resonance Imaging (MRI) scans alongside basic covariates (demographics, APOE4 status, and baseline diagnosis). From this forecaster, we extract three scalar scores for each subject:

- *Predicted change* ($s_\mu$): The size of the predicted change, computed as the root-mean-square difference between the forecast and current measurements across 200 cortical regions.
- *Residual scale* ($s_\sigma$): The average per-region standard deviation assigned by the model's likelihood head, intended to represent the subject-specific prediction error anticipated by the model.
- *Model variability* ($s_{\mathrm{PW}}$): The fluctuation of the predicted mean across retrainings, dropout passes, or sampled posterior weights, serving as a sensitivity diagnostic rather than a biological indicator.

We then test whether each score, evaluated directly as a ranking statistic, can identify subjects who will convert by progressing from cognitively normal (CN) to mild cognitive impairment (MCI) or Alzheimer's disease (AD), or from MCI to AD. Conversion is a sparse, delayed clinical label, whereas imaging measurements provide dense, early supervision. Because the forecaster is trained without future conversion labels, downstream ranking provides an independent assessment of the representations induced by imaging supervision.

This separation is critical because forecast accuracy and conversion ranking measure fundamentally different properties. A forecaster can predict cortical thickness with low mean absolute error while its predicted-change score carries little diagnostic ranking information. Conversely, an output that lacks calibrated uncertainty can still separate future converters from non-converters. This setup resembles a dense-to-sparse transfer problem (Raghu et al., 2019): abundant imaging trajectories guide training,

sparse clinical events define evaluation, and representations that optimize one task do not necessarily transfer to the other.

We train models on the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005) and evaluate frozen predictors on the Australian Imaging, Biomarkers and Lifestyle study of ageing (AIBL), which introduces discrepancies in visit intervals, scanner hardware, and cohort composition. The primary audit fits no conversion classifier; each score is ranked directly against future diagnostic worsening. A separate set of source-supervised adapters, fitted on ADNI conversion labels and frozen prior to target evaluation, provides a complementary benchmark.

The central empirical finding is that the most informative output depends on the setting. On the external AIBL cohort (139 subjects, 14 converters), predicted change performs near chance (ROC-AUC $0.534$), whereas residual scale retains discriminative power ($0.695$). Within ADNI and without cohort shift, a separate forecaster of six regional brain volumes shows the opposite ordering on a matched hold-out 127 subjects, 26 converters): predicted change is informative ($0.755$), residual scale is weaker ($0.656$), and model variability is near chance. No single output is universally superior; transferability depends on how a specific cohort shift interacts with the mechanism generating the score. With only 14 AIBL converters at 24 months, the confidence interval for the difference between residual scale and predicted change includes zero. At 36 and 54 months, the corresponding intervals exclude zero, but consistent results across training seeds do not make up for the small number of converters. The goal of this paper is therefore to delineate the conditions under which each forecast output remains informative, explaining why specific outputs fail or survive rather than asserting an intrinsic advantage for uncertainty over point predictions.

**Contributions.** We make three contributions:

1. **An explicit output-wise evaluation protocol.** We separate imaging supervision, score construction, and clinical evaluation so that each forecast output can be assessed without conflating it with a fitted conversion classifier.
2. **Evidence across cohorts and MRI representations.** Matched-subject comparisons, training-objective controls, and sample-sensitivity checks determine which ranking differences are supported by the external and volumetric experiments.
3. **Conditional explanations for different rankings.** We distinguish mean-score severity alignment, class-conditional score stability, and posterior-weight sensitivity. The analysis specifies sufficient conditions, rather than assuming that the observed cohort difference has a unique mechanism.

We present the protocol, empirical comparisons, and theoretical interpretation in that order.

## 2 Related work

**Imaging models and clinical outcomes.** AD progression has been studied using event-based and spatiotemporal models (Fonteijn et al., 2012; Lorenzi et al., 2019), direct classifiers (Cuingnet et al., 2011; Liu et al., 2018), and multimodal or graph-based predictors (Spasov et al., 2019; Parisot et al., 2018; Schneuwly Diaz et al., 2025; Baghirova et al., 2025). Our question differs from predicting conversion directly: we evaluate explicit outputs of a forecaster trained on imaging, rather than learning the main predictor from future clinical labels.

**Uncertainty and transfer.** Heteroscedastic regression learns conditional residual scale (Kendall & Gal, 2017); ensembles and approximate Bayesian methods characterize model-related variability (Gal, 2016; Lakshminarayanan et al., 2017; Daxberger et al., 2021; Liu et al., 2020). Time-series interval methods primarily assess uncertainty about the forecast target (Lin et al., 2022; Auer et al., 2023; Sun & Yu, 2025). Here, an uncertainty-derived score is evaluated for a different outcome. Related transfer approaches learn reusable representations (Raghu et al., 2019; Li et al., 2025); we instead audit explicit forecast summaries. Our conditional analysis complements distribution-shift theory (Shimodaira, 2000; Sugiyama et al., 2007; Ben-David et al., 2010) without claiming a general domain-adaptation guarantee.

# 3 METHODOLOGY

## 3.1 FORECASTING AND SCORE CONSTRUCTION

For subject $i$, let $x_{i,t} \in \mathbb{R}^R$ contain cortical thickness in the $R = 200$ Schaefer regions (Schaefer et al., 2018). A physics-informed graph forecaster predicts future mean thickness and regional conditional standard deviations:

$$(\hat{x}_{i,t+\tau}, \sigma_i) = f_\theta(x_{i,t-h}, x_{i,t}, c_i, h, \tau).$$

Here, $h$ is the history interval, $\tau$ the imaging forecast horizon, and $c_i$ contains baseline diagnosis, age, sex, APOE4 status, and mean thickness. With no earlier scan, the current scan is reused, giving zero historical slope. Baseline diagnosis is an input; future conversion is not a training target.

The forecaster is a physics-informed graph network that predicts future mean thickness and regional conditional standard deviations. Full architectural details, including the reaction-diffusion prior and graph-Laplacian diffusion, are provided in Section F.

The primary Gaussian head learns an input-dependent residual scale alongside the mean; misspecified means can also affect learned scale, so we treat it as a residual-scale score rather than validated uncertainty. Separately, we evaluate MC dropout, deep ensembles, last-layer Laplace, and SNGP as model-variability controls; these are not all exact Bayesian posteriors.

**Three scalar outputs.** After converting cortical predictions back to millimetres, we compute

$$s_{\mu\, i} = \left[ \frac{1}{R} \sum_{r=1}^{R} (\hat{x}_{i,t+\tau,r} - x_{i,t,r})^2 \right]^{1/2}, \qquad s_{\sigma\, i} = \frac{1}{R} \sum_{r=1}^{R} \sigma_{ir}. \tag{1}$$

The first measures change magnitude, not thinning direction. The second is an average of regional standard deviations, not the standard deviation of an average. Equal weighting avoids outcome-fitted region selection; it does not assume a common latent severity factor.

For $K$ sampled predicted means $\hat{x}_{ir}^{(k)}$, with $\bar{x}_{ir} = K^{-1} \sum_k \hat{x}_{ir}^{(k)}$, model variability is

$$s_{\mathrm{PW},i} = \frac{1}{R} \sum_{r=1}^{R} \left[ \frac{1}{K} \sum_{k=1}^{K} (\hat{x}_{ir}^{(k)} - \bar{x}_{ir})^2 \right]^{1/2}.$$

Variability is thus computed across predictions within each region before averaging. Ensembles use five independent fits and dropout uses 30 passes; analytic approximations average regional model-variability standard deviations. This score excludes separately predicted residual variance.

**Raw-score evaluation and supervised adapters.** We evaluate each raw score by how well it ranks patients who later worsen; we do not interpret the score as a calibrated probability. In a separate analysis, we fit logistic adapters on ADNI conversion labels and then freeze them. The adapters use either $[s_\mu, s_\sigma]$ or a quadratic expansion of these features. The raw-score and adapter analyses are alternative pathways, not sequential stages (Figure 1a).

## 3.2 VOLUMETRIC REPRESENTATION CONTROL

To examine whether our conclusions generalize beyond cortical surfaces and graph architectures, we evaluate an alternative, clinically standard representation based on regional brain volumes. Tracking six predefined brain regions (e.g., hippocampus, ventricles) normalized by intracranial volume, we train an ensemble of heteroscedastic MLPs to predict volumetric trajectories. This setup extracts the identical three scalar outputs: the change magnitude, residual scale, and model variability, all in standardized volumetric units (Section J.1). Because harmonized longitudinal volume segmentations are available only in ADNI, this serves as an in-domain representation control rather than an external transfer benchmark. Evaluation is restricted to the exact subset of held-out ADNI subjects with complete 24-month scans, ensuring a direct head-to-head comparison with cortical graph predictions under the same label-free evaluation protocols.

Table 1: Distinct 24-month risk-evaluation protocols. Counts are people, not scans. AIBL uses the ADNI external-transfer source model; it contributes no training subjects.

| Protocol | Train | Evaluate | Converters |
|---|---|---|---|
| ADNI: exact-scan cortical benchmark | 814 | 204 | 40 |
| ADNI: external-transfer source run | 837 | 209 | 43 |
| AIBL: external evaluation | — | 139 | 14 |
| ADNI: volumetric control | 507 | 127 | 26 |

# 4 Experiments and results

## 4.1 Protocols and statistical interpretation

**Cohorts, eligibility, and horizons.** Our in-domain ADNI benchmarks and external AIBL transfer experiments use distinct training splits and eligibility criteria (Table 1). To evaluate the forecasters' accuracy, we restrict our study population to individuals who have both baseline scans and longitudinal follow-up imaging. All candidate scores are computed from inputs available at the prediction visit; future imaging is used only as the ground-truth target. Real-world clinical studies also follow disparate visit schedules, creating cross-cohort timing discrepancies. While ADNI conducts imaging follow-ups at 12 and 24 months, AIBL protocols yield follow-up scans at approximately 18 months for all 139 target subjects. To accommodate this reality, we condition the image forecaster on an 18-month horizon for AIBL, while evaluating clinical conversion over the standard 24-month clinical follow-up window.

**Metrics and resampling.** The primary metric is ROC-AUC, and paired comparisons use the subject-level difference $\Delta_{\sigma-\mu} = \mathrm{AUC}(s_\sigma) - \mathrm{AUC}(s_\mu)$ on identical individuals, where positive values favor residual scale. We estimate confidence intervals using 10,000 class-stratified subject-bootstrap resamples for external transfer and 2,000 for the within-ADNI volumetric analysis and source-supervised adapters. Training-seed variability is reported separately, as retraining does not increase the effective number of target subjects. Because the primary AIBL transfer benchmark involves few converters, we report extensive estimator-calibration and small-sample validity checks in Section G.1.

## 4.2 Primary transfer: ADNI to AIBL

On 139 AIBL subjects, including 14 converters, predicted change has AUC $0.534$ and residual scale $0.695$ (Table 2), an observed paired difference of $+0.162$ with 95% interval $[-0.095, 0.411]$ (DeLong $p = 0.230$). On the 209-person ADNI holdout, raw-score AUCs are $0.661$ and $0.767$, respectively. This comparison is distinct from the exact-scan 204-person benchmark. The external result shows retained residual-scale discrimination, not proof that it is superior in the target population.

**Target-sample uncertainty.** All paired inference, window analyses, and seed checks in this paper use the primary seeded run, with the score ordering consistent across multiple training runs. All 20 reported training seeds and all 14 leave-one-converter-out checks favor scale. Yet 100 training-set bootstrap retrainings, with independent AIBL test resampling, give a difference interval $[-0.100, 0.369]$. Training consistency therefore coexists with imprecise target-population inference; additional seeds do not replace additional converters.

**Extended outcome windows.** The 14-converter limit is a property of the strict 24-month window. Extending the clinical evaluation window to 36 and 54 months on the same frozen forecasts increases the converter count to 21 and 29, respectively. This extension further widens the performance gap in favor of residual scale (paired intervals exclude zero; see Table 3). Part of the widening reflects below-chance performance for predicted change: 0.435 at 36 months and 0.403 at 54 months. Furthermore, cumulative gain analysis at 36 months (Figure 2) shows that residual scale places more converters near the top of the ranking, capturing 57% of converters in the top quintile compared to 19% for predicted change.

Table 2: Primary AIBL evaluation: 139 subjects, 14 converters. Raw rows are follow-up point estimates; their paired difference CI is shown in-table (DeLong $p = 0.230$). Source-supervised rows show 95% intervals from 2,000 subject-bootstrap resamples. Only the raw rows require no conversion-label fitting. All other models use ADNI labels and no target tuning. Single-score logistic evaluations give the same AUCs as direct raw ranking.

| Predictor | AIBL AUC [95% CI] |
|---|---|
| *Raw scores* | |
| Predicted change, $s_\mu$ | $0.534$ |
| Residual scale, $s_\sigma$ | $0.695$ |
| Paired difference, $\Delta_{\sigma-\mu}$ | $+0.162\ [-0.095, 0.411]\ (p = 0.230)$ |
| *Source-supervised models* | |
| Linear mean–scale adapter | $0.746\ [0.610, 0.868]$ |
| Quadratic mean–scale adapter | $0.775\ [0.665, 0.879]$ |
| Clinical covariates | $0.701\ [0.538, 0.851]$ |
| Clinical covariates and mean | $0.742\ [0.573, 0.887]$ |
| Clinical covariates and scale | $0.726\ [0.568, 0.870]$ |

Table 3: AIBL outcome-window extension. Same frozen forecasts; only the outcome window changes, so rows are not independent cohorts. Paired 95% intervals use 10,000 subject-bootstrap resamples. Final column: 20-seed ensemble difference and constant-score control.

| Outcome window | Subj. | Conv. | AUC($s_\sigma$) | AUC($s_\mu$) | $\Delta_{\sigma-\mu}$ [95% CI] | 20-seed / const. |
|---|---|---|---|---|---|---|
| 24 months (primary) | 139 | 14 | 0.695 | 0.534 | $+0.162\ [-0.095, 0.411]$ | $+0.138$ / $-0.042$ |
| 36 months | 139 | 21 | 0.712 | 0.435 | $+0.278\ [\ 0.038, 0.498]$ | $+0.256$ / $+0.051$ |
| 54 months | 119 | 29 | 0.769 | 0.403 | $+0.366\ [\ 0.167, 0.551]$ | $+0.339$ / $+0.079$ |

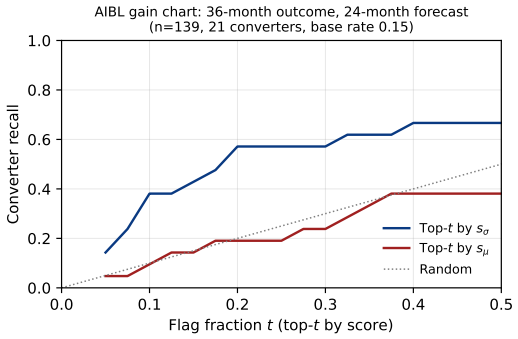



Figure 2: Cumulative gain on AIBL (36-month window, $n = 139$, 21 converters): converter recall among top-$t$ flagged subjects for residual scale (blue) versus predicted change (red); dotted line is random flagging.

**Supervised combination.** The quadratic adapter reaches AUC $0.775$, improving over the mean score by $+0.258$ with paired interval $[0.070, 0.438]$. Its gain over scale alone is $+0.079$ with interval $[-0.024, 0.198]$, so improvement over the weak mean score is established, but improvement over residual scale alone is not. These adapters use source conversion labels and are not part of the label-free score audit.

### 4.3 Within-ADNI forecasting and risk

**Forecasting accuracy versus clinical ranking.** In the separate exact-scan ADNI benchmark, deterministic MLP and heteroscedastic Gaussian heads both achieve approximately $0.082$ mm 12-month MAE on the cortical forecasting task. When used for 24-month conversion risk ranking on 204 held-out subjects, the Gaussian model yields an AUC of $0.784$ for predicted mean change ($s_\mu$) and $0.778$ for learned residual scale ($s_\sigma$). Residual scale is therefore not always the superior

score in-domain. Across ten subject splits and three seeds per split, the average AUCs are $0.750$ for Gaussian mean change, $0.775$ for Gaussian residual scale, and $0.765$ for MLP change (Table 19): similar imaging forecast errors do not determine which downstream output is most discriminative, and the label-free raw scores perform comparably to supervised models trained on a small number of conversion labels(Table 10).

### 4.4 Volumetric control: utility depends on the setting

On the matched 127-subject ADNI holdout (26 converters; Figure 3), volumetric and cortical/graph change reach AUCs $0.755$ and $0.776$. Their paired difference is $-0.021$ with interval $[-0.096, 0.049]$, leaving relative performance unresolved but showing that informative mean-change ranking extends beyond the cortical graph. Volumetric scale is weaker than cortical/graph scale: AUC $0.656$ versus $0.783$, paired difference $-0.127$ with interval $[-0.228, -0.020]$. In the baseline-MCI subset, volumetric change remains informative (AUC $0.722$) while residual scale ($0.548$) and model variability ($0.494$) fall to chance (Table 23), suggesting residual scale partly distinguishes baseline stages rather than forecasting progression within high-risk individuals.

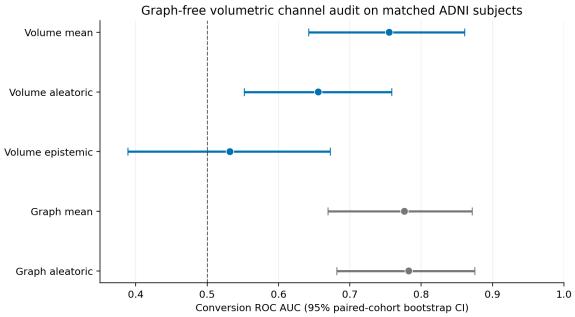



Figure 3: Matched within-ADNI volumetric audit (127 subjects; 26 converters). Raw-score conversion AUCs with 95% intervals (2,000 resamples); dashed line marks chance. Gray graph scores use the same subjects; models differ in representation and training size.

**Training objectives.** NLL-trained residual-scale outputs exceed their own mean scores while MSE-trained model-variability outputs do not; full per-method seed counts are in Section H.5.

### 4.5 Robustness across cohorts and shifts

We examined AIBL-to-ADNI transfer, shifts between ADNI enrollment phases, and 36-, 54-, and 72-month outcomes in OASIS-3 (Section H.9). Residual scale had a higher point-estimate AUC throughout, but the gaps varied: +0.275 for later-to-original ADNI versus +0.003 in reverse, using diagnostic groups represented in both phases. The AIBL-to-ADNI gap was +0.102 at 36 months but varied across training runs. OASIS-3 gaps were +0.121, +0.021, and +0.007, respectively, with all three confidence intervals including zero. These results do not establish a universal advantage for residual scale.

### 4.6 Risk ranking is not uncertainty calibration

A critical conceptual distinction in our evaluation is that diagnostic risk ranking does not imply valid forecast uncertainty. On AIBL, the correlation between predicted residual scale and actual cortical forecasting error is essentially zero (Spearman $\rho = 0.027$, 95% CI $[-0.126, 0.177]$). This decorrelation reveals a key insight: an output can serve as a useful diagnostic ranking score while carrying no detectable information about individual forecast error. Downstream clinical separation and point-forecast residual accuracy are simply two distinct quantities.

# 5 WHEN DO FORECAST-DERIVED RANKINGS TRANSFER?

The empirical comparisons concern rankings, not only forecast errors. For independent converter and non-converter scores $S_1, S_0$, $\mathrm{AUC}(S) = \Pr(S_1 > S_0) + \frac{1}{2}\Pr(S_1 = S_0)$, with orientation held fixed. The following simplified models distinguish failure and stability; the parameter-sensitivity explanation is in Section B.5. They do not identify the cause of a cohort difference from AUC alone.

## 5.1 MEAN-SCORE SEVERITY ALIGNMENT

To see how cohort shift can break a mean score, we track its alignment with unobserved disease severity $Z$. Let $Y \in \{0, 1\}$ denote future conversion with class-conditional severity $Z \mid Y = y \sim \mathcal{N}(\mu_y, \sigma_Z^2)$ and $\Delta\mu = \mu_1 - \mu_0 > 0$. On the target cohort, $\widetilde{M} = b + \eta Z + \varepsilon_\perp$ with $\varepsilon_\perp \sim \mathcal{N}(0, \sigma_\perp^2)$: $\eta$ measures severity tracking while $\sigma_\perp^2$ captures unrelated score variation.

**Proposition 1** (Mean-score AUC)**.** *Under these assumptions, if the score variance within classes is non-zero ($\eta^2\sigma_Z^2 + \sigma_\perp^2 > 0$), the ranking performance on the target cohort is given analytically by:*

$$\mathrm{AUC}(\widetilde{M}) = \Phi\left(\frac{\eta\Delta\mu}{\sqrt{2(\eta^2\sigma_Z^2 + \sigma_\perp^2)}}\right), \tag{2}$$

*where $\Phi$ denotes the standard normal cumulative distribution function. The resulting AUC is greater than, equal to, or below $0.5$ (chance level) depending entirely on whether the severity slope $\eta$ is positive, zero, or negative.*

Pairwise ranking cancels any common offset $b$ (Section B.2). In Equation (2), class separation competes with within-class variability: negative alignment ($\eta < 0$) reverses ranks, while decorrelation ($\eta \approx 0$) drops discrimination to chance.

To illustrate how transfer shift can drive ranking performance down to chance, consider a source model score $M = \alpha Z + \varepsilon_M$ (with $\alpha > 0$) subjected to a cohort perturbation:

$$\widetilde{M} = (1 - \rho)M + \rho(\gamma Z + \delta), \qquad \delta = \kappa(Z - m_Z) + \xi, \qquad \rho \in [0, 1],$$

where $\rho$ controls the shift severity and $\delta$ models cohort-specific discrepancies that scale with disease severity. The effective relationship between the target score and disease severity becomes:

$$\eta = (1 - \rho)\alpha + \rho(\gamma + \kappa), \qquad \sigma_\perp^2 = (1 - \rho)^2\sigma_M^2 + \rho^2\sigma_\xi^2.$$

With an opposing cohort trend ($\gamma + \kappa < 0$), the slope cancels at $\rho^\star = \alpha/[\alpha - (\gamma + \kappa)]$, where $\mathrm{AUC}(\widetilde{M}) = 0.5$ regardless of source performance (sign regimes: Sections B.3 and J.2).

While this Gaussian model is an analytical idealization of non-negative change magnitudes, a retrospective plug-in check using baseline diagnosis as a severity proxy reproduces the near-chance mean-score AUC. Because that check uses target conversion labels for its class-conditional moments, the numerical agreement does not validate the Gaussian assumptions or identify the cancellation mechanism.

## 5.2 CLASS-CONDITIONAL STABILITY

A score's ranking survives shift if its distribution within each outcome class barely changes, quantified by total variation, $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$.

**Lemma 1** (AUC perturbation under total variation)**.** *For any fixed scoring function $S$, the change in AUC between the source and target cohorts is bounded by the total variation distance between their class-conditional score distributions:*

$$|\mathrm{AUC}^{\mathrm{tgt}}(S) - \mathrm{AUC}^{\mathrm{src}}(S)| \leq \sum_{y \in \{0,1\}} \mathrm{TV}(P_{S|Y=y}, \widetilde{P}_{S|Y=y}).$$

This bound holds universally without requiring Gaussian distributions or a latent severity structure. To connect it directly to clinical severity, we assume that the learned residual scale score $U$ follows a consistent conditional distribution given a patient's true underlying severity $z$:

$$P(U \in A \mid Z = z, Y = y) = \widetilde{P}(U \in A \mid Z = z, Y = y) = K(A \mid z).$$

Frozen weights do not guarantee this, since cross-site acquisition differences can shift inputs at fixed disease severity.

**Proposition 2** (Scale stability)**.** *Under this shared conditional mapping $K(A \mid z)$, if the shift in the severity distribution within each outcome class is bounded by $\mathrm{TV}(P_{Z|Y=y}, \widetilde{P}_{Z|Y=y}) \le \varepsilon_y$ for $y \in \{0, 1\}$, then the transfer gap in ranking performance satisfies:*

$$|\mathrm{AUC}^{\mathrm{tgt}}(U) - \mathrm{AUC}^{\mathrm{src}}(U)| \le \varepsilon_1 + \varepsilon_0.$$

By the data-processing inequality, the score contracts latent-severity perturbations (Section B.4), but stability must hold *class-conditionally*: marginal stability is insufficient, since converter and non-converter distributions could invert under an unchanged population average. This result provides a sufficient-condition explanation consistent with the retained AIBL ranking of residual scale: if its severity mapping within each class stays relatively invariant, the ranking gap stays bounded.

**Posterior-weight sensitivity.** A third candidate score is model variability, measured across ensemble members or posterior weight samples. Under standard asymptotics, this variability reflects parameter sensitivity, that is, how sharply predictions change under small weight perturbations, rather than biological uncertainty about disease progression. Nothing guarantees that this sensitivity tracks a specific patient's clinical risk, and it has no theoretical reason to preserve conversion rankings under cohort shift (Sections B.5 and B.6).

## 6 Discussion and conclusion

We demonstrate this divergence empirically. When transferred from ADNI to AIBL, learned residual scale retains its ability to rank conversion risk, while predicted change performs near chance. In a separate in-domain ADNI control using regional brain volumes, the ordering reverses: predicted change is more discriminative than residual scale. Residual scale is therefore not an intrinsically superior measure of uncertainty.

Our theoretical analysis formalizes these behaviors as conditional transfer mechanisms rather than uniform failure rules. In contrast, auxiliary outputs such as residual scale transfer effectively if their class-conditional distributions remain stable under shift, whereas parameter-sensitivity scores carry no such guarantee.

**Limitations.** *Statistical precision:* To address the event sparsity inherent in the strict 24-month window (14 AIBL converters), we extended clinical evaluation to 36 and 54 months, observing 21 and 29 converters with paired confidence intervals that consistently exclude zero in favor of residual scale. Even with these extensions, however, statistical precision remains bounded: the longitudinal evaluations share the same cohort, inverse-direction transfer disagrees in sign across training seeds, and OASIS-3 features low event rates.

*Measurement and representation:* Requiring follow-up MRI scans selects a specific longitudinal cohort, while visit timing and ADNI source runs differ. Furthermore, the volumetric control alters inputs, network architecture, and training size simultaneously without an external volume test. It demonstrates that the audit protocol generalizes beyond cortical graphs, not that any isolated architectural feature drives performance.

*Clinical and theoretical scope:* Conversion ranking establishes neither calibrated clinical risk nor validated forecast uncertainty; moreover, target recalibration did not restore nominal error coverage. Larger independent cohorts, prospective validation, and formal decision-utility assessments remain necessary before potential clinical application.

**Conclusion.** We do not claim our results establish a universal advantage of uncertainty, a calibrated risk model, or a clinical recommendation. They present a narrower practice: evaluate each forecast output against its intended outcome and population, and report the conditions under which each output remains informative.

## REFERENCES


Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern Hopfield networks. In *Advances in Neural Information Processing Systems*, 2023.

Narmina Baghirova, Duy-Thanh Vũ, Duy-Cat Can, Christelle Schneuwly Diaz, Julien Bodlet, Guillaume Blanc, Georgi Hrusanov, Bernard Ries, and Oliver Y. Chén. Explainable graph-theoretical machine learning: with application to alzheimer's disease prediction, 2025. URL https://arxiv.org/abs/2503.16286.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

Rémi Cuingnet, Emilie Gerardin, Jérôme Tessieras, Guillaume Auzias, Stéphane Lehéricy, Marie-Odile Habert, Marie Chupin, Habib Benali, and Olivier Colliot. 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.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux – effortless bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

B. Fischl. Freesurfer. *NeuroImage*, 62(2):774–781, 2012. doi: 10.1016/j.neuroimage.2012.01.021.

Hubert M. Fonteijn, Marc Modat, Matthew J. Clarkson, Josephine Barnes, Manja Lehmann, Nicola Z. Hobbs, Rachael I. Scahill, Sarah J. Tabrizi, Sebastien Ourselin, Nick C. Fox, and Daniel C. Alexander. 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.

Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Clifford R Jack, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. 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.

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

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Hao Li, Bowen Deng, Chang Xu, ZhiYuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, and Jiang Bian. Mira: Medical time series foundation model for real-world health data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://neurips.cc/virtual/2025/loc/san-diego/poster/119424.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Conformal prediction with temporal quantile adjustments. In *Advances in Neural Information Processing Systems*, 2022.

Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. Landmark-based deep multi-instance learning for brain disease diagnosis. *Medical Image Analysis*, 43:157–168, 2018.

Marco Lorenzi, Maurizio Filippone, Giovanni B. Frisoni, Daniel C. Alexander, and Sebastien Ourselin. Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease. *NeuroImage*, 190:56–68, 2019.

Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's disease neuroimaging initiative. *Alzheimer's & Dementia*, 1(1):55–66, 2005.

Sebastian Palmqvist, Michael Schöll, Olof Strandberg, Niklas Mattsson, Erik Stomrud, Henrik Zetterberg, Kaj Blennow, Susan Landau, William Jagust, and Oskar Hansson. Earliest accumulation of $\beta$-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.

Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero, Ben Glocker, and Daniel Rueckert. Disease prediction using graph convolutional networks: Application to autism spectrum disorder and alzheimer's disease. *Medical Image Analysis*, 48:117–130, 2018.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, 2019.

Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Avram J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018.

C. Schneuwly Diaz, D. T. Vũ, J. Bodelet, D-C. Can, G. Blanc, H. Jiang, L. Yao, G. Pantaleo, and O. Y. Chén. Optimus: Predicting multivariate outcomes in alzheimer's disease using multi-modal data amidst missing values. *IEEE Transactions on Biomedical Engineering*, 2025.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Simeon Spasov, Luca Passamonti, Andrea Duggento, Pietro Liò, and Nicola Toschi. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease. *Neuroimage*, 189:276–287, 2019.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Sophia Huiwen Sun and Rose Yu. Conformal prediction for time-series forecasting with change points. In *Advances in Neural Information Processing Systems*, 2025.

## AI USE STATEMENT

Generative AI tools assisted with revising portions of the manuscript, editorial analysis, manuscript restructuring, clarity revisions, LaTeX formatting. They also provided feedback on how results and limitations were presented. Their use was not limited to sentence-level grammar correction, but also phrase suggestions. The authors reviewed the AI-assisted material and take responsibility for the final text, numerical results, interpretations, and submission artifacts.

## ETHICS STATEMENT

This work is a secondary analysis of existing de-identified neuroimaging cohorts. It involved no new participant recruitment, data collection, or intervention. Data were accessed under the cohorts' established data-use agreements, and participant-level imaging data are not redistributed. The models are presented as research tools, not clinically deployable systems.

## REPRODUCIBILITY STATEMENT

The appendix describes the cohort definitions, forecasting models, score construction, and additional analyses. Results from different subject splits, forecast horizons, training runs, and resampling procedures are identified separately because they are not interchangeable. Code, trained checkpoints, and exact subject-level splits are planned for release upon acceptance, subject to the applicable cohort data-use agreements.

## CODE AVAILABILITY

An anonymized repository for this work is available at `https://anonymous.4open.science/r/Uncertainty-Audit/README.md`.

# SUPPLEMENTARY MATERIAL

This supplement contains: (A) a notation reference; (B) full proofs of Lemma 1 and Propositions 1 and 2, including the additional Gaussian and class-conditional stability assumptions needed for the stated AUC identities to be exact, and a sensitivity-to-assumptions analysis; (C) the controlled cohort-shift simulation that produces Figure 1b (D) the retrospective AIBL moment check (Table 5); (E) cohort, preprocessing, and labelling details; (F) architecture, training, and adapter hyperparameters; (G) statistical-testing methodology; (H) additional empirical results that did not fit in the main paper (in-domain head comparison, posterior-predictive controls, label-scarcity sweep, inverse-direction transfer, OASIS-3 multi-horizon probe, backbone ablations, repeated-split robustness, calibration–ranking decomposition, and the pooled MCI analysis); (I) limitations, broader impact, and reproducibility.

**How to read this supplement.** The core supporting material is (F) architecture and training, (G) statistical protocol, and (B) exact theorem statements; (H) collects secondary and exploratory controls that qualify but do not alter the main-text conclusions. Throughout, $s_\mu$ denotes predicted change, $s_\sigma$ residual scale, and $s_{\mathrm{PW}}$ model variability.

## A NOTATION

Table 4 summarises the symbols used in the theoretical analysis (Section 5) and below. We work with a scalar latent severity $Z$ purely as an analytical surrogate for the real disease state; the empirical results in Section 4 do not require this scalar reduction. The proofs below use the centered residual decomposition required for the exact Gaussian AUC formula:

$$\delta = \kappa(Z - m_Z) + \xi, \qquad \xi \perp\!\!\!\perp (Z, Y, \varepsilon_M),$$

where $m_Z$ is a fixed centring constant. The constant term $-\rho\kappa m_Z$ induced in $\widetilde{M}$ does not affect ranks and therefore does not affect AUC.

Table 4: Notation used in the theoretical analysis.

| Symbol | Meaning |
|---|---|
| $Y \in \{0, 1\}$ | 24-month conversion indicator |
| $Z \in \mathbb{R}$ | scalar latent severity |
| $\mu_y, \sigma_Z^2$ | class-conditional mean and variance of $Z$ |
| $\Delta\mu$ | class-mean separation $\mu_1 - \mu_0 > 0$ |
| $M$ | source mean-change score $\alpha Z + \varepsilon_M$ |
| $\widetilde{M}$ | target mean-change score under cohort shift |
| $\rho \in [0, 1]$ | cohort-shift strength |
| $\gamma$ | severity-aligned component of cohort perturbation |
| $\delta$ | cohort residual term in the target mean score |
| $c$ | common within-class covariance $\mathrm{Cov}(\delta, Z \mid Y = y)$ |
| $\kappa$ | within-class slope $c/\sigma_Z^2$ in $\delta = \kappa(Z - m_Z) + \xi$ |
| $\xi$ | centred Gaussian residual independent of $(Z, Y, \varepsilon_M)$ |
| $\eta$ | effective slope $(1 - \rho)\alpha + \rho(\gamma + \kappa)$ |
| $\sigma_\perp^2$ | residual variance $(1 - \rho)^2\sigma_M^2 + \rho^2\sigma_\xi^2$ |
| $U$ | uncertainty score $u(Z) + \varepsilon_U$ |
| $\mathrm{TV}(P, Q)$ | total variation distance, $\sup_A \lvert P(A) - Q(A)\rvert$ |
| $\Phi$ | standard normal cumulative distribution function |

## B PROOFS AND THEORETICAL COMPLEMENTS

This section gives the mathematically exact versions of the theoretical claims used in the main paper. Two points are important. First, the closed-form AUC identity for the mean score is exact only under Gaussian residual assumptions; without them it is a Gaussian projection approximation. Second, the stability of the residual-scale score requires class-conditional control of the latent severity distribution. A marginal bound on $P_Z$ alone does not control AUC.

### B.1 Proof of Lemma 1

*Proof.* Write $P_y := P_{S|Y=y}$ and $Q_y := \widetilde{P}_{S|Y=y}$ for $y \in \{0,1\}$. The Mann–Whitney representation of AUC for a score $S$ is

$$\mathrm{AUC}(S) \;=\; \mathbb{E}\big[\, f(S_1, S_0)\,\big], \qquad f(s_1, s_0) \;:=\; \mathbf{1}\{s_1 > s_0\} + \tfrac{1}{2}\mathbf{1}\{s_1 = s_0\},$$

with $S_y \sim P_{S|Y=y}$ independently. The function $f$ is bounded in $[0,1]$.

Using a telescoping decomposition of product measures,

$$Q_1 \otimes Q_0 - P_1 \otimes P_0 = (Q_1 - P_1) \otimes Q_0 + P_1 \otimes (Q_0 - P_0).$$

Therefore

$$\big|\mathrm{AUC}^{\mathrm{tgt}}(S) - \mathrm{AUC}^{\mathrm{src}}(S)\big| \le \Big| \textstyle\int f \, d\big[(Q_1 - P_1) \otimes Q_0\big] \Big| + \Big| \textstyle\int f \, d\big[P_1 \otimes (Q_0 - P_0)\big] \Big|.$$

For the first term, Fubini's theorem and the variational characterisation of total variation give

$$\Big| \textstyle\int f \, d\big[(Q_1 - P_1) \otimes Q_0\big] \Big| \le \mathrm{TV}(P_1, Q_1),$$

because $f(\cdot, s_0) \in [0,1]$ for every $s_0$. The second term is bounded analogously by $\mathrm{TV}(P_0, Q_0)$. Hence

$$\big|\mathrm{AUC}^{\mathrm{tgt}}(S) - \mathrm{AUC}^{\mathrm{src}}(S)\big| \le \mathrm{TV}(P_1, Q_1) + \mathrm{TV}(P_0, Q_0).$$

□

**Remark 1** (Range of the bound)**.** *No AUC gap can exceed one, so the additive bound can be capped at one. We use it as a sufficient stability bound and make no sharpness claim.*

### B.2 Proof of Proposition 1

We state the exact form of the assumptions used in the proof. Assume

$$Z \mid Y = y \sim \mathcal{N}(\mu_y, \sigma_Z^2), \qquad \mu_1 > \mu_0,$$

and

$$M = \alpha Z + \varepsilon_M, \qquad \varepsilon_M \sim \mathcal{N}(0, \sigma_M^2), \qquad \varepsilon_M \perp\!\!\!\perp (Z, Y).$$

For the cohort residual, assume the centred Gaussian decomposition

$$\delta = \kappa(Z - m_Z) + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma_\xi^2), \qquad \xi \perp\!\!\!\perp (Z, Y, \varepsilon_M),$$

where $m_Z$ is fixed. Equivalently, within each class $\mathrm{Cov}(\delta, Z \mid Y = y) = \kappa\sigma_Z^2$. If one writes $\sigma_\delta^2 := \mathrm{Var}(\delta \mid Y = y)$, then $\sigma_\xi^2 = \sigma_\delta^2 - \kappa^2\sigma_Z^2$.

*Proof.* Substitute $M = \alpha Z + \varepsilon_M$ and $\delta = \kappa(Z - m_Z) + \xi$ into the target score $\widetilde{M} = (1-\rho)M + \rho(\gamma Z + \delta)$. This gives

$$\widetilde{M} = \underbrace{\big[(1-\rho)\alpha + \rho(\gamma + \kappa)\big]}_{\eta} Z + (1-\rho)\varepsilon_M + \rho\xi - \rho\kappa m_Z.$$

The last term is a constant and therefore has no effect on AUC. Define

$$R := (1-\rho)\varepsilon_M + \rho\xi, \qquad \sigma_\perp^2 := (1-\rho)^2\sigma_M^2 + \rho^2\sigma_\xi^2.$$

By the independence assumptions, $R \sim \mathcal{N}(0, \sigma_\perp^2)$ and $R \perp\!\!\!\perp (Z, Y)$. Thus, conditional on $Y = y$,

$$\widetilde{M} \mid Y = y \sim \mathcal{N}\big(\eta\mu_y - \rho\kappa m_Z,\, \eta^2\sigma_Z^2 + \sigma_\perp^2\big)\,.$$

The constant $-\rho\kappa m_Z$ is the same in both classes. Therefore, for independent draws $\widetilde{M}_1 \sim \widetilde{M} \mid Y = 1$ and $\widetilde{M}_0 \sim \widetilde{M} \mid Y = 0$,

$$\widetilde{M}_1 - \widetilde{M}_0 \sim \mathcal{N}\big(\eta\Delta\mu,\, 2(\eta^2\sigma_Z^2 + \sigma_\perp^2)\big)\,.$$

Since the distributions are continuous, ties have probability zero and

$$\mathrm{AUC}(\widetilde{M}) = \mathbb{P}(\widetilde{M}_1 > \widetilde{M}_0) = \Phi\left(\frac{\eta\Delta\mu}{\sqrt{2(\eta^2\sigma_Z^2 + \sigma_\perp^2)}}\right).$$

Because $\Phi$ is strictly increasing and $\Delta\mu > 0$, $\mathrm{AUC}(\widetilde{M}) = \frac{1}{2} \iff \eta = 0 \iff (1-\rho)\alpha + \rho(\gamma + \kappa) = 0$. □

**Remark 2** (Why the Gaussian residual assumptions are needed)**.** *A linear projection of $\delta$ on $Z$ only gives an uncorrelated residual; it does not give independence. The closed-form AUC identity above is exact because the residuals are assumed Gaussian and independent after the centred decomposition. Without these assumptions, the same calculation should be described as a Gaussian working approximation rather than as an exact AUC identity.*

### B.3 THREE REGIMES

Let $\eta(\rho) = \alpha + \rho\big[(\gamma + \kappa) - \alpha\big], \rho \in [0, 1]$. The sign of $\eta$ controls the orientation of the raw source-oriented mean score, while the actual AUC also depends on the residual variance $\sigma_\perp^2$.

**(i) Aligned shift** $(\gamma + \kappa \geq 0)$**.** The effective slope does not change sign on $[0, 1]$, so the raw mean score does not reverse orientation. This does not imply that its AUC remains close to the source AUC: the signal-to-noise ratio may still decrease if $\eta$ shrinks or if $\sigma_\perp^2$ grows.

**(ii) Orthogonal shift** $(\gamma + \kappa = 0)$**.** Here $\eta(\rho) = (1 - \rho)\alpha$, which reaches zero only at $\rho = 1$. The mean-score AUC degrades through both the numerator $\eta\Delta\mu$ and the denominator $\sqrt{2(\eta^2\sigma_Z^2 + \sigma_\perp^2)}$.

**(iii) Anti-aligned shift** $(\gamma + \kappa < 0)$**.** There is a unique zero $\rho^\star = \alpha/[\alpha - (\gamma + \kappa)] \in (0, 1)$. At $\rho = \rho^\star$, the raw target mean score is exactly at chance. For $\rho > \rho^\star$, $\eta < 0$ and the raw score is anti-oriented: $\mathrm{AUC}(\widetilde{M}) < 1/2$. This is rank reversal, not absence of rank information; changing the sign of the score would recover an AUC above $1/2$. In the intended application, however, the biological orientation of $s_\mu$ is fixed in advance, so the raw source-oriented score is the relevant transfer quantity.

### B.4 PROOF OF PROPOSITION 2

The stability statement is class-conditional. Assume the score function is fixed and, separately, that the conditional law of the score given severity is the same kernel in both cohorts. Assume also that for $y \in \{0, 1\}$, $\mathrm{TV}(P_{Z|Y=y}, \widetilde{P}_{Z|Y=y}) \leq \varepsilon_y$. Then $|\mathrm{AUC}^{\mathrm{tgt}}(U) - \mathrm{AUC}^{\mathrm{src}}(U)| \leq \varepsilon_1 + \varepsilon_0$. In particular, if $\varepsilon_0, \varepsilon_1 \leq \varepsilon$, the gap is at most $2\varepsilon$.

*Proof.* Let $K$ be the Markov kernel $z \mapsto u(z) + \varepsilon_U$: for every Borel set $A \subseteq \mathbb{R}$, $K(A \mid z) = \mathbb{P}\{u(z) + \varepsilon_U \in A\}$. By the conditional-invariance assumption, the same kernel $K$ is used in source and target cohorts; frozen weights alone do not imply this. The data-processing inequality for total variation gives $\mathrm{TV}(K \circ P, K \circ Q) \leq \mathrm{TV}(P, Q)$ for any probability measures $P, Q$. Applying this with $P = P_{Z|Y=y}$ and $Q = \widetilde{P}_{Z|Y=y}$ yields $\mathrm{TV}(P_{U|Y=y}, \widetilde{P}_{U|Y=y}) \leq \varepsilon_y$. Combining with Lemma 1 gives the result. □

**Remark 3** (A marginal TV bound is insufficient)**.** *The condition $\mathrm{TV}(P_Z, \widetilde{P}_Z) \leq \varepsilon$ by itself does not control AUC. With equal class priors let the source satisfy $Z \mid Y = 1 = \delta_1$ and $Z \mid Y = 0 = \delta_0$, and let the target swap the two class-conditionals. The marginal distribution of $Z$ is unchanged, so $\mathrm{TV}(P_Z, \widetilde{P}_Z) = 0$, but the AUC of the score $U = Z$ changes from $1$ to $0$.*

### B.5 POSTERIOR-WEIGHT RANKING

This section states the third theoretical explanation summarized in Section 5: in a regular parametric regime, posterior-weight variability tracks local parameter sensitivity rather than conditional outcome variance.

For a scalar mean predictor $g_\theta(x)$ trained on $n$ independent forecasting pairs, define $\hat{\sigma}_{\mathrm{PW}}(x) = [\mathrm{Var}_{\theta|\mathcal{D}_n}\{g_\theta(x)\}]^{1/2}$.

**Proposition 3** (Ranking by parameter sensitivity)**.** *In a regular, well-specified parametric model with consistent estimation and a valid mean-square posterior delta method, assume $\hat{\theta}_n \to \theta^\star$ and*

$$\sqrt{n}(\theta - \hat{\theta}_n) \mid \mathcal{D}_n \implies \mathcal{N}(0, I(\theta^\star)^{-1})$$

*with second-moment convergence. Assume the mean predictor is continuously differentiable near $\theta^\star$ and its posterior mean-square Taylor remainder is $o_P(n^{-1})$. Then*

$$\hat{\sigma}_{\mathrm{PW}}(x) = n^{-1/2} v(x) + o_P(n^{-1/2}), \qquad v(x) = [\nabla g_{\theta^\star}(x)^\top I(\theta^\star)^{-1} \nabla g_{\theta^\star}(x)]^{1/2}.$$

*This holds jointly on a fixed finite in-support test set. If its $v$ values are distinct, the posterior-scale rankings agree with the $v$ rankings with probability tending to one.*

Here $I$ is nonsingular per-observation Fisher information. The function $v$ measures local parameter sensitivity, which may itself align with risk. Shrinking scale does not imply shrinking AUC, and tied leading terms do not determine rankings. In contrast, a correctly specified heteroscedastic likelihood targets conditional outcome variance. Neither quantity is guaranteed to rank conversion. The proposition is not a theorem for arbitrary deep-network approximations or out-of-support inputs.

The finite-set ranking conclusion additionally requires distinct leading sensitivity values; ties leave lower-order terms to decide rank. Neither small magnitudes nor flat leading terms imply chance AUC. Parameter sensitivity and conditional residual variance may each align with risk, so their empirical ranking is not a consequence of contraction alone.

### B.6 Proof of Proposition 3

We prove the posterior-weight statement in the regular parametric regime used in Proposition 3. The result is a posterior delta-method calculation followed by a rank-continuity argument. It should not be read as a theorem for arbitrary deep-network posteriors; rather, it describes the regime in which posterior-weight variability behaves as a local parametric sensitivity score.

*Proof.* Let

$$Z_n := \sqrt{n}(\theta - \hat{\theta}_n), \qquad \Pi_n(\cdot) := \mathcal{L}(Z_n \mid \mathcal{D}_n)$$

denote the centred and rescaled posterior. By assumption, $\Pi_n$ converges to $\mathcal{N}(0, I(\theta^\star)^{-1})$ with convergence of second moments. Equivalently, for every fixed vector $a$,

$$\mathrm{Var}_{\Pi_n}(a^\top Z_n) \to a^\top I(\theta^\star)^{-1} a$$

in $P^\star$-probability, and the posterior second moments remain asymptotically bounded.

Fix a test point $x$ in the training support and write

$$a_n(x) := \nabla_\theta g_{\hat{\theta}_n}(x), \qquad a(x) := \nabla_\theta g_{\theta^\star}(x).$$

By the differentiability assumption and consistency of $\hat{\theta}_n$, $a_n(x) \to a(x)$ in probability. Taylor expansion around $\hat{\theta}_n$ gives, conditionally on $\mathcal{D}_n$,

$$g_\theta(x) - g_{\hat{\theta}_n}(x) = a_n(x)^\top (\theta - \hat{\theta}_n) + r_n(\theta, x),$$

where the posterior $L^2$-remainder satisfies

$$\mathbb{E}_{\theta|\mathcal{D}_n}\left[r_n(\theta, x)^2\right] = o_P(n^{-1}).$$

Since $\theta - \hat{\theta}_n = n^{-1/2} Z_n$, we obtain

$$\mathrm{Var}_{\theta|\mathcal{D}_n}\{g_\theta(x)\} = n^{-1} \mathrm{Var}_{\Pi_n}\{a_n(x)^\top Z_n\} + o_P(n^{-1}).$$

The second-moment form of the Bernstein–von Mises/delta-method assumption therefore yields

$$\mathrm{Var}_{\theta|\mathcal{D}_n}\{g_\theta(x)\} = n^{-1} a(x)^\top I(\theta^\star)^{-1} a(x) + o_P(n^{-1}).$$

Taking square roots gives

$$\hat{\sigma}_{\mathrm{PW}}(x) = n^{-1/2} \left[a(x)^\top I(\theta^\star)^{-1} a(x)\right]^{1/2} + o_P(n^{-1/2}) = n^{-1/2} v(x) + o_P(n^{-1/2}).$$

For a fixed finite test set $\{(X_j, Y_j)\}_{j=1}^m$, the same argument applied to each $X_j$, followed by a union bound, gives the joint expansion

$$\max_{1 \le j \le m} \left|\sqrt{n}\, \hat{\sigma}_{\mathrm{PW}}(X_j) - v(X_j)\right| \xrightarrow{P} 0.$$

It remains to pass from pointwise convergence to rank convergence. Suppose that $v(X_1), \dots, v(X_m)$ are pairwise distinct and define

$$\Delta_m := \min_{j \neq k} |v(X_j) - v(X_k)| > 0.$$

With probability tending to one,

$$\max_j \left| \sqrt{n}\, \hat{\sigma}_{\mathrm{PW}}(X_j) - v(X_j) \right| < \Delta_m / 3.$$

On this event, the ordering of $\{\hat{\sigma}_{\mathrm{PW}}(X_j)\}_{j=1}^m$ is exactly the ordering of $\{v(X_j)\}_{j=1}^m$, because multiplication by the positive constant $\sqrt{n}$ does not change ranks. Hence the Spearman correlation with the fixed labels $\{Y_j\}_{j=1}^m$ is eventually identical to the Spearman correlation obtained from $v$:

$$\rho_{\mathrm{S}}\big(\{\hat{\sigma}_{\mathrm{PW}}(X_j)\}_{j=1}^m, \{Y_j\}_{j=1}^m\big) \xrightarrow{P} \rho_{\mathrm{S}}\big(\{v(X_j)\}_{j=1}^m, \{Y_j\}_{j=1}^m\big).$$

Finally, $v$ is determined by the forecasting problem: the population parameter $\theta^\star$, the Fisher information $I(\theta^\star)$, and the gradient of the mean head $g_\theta$. These are determined by $(P_{X,X'}, \ell, \text{architecture})$, not by supervised optimisation against the downstream conversion label $Y$. Therefore posterior-weight uncertainty asymptotically ranks subjects by local parametric sensitivity of the mean forecaster. If this sensitivity score is rank-uncorrelated with conversion on the target cohort, then the limiting rank correlation is zero; if it happens to align with conversion, nonzero limiting rank correlation is possible. □

**Remark 4** (No claim of $n^{-1/2}$ rank decay)**.** *The proposition concerns the scale and the limiting ordering of posterior-weight variability. It does not imply that the rank correlation itself is $O(n^{-1/2})$. Spearman correlation is invariant under positive rescaling, so the factor $n^{-1/2}$ disappears from the ranking. The limiting rank correlation is controlled by the deterministic sensitivity score $v(X)$, not by the posterior scale.*

**Remark 5** (Ties in the sensitivity score)**.** *The pairwise-distinctness condition on $v(X_1), \dots, v(X_m)$ is used only to make the rank map continuous. If $v$ has ties on the test set, the leading-order expansion still holds, but the limiting ranks may be determined by lower-order terms. In particular, a constant leading sensitivity score does not by itself determine the limiting Spearman correlation of the finite-sample posterior-weight scores. This is why Proposition 3 states the rank limit under a no-ties condition and phrases the zero-correlation conclusion in terms of $v(X)$ being rank-uncorrelated with $Y$.*

**Remark 6** (Relation to aleatoric uncertainty)**.** *The result separates posterior-weight variability from heteroscedastic residual uncertainty. Posterior-weight variability ranks local sensitivity of the mean head. A correctly specified heteroscedastic likelihood, by contrast, targets the conditional residual variance $\mathrm{Var}(X' \mid X = x)$. Thus a residual-scale head provides a designed route to residual-variance ranking, whereas posterior-weight variability provides such a ranking only when local sensitivity happens to align with severity.*

**Remark 7** (Deep models)**.** *The assumptions above are classical regular-parametric assumptions. They are reasonable idealisations for last-layer Laplace, SNGP-type linearised regimes, or other posterior approximations whose purpose is to approximate a local Gaussian posterior. They need not hold for fully Bayesian deep networks, multimodal posteriors, or ensembles. Failure of the assumptions does not imply better severity ranking; it simply means that the posterior-weight score is no longer characterised by the parametric sensitivity limit above.*

### B.7 Sensitivity to assumption violations

The exact mean-score identity uses Gaussian class-conditionals, Gaussian residuals, and an affine target perturbation. The residual-scale bound uses a frozen score kernel and class-conditional stability of the latent severity distribution.

**Non-Gaussian severity.** If $Z \mid Y = y$ is not Gaussian, $\widetilde{M} \mid Y = y$ need not be Gaussian. Equation (2) should then be interpreted as a Gaussian projection or moment-matched approximation. The orientation statement remains informative: when $\eta = 0$, the class-dependent linear component of the mean-score score vanishes, leaving only residual terms.

**Non-independent residuals.** If the residual in $\delta$ is only uncorrelated with $Z$, not independent of it, Step 3 of the mean-score proof fails. The variance expression remains a projection identity, but the AUC may differ from the Gaussian closed form.

**Non-linear mean score.** If the source mean score is $M = g(Z) + \varepsilon_M$ with $g$ smooth and monotone, the linear theory applies locally after replacing $g$ by its first-order expansion. A global closed form would require assumptions on the full shape of $g$ and on the target perturbation.

**Non-linear cohort shift.** For a non-linear target perturbation, the coefficient of the projection onto $Z$ still determines the leading orientation of the raw mean score, while the non-linear remainder contributes to the residual variance and higher-order shape of the score distribution.

**Non-monotone uncertainty.** The data-processing proof does not require $u$ to be monotone; it only requires the same frozen kernel in both cohorts. Monotonicity matters for source usefulness: if $u$ is non-monotone in a way that makes $\mathrm{AUC}^{\mathrm{src}}(U) \approx 1/2$, the TV bound merely says that a useless score remains stable.

# C Cohort-shift simulation and quantitative bounds on real shift

## C.1 Simulation that produces Figure 1b

The parameters below are chosen so that the analytic AUC formula gives a source mean-score AUC near $0.829$, a target mean-score AUC near $0.575$ at $\rho = 0.8$, and chance-level performance at $\rho = 1$, while the residual-scale AUC remains near $0.747$ throughout the shift sweep.

**Generative process.** For each subject $i \in \{1, \dots, N_{\mathrm{sim}}\}$ with $N_{\mathrm{sim}} = 4000$ and converter base rate $\pi = 0.25$, sample

$$Y_i \sim \mathrm{Bernoulli}(\pi), \qquad Z_i \mid Y_i \sim \mathcal{N}(\mu_{Y_i}, \sigma_Z^2),$$
$$\varepsilon_{M,i} \sim \mathcal{N}(0, \sigma_M^2), \qquad \xi_i \sim \mathcal{N}(0, \sigma_\xi^2),$$
$$\varepsilon_{U,i} \sim \mathcal{N}(0, \sigma_U^2).$$

All residual terms are mutually independent and independent of $(Z_i, Y_i)$. The source and target mean-score scores are $M_i = \alpha Z_i + \varepsilon_{M,i}$, $\widetilde{M}_i = (1-\rho)M_i + \rho\{\gamma Z_i + \kappa(Z_i - m_Z) + \xi_i\}$. The uncertainty score is $U_i = u(Z_i) + \varepsilon_{U,i}$. We set $m_Z = \mathbb{E}[Z]$; changing $m_Z$ only adds a constant to $\widetilde{M}$ and therefore leaves AUC unchanged.

**Parameter values.** $\mu_0 = 0$, $\mu_1 = 1.5$, $\sigma_Z^2 = 1$, $\alpha = 1$, $\sigma_M^2 = 0.25$, $\sigma_\xi^2 = 1.83$. For the target perturbation $\gamma = -0.404$, $\kappa = 0.4$, so that $\gamma + \kappa = -0.004$ and the zero crossing is $\rho^\star = \alpha/[\alpha - (\gamma + \kappa)] \approx 0.996$. This places the simulation in a mildly anti-aligned regime: the raw mean score approaches chance by $\rho = 1$ without becoming strongly anti-predictive. For the residual-scale score we use the monotone score $u(z) = 0.4 + 0.4z$ and $\sigma_U^2 = 0.25$.

**Analytic AUC values.** With the values above, Equation (2) gives $\mathrm{AUC}(\widetilde{M}_{\rho=0}) = 0.829$, $\mathrm{AUC}(\widetilde{M}_{\rho=0.8}) = 0.575$, $\mathrm{AUC}(\widetilde{M}_{\rho=1}) = 0.499$. The residual-scale score has the constant analytic AUC $\mathrm{AUC}(U) = \Phi\big(0.4(\mu_1 - \mu_0)/\sqrt{2(0.4^2\sigma_Z^2 + \sigma_U^2)}\big) = 0.746$. Monte Carlo estimates fluctuate around these analytic values.

**Evaluation.** $\rho$ is swept over $\{0, 0.1, \dots, 1.0\}$. For each $\rho$ we draw 30 independent Monte Carlo realisations. Source AUC is computed on the $\rho = 0$ samples and target AUC on the $\rho$-shifted samples. Reported values in Figure 1b are medians across realisations; bands are 5th–95th percentiles.

**Interpretation.** The simulation is an illustration of the mechanism, not a fitted model of ADNI–AIBL shift. Its purpose is to show that a source-oriented mean score can approach chance under a coherent cohort perturbation while a frozen severity-based uncertainty score remains stable.

## D RETROSPECTIVE AIBL MOMENT CHECK

A natural concern with Proposition 1 is that the closed-form AUC formula relies on Gaussian class-conditionals and Gaussian residuals that need not hold on AIBL. As a retrospective check, we plug target-cohort quantities into Equation (2) and compare the resulting number against the empirical AIBL $s_\mu$-AUC. The only quantity entering Equation (2) is the effective severity slope $\eta(\rho) = (1-\rho)\alpha + \rho(\gamma+\kappa)$, and $\eta$ is identifiable from target-cohort data alone as the OLS slope of $s_\mu$ on a severity proxy $Z$: the source slope $\alpha$, the alignment $\gamma$, the cross-term $\kappa$, and the mixing weight $\rho$ are not separately estimated. This check uses the initial-run raw-score predictions, and its class-conditional moments use target conversion labels, so it is retrospective rather than a label-free prediction of target performance.

**Severity proxy.** We take $Z$ to be baseline diagnosis (CN$=0$, MCI$=1$) on AIBL, which uses only baseline inputs. As a sensitivity check we also use a composite $Z$ formed by standardising and combining baseline diagnosis, $-$baseline hippocampal-region thickness, age, and APOE4 carrier status. Both proxies give the same qualitative answer.

**Estimation protocol.** On each AIBL evaluation subject we compute $s_{\mu_i}$ from the frozen ADNI-trained head; we regress $s_\mu$ on $Z$ by OLS to get $\widehat{\eta}$; and we plug $\widehat{\eta}$ together with the target class-conditional severity statistics $(\widehat{\Delta\mu}_Z, \widehat{\sigma}_Z)$ and the target residual variance $\widehat{\sigma}^2_\perp$ into Equation (2). Confidence intervals are nonparametric bootstrap over AIBL subjects with $2{,}000$ resamples, propagating the OLS and moment estimates jointly.

Table 5: Retrospective AIBL moment check. The closed-form mean-score AUC of Proposition 1, evaluated with target-cohort moments from the initial-run predictions, agrees numerically with the empirical AIBL $s_\mu$-AUC; this agreement does not validate the Gaussian assumptions.

| Quantity | | Model-based quantity | Empirical value |
|---|---|---|---|
| Mean-score AUC ($s_\mu$, AIBL) | | $0.514\ [0.483, 0.555]$ | $0.519$ |
| Residual-scale $\lvert\Delta\mathrm{AUC}\rvert$ | $0.045$ | (first-order directional estimate) | $0.117$ |

**Findings.** The plug-in value for the AIBL $s_\mu$ score is $0.514$ with bootstrap CI $[0.483, 0.555]$, against an empirical AIBL $s_\mu$ AUC of $0.519$ (Table 5). This numerical agreement does not validate the Gaussian moment assumptions of Proposition 1, nor does it identify cancellation: the source and perturbation components are not separately estimated, and the class-conditional moments use target conversion labels. The fitted $\widehat{\eta} \approx 0$ is consistent with the anti-aligned regime, without establishing it as the mechanism of the observed mean-score collapse.

## E COHORT AND LABELLING DETAILS

Table 6 gives the per-cohort breakdown used throughout the paper.

Table 6: Evaluation cohorts. ADNI is the source cohort; AIBL is the primary external cohort. OASIS-3 is non-evaluable under the strict 24-month conversion-window definition used in the headline results (only five subjects with zero converters survive the filter); we therefore report it here only as a filtering-pipeline diagnostic and evaluate it under a multi-horizon protocol in Section H.8.

| 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 (24m) | 24m exact | 5 | 0 |

**Conversion labels.** A subject is positive if any clinical follow-up within the conversion window strictly worsens the baseline diagnosis (CN→MCI, CN→AD, or MCI→AD). For ADNI we require the target structural scan at exactly 24 months. For AIBL we use a $\pm 6$-month tolerance because the AIBL imaging protocol places visits at 18, 36, 54, and 72 months.

**Preprocessing.** Cortical thickness is extracted on the 200-region Schaefer parcellation (Schaefer et al., 2018) from T1-weighted MRI using the standard FreeSurfer pipeline (Fischl, 2012) . We $z$-score each region's thickness using ADNI training-split statistics; the same statistics are applied unchanged to AIBL.

**Subject splitting.** The ADNI cohort is split into train, validation, and test at the subject level (no subject appears in more than one split) with stratification on baseline diagnosis. We use seed 42 to define the split used for all main-text numbers; repeated-split results in Section H.11 use seeds $\{0, 1, \ldots, 9\}$ with three training random seeds each (30 runs total).

**No conversion labels seen during forecaster training.** The forecaster is trained purely on the dense imaging task (12-month future prediction); conversion labels are only used at the downstream risk-evaluation stage. This is what makes the raw $s_\mu$ and $s_\sigma$ scores label-free.

# F Architecture, training, and adapters

This section gives the full specification behind the main-text architecture summary (Section 3).

## F.1 Backbone

The backbone is a reaction–diffusion graph forecaster on the 200-node Schaefer atlas. The diffusion term uses the symmetric graph Laplacian $L = D - A$ with $A$ the adjacency from a symmetric $k$-nearest-neighbour rule ($k = 10$) on parcel-centroid Euclidean distances, with radial-basis edge weights. The reaction term is a graph-attention network with spectral positional features (the first 16 Laplacian eigenvectors) and FiLM/AdaIN-style conditioning on baseline covariates. A small learned latent graph (top-$k = 10$, latent dim 32) is added as a residual adjacency. The backbone outputs a physics-prior forecast $\hat{x}^{\text{phys}}_{i,t+\tau}$, which the residual head then refines.

## F.2 Heads and loss

Five heads share the backbone: **none** (the physics-prior forecast is taken directly); **MLP** (deterministic refinement); **Gaussian** (predicts a mean and a per-region log-variance); **flow** (a small affine-coupling normalising flow on residuals); **diffusion** (a few-step denoiser on residuals). For the Gaussian head, the predictive likelihood term is the heteroscedastic Gaussian negative log-likelihood,

$$\mathcal{L}_{\text{nll}} \;=\; \tfrac{1}{2}\Big(\log \sigma_{ij}^2 + \tfrac{(x_{ij,t+\tau} - \hat{x}_{ij,t+\tau})^2}{\sigma_{ij}^2}\Big),$$

averaged over regions $j$ and subjects $i$. The full training loss combines this with an MSE forecasting loss on the backbone output, a Pearson-correlation loss on predicted vs. observed regional change, and a margin-ranking loss on the magnitudes of regional change: $\mathcal{L} = \lambda_{\text{nll}}\mathcal{L}_{\text{nll}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}} + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}}$.

## F.3 Hyperparameters

Table 7 lists the hyperparameters used for the Gaussian biomarker model. Other heads share the backbone settings; the head-specific parameters (e.g. flow coupling depth, diffusion timesteps) are listed in the released code.

## F.4 Downstream adapters

Linear adapters are balanced logistic-regression models on standardised $[s_\mu, s_\sigma]$ features. The quadratic adapter applies a degree-two polynomial expansion $[s_\mu, s_\sigma, s_\mu^2, s_\mu s_\sigma, s_\sigma^2]$ and fits a balanced

Table 7: Main hyperparameters for the Gaussian biomarker model.

| Category | Setting |
|---|---|
| Atlas | 200-node Schaefer; ADNI source; AIBL transfer |
| Forecast horizon | 12-month cortical-thickness prediction |
| Conversion label | diagnosis worsening within conversion window |
| History | baseline + previous scan; fallback gap 12 months |
| Subject split | stratified, train fraction 0.8, seed 42 |
| Backbone | 3 layers, hidden 128, 4 attention heads, dropout 0.1 |
| Spectral features | 16 Laplacian eigenvectors, dim 16 |
| Latent graph | dim 32, top-$k = 10$ |
| Optimiser | AdamW, lr $2\times10^{-4}$, weight decay $10^{-4}$ |
| Schedule | cosine with warmup, batch 32 |
| Training | 80 epochs, early-stopping patience 12 |
| Input noise | $\mathcal{N}(0, 0.02^2)$ at training time |
| Loss weights | $\lambda_{\text{phys}}$=0.5, $\lambda_{\text{corr}}$=0.5, $\lambda_{\text{rank}}$=0.2, $\lambda_{\text{nll}}$=0.7 |
| Parameters | 753,557 |

logistic regression with $C = 0.1$. Clinical-covariate baselines use baseline diagnosis, age, sex, APOE4 status, and baseline mean cortical thickness. Adapters are always trained on ADNI and applied without retuning to AIBL.

### F.5 Compute resources

Each forecaster trains on a single NVIDIA RTX 6000 Ada GPU in approximately 35 minutes for 80 epochs. The full main-text experiment matrix (5 heads $\times$ 30 repeated splits, plus the ablations of Table 18, the deep-ensemble run with 5 members, and the MC-dropout run with 30 stochastic forward passes) fits within roughly 80 GPU-hours. The downstream adapter fits and bootstrap intervals are CPU-bound and take seconds per configuration.

## G Statistical methodology

**Bootstrap confidence intervals.** For a single AUC reported on $n$ validation subjects, we resample subject indices with replacement 2,000 times. For each resample, the AUC is recomputed; resamples that contain only one class are discarded. Reported intervals are the 2.5th and 97.5th percentiles of the kept resamples.

**Paired bootstrap.** For comparing two scores on the same validation subjects, we use a paired bootstrap on the subject index: for each resample, both scores are evaluated on the same set of resampled subjects. The reported $p$-value is the one-sided proportion of resamples in which the gap has the opposite sign of the observed gap (with the usual $+1$ Laplace smoothing in numerator and denominator).

**Repeated subject-level splits.** The repeated AUC values in Table 19 use 10 subject-level splits ($\times$3 training seeds each), totalling 30 runs per row. Reported confidence intervals are 2.5/97.5 percentiles across runs.

**High-risk threshold.** The high-risk threshold used in the quartile analysis (Section H.13) is the 75th percentile of the training-split predicted $s_\sigma$, applied unchanged to validation. Validation conversion labels are not used to set the threshold.

### G.1 Validity of the inference at small event counts

The primary external comparison rests on 14 conversion events, where both the percentile bootstrap and the asymptotic DeLong test could in principle be unreliable. We therefore checked the inference procedures themselves. These are diagnostics of the estimators, not additional evidence about the scores, and they do not change any reported estimate.

**Exact paired test.** DeLong's test relies on a normal approximation. As an assumption-free alternative we use a within-subject permutation test: both scores are rank-normalised, which leaves every AUC unchanged, and each subject's pair of ranks is exchanged with probability $1/2$ over $10{,}000$ draws, giving an exact null for $\Delta_{\sigma-\mu}$ under within-subject exchangeability of the two scores. On AIBL the exact two-sided $p$ is $0.235$, against $0.230$ for DeLong; the two agree to within $0.04$ on every single-model comparison (largest gap at OASIS-3 36 months, $0.390$ versus $0.349$). The failure to resolve a difference on AIBL is therefore a property of the sample, not an artifact of the normal approximation. Rejection of this null indicates that the scores differ in ranking information; non-rejection does not establish that they are equivalent.

**Interval shape.** Percentile intervals can be miscalibrated for skewed statistics. Recomputing bias-corrected and accelerated (BCa) intervals on the *same* bootstrap replicates, with the acceleration constant from a class-stratified jackknife, changes no reported interval by more than $0.013$ and leaves every zero-crossing conclusion unchanged (bias correction $|z_0| \leq 0.08$, acceleration $|a| \leq 0.04$). We therefore report percentile intervals throughout.

**Interval coverage and attainable precision.** To ask whether a nominal $95\%$ paired interval is calibrated at these cohort sizes, we treat each inverse-direction score file (494 subjects, 88 converters) as a finite population with a known difference, draw 1,000 sub-cohorts of a given size without replacement, and record how often the interval covers the population value. At the AIBL shape (14 converters, 125 non-converters) coverage is $0.94$ and the mean bootstrap standard error matches the sampling standard deviation to within $7\%$. At the OASIS-3 36-month shape (7 converters) coverage falls to $0.90$ and the bootstrap standard error is about $6\%$ too small; BCa does not repair this. The 7-converter intervals in Table 16 should accordingly be read as approximately $90\%$ intervals. The same simulation gives the precision attainable at a given event count: the median interval width for $\Delta_{\sigma-\mu}$ is $0.28$ at 14 converters, $0.24$ at 21, and $0.20$ at 29, and obtaining an interval excluding zero with probability $0.8$ requires roughly 14 converters when the population difference is $+0.36$ and roughly 29 when it is $+0.19$. This quantifies the sample sizes at which the present comparison could be resolved.

# H Additional empirical results

## H.1 In-domain residual-head comparison

Table 8 reports the seed-42 split that backs the in-domain discussion in the main paper. The deterministic MLP and the heteroscedastic Gaussian residual heads are tied on 12-month forecasting MAE ($0.0818$ vs. $0.0819$); the Gaussian-mean and MLP-mean 24-month risk AUCs differ by a paired-bootstrap gap of $+0.029$ on the same validation subjects, $95\%$ interval $[-0.004, 0.065]$ and one-sided probability $0.044$. Generative residual heads (flow, diffusion) are worse on both MAE and AUC.

Table 8: Residual forecasting heads on the seed-42 ADNI split.

| Head | 12m MAE ↓ | AUC all ↑ | AUC MCI ↑ |
|---|---|---|---|
| None | 0.0828 | 0.618 | 0.574 |
| MLP | 0.0818 | 0.755 | 0.702 |
| Gaussian | **0.0818** | **0.784** | **0.723** |
| Flow | 0.0869 | 0.652 | 0.516 |
| Diffusion | 0.0883 | 0.518 | 0.516 |

Table 9: Approximate posterior-predictive controls on the 24-month ADNI biomarker split. They are competitive with the deterministic MLP but do not surpass the heteroscedastic Gaussian head.

| Forecaster | 24m MAE ↓ | AUC all ↑ | AUC MCI ↑ |
|---|---|---|---|
| Deterministic MLP | 0.09036 | 0.755 | 0.702 |
| Deep ensemble | 0.09013 | 0.751 | 0.694 |
| MC dropout | 0.09038 | 0.755 | 0.700 |
| Gaussian | 0.09041 | **0.784** | **0.723** |

Table 10: Selected all-subject ADNI label-scarcity results. Budgets are numbers of labelled converters available to downstream models. Raw forecast scores are label-free, so AUCs repeat across budgets.

| Model | 10 labels | 40 labels | 80 labels |
|---|---|---|---|
| Clinical covariates | 0.710 | 0.759 | 0.782 |
| Direct imaging logreg | 0.680 | 0.708 | 0.699 |
| $s_\mu$ raw | 0.784 | 0.784 | 0.784 |
| $s_\mu + s_\sigma$ logistic | 0.766 | 0.788 | 0.786 |

### H.2 Approximate posterior-predictive controls (in domain)

### H.3 Label-scarcity sweep

### H.4 AIBL→ADNI: the dissociation reproduces in the inverse direction

We trained the full set of 9 UQ scores on AIBL and evaluated them on ADNI ($n$=494 subject-horizons, 88 converters at the 36m horizon; 5 training seeds each), with paired subject-level 95% bootstrap CIs (Table 11).

The inverse cohort shift is more benign than ADNI→AIBL: $s_\mu$ does not collapse on ADNI (AUC 0.676–0.689 across scores, vs. 0.534 on AIBL forward), so the absolute dissociation is smaller. The qualitative pattern nonetheless reproduces. Every NLL-trained $\hat{\sigma}$ head ranks ADNI converters at or above the corresponding $s_\mu$ score: the $\sigma$-aware ensemble has $\Delta_{\sigma-\mu} = +0.047$ $[+0.014, +0.081]$; last-layer Laplace has $+0.042$ $[+0.012, +0.074]$; the heteroscedastic SNGP and Gaussian rows straddle zero with overlapping CIs, consistent with the more benign shift. The frozen-$\sigma$ control collapses to $\hat{\sigma}$-AUC$= 0.500$ exactly as in the forward direction. The supervised-confound prediction holds in the inverse direction too: the $\sigma$-aware ensemble, trained with the same NLL loss as the Gaussian head, still does not match the forward direction's $\hat{\sigma}$-AUC, indicating that target-cohort population heterogeneity, not training-objective alignment, is the load-bearing factor in the AIBL-source case.

Table 11: Inverse-direction evaluation: AIBL-source UQ checkpoints scored on ADNI (36m horizon, $n$=494, 88 converters). 5 training seeds; subject-level paired bootstrap, 2,000 resamples; pooled across seeds for $n$=2,470 observations / 440 converters in the CI. $\Delta_{\sigma-\mu} = \mathrm{AUC}(\hat{\sigma}) - \mathrm{AUC}(s_\mu)$.

| Method | $\mathrm{AUC}(\hat{\sigma})$ | $\mathrm{AUC}(s_\mu)$ | $\Delta_{\sigma-\mu}$ |
|---|---|---|---|
| Gaussian (residual-likelihood) | 0.692 [.664,.719] | 0.676 [.649,.703] | +0.016 [−.017,+.052] |
| $\sigma$-aware ensemble (NLL) | 0.723 [.698,.748] | 0.676 [.649,.703] | +0.047 [+.014,+.081] |
| Heteroscedastic SNGP (NLL) | 0.691 [.666,.717] | 0.689 [.662,.715] | +0.002 [−.029,+.036] |
| Last-layer Laplace | 0.723 [.699,.748] | 0.681 [.655,.708] | +0.042 [+.012,+.074] |
| MC dropout | 0.702 [.678,.728] | 0.678 [.651,.705] | +0.024 [−.009,+.059] |
| Vanilla MSE ensemble | 0.741 [.717,.765] | 0.676 [.648,.702] | +0.065 [+.034,+.097] |
| Evidential (aleatoric) | 0.682 [.653,.710] | 0.686 [.659,.712] | −0.004 [−.037,+.033] |
| Evidential (epistemic) | 0.699 [.672,.727] | 0.686 [.659,.712] | +0.014 [−.018,+.049] |
| Frozen-$\sigma$ control | 0.500 [.470,.529] | 0.682 [.655,.708] | −0.181 [−.222,−.142] |

### H.5 Training objectives and uncertainty components

NLL-trained scale and MSE-trained model variability differ in both objective and extracted quantity. A Gaussian head learns a conditional residual variance through its likelihood; an MSE ensemble instead measures variation between fitted means. A performance gap can therefore reflect either difference or their combination. In the 20-seed AIBL method comparison, heteroscedastic ensemble and SNGP outputs exceed their corresponding means by average AUC differences $+0.139$ and $+0.097$ (20/20 and 18/20 positive seeds). MSE ensemble, dropout, and Laplace differences average $-0.047$, $-0.148$, and $-0.160$ (Table 12). The heteroscedastic outputs combine residual and model-related uncertainty, so this is not an isolated causal test of NLL or pure residual variance.

The positive-seed counts in Table 12 describe training variation on a shared target cohort. They are not independent population replications, and the deep ensemble's negative average does not imply a negative result in every seed. Likewise, a constant scale has AUC $0.5$ because every subject is tied. That control cannot rule out memorization or clinical confounding in an input-dependent score. These distinctions limit the conclusion to the specified outputs and training protocols, not uncertainty estimation in general.

Table 12: AIBL 24-month objective controls. $\Delta$ is uncertainty-score AUC minus the same method's mean-score AUC, averaged across 20 seeds. Heteroscedastic outputs combine residual scale with model-related uncertainty; counts describe seeds, not independent target cohorts.

| Method | Objective | Mean $\Delta$ | Positive seeds |
|---|---|---|---|
| Heteroscedastic ensemble | NLL | $+0.139$ | $20/20$ |
| Heteroscedastic SNGP | NLL | $+0.097$ | $18/20$ |
| Deep ensemble | MSE | $-0.047$ | $7/20$ |
| MC dropout | MSE | $-0.148$ | $0/20$ |
| Last-layer Laplace | MSE | $-0.160$ | $0/20$ |

### H.6 AIBL outcome-window extension: per-method detail

Table 3 reports the single main-Gaussian model under the 24-, 36-, and 54-month outcome windows. Applying the same relabelling to every residual-scale score gives Table 13.

**Longer outcome windows on the same target subjects.** The 14-converter limit is a property of the 24-month outcome window, not of the cohort. Extending the window to 36 and 54 months raises the converter count to 21 and 29 (Table 3). Under the 36- and 54-month windows the paired interval excludes zero, and the 20-seed heteroscedastic-ensemble differences are $+0.256$ $[0.205, 0.305]$ and $+0.339$ $[0.295, 0.383]$, positive in 20/20 seeds at every window.

Two features of the table limit its interpretation. The 54-month window drops 20 subjects who lack sufficient follow-up, so its population is not identical to the other two rows. More importantly, part of the increase in $\Delta_{\sigma-\mu}$ reflects predicted change falling below chance ($0.435$ and $0.403$) rather than residual scale improving: a constant score, which ranks every subject equally, attains a $+0.051$ and $+0.079$ average difference against the same mean score at these windows. The residual-scale AUCs of $0.712$ and $0.769$ are therefore the quantity to read alongside $\Delta_{\sigma-\mu}$, and the longer windows should be understood as increasing event count rather than as independent cohorts.

**What would this ranking look like in practice?** Figure 2 gives a budgeted view of the 36-month window: if only a small fraction of people could be selected for closer review, how many of those who later convert would be found? Selecting the 10% of subjects with the highest scores, residual scale identifies $38\%$ of converters and predicted change $10\%$; random selection of 10% of subjects would identify about 10%. At a 20% budget the recalls are $57\%$ and $19\%$. In this sample, residual scale therefore places more converters near the top of the ranking, which is the operational counterpart of the AUC difference. With 21 conversion events these percentages describe this dataset and are not guaranteed performance in a larger population.

The four heteroscedastic-NLL scores rank between $0.695$ and $0.775$ at every window and favor residual scale in every seed except the heteroscedastic SNGP at 24 months (18/20). The posterior-

Table 13: Per-method AIBL outcome-window extension. $\mathrm{AUC}(s_\sigma)$ is the mean across training seeds with the across-seed standard deviation; $\Delta_{\sigma-\mu}$ is the mean paired difference against the same method's mean score, and "+seeds" counts seeds with a positive difference. All rows use the same 139-subject evaluation (119 at 54 months); only the outcome window changes.

| Method | Objective | 24 months | 36 months | 54 months |
|---|---|---|---|---|
| $\mathrm{AUC}(s_\sigma)$ | | | | |
| Heteroscedastic ensemble (20 seeds) | NLL | $0.697 \pm .011$ | $0.714 \pm .006$ | $0.772 \pm .006$ |
| Heteroscedastic SNGP (20 seeds) | NLL | $0.696 \pm .031$ | $0.712 \pm .027$ | $0.775 \pm .023$ |
| Evidential aleatoric (5 seeds) | NLL | $0.695 \pm .009$ | $0.716 \pm .005$ | $0.770 \pm .006$ |
| Evidential epistemic (5 seeds) | NLL | $0.700 \pm .007$ | $0.716 \pm .004$ | $0.770 \pm .005$ |
| Deep ensemble (20 seeds) | MSE | $0.516 \pm .087$ | $0.497 \pm .075$ | $0.466 \pm .105$ |
| MC dropout (20 seeds) | MSE | $0.416 \pm .014$ | $0.391 \pm .016$ | $0.365 \pm .023$ |
| Last-layer Laplace (20 seeds) | MSE | $0.403 \pm .042$ | $0.353 \pm .037$ | $0.297 \pm .037$ |
| SNGP (5 seeds) | MSE | $0.448 \pm .089$ | $0.408 \pm .062$ | $0.357 \pm .065$ |
| Constant-scale control (20 seeds) | — | 0.500 | 0.500 | 0.500 |
| $\Delta_{\sigma-\mu}$ *(+seeds)* | | | | |
| Heteroscedastic ensemble | NLL | $+0.139$ (20/20) | $+0.256$ (20/20) | $+0.338$ (20/20) |
| Heteroscedastic SNGP | NLL | $+0.097$ (18/20) | $+0.197$ (20/20) | $+0.269$ (20/20) |
| Evidential aleatoric | NLL | $+0.140$ (5/5) | $+0.258$ (5/5) | $+0.330$ (5/5) |
| Evidential epistemic | NLL | $+0.144$ (5/5) | $+0.258$ (5/5) | $+0.330$ (5/5) |
| Deep ensemble | MSE | $-0.047$ (7/20) | $+0.034$ (12/20) | $+0.022$ (10/20) |
| MC dropout | MSE | $-0.148$ (0/20) | $-0.073$ (0/20) | $-0.084$ (0/20) |
| Last-layer Laplace | MSE | $-0.160$ (0/20) | $-0.108$ (1/20) | $-0.145$ (0/20) |
| SNGP | MSE | $-0.095$ (1/5) | $-0.047$ (1/5) | $-0.073$ (1/5) |
| Constant-scale control | — | $-0.042$ (4/20) | $+0.051$ (17/20) | $+0.079$ (19/20) |

weight scores rank between $0.297$ and $0.516$; only the deep ensemble at 24 months exceeds chance, and it does so by less than its across-seed standard deviation ($0.516 \pm 0.087$). The constant-scale control is exactly $0.5$ by construction, and its $\Delta_{\sigma-\mu}$ row makes the caveat of Table 3 concrete: at the longer windows an uninformative score attains a positive paired difference in 17/20 and 19/20 seeds, purely because the mean score is anti-predictive there. Positive-seed counts for the MSE-trained scores should be read against that baseline rather than against zero.

### H.7 Within-ADNI acquisition and enrolment-era shift

ADNI-1 (2004–2009) acquired predominantly at 1.5 T and enrolled cognitively normal, late-MCI, and AD participants. ADNI-GO/2/3 (2009–) acquired at 3 T and additionally enrolled early MCI (ADNI-GO consists entirely of early-MCI participants) and significant memory concern. Splitting ADNI by enrolment phase therefore yields a transfer problem inside a single study and a single processing pipeline, with 61–113 conversion events in the target instead of the 14 available on AIBL. For each direction and each of five seeds we train the heteroscedastic Gaussian forecaster from scratch on one phase using the protocol of Section F and score it frozen on the other, with no target-label fitting.

Because the two phases differ in baseline diagnostic composition as well as acquisition, we report each direction twice: once using all enrolled subjects, and once with both sides restricted to the baseline strata present in both phases (CN, late MCI, AD), which removes the early-MCI and significant-memory-concern asymmetry. Table 14 reports both.

Three observations follow. First, the direction that trains on the later phases and scores ADNI-1 reproduces the AIBL pattern at eight times the event count: predicted change falls to near chance while residual scale is retained, in 5/5 seeds. Second, the opposite direction shows no separation at all; both scores remain informative and their difference is within $0.01$ of zero. A shift in acquisition and enrolment era is therefore neither sufficient to produce the dissociation nor sufficient to destroy both scores. Third, restricting to shared strata raises *both* scores by about $0.07$ in the first direction while leaving their difference unchanged, so the composition gap between phases depresses the two scores together rather than creating the asymmetry. Consistent with the external results, the across-seed spread of $\mathrm{AUC}(s_\sigma)$ is 0.006–0.013 in every row, whereas $\mathrm{AUC}(s_\mu)$ varies by up to $0.075$.

Table 14: Within-ADNI transfer by enrolment phase. Each row is five training seeds; AUC columns are means across seeds with the across-seed standard deviation. $\Delta_{\sigma-\mu}$ uses the hierarchical seed-and-subject bootstrap (10,000 resamples); $p$ is an exact within-subject score-swap permutation test on the seed-averaged difference, conditional on the trained models. "Matched" restricts both phases to the baseline strata present in both.

| Direction | Strata | Subjects | Conv. | AUC($s_\sigma$) | AUC($s_\mu$) | $\Delta_{\sigma-\mu}$ [95% CI] ($p$) |
|---|---|---|---|---|---|---|
| ADNI-1 → GO/2/3 | all | 559 | 86 | $0.692 \pm 0.006$ | $0.688 \pm 0.013$ | $+0.004\ [-0.015, 0.024]$ (0.80) |
| ADNI-1 → GO/2/3 | matched | 254 | 61 | $0.760 \pm 0.008$ | $0.756 \pm 0.003$ | $+0.003\ [-0.016, 0.024]$ (0.86) |
| GO/2/3 → ADNI-1 | all | 459 | 113 | $0.758 \pm 0.013$ | $0.528 \pm 0.075$ | $+0.230\ [\ 0.164, 0.301]$ ($< 10^{-4}$) |
| GO/2/3 → ADNI-1 | matched | 459 | 113 | $0.794 \pm 0.012$ | $0.519 \pm 0.066$ | $+0.275\ [\ 0.217, 0.333]$ ($< 10^{-4}$) |

These comparisons are not an independent cohort: they share one study, one set of diagnostic criteria, and one processing pipeline, and both directions reuse overlapping populations. The matched GO/2/3 source also leaves only 51 held-out subjects with 12 events, so its in-domain validation AUC ($0.695 \pm 0.108$) is too imprecise to serve as a source-domain reference; we therefore compare directions to each other rather than to their own held-out splits. What the experiment establishes is narrower than the AIBL result and in one respect stronger: with 113 events and no target-label fitting, the ordering of the two scores depends on which phase supplies the training data, which is the setting-dependence the audit is intended to expose.

### H.8 OASIS-3: A MEANINGFUL BUT UNDERPOWERED SECOND EXTERNAL PROBE

OASIS-3 is not a duplicate of the ADNI/AIBL experiment: its visit schedule and community-based recruitment make a strict 24-month exact-scan endpoint too sparse to evaluate (5 subjects, 0 converters in the strict 24m window; Table 6). We therefore treat OASIS-3 as a multi-horizon external transfer probe rather than a 24-month replication of AIBL: Table 15 reports 36/54/72-month windows with the same tolerance and history-gap rules as the 24-month protocol, which yields evaluable cohorts of 148–162 subjects and 7–14 converters. This is meaningful for the machine-learning question because the ADNI-trained forecaster is frozen and no OASIS-3 labels are used for training, tuning, or adapter fitting.

Forecasting becomes harder on OASIS-3: MAE rises to $0.2384$ versus $0.0867$ on ADNI, consistent with scanner, protocol, and sampling shift. The risk signal remains positive but underpowered. At 36 months, the residual-scale score clearly exceeds the mean score in point estimate ($0.726$ vs. $0.606$ AUC). At 54 and 72 months, both scores remain above chance but the differences are small and the confidence intervals overlap. We treat OASIS-3 as supportive external evidence for the score-wise transfer story, not as a confirmatory clinical validation cohort.

Table 15: OASIS-3 transfer under the multi-horizon external-probe protocol. The forecaster is trained on ADNI and evaluated on OASIS-3 without target tuning. Risk AUCs are exploratory because only 7–14 converters are available across horizons, but the 36-month window provides a meaningful positive signal for the residual-scale score.

| Endpoint | Subjects / converters | $s_\mu$ AUC [95% CI] | $s_\sigma$ AUC [95% CI] |
|---|---|---|---|
| 36m risk | 162 / 7 | 0.606 [0.299, 0.893] | **0.726 [0.537, 0.883]** |
| 54m risk | 148 / 13 | 0.618 [0.446, 0.784] | **0.638 [0.501, 0.771]** |
| 72m risk | 131 / 14 | 0.612 [0.441, 0.764] | **0.618 [0.477, 0.751]** |

### H.9 EXTERNAL ROBUSTNESS CHECKS

This section reports the inverse-direction, within-study, and OASIS-3 checks summarized in the main text (Section 4.5), complementing the dedicated analyses in Sections H.4, H.7 and H.8.

**Inverse-direction transfer.** For the inverse AIBL-to-ADNI evaluation, we use a 36-month conversion endpoint comprising 494 ADNI subjects, including 88 converters. Across 20 training seeds, the

mean residual-scale-minus-predicted-change AUC difference is $+0.102$, with a hierarchical seed–subject interval of $[0.052, 0.155]$; 15 of 20 seeds favor residual scale (median difference $+0.084$). The direction of the difference is not uniform across runs: one seed yields $-0.034$ with interval $[-0.078, 0.012]$. Thus, the average favors residual scale under this protocol, but the training runs reuse one target cohort and do not constitute independent external replications. Because this evaluation also changes the outcome window relative to the primary ADNI-to-AIBL experiment, the difference cannot be attributed to transfer direction alone.

**Within-ADNI acquisition shift.** ADNI changed scanner field strength and enrolment criteria between its first phase and its later phases, which supports a higher-powered within-study transfer test: a forecaster trained on one enrolment phase and scored frozen on the other is evaluated on 61–113 conversion events rather than 14 (Section H.7). The two directions behave differently. Training on the later phases and scoring the first gives $\Delta_{\sigma-\mu} = +0.275$ $[0.217, 0.333]$ (459 subjects, 113 converters, 5/5 seeds), with predicted change near chance ($0.519$) and residual scale retained ($0.794$): the AIBL pattern at eight times the event count. The reverse direction gives $+0.003$ $[-0.016, 0.024]$ (254 subjects, 61 converters), with both scores informative and neither favored. Because the first phase enrolled no early-MCI or significant-memory-concern participants, we repeat both directions with each side restricted to the baseline strata present in both. This raises both scores in the first direction by about $0.07$ and leaves their difference unchanged, so the composition gap depresses the two scores together rather than producing the dissociation. These comparisons share one study and one processing pipeline, so they probe acquisition and enrolment era rather than an independent cohort, and the asymmetry between directions is not explained by field strength alone.

**OASIS-3 and pooled analyses.** OASIS-3 uses 36/54/72-month probes with 7/13/14 converters. Its paired differences are $+0.121$, $+0.021$, and $+0.007$, all with intervals crossing zero (Table 16). At 7 events the paired interval is slightly liberal, covering about $90\%$ rather than $95\%$ in a resampling study of the same cohort shape (Section G.1), so the 36-month row is the weakest of the three. These are exploratory checks, not equivalent independent confirmations. Pooled MCI observations across cohorts, horizons, and runs likewise do not create new independent events or establish a target-only effect.

Table 16: Exploratory OASIS-3 probes. Paired 95% intervals use 10,000 class-stratified subject-bootstrap resamples. People can recur across horizons; these rows are not independent cohorts.

| Horizon | Subjects | Converters | $\Delta_{\sigma-\mu}$ [95% CI] |
|---|---|---|---|
| 36 months | 162 | 7 | $+0.121$ $[-0.123, 0.349]$ |
| 54 months | 148 | 13 | $+0.021$ $[-0.134, 0.182]$ |
| 72 months | 131 | 14 | $+0.007$ $[-0.136, 0.154]$ |

### H.10 Backbone ablations

The backbone-ablation results in Table 17 support a narrower component-level claim than "every physics-motivated piece is essential." Removing the fixed atlas graph drops the overall AUC from $0.784$ to $0.700$ and removing spectral features drops it to $0.719$. In contrast, removing the latent graph is essentially neutral and removing temporal-gap conditioning slightly improves AUC. Structured graph priors and spectral structure dominate the downstream gain at the current data scale, while some additional physics-motivated mechanisms appear optional. Table 18 extends this with repeated-split intervals across $5$ subject-level splits: only removing the diffusion term ("Reaction only") produces a consistent, large drop. The residual-scale advantage is therefore localised to the heteroscedastic Gaussian head and is not an artefact of any specific physics component. Across $28$ matched seeds (full model vs. no-physics GCN with a heteroscedastic Gaussian head and matched parameter count), the $s_\sigma$ AUC differs by $\Delta = 0.000$ ($p = 0.97$, paired bootstrap), so the deterministic-MLP-vs-Gaussian gap is attributable to the heteroscedastic head itself, not to the physics-informed backbone.

### H.11 Repeated-split robustness

Table 19 repeats Tables 8 and 9 across 30 runs (10 subject-level splits $\times$ 3 training seeds). The Gaussian head's $s_\sigma$ remains the strongest score on average but the $95\%$ interval substantially overlaps

Table 17: Gaussian backbone ablations on the 24-month ADNI biomarker split (seed 42).

| Configuration | 24m MAE | AUC all | AUC MCI |
|---|---|---|---|
| Full model | 0.09041 | 0.784 | 0.723 |
| No temporal-gap conditioning | 0.09049 | 0.799 | 0.756 |
| No latent graph | 0.09049 | 0.786 | 0.723 |
| Diffusion only | 0.09098 | 0.772 | 0.706 |
| No spectral features | 0.09029 | 0.719 | 0.673 |
| No fixed atlas graph | 0.09017 | 0.700 | 0.675 |
| Reaction only | 0.08986 | 0.692 | 0.626 |

Table 18: Backbone ablations with the Gaussian head (5 random subject splits; mean AUC $\pm$ 95% CI from $t_4$). Only removing the diffusion term 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 conditioning | $0.754\ [0.692,\ 0.816]$ |
| Diffusion only | $0.764\ [0.692,\ 0.836]$ |
| Reaction only | $0.689\ [0.606,\ 0.773]$ |

with the Gaussian-mean and deterministic-MLP intervals, which is why the main paper uses a paired bootstrap rather than an unpaired comparison.

Table 19: Repeated-split AUCs on ADNI (30 runs). Columns are mean and 95% percentile interval. Mean AUC rankings are noisy in absolute terms; the within-pair Gaussian-vs-MLP gap is more stable (Table 8).

| Score | Repeated AUC (mean [95% CI]) |
|---|---|
| MLP (mean) | $0.765\ [0.749,\ 0.779]$ |
| Gaussian (mean) | $0.750\ [0.736,\ 0.763]$ |
| Gaussian ($s_\sigma$, resid.) | $\mathbf{0.775}\ [\mathbf{0.762},\ \mathbf{0.789}]$ |
| Deep ensemble (mean) | $0.772\ [0.730,\ 0.808]$ |
| Deep ensemble ($s_\sigma$, epi.) | $0.712$ † |
| MC dropout (mean) | $0.772\ [0.736,\ 0.803]$ |
| MC dropout ($s_\sigma$, epi.) | $0.573$ † |

† 5-split mean for the model-variability $s_\sigma$ proxies (computing model variability requires $K \geq 5$ trained members per split).

### H.12 $s_\sigma$ AS COMPRESSED TRAJECTORY ATYPICALITY

A natural worry is that the predicted $s_\sigma$ is uninformative beyond what the directly observed within-subject trajectory standard deviation already captures. We falsify this directly. The Spearman correlation between predicted $s_\sigma$ and the observed within-subject trajectory standard deviation is $\rho = 0.287$ ($p < 10^{-4}$): moderate, not high. Despite the moderate correlation, $s_\sigma$ substantially outperforms the raw signal as an in-domain conversion discriminator: AUC is 0.777 for predicted $s_\sigma$ versus 0.656 for the directly measured trajectory standard deviation. The model has therefore learned a portable representation of trajectory difficulty, not merely a noisy copy of an existing measurement.

### H.13 QUARTILE-STRATIFIED CONVERSION RATES

Figure 4 stratifies the ADNI validation cohort by predicted-$s_\sigma$ quartile. The top quartile (Q4) converts at a rate roughly an order of magnitude above the bottom three quartiles, with a median odds ratio across 5 training seeds of OR = 10.5 (range 5.0–12.3). Quartile thresholds are derived purely from

the forecaster's training-split predicted $s_\sigma$ and applied to the held-out validation set; no conversion labels are used to define them.

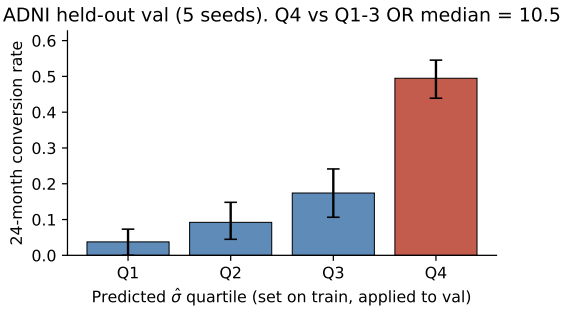



Figure 4: Conversion rate by predicted-$s_\sigma$ quartile (ADNI held-out validation, 5 training seeds). Q4 vs. Q1–3 median OR $= 10.5$ (5.0–12.3). Thresholds are set on training-split $s_\sigma$ alone. Bars are seed means; error bars span the seed min/max.

### H.14 Per-region $\sigma$ topography

Figure 5 maps the per-region Spearman correlation $\rho(\sigma_{ij}, Y_i)$ between the predicted regional uncertainty and 24-month conversion across the cortex. After Benjamini–Hochberg FDR correction at $q = 0.05$, all 200 Schaefer parcels are significant, with $\rho$ in the range 0.26–0.42. This indicates that elevated regional $\sigma$ is a cortex-wide signature of unstable subjects rather than a focal hotspot; nevertheless, a clear gradient emerges. The strongest correlations concentrate in the inferior parietal lobule and salience network ($\rho \approx 0.42$), default-mode prefrontal cortex ($\rho \approx 0.42$), and a temporal-parietal extension ($\rho \approx 0.38$). These are precisely the regions implicated in earliest amyloid accumulation and default-mode degeneration in Alzheimer's disease (Palmqvist et al., 2017; Jack et al., 2010), recovered without any conversion supervision during forecaster training. This cortex-wide severity alignment justifies collapsing the regional $\sigma$ field into the scalar $s_\sigma$ used in the main paper.

### H.15 Calibration–ranking decomposition (PR-AUC and Brier)

Table 20 reports ROC AUC, ECE, empirical 90% Gaussian coverage, conformal coverage, and the Brier score of a logistic regression of $\hat{\sigma}$ onto 24m conversion. The decomposition shows that ranking and calibration are dissociable: the Gaussian and evidential-epistemic scores are simultaneously the best-ranking and best-calibrated; vanilla MSE-trained ensembles, last-layer Laplace, MC dropout, and SNGP have 0 Gaussian coverage because their post-hoc $\hat{\sigma}$ is on the wrong scale. Brier differences across methods are within 0.005 in domain because the sigma-to-probability logistic regression is fit per-method on the source cohort, soaking up scale; this is why the calibration argument is carried by ECE/coverage rather than by Brier. A related but distinct failure is visible in domain for the NIG residual-scale score: its empirical Gaussian 90% coverage on ADNI val is only 0.286 despite a competitive ranking AUC of 0.767. We read this as a calibration defect of the NIG parameterisation rather than a transfer failure, which is why the paper reports the calibration–ranking decomposition rather than a single 90% coverage number.

### H.16 Target-cohort recalibration

The transferred residual-scale score preserves subject ranking on AIBL, but its Gaussian predictive intervals are not perfectly calibrated under cohort shift. We therefore evaluate a one-parameter recalibration that rescales the source-trained predictive dispersion by a multiplicative factor estimated from a small target-site calibration set:

$$\hat{\sigma}_{\text{hybrid}}(x) = \hat{\sigma}_{\text{aleat}}(x)\, c_{\text{tgt}}.$$

For a common positive factor, this transformation changes interval width but preserves rank order for any score. The reported coverage in Table 21 moves from 0.920 to 0.971 with 50 calibration subjects, farther from the nominal 0.900 target. It therefore does not demonstrate improved nominal coverage.

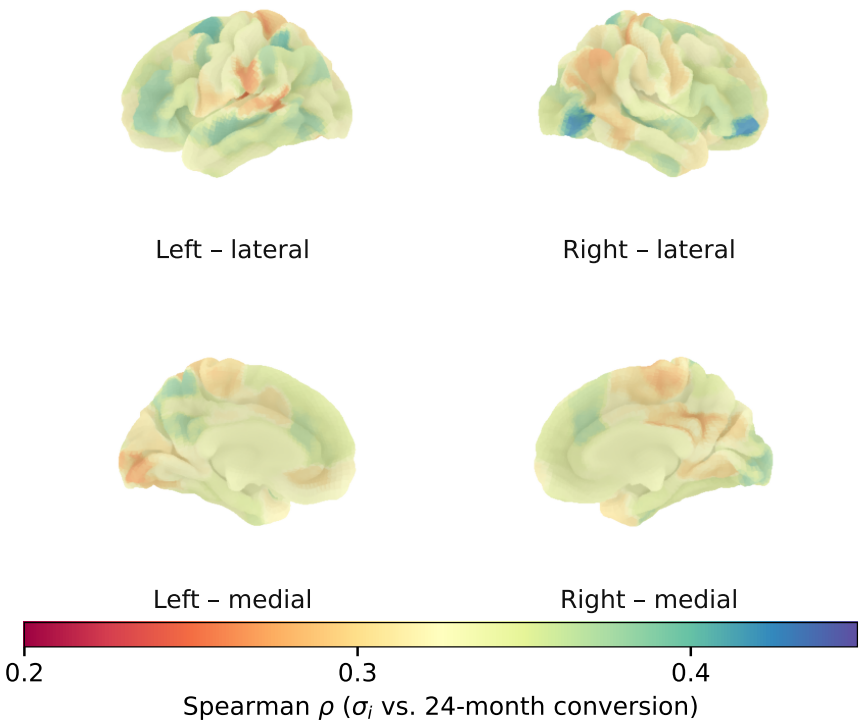



Figure 5: Per-region Spearman $\rho$ between predicted regional uncertainty $\sigma_{ij}$ and 24-month conversion (ADNI, $n = 204$; FDR-corrected at $q = 0.05$). All 200 regions significant; gradient peaks in temporal-parietal cortex and default-mode regions ($\rho \approx 0.42$).

Table 20: Calibration–ranking decomposition. ↓ = lower better.

| Method | Cohort | AUC | ECE ↓ | $\mathrm{Cov}_{90}^{\mathrm{Gauss}}$ | $\mathrm{Cov}_{90}^{\mathrm{conf}}$ | Brier ↓ |
|---|---|---|---|---|---|---|
| Gaussian (reference) | ADNI val | 0.792 | 0.006 | 0.971 | 0.907 | 0.157 |
| | AIBL | 0.660 | 0.006 | 0.960 | 0.881 | 0.110 |
| Evidential (epistemic) | ADNI val | 0.768 | 0.013 | 0.955 | 0.908 | 0.156 |
| | AIBL | 0.667 | 0.016 | 0.911 | 0.829 | 0.070 |
| Evidential (aleatoric) | ADNI val | 0.767 | 0.061 | 0.286 | 0.883 | 0.156 |
| | AIBL | 0.650 | 0.067 | 0.086 | 0.681 | 0.070 |
| Deep ensemble (MSE) | ADNI val | 0.722 | 0.114 | 0.000 | 0.907 | 0.158 |
| | AIBL | 0.577 | 0.108 | 0.000 | 0.949 | 0.072 |
| SNGP | ADNI val | 0.608 | 0.114 | 0.000 | 0.907 | 0.158 |
| | AIBL | 0.574 | 0.092 | 0.000 | 1.000 | 0.072 |
| Last-layer Laplace | ADNI val | 0.605 | 0.114 | 0.000 | 0.907 | 0.158 |
| | AIBL | 0.565 | 0.092 | 0.000 | 1.000 | 0.072 |
| MC dropout | ADNI val | 0.572 | 0.112 | 0.000 | 0.907 | 0.158 |
| | AIBL | 0.544 | 0.105 | 0.000 | 0.955 | 0.072 |

### H.17 MCI SUBGROUP POOLED ACROSS COHORTS AND HORIZONS

The baseline-MCI subgroups are small when evaluated separately: the 24-month ADNI subgroup contains 112 subjects with 34 converters, and the 24-month AIBL subgroup contains 19 subjects with 5 converters. To describe score behavior in a larger set of MCI evaluations, we pool ADNI validation subjects with AIBL subjects evaluated at the 24-, 36-, and 54-month outcome windows.

Table 21: **Target-cohort recalibration.** A single multiplicative scale factor $c_{\text{tgt}}$ on the source-trained $\hat{\sigma}$, estimated on $n_{\text{cal}}$ AIBL subjects, changes empirical coverage of nominal 90% intervals. Means across 20 seeds. These reported coverages do not demonstrate successful recalibration; positive common rescaling cannot alter the ranking of any score.

| Estimator | $n_{\text{cal}}$ | AIBL AUC | 90% coverage | $c_{\text{tgt}}$ |
|---|---|---|---|---|
| $s_\sigma$ (source-trained, baseline) | — | 0.698 | 0.920 | 1.000 |
| $s_{\sigma\,\text{hybrid}}$ | 20 | 0.696 | 0.962 | 1.171 |
| $s_{\sigma\,\text{hybrid}}$ | 50 | 0.698 | 0.971 | 1.196 |
| $s_{\sigma\,\text{hybrid}}$ | 100 | 0.697 | 0.976 | 1.210 |
| Posterior-weight $\sigma$ (recal.) | 50 | 0.479 (unchanged) | 0.000 (degenerate) | 0.986 |

For each method, we report the AUC of its uncertainty-derived score and mean score, together with their paired difference $\Delta_{\sigma-\mu} = \text{AUC}(\hat{\sigma}) - \text{AUC}(s_\mu)$ (Table 22).

This pool contains approximately 172 MCI subject–horizon records and 44 converter records per training seed. Subjects can recur across AIBL outcome windows, and the same subjects are scored across training seeds. These counts therefore do not represent distinct people or independent conversion events. The pooled estimates combine source- and target-cohort observations and should not be interpreted as AIBL-specific transfer effects.

In this pooled analysis, all four NLL-trained methods in Table 22 have positive uncertainty-minus-mean AUC differences. The vanilla MSE ensemble also has a positive difference ($+0.081$); thus, these results do not support a division in which all NLL-trained methods favor their uncertainty-derived scores and all MSE-trained methods favor their mean scores. The SNGP, MC-dropout, and last-layer Laplace rows instead have negative differences. These are method-specific associations: the methods differ in both training objective and the quantity used as $\hat{\sigma}$, so this comparison does not isolate the effect of either factor.

Pooling changes the evaluation population and repeats some subjects across horizons. Consequently, the positive pooled differences neither contradict nor resolve the imprecise AIBL-MCI-only comparison based on 19 subjects. The intervals in the table should be interpreted in light of this repeated-observation structure, rather than as evidence from thousands of independent subjects.

Table 22: Pooled MCI within-method comparison (ADNI validation MCI and AIBL MCI at 24, 36, and 54 months). The paired gap is $\Delta_{\sigma-\mu} = \text{AUC}(\hat{\sigma}) - \text{AUC}(s_\mu)$. Intervals are from 2,000 cohort-stratified subject-bootstrap resamples.

| Method | $\text{AUC}(\hat{\sigma})$ | $\text{AUC}(s_\mu)$ | $\Delta_{\sigma-\mu}$ |
|---|---|---|---|
| $\sigma$-aware ensemble (NLL) | 0.755 [.736,.775] | 0.605 [.584,.625] | $+0.150$ [+.128,+.174] |
| Heteroscedastic SNGP (NLL head) | 0.760 [.740,.779] | 0.605 [.585,.626] | $+0.154$ [+.131,+.178] |
| Vanilla MSE ensemble | 0.686 [.665,.707] | 0.605 [.584,.625] | $+0.081$ [+.058,+.104] |
| Evidential (epistemic, 5 seeds) | 0.741 [.718,.763] | 0.637 [.611,.663] | $+0.104$ [+.080,+.131] |
| Evidential (aleatoric, 5 seeds) | 0.729 [.707,.752] | 0.637 [.611,.663] | $+0.092$ [+.067,+.119] |
| SNGP (5 seeds) | 0.558 [.509,.607] | 0.646 [.599,.693] | $-0.088$ [-.146,-.033] |
| MC dropout | 0.517 [.494,.538] | 0.606 [.585,.627] | $-0.089$ [-.117,-.061] |
| Last-layer Laplace | 0.508 [.488,.528] | 0.609 [.587,.632] | $-0.101$ [-.129,-.071] |
| Frozen-$\sigma$ control | 0.500 [.478,.523] | 0.604 [.583,.625] | $-0.104$ [-.134,-.074] |

# I LIMITATIONS, BROADER IMPACT, AND REPRODUCIBILITY

## I.1 LIMITATIONS

The empirical evidence rests on one source cohort and one external cohort with sufficient positive events for AUC inference. AIBL has 14 converters under the strict 24-month definition. Single-cohort subgroup AUCs (CN: 9 converters; MCI: 5 at AIBL 24m alone) are underpowered; we therefore use cohort-pooled subgroup analyses (Table 22 for MCI) rather than per-cohort subgroup claims.

OASIS-3 is non-evaluable under the strict 24-month definition; we report it as a multi-horizon external probe (Section H.8) and a pipeline diagnostic (Table 6). The residual-scale score as we use it is a heteroscedastic predictive-dispersion signal, not a full Bayesian decomposition of model variability and residual uncertainty; Proposition 3 discusses why model-variability approximations transfer poorly, but it does not exclude last-layer Laplace or SWAG, which we leave to future work. Conversion labels in both cohorts are derived from clinical assessments rather than biomarker-confirmed AD pathology and are visit-schedule dependent.

### I.2 Broader impact

The strongest claim this paper makes is methodological: distributional form, not point error, governs which score of a forecast survives cohort transfer. The clinical claim is far narrower. The external-cohort AUC of $0.695$ on AIBL is well below any threshold that would justify individual screening or treatment decisions. We position the model as a research tool for cohort-level risk stratification only. Before any clinical deployment the following are needed: prospective evaluation under the intended deployment protocol; calibration analysis on each target cohort; fairness analysis across age, sex, race, and APOE4 status (we do not have the demographic granularity in the public-release versions of ADNI/AIBL to do the fairness analysis at the level a clinical deployment would require); and regulatory review. A research-tool deployment without these steps is inappropriate.

### I.3 Data and code availability

ADNI and AIBL data access requires institutional data-use agreements; see `https://adni.loni.usc.edu` and `https://aibl.com.au`. We do not redistribute the imaging data. Code, trained model checkpoints, the exact subject-id splits used for every reported number, and the scripts that reproduce Figure 1b from Section C will be released upon acceptance under an OSI-approved licence.

### I.4 Reproducibility statement

Every empirical number in the main paper is paired with sufficient methodological detail to reproduce it. Section E specifies the cohorts, splits, and labelling rule; Section F specifies the architecture, loss, hyperparameters, and adapter fitting procedure; Section G specifies the bootstrap and paired-bootstrap protocols; Section C specifies the simulation that generates Figure 1b, including all parameter values, seeds, and the sweep over $\rho$. The single piece of additional information required for byte-exact reproduction, the random subject-id split for seed $42$, is included in the code release.

# J EXTENDED PROTOCOLS AND SENSITIVITY ANALYSES

## J.1 VOLUMETRIC REPRESENTATION, TRAINING, AND AGGREGATION

**Purpose and representation.** To examine whether the score comparison extends beyond the cortical graph representation, we conduct an additional within-ADNI experiment using six longitudinal FreeSurfer volume measurements: ventricles, hippocampus, whole brain, entorhinal cortex, fusiform gyrus, and middle temporal cortex. This experiment changes both the imaging representation and the forecasting architecture. It therefore tests the score-evaluation protocol in a different setting, rather than isolating the effect of the graph architecture alone.

For volume measurement $V_{i,t,r}$, define the intracranial-volume (ICV) normalized value

$$v_{i,t,r} = 1000\,\frac{V_{i,t,r}}{\mathrm{ICV}_{i,t}}.$$

Normalization uses ICV from the corresponding visit. The model predicts baseline-to-24-month change,

$$d_{ir} = \frac{v_{i,24,r} - v_{i,0,r}}{a_r},$$

where $a_r$ is the standard deviation of baseline normalized volume $r$ in the training subjects. Baseline volume inputs are also centered and scaled using training-set statistics. No held-out subjects are used to estimate these transformations.

**Forecaster and uncertainty decomposition.** A graph-free heteroscedastic MLP predicts the six standardized changes from baseline volumes, age, sex, baseline diagnosis, APOE4 status, and log baseline ICV. We train five independently initialized models with a Gaussian negative log-likelihood. Training duration is selected using forecasting loss on an inner training-set split, followed by refitting on all available training subjects. Conversion labels are not used in this selection or in the forecasting loss.

Let $\hat{d}_{ir}^{(k)}$ and $\sigma_{ir}^{(k)}$ denote the predicted mean change and conditional standard deviation from member $k$, with $K = 5$. We form

$$\bar{d}_{ir} = \frac{1}{K}\sum_{k=1}^{K} \hat{d}_{ir}^{(k)}, \qquad a_{ir}^{\mathrm{vol}} = \left[\frac{1}{K}\sum_{k=1}^{K}\big(\sigma_{ir}^{(k)}\big)^2\right]^{1/2},$$

$$e_{ir}^{\mathrm{vol}} = \left[\frac{1}{K}\sum_{k=1}^{K}\big(\hat{d}_{ir}^{(k)} - \bar{d}_{ir}\big)^2\right]^{1/2}.$$

The volumetric scores are then

$$s_{\mu,i}^{\mathrm{vol}} = \left[\frac{1}{6}\sum_{r=1}^{6}\bar{d}_{ir}^{2}\right]^{1/2}, \qquad s_{\sigma,i}^{\mathrm{vol}} = \frac{1}{6}\sum_{r=1}^{6} a_{ir}^{\mathrm{vol}}, \qquad s_{\mathrm{PW},i}^{\mathrm{vol}} = \frac{1}{6}\sum_{r=1}^{6} e_{ir}^{\mathrm{vol}}.$$

Residual variances are therefore averaged across ensemble members before taking the square root. Model variability is computed separately across their means. All volumetric scores are expressed in training-standardized volume-change units, not millimetres.

**Matched evaluation and scope.** We retain the original ADNI subject split and restrict the volumetric experiment to subjects with complete baseline and 24-month measurements. This yields 507 training subjects and 127 held-out subjects, including 26 converters. Cortical and volumetric scores are compared on the same 127 held-out subjects. We also report the baseline-MCI subgroup separately (73 subjects, 23 converters) to distinguish within-stage progression ranking from associations across baseline diagnostic stages.

The matched evaluation controls the test population, but does not equalize the training-set size or the information available to the two models. Moreover, this is an *in-domain* representation-and-model control. Matching longitudinal volume inputs are not available in the external-cohort data used here, so we do not report volumetric cross-cohort transfer or use this experiment to validate a cross-cohort stability claim.

Table 23: Matched within-ADNI raw-score comparison underlying Figure 3. Intervals use 2,000 class-stratified subject-bootstrap resamples. Baseline MCI is a subgroup, not an independent replication.

| Score | All: 127 subjects, 26 events AUC [95% CI] | MCI: 73 subjects, 23 events AUC [95% CI] |
|---|---|---|
| Volume change | $0.755\ [0.642, 0.861]$ | $0.722\ [0.599, 0.842]$ |
| Volume residual scale | $0.656\ [0.552, 0.759]$ | $0.548\ [0.414, 0.686]$ |
| Volume model variability | $0.532\ [0.389, 0.673]$ | $0.494\ [0.340, 0.657]$ |
| Cortical change | $0.776\ [0.669, 0.871]$ | $0.716\ [0.584, 0.838]$ |
| Cortical residual scale | $0.783\ [0.682, 0.875]$ | $0.748\ [0.629, 0.863]$ |

### J.2 Severity-dependent cohort perturbations

**Cancellation as one possible shift mechanism.** The effective model includes the perturbation construction

$$M = \alpha Z + \varepsilon_M, \qquad \widetilde{M} = (1-\rho)M + \rho(\gamma Z + \delta),$$

where $\alpha > 0$, $\rho \in [0,1]$, and $\delta = \kappa(Z - m_Z) + \xi$ for a fixed centering constant $m_Z$. Assume independent Gaussian residuals $\varepsilon_M$ and $\xi$, also independent of $(Z, Y)$. Then

$$\eta = (1-\rho)\alpha + \rho(\gamma + \kappa), \qquad b = -\rho\kappa m_Z,$$

$$\sigma_\perp^2 = (1-\rho)^2\sigma_M^2 + \rho^2\sigma_\xi^2.$$

Here, $\kappa$ allows the perturbation to depend on severity; when $\sigma_Z^2 > 0$, it equals the common within-class slope $\mathrm{Cov}(\delta, Z \mid Y = y)/\sigma_Z^2$.

If $\gamma + \kappa < 0$, the perturbation opposes the source association, and the effective slope vanishes at

$$\rho^\star = \frac{\alpha}{\alpha - (\gamma + \kappa)}.$$

Provided the score remains nondegenerate, its AUC then equals $1/2$. Cancellation is one way to lose discrimination; increasing residual variation relative to class separation can also move AUC towards chance. A cohort-wide additive offset alone cannot do so. Further regimes are described in Section B.3. Outside the stated distributional assumptions, the closed-form expression is a working moment-based approximation, not an identity.

### J.3 Retrospective AIBL moment diagnostic

**Retrospective diagnostic on AIBL.** Using baseline diagnosis as an observed severity proxy, we fit a linear score–proxy relation and estimate the class-conditional moments on AIBL, using the initial-run raw-score predictions. The resulting plug-in AUC is $0.514$ with interval $[0.483, 0.555]$, compared with an empirical mean-score AUC of $0.519$ (Section D). This calculation uses target conversion labels to estimate class-conditional moments: it is a retrospective diagnostic, not a label-free prediction of target performance. Agreement on this sample does not validate the Gaussian assumptions or identify cancellation, because the source and perturbation components are not separately estimated.

### J.4 Risk ranking and forecast uncertainty are distinct

**Association with realized forecast error.** A useful conversion-ranking score need not accurately identify which imaging forecasts will be wrong. In the follow-up AIBL analysis, the Spearman correlation between predicted residual scale and realized forecast error is $0.027$, with subject-bootstrap interval $[-0.126, 0.177]$. Thus, the observed conversion association does not establish that the transferred scale estimates individual forecast difficulty on AIBL.

This distinction also limits the interpretation of the regional and trajectory analyses (Section H.14). Associations between predicted scale, diagnostic outcomes, and observed trajectory variability describe different quantities and should not be treated as interchangeable validation of residual uncertainty.

Table 24: Exploratory OASIS-3 comparisons. $\Delta_{\sigma-\mu}$ is evaluated within each horizon using 10,000 class-stratified paired subject-bootstrap resamples. Subjects may appear at multiple horizons; the rows are not independent replication cohorts.

| Horizon | Subjects | Converters | $\Delta_{\sigma-\mu}$ [95% CI] |
|---|---|---|---|
| 36 months | 162 | 7 | $+0.121\ [-0.123, 0.349]$ |
| 54 months | 148 | 13 | $+0.021\ [-0.134, 0.182]$ |
| 72 months | 131 | 14 | $+0.007\ [-0.136, 0.154]$ |

**Exploratory target-scale recalibration.** The reported scale-recalibration experiment does not demonstrate restoration of nominal coverage. For nominal 90% Gaussian intervals, reported coverage changes from $0.920$ before recalibration to $0.971$ with 50 target-calibration subjects (Section H.16). The latter is farther from the nominal level, so this result cannot support a claim of improved 90% coverage.

Multiplication by a common positive scale factor preserves the AUC of any scalar score. Consequently, unchanged ranking after this transformation is not evidence that calibration has been solved, nor is rank preservation specific to the residual-scale score. This target-data analysis is separate from the frozen, no-target-fitting transfer evaluation.

### J.5 Additional external analyses

**Inverse-direction transfer.** Training on AIBL and evaluating on ADNI at 36 months changes both transfer direction and endpoint horizon relative to the primary experiment (Section H.4). The follow-up analysis includes 494 ADNI subjects and 88 converters. Across five training seeds, the reported average residual-scale advantage is $+0.101$ AUC, with hierarchical seed-and-subject interval $[-0.025, 0.250]$; three of the five seeds favor residual scale. The inverse analysis therefore does not establish a uniformly positive effect across training runs.

**OASIS-3 protocol-shift probe.** OASIS-3 provides an exploratory external evaluation at 36, 54, and 72 months (Section H.8). At 36 months, the mean and residual-scale AUCs are $0.606$ and $0.726$, respectively. The paired follow-up differences and their uncertainty are reported in Table 16. All three intervals include zero, and the longer-horizon point differences are small.

**Pooled subgroup analyses.** The pooled MCI analysis combines observations across cohorts, horizons, and training runs (Section H.17). These repeated predictions do not constitute additional independent subjects or conversion events. Moreover, pooling source-domain and external observations cannot establish a target-only effect. We therefore treat this analysis as a sensitivity check, rather than an independent confirmation of transfer.