Toward Robust Neuroanatomical Normative Models: Influence of Sample Size and Covariates Distributions

Authors

Camille Elleaume[1,2], Bruno Hebling Vieira[1], Dorothea L. Floris[1], Nicolas Langer[1,2] and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing[*]

Affiliations

[1]Methods of Plasticity Research, Department of Psychology, University of Zürich, Zürich, Switzerland

[2]Neuroscience Center Zürich (ZNZ), Zürich, Switzerland

[*]Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au.

Abstract

Normative models are increasingly used to characterize individual-level brain deviations in neuroimaging studies, but their performance depends heavily on the reference sample used for training or adaptation. In this study, we systematically investigated how sample size and covariate composition of the reference cohort influence model fit, deviation estimates, and clinical readouts in Alzheimer's disease (AD). Using a discovery dataset (OASIS-3, n = 1032), we trained models on healthy control (HC), subsamples ranging from 5 to 600 individuals, while varying age and sex distributions to simulate biases in reference populations. We further assessed the use of adaptive transfer learning by pre-training models on the UK Biobank (n = 42,747) and adapting them to the clinical dataset applying the same sub-sampling strategies. We evaluated model performance on a fixed HC test set and quantified deviation score errors, outlier detection, and classification accuracy in both the HC test set and the AD cohort. The findings were replicated in an external validation sample (AIBL, n = 463). Across all settings, model performance improved with increasing sample size, but demographic alignment of the covariates, particularly in age, was essential for reliable deviation estimates. Models trained directly within the dataset achieved stable fit with approximately 200 HCs, while adapted models

reached comparable performance with as few as 50 individuals when pre-trained on large-scale data. These results show that robust individual-level modeling can be achieved using moderately sized but demographically matched cohorts, supporting broader application of normative modeling in ageing and neurodegeneration research.

1. Introduction

Normative modeling is a statistical framework for quantifying individual deviations from typical brain structure or function (Marquand et al., 2016). It estimates covariate-adjusted percentiles of a given measurement in a reference population. These percentiles serve as a basis for quantifying individual deviations from population norms, typically accounting for covariates such as age and sex. Unlike traditional case–control analyses that rely on group-level means, normative models generate subject-specific deviation scores that preserve inter-individual variability. This allows the characterization of brain atypicality at the individual level, without relying on the assumption that individuals within the same group share common patterns (Marquand et al., 2016; Rutherford, Fraza, et al., 2022; Rutherford et al., 2023). Warped Bayesian Linear Regression (Fraza et al., 2021) is a widely used normative modeling approach that combines flexibility and scalability, making it particularly well-suited for large neuroimaging datasets. (Corrigan et al., 2024; Holz et al., 2023; Meijer et al., 2024; Rutherford et al., 2023; Savage et al., 2024; Verdi et al., 2024). Long-term, these models can be envisioned as interpretable and quantitative neuroradiology tools to support clinical decision making (Bozek et al., 2023; Goodkin et al., 2019).

Among clinical applications of normative models, Alzheimer's disease (AD) represents a particularly relevant use case. AD is the most common cause of dementia and remains challenging to characterize due to its marked heterogeneity in clinical presentation and neuroanatomical patterns (Duara & Barker, 2022; Lam et al., 2013; Rajan et al., 2021). In clinical practice, structural imaging already plays a central role in diagnosis, and hippocampal volume is routinely evaluated in patients relative to expectations built from reference data (Vernooij et al., 2019). Normative modeling builds on this approach by enabling a more fine-grained characterization of neuroanatomical variability, supporting more precise and personalized assessments in AD. Recent applications to AD research had demonstrated its potential to reveal individualized patterns of cortical atrophy that correlate with cognitive performance and key biomarkers, including CSF $A\beta_{42}$ and phosphorylated tau (Loreto et al., 2024; Verdi et al., 2021, 2023, 2024).

However, the performance of normative models depends critically on the reference population used to fit them (Bethlehem et al., 2022; Bozek et al., 2023). The size of the training data, and the extent to which relevant covariates (e.g., age and sex) are adequately represented can substantially influence model fit and the accuracy of the resulting deviation scores. Large-scale samples have been recommended to accurately estimate outlying percentiles for clinical

use (Bozek et al., 2023). However, collecting sufficiently large neuroimaging datasets is particularly challenging, as it often requires multi-site data aggregation and extensive efforts to harmonize differences in acquisition protocols, scanner hardware. To address the challenge of limited sample sizes, studies often employ adaptive transfer learning strategies that allows to recalibrate pre-trained models on smaller samples (Bayer et al., 2022; Gaiser et al., 2023; Kia et al., 2021). But transfer learning applications face a similar challenge. While it offers a pragmatic solution, its success ultimately depends on the quality and representativeness of the available adaptation sample used for re-calibration. Whether models are directly trained within cohorts or adapted from large-scale references, both strategies raise concerns about robustness of clinical interpretations, particularly when the available data are limited, as is commonly observed in neuroimaging studies. Beyond sample size, careful consideration of covariates is critical, especially in applications to AD. The disease typically emerges in older adulthood and is more commonly diagnosed in women (Ferretti et al., 2018; Riedel et al., 2016), making accurate modeling of age and sex covariates especially important in this context. Despite growing adoption of normative modeling, systematic evaluations of how performance is affected by the composition and size of the training or adaptation sample remain limited.

To address this gap, we conducted a systematic investigation of how reference sample size and covariate composition affect normative model performance and clinical readouts. Using a single-site dataset (OASIS-3), we varied the size of the healthy control (HC) reference cohort (from 5 to 600 individuals) and manipulated age and sex distributions to simulate biases in reference populations. We replicated the approach in a two-site cohort (AIBL) and further assessed whether models pre-trained on a large external dataset (UK Biobank; n = 42,747) could be effectively adapted through transfer learning (Rutherford, Kia, et al., 2022), using the same sub-sampling strategies for the adaptation set. Adapted models were then applied to the same fixed test set, allowing direct comparison between models trained within each cohort and UKB-adapted models across all sampling conditions. Models' performances were evaluated using standard normative modeling evaluation metrics in an independent fixed HC test set. We assessed the error in deviation scores relative to those obtained using the full, theoretically optimal, reference sample in both the HC test set and the clinical sample diagnosed with AD. To assess clinical implications at the application level, we further quantified differences in outlier detection and classification performance when distinguishing AD from HC groups.

This work provides a systematic analysis of how reference sample characteristics influence model fit, deviation estimates, and their clinical interpretability, offering practical

guidance for applying normative modeling in ageing and neurodegeneration research and for maximizing the value of existing, deeply phenotyped cohorts when assembling reference datasets.

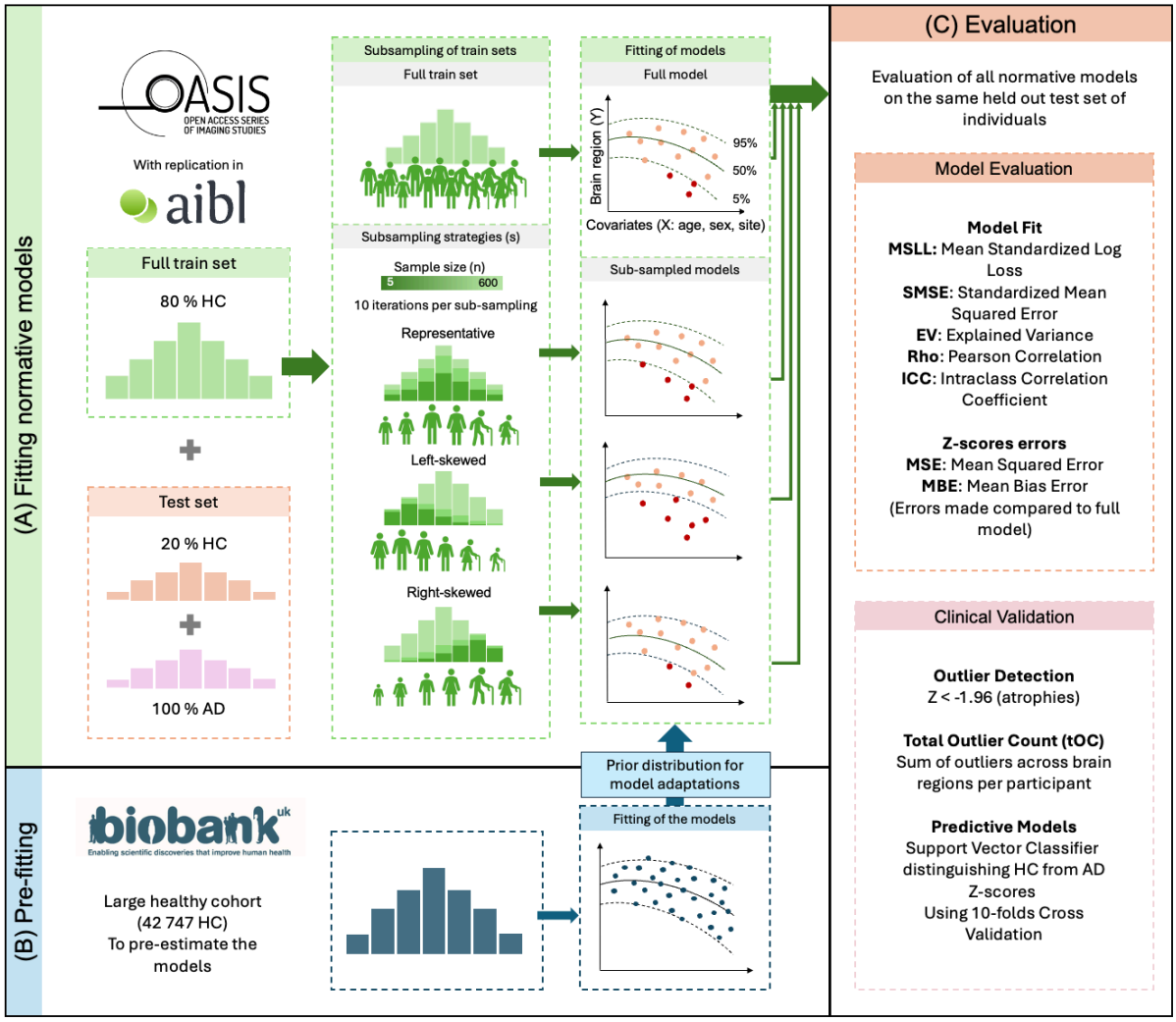

Figure 1. Methodology diagram for evaluating normative model estimation under different sampling scenarios. A: Analyses were conducted on the OASIS-3 dataset, with replication in AIBL. Normative models were first fitted using the entire training set (i.e., 80% of HC from the cohort). For each sampling strategy (s), models were re-fitted on randomly drawn subsamples of the training set, with sample sizes (n) ranging from 5 to the maximum available (n = 692 for OASIS, n = 322) and repeated across 10 iterations per sample size. Representative sampling: subsamples preserved the original age distribution using 10 age bins with an equal number of individuals per bin, while also ensuring balanced sex distributions. Left-skewed sampling: overrepresented younger age ranges by applying a beta distribution (α=2, β=5), across 10 equally spaced bins. Right-skewed sampling: overrepresented older age ranges using a beta distribution (α=5, β=2). Each model was evaluated on a fixed test set composed of 20% of HC from the same cohort and all AD individuals. Example normative fits illustrate how identical test values are interpreted under different models: red dots represent values outside the 95% centile range (outliers), while orange dots fall within the normative range. B: In a parallel analysis, normative models were pre-trained on the UK Biobank dataset. Adaptive transfer learning was used to adapt these pre-trained models to the clinical cohorts. The same sub-sampling strategies and sample sizes described in panel A were used here to define the adaptation sets. Adapted models were then applied to the same fixed

test set as in panel A, allowing direct comparison between within-cohort models and adapted UKB models across all sampling conditions. C: Each resulting model was evaluated using standard performance metrics to assess model fit. Z-score errors were additionally computed and compared to those obtained from the model fitted using the full training data. Clinical validation was then performed by analyzing outlier detection and predictive performance.

2. Results

We first evaluated how reference population characteristics influenced normative model accuracy. Using Warped Bayesian Linear Regression (Fraza et al., 2021), we trained normative models for 167 neuroanatomical regions of interest (ROI) using two independent datasets, OASIS-3 and AIBL (Figure 1, Table 1, see Supplementary Figures S17–S33 for the replication in AIBL). Training subsets ranged from 5 participants to the entire training cohort (n = 692), obtained by subsampling to reproduce realistic age- and sex-related skews. For each subsampling scenario, we generated 10 random draws and benchmarked their performance against the model trained on the full training sample.

		HEALTHY CONTROLS GROUP			**ALZHEIMER'S DISEASE GROUP**		
DATASET	Site	N	Age range (mean ± sd)	Sex ratio F/M (%)	N	Age range (mean ± sd)	Sex ratio F/M (%)
UKB	Full dataset	42 747	44.61 - 82.79 (64.50 ± 7.67)	52.75/47.25			
	UKB 1	25 536	44.61 - 82.34 (63.68 ± 7.57)	51.78/48.22			
	UKB 2	10 914	48.68 - 82.79 (66.33 ± 7.87)	54.20/45.80			
	UKB 3	6 297	47.92 - 81.93 (65.36 ± 7.52)	54.19/45.81			
OASIS-3	Full dataset	865	45.41 – 82.69 (68.21 ± 7.83)	58.96/41.04	167	50.79 - 82.67 (73.79 ± 5.99)	44.91/55.09
AIBL	Full dataset	403	60.13 - 82.61 (72.24 ± 5.40)	56.67/43.33	60	58.69 - 82.75 (73.59 ± 6.62)	56.67/43.33
	AIBL 1	274	60.99 - 82.61 (72.64 ± 5.26)	58.03/41.97	39	60.33 - 82.75 (73.05 ± 6.35)	53.85/46.15
	AIBL 2	129	60.13 - 82.56 (71.39 ± 5.63)	60.16/39.84	21	58.69 - 81.43 (74.60 ± 7.14)	61.90/38.10

Table 1. Demographics of HC and AD participants across datasets and sites summary of age ranges, mean ages, standard deviations, and sex ratios (F/M) for HC and AD groups within the UKB, OASIS-3, and AIBL datasets, including site-specific subgroups

2.1. Model fit evaluation

First, to quantify training-set sample size effects, we trained normative models on HC subsamples with representative age distributions (mirroring the full training set age distribution) and balanced sex ratios. We evaluated these models on a fixed HC test set and benchmarked their performance against corresponding full training sample models (Figure 1A). Normative model fits were assessed using *Mean Standardized Log Loss (MSLL), Standardized Mean Squared*

Error (SMSE), Explained Variance (EV), Pearson Correlation (Rho), and *Intraclass Correlation Coefficient (ICC)* (Figure 1C).

Sample size demonstrated a strong influence on model fit for cortical thicknesses and subcortical volumes (Figure 2). Specifically, model performance improved with larger sample sizes, as reflected by reductions in *MSLL* (β = -0.496, p < .001) and *SMSE* (β = -0.427, p < .001) (Figure 2A-B) and increases in *EV* (β = 0.499, p < .001), *Rho* (B = 0.527, p < .001), and *ICC* (β = 0.728, p < .001) (Figure 2C-E, Table 2). The standardized effect sizes (β values) allow for direct comparisons across metrics and sampling strategies. Performance improved consistently from n = 10 to the full training sample size (n = 692), with n = 5 showing unstable values and falling outside the overall trend. Rapid improvements were observed between n = 10 and 50, capturing ~70–80% of total gains across metrics. Gains continued more gradually between n = 50 and 200, accounting for most of the remaining improvements (~15–25%). Beyond n = 200, all metrics reached ~92–95% of their final values. After n = 300, changes became minimal, and metrics plateaued near 97–99% (Figure 2). Notably, ICC reaches excellent reliability from n = 50 across all lobes (Koo & Li, 2016).

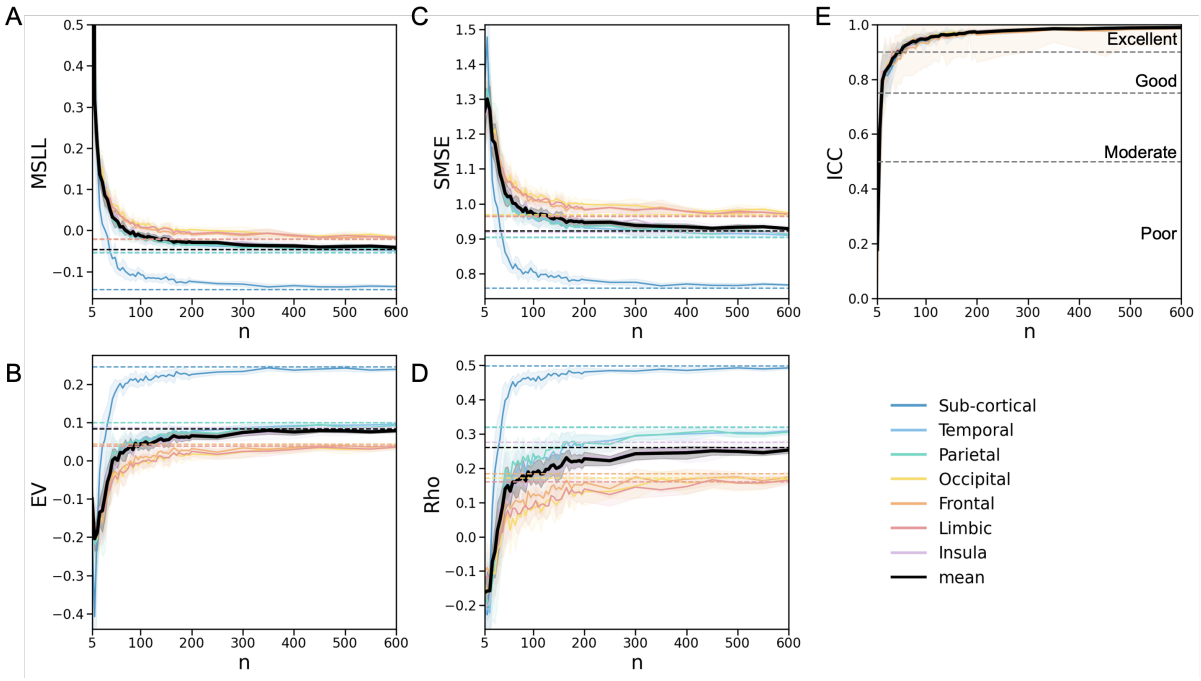

Figure 2. Model fit evaluation in the HC test set. Performance evaluation of models as a function of the number of participants used for fitting the models (n), calculated using the HC test set from the OASIS-3 dataset. Depicted are in A the MSLL; B: SMSE; C: EV; D: Rho; E: ICC. For all plots, solid lines represent the mean performance of the evaluation metric across all cortical region models grouped by lobe and the mean performance across subcortical region models. The thick black line indicates the overall mean performance across all cortical and subcortical models. In Panels A-D, dashed lines denote the performance using the full training set (n=692) and the shaded areas indicate the standard

deviation, reflecting variability across 10 iterations of same sample size. In Panel E, grey dotted lines indicate commonly accepted reliability thresholds: below 0.5 (poor), up to 0.75 (moderate), up to 0.9 (good) and above 0.9 (excellent) (Koo & Li, 2016).

Next, to isolate the effect of age distribution, we compared models trained on HC subsamples with representative age distributions (as used above) to those trained using left-skewed (younger-biased) and right-skewed (older-biased) age distributions, while maintaining balanced sex ratios. All models used the same incremental sample sizes and were evaluated on the same fixed HC test set (see Figure 1A, C, and Table 2).

Left-skewed sampling (i.e., oversampling of younger subjects) had the most pronounced adverse effect on model fit, with a marked increase in *SMSE* ($\beta = 1.323$, $p < .001$), increase in *MSLL* ($\beta = 0.219$, $p < .001$), decreases in *EV* ($\beta = -0.927$, $p < .001$), *Rho* ($\beta = -0.642$, $p < .001$), and in *ICC* ($\beta = -0.300$, $p < .001$). Right-skewed sampling (i.e., oversampling of older subjects) had a smaller effect on model fits compared to left-skewed sampling. It increased *SMSE* ($\beta = 0.585$, $p < .001$) and *MSLL* ($\beta = 0.115$, $p < .001$) and decreased *EV* ($\beta = -0.692$, $p < .001$) and *Rho* ($\beta = -0.567$, $p < .001$). The effect on *ICC* was minimal ($\beta = -0.024$, $p = .026$).

Model fit generally improved with increasing sample size across all metrics and sampling strategies. Interaction effects, however, revealed that the rate of improvement varied depending on the sampling strategy. Under left-skewed sampling, SMSE improved substantially with larger sample sizes (interaction $\beta = -0.636$, $p < .001$), suggesting a steeper recovery. Rho and ICC improved, but at slower rate of improvement than under representative sampling (interaction $\beta = -0.112$, $p < .001$ and $\beta = -0.032$, $p = 0.003$). Right-skewed sampling showed uniform improvements across metrics, with interaction effects indicating modest gains in EV ($\beta = 0.047$, $p < .001$), Rho ($\beta = -0.034$, $p < .001$), and ICC ($\beta = -0.048$, $p < .001$), and reductions in MSLL ($\beta = -0.084$, $p < .001$) and SMSE ($\beta = -0.139$, $p < .001$).

These results highlight that the effect of increasing sample size varies depending on the sampling strategy, with less consistent improvements observed under left-skewed conditions. All results for age-skewed samplings are summarized in Table 2.

Finally, to examine the effect of sex ratio imbalance, we trained models on HC subsamples with representative age distributions and varying sex ratios: balanced (1F:1M), 1F:10M, 1F:4M, 4F:1M, and 10F:1M, where *F* denotes females and *M* denotes males. As before, all models used the same incremental sample sizes and were evaluated on the fixed HC test set.

Sex distribution imbalances had a comparatively smaller effect on model fit than age distribution shifts (Table 3, Figure 3). Across metrics, the extent of deviation increased with the degree of imbalance. More extreme ratios (1F:10M and 10F:1M) were associated with moderate increases in *SMSE* ($\beta = 0.166$ and $\beta = 0.076$, respectively; both $p < .001$) and reductions in *EV* ($\beta = -0.124$ and $\beta = -0.033$, respectively; both $p < .001$), while more moderate imbalances (1F:4M and 4F:1M) produced minimal changes. *ICC* was reduced in all imbalanced conditions, with the largest effects observed under the most extreme ratios ($\beta = -0.152$ for 1F:10M and $\beta = -0.162$ for 10F:1M; both $p < .001$). Interaction effects with sample size were statistically significant in some cases but remained of small magnitude. All results for sex-unbalance samplings are summarized in Table 3.

In summary, these findings demonstrate that representative sampling generally provided the best performance in terms of model fit across metrics on the test set, while skewed distributions introduced substantial model fit degradation. The left-skewed sampling particularly exacerbated errors, as shown by the larger effect sizes. Increasing the sample size improved fits in all configurations. Sex ratio imbalance had a statistically significant but limited effect on model fit, with only minor deviations observed across metrics. These findings underscore the dominant influence of age distribution and sample size in model fit. Individual effects of sample size for the different age and sex distributions are presented in supplementary material (Figures S1-3) and importantly we replicated these findings in an independent dataset (AIBL) (Figures S17-20, Tables S5-6).

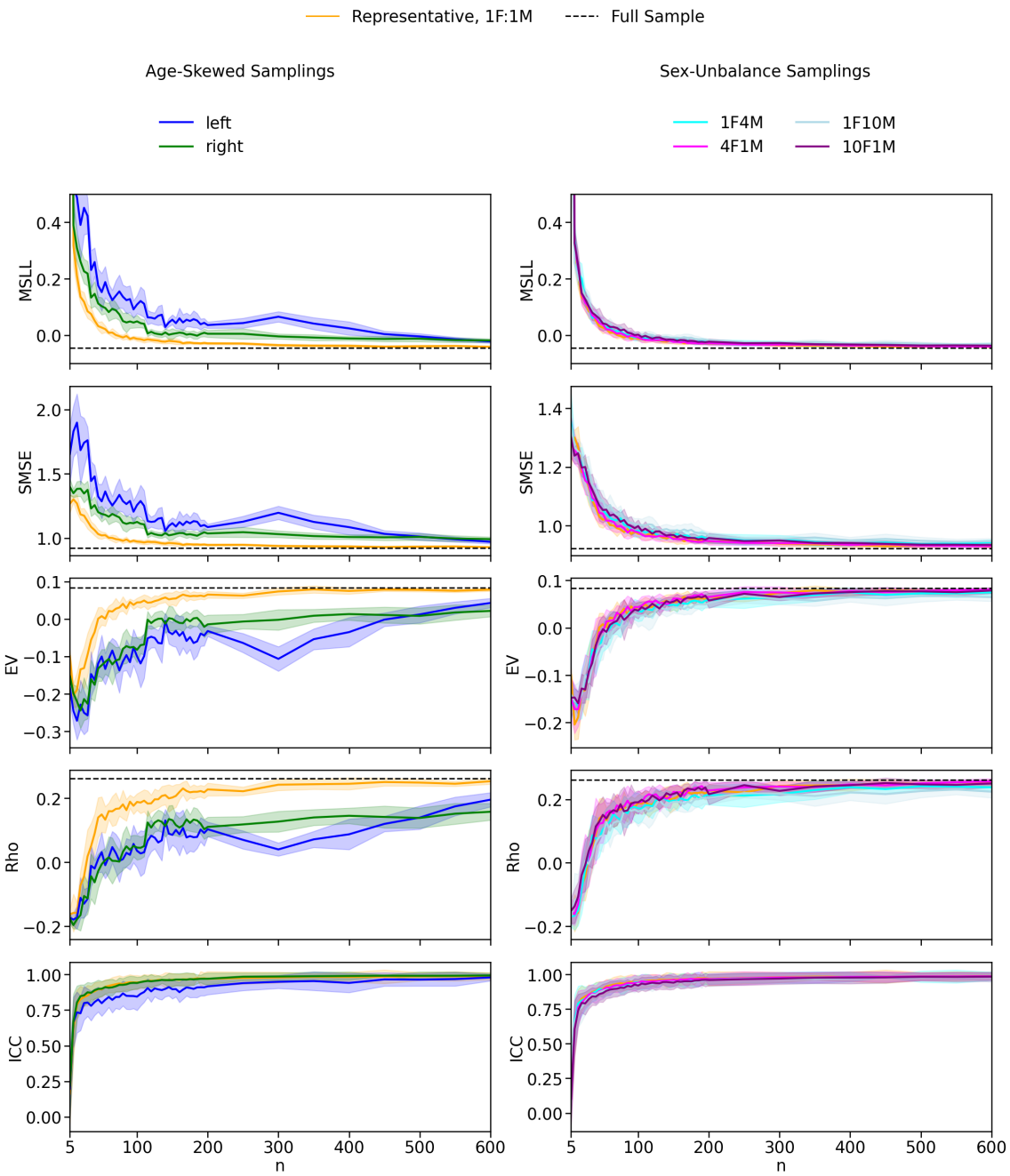

Figure 3. Evaluation of model fits in the HC test set across different sampling strategies of the training set, shown for varying sample sizes (n). Age-skewed sampling strategies include representative (matching the initial age distribution with balanced sex, 1F:1M), left-skewed (favoring younger individuals), and right-skewed (favoring older individuals), each with balanced sex distributions. Sex-imbalanced sampling strategies include female-to-male ratios of 1:1, 1:4, 1:10, 4:1, and 10:1, all with representative age distributions. Solid lines represent the mean metric values across ROIs and iterations, with shaded areas indicating the standard deviation across iterations for MSLL, SMSE, EV, and Rho. For ICC, shaded areas represent variation across ROIs, as ICC already reflects variability across iteration. Dashed lines indicate the mean metric values for models trained with the full sample for MSLL, SMSE, EV, and Rho.

	MSLL	***SMSE***	***EV***	***Rho***	***ICC***
	β (p-value)	***β (p-value)***	***β (p-value)***	***β (p-value)***	***β (p-value)***
Intercept	-0.031	-0.318	0.260	0.183	0.106
(Representative)	(p = .022)	(p < .001)	(p < .001)	(p < .001)	(p < .001)
Log(n)	-0.496	-0.427	0.499	0.527	0.728
	(p < .001)	(p < .001)	(p < .001)	(p < .001)	(p < .001)
Left-skewed	0.219	1.323	-0.927	-0.642	-0.300
	(p < .001)	(p < .001)	(p < .001)	(p < .001)	(p < .001)
Right-skewed	0.115	0.585	-0.692	-0.567	-0.024
	(p < .001)	(p < .001)	(p < .001)	(p < .001)	(p = .026)
Log(n):Left-skewed	-0.064	-0.636	0.019	-0.112	-0.032
	(p < .001)	(p < .001)	(p = .047)	(p < .001)	(p = .003)
Log(n):Right-skewed	-0.084	-0.139	0.047	-0.034	-0.048
	(p < .001)	(p < .001)	(p < .001)	(p < .001)	(p < .001)

Table 2. Linear mixed model results for evaluation metrics under age-skewed sampling conditions. Models assess the influence of standardized and log-transformed sample size (n) and age sampling strategy (Representative, Left-skewed, Right-skewed) on model performance metrics (MSLL, SMSE, EV, Rho, and ICC). All variables were standardized to allow comparison of effect sizes. Sample size was log-transformed to account for the non-linear association between sample size and model performance. Representative sampling serves as the reference level. Reported β coefficients and corresponding p-values indicate the direction and significance of each effect.

	MSLL	***SMSE***	***EV***	***Rho***	***ICC***
	β (p-value)	***β (p-value)***	***β (p-value)***	***β (p-value)***	***β (p-value)***
Intercept	-0.031	-0.318	0.260	0.183	NA
(1F1M)	(p = .005)	(p < .001)	(p < .001)	(p < .001)	
Log(n)	-0.496	-0.427	0.498	0.527	0.728
	(p < .001)	(p < .001)	(p < .001)	(p < .001)	(p < .001)
10F1M	0.001	0.076	-0.023	0.044	-0.152
	(p = .965)	(p < .001)	(p < .001)	(p < .001)	(p < .001)
4F1M	-0.015	0.010	0.010	0.042	-0.055
	(p = .252)	(p = .030)	(p = .108)	(p < .001)	(p < .001)
1F4M	-0.039	0.068	-0.062	-0.062	-0.048
	(p = .003)	(p < .001)	(p < .001)	(p < .001)	(p < .001)
1F10M	-0.062	0.166	-0.121	-0.098	-0.162
	(p < .001)	(p < .001)	(p < .001)	(p < .001)	(p < .001)
Log(n):10F1M	0.057	-0.018	0.005	-0.034	0.073
	(p < .001)	(p < .001)	(p = .428)	(p < .001)	(p < .001)
Log(n):4F1M	0.060	-0.009	0.015	-0.011	0.048
	(p < .001)	(p = .051)	(p = .018)	(p = .114)	(p < .001)

Log(n):1F4M	0.159	-0.020	0.001	-0.011	-0.013
	(p < .001)	(p < .001)	(p = .874)	(p = .134)	(p = .221)
Log(n):1F10M	0.279	-0.032	-0.019	-0.049	-0.017
	(p < .001)	(p < .001)	(p = .002)	(p < .001)	(p = .120)

Table 3. Linear mixed model results for evaluation metrics under sex-imbalanced sampling conditions. Models assess the influence of standardized and log-transformed sample size (n) and sex ratio in the training set (1F:1M, 1F:4M, 1F:10M, 4F:1M, 10F:1M; F = female, M = male) on model performance metrics (MSLL, SMSE, EV, Rho, and ICC). All variables were standardized to allow comparison of effect sizes. Sample size was log-transformed to account for the non-linear association between sample size and model performance. Representative sampling (1F:1M) serves as the reference level. Reported β coefficients and corresponding p-values indicate the direction and significance of each effect.

2.2. Z-scores errors

Subsequently, we assessed the direct impact of age and sex distributions and sample size on normative models' outcomes, quantified by the *mean squared error* (*MSE*) and *mean bias error* (*MBE*) in Z-scores relative to models trained with the full training set.

Increasing the sample size significantly reduced MSE in both HC and AD groups (β = –0.443, p < .001), with no significant interaction, indicating similar effects across groups. Age distribution sampling (Figure 4A) had a significant effect on Z-score errors: left-skewed sampling (i.e., oversampling of younger individuals) led to a larger increase in MSE (β = 1.045, p < .001) compared to right-skewed sampling (i.e., oversampling of older individuals) (β = 0.101, p < .001) (Figure 4B, Table 4).

Because the MSE quantifies the magnitude of errors but does not provide information about their direction, we additionally computed the MBE to assess whether errors reflected systematic over- or underestimation of deviation scores. Left-skewed sampling consistently resulted in negative *MBE* values, indicating overestimation of deviations, while right-skewed sampling produced positive MBE values, indicating underestimation, particularly at smaller sample sizes (Figure 4B).

Cubic regression analyses on age revealed that, under the representative distribution, Z-score errors were lowest in mid-range ages and increased toward both extremes. Left-skewed sampling amplified errors in older individuals, with deviations persisting even at larger sample sizes (e.g., n = 100). In contrast, right-skewed sampling led to elevated errors in younger individuals, which progressively decreased with increasing sample size (Figure 4D).

Age-related trends in MSE were exemplified in centile estimation maps for the Left-Hippocampus (Figure 4E), where left-skewed sampling led to higher centiles in older individuals (indicating overestimation), while right-skewed sampling produced lower centiles in younger individuals (indicating underestimation) (Figure 4E).

Sex imbalances in the training set had a smaller but statistically significant effect on Z-score errors (Table 5). Extreme imbalances led to higher MSE compared to the representative configuration, with the largest increases observed for the 10F:1M ($\beta = 0.089$, $p < .001$) and 1F:10M ($\beta = 0.083$, $p < .001$) ratios and smaller increases for 4F:1M ($\beta = 0.031$, $p < .001$) and 1F:4M ($\beta = 0.046$, $p < .001$). Errors did not statistically differ for the HC and AD groups ($\beta = 0.021$, $p = .137$). Errors tended to be lower for individuals matching the overrepresented sex in the training set, with discrepancies increasing with the degree of imbalance (Figure 4F).

Highly similar results were found for the AIBL dataset, with the exception that the right-skewed distribution led to a larger effect than in the OASIS-3 dataset (Figure S21, Tables S5–S6).

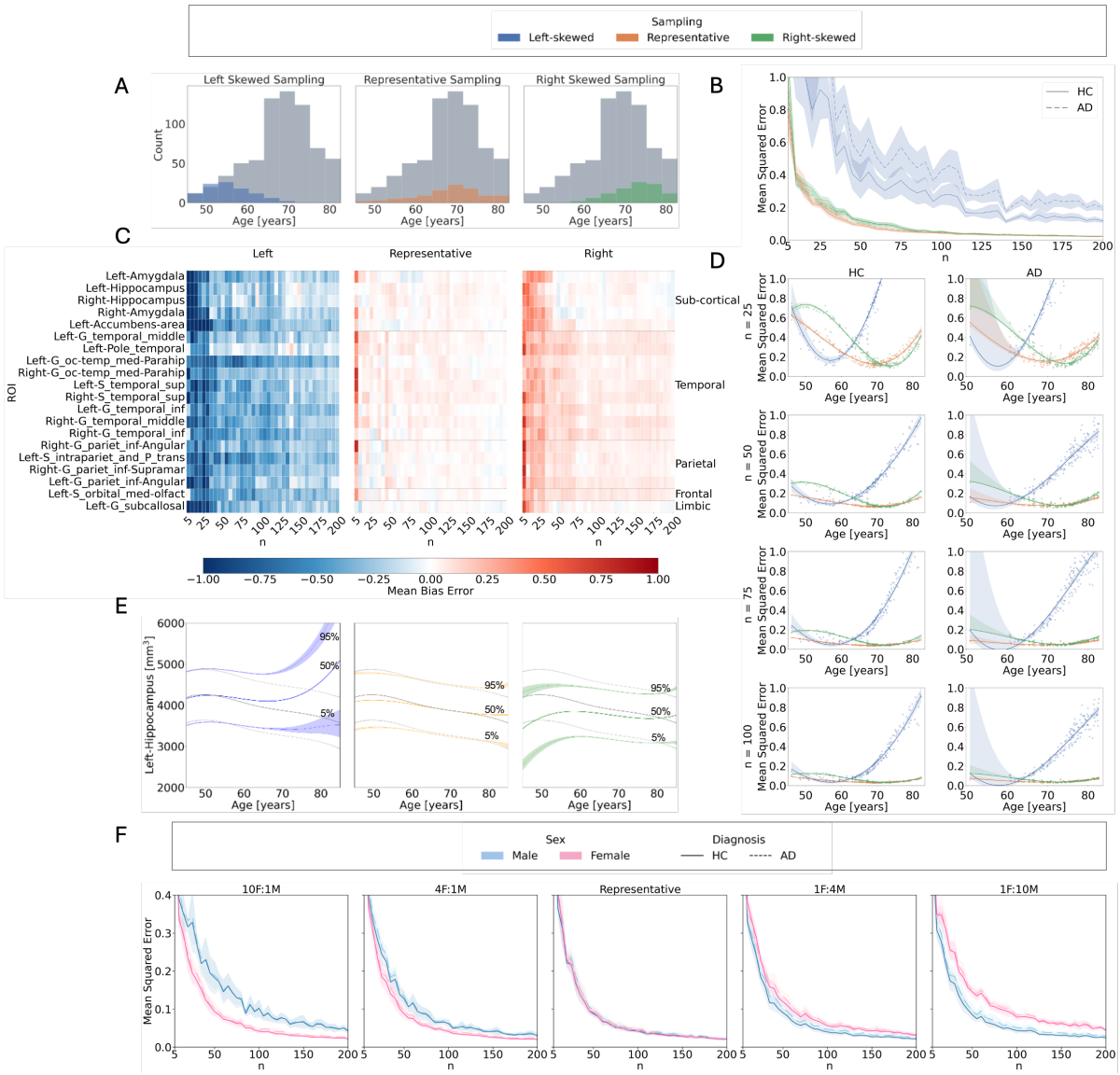

Figure 4. Z-score errors: Influence of sample size, age distribution, and sex imbalance on normative model outcomes. A: Age distributions for left-skewed (younger-biased), representative, and right-skewed (older-biased) sampling strategies, compared to the full training set in the OASIS-3 dataset. B: Mean squared error (MSE) of Z-scores relative to the full training set model across sample sizes, age distributions, and diagnostic groups. C: Mean bias error (MBE) per region across sample sizes and age distributions in the test set. Shown are the 20 brain regions with the highest Cohen's d effect sizes distinguishing between HC and AD calculated from models trained on the full training set. From left to right: results for left-skewed, representative, and right-skewed training sets. Blue indicates negative MBE (underestimation), red indicates positive MBE (overestimation), and white indicates close alignment with the full training set model. D: Cubic regression of MSE as a function of age, across sample sizes and sampling strategies. Left-skewed sampling shows increased errors in older individuals; right-skewed sampling shows increased errors in younger individuals. E: Centile curves for the Left Hippocampus as a function of age, derived from models trained on left-skewed, representative, and right-skewed sampling (n = 100). Colored lines represent the 5th, 50th, and 95th percentiles; grey lines show centiles from the full training set model. F: MSE across sample sizes and test set sex, obtained using sex-imbalanced training sets (female-to-male ratios: 10:1, 4:1, 1:1, 1:4, 1:10), all with representative age distributions.

	MSE	***tOC***
	β (p-value)	***β (p-value)***
Intercept (HC, Representative)	-0.245 (p < .001)	-0.443 (p < .001)
Log(n)	-0.443 (p < .001)	-0.004 (p = 0.405)
AD	0.021 (p = 0.570)	0.665 (p < .001)
Left-skewed	1.045 (p < .001)	0.461 (p < .001)
Right-skewed	0.101 (p < .001)	-0.021 (p = 0.004)
Log(n):AD	-0.007 (p = 0.530)	0.096 (p < .001)
Log(n):Left-skewed	-0.752 (p < .001)	-0.323 (p < .001)
Log(n):Right-skewed	-0.148 (p < .001)	0.019 (p = 0.008)
Left-skewed:AD	0.575 (p < .001)	0.721 (p < .001)
Right-skewed:AD	-0.077 (p < .001)	0.016 (p = 0.136)
Log(n):Left-skewed:AD	-0.234 (p < .001)	-0.290 (p < .001)
Log(n):Right-skewed:AD	0.100 (p < .001)	-0.004 (p = 0.674)

Table 4. Linear mixed model results for MSE and total outlier count (tOC) under age-skewed sampling. Models evaluate the influence of diagnosis (HC, AD), log-transformed and standardized sample size (n), and age sampling strategy (Representative, Left-skewed, Right-skewed) on standardized deviation score outcomes. Continuous variables were standardized to allow comparison of effect sizes. Sample size was log-transformed to account for the non-linear association between sample size and model performance. Representative sampling and HC are used as reference levels. Reported β coefficients and corresponding p-values indicate the direction and significance of the effects. Age was not included as an additional predictor to keep the modeling approach aligned with the sex-based analysis (Table 5) and to limit the complexity of the models.

	MSE	***tOC***
	β (p-value)	***β (p-value)***
Intercept (HC, 1F 1M)	-0.245 (p < .001)	-0.443 (p < .001)
Log(n)	-0.443 (p < .001)	-0.004 (p = 0.084)
AD	0.021 (p = 0.137)	0.665 (p < .001)
10F1M	0.089 (p < .001)	0.005 (p = 0.173)
4F1M	0.031 (p < .001)	0.006 (p = 0.076)
1F4M	0.046 (p < .001)	0.004 (p = 0.212)
1F10M	0.083 (p < .001)	0.005 (p = 0.170)
Log(n):AD	-0.007 (p = 0.153)	0.096 (p < .001)
Log(n):10F1M	-0.021 (p < .001)	-0.010 (p = 0.005)
Log(n):4F1M	-0.012 (p = 0.018)	-0.012 (p = 0.001)
Log(n):1F4M	-0.015 (p = 0.003)	-0.005 (p = 0.148)
Log(n):1F10M	0.037 (p < .001)	0.014 (p < .001)
10F1M:AD	0.025 (p < .001)	-0.006 (p = 0.258)
4F1M:AD	0.013 (p = 0.073)	0.009 (p = 0.079)
1F4M:AD	-0.011 (p = 0.133)	-0.002 (p = 0.718)
1F10M:AD	-0.013 (p = 0.072)	-0.013 (p = 0.012)
Log(n):10F1M:AD	-0.013 (p = 0.078)	0.009 (p = 0.092)

Log(n):4F1M:AD	-0.013 (p = 0.081)	0.001 (p = 0.900)
Log(n):1F4M:AD	0.013 (p = 0.081)	-0.006 (p = 0.220)
Log(n):1F10M:AD	-0.002 (p = 0.750)	0.000 (p = 0.930)

Table 5. Linear mixed model results for MSE and total outlier count (tOC) under sex-imbalanced sampling. Models evaluate the influence of diagnosis (HC, AD), log-transformed and standardized sample size (n), and sex ratio in the training set (1:1, 1F:4M, 1F:10M, 4F:1M, 10F:1M) on standardized deviation score outcomes. Continuous variables were standardized to allow comparison of effect sizes. Sample size was log-transformed to account for the non-linear association between sample size and model performance. The 1:1 ratio and HC are used as reference levels. Reported β coefficients and corresponding p-values indicate the direction and significance of the effects. Sex was not included as an additional predictor to keep the modeling approach aligned with the age-based analysis (Table 4) and to limit the complexity of the models.

2.3. Clinical validation

Clinical validity was evaluated for every sampling strategy and sample size using three metrics: (1) the proportion of subjects with atrophy outliers in each ROIs ($Z < -1.96$); (2) the *total outlier count* (*tOC*) per subject, defined as the number of ROIs in which a subject's deviation score exceeds the Z threshold; and (3) the *ROC-AUC* of an support vector classifiers (SVC) distinguishing HC versus AD based on deviation scores (Figure 1C).

In the model fitted using the full training set, the highest percentages of volume outliers in the AD group were observed in subcortical regions such as the left hippocampus (30.5%) and left amygdala (28.1%). Among cortical areas, the middle temporal gyrus and parahippocampal gyrus showed also frequent deviations (e.g., 22.8% in the right middle temporal gyrus, 19.2 % in parahippocampal gyrus).

To illustrate how outlier estimation varies with training sample size, example iterations are shown in Figure 5A. At n = 25, outlier percentages were substantially lower than in the full sample model, for example, 8.4% in the left hippocampus (–22.1% compared to the full sample) and 0% in the left amygdala (–28.1%). At n = 50, estimates became closer to the full model in some regions (e.g., 24.0% in the left hippocampus, –6.5%), while others showed higher deviation rates (e.g., 41.3% in the left amygdala, +13.2%). Outlier estimates in the example at n = 75 closely matched those of the full model across many regions; for example, the left hippocampus showed 30.5% (-0.0%). At n = 100, values showed minor variability, with the left hippocampus slightly above the reference at 34.7% (+4.2%). This residual fluctuation across runs highlight some instability in outlier estimation even at higher sample sizes. Similar variability was observed in cortical areas. For instance, the percentage of outliers in the right middle temporal gyrus increased from 12.6% at n = 25 (–10.2%) to 18.0% at n = 100 (–4.8%), compared to 22.8% in the

full model. In the right parahippocampal gyrus, values rose from 3.6% at n = 25 (–15.6%) to 18.0% at n = 100 (–1.2%), relative to 19.2% in the full model.

tOC analysis (Figure 5B, Tables 3-4) showed consistently higher estimated outlier counts in AD compared to HC. Models trained on the full training set estimated an average of 14 outliers per individual in AD compared to 4 in HC. This group difference was confirmed by the statistical models (Tables 3-4). which reported a significant main effect of diagnosis ($\beta = 0.665$, $p < .001$), independent of sample size or sampling distributions.

No significant main effect of sample size was found in HC, as outlier counts quickly converged to full-model estimates. In AD, a significant interaction with sample size ($\beta = 0.096$, $p < .001$) indicated that tOC was underestimated at small sample sizes and increased with larger sample sizes, with representative and right-skewed sampling converging to full-model estimates around n = 100 (Figure 5B).

Left-skewed sampling strongly increased tOC ($\beta = 0.461$, $p < .001$), an effect amplified in AD (interaction $\beta = 0.721$, $p < .001$). Under this sampling strategy, tOC only slowly decreased with larger sample sizes ($\beta = -0.323$, $p < .001$ for the interaction of sample size with left-skewed sampling; $\beta = -0.290$, $p < .001$ for the three-way interaction with AD).

Sex-imbalanced samplings showed no significant main effects on tOC (all $p > .05$). Several interactions with sample size were significant for unbalanced sex ratios (e.g., 4F1M $\beta = -0.012$, $p = .001$; 1F10M $\beta = 0.014$, $p < .001$), but these effects were very small compared to those observed for age distributions. The only interaction with AD reaching significance was under the extreme male-dominated condition (1F10M:AD $\beta = -0.013$, $p = .012$), indicating a slight moderation of tOC in AD group in that scenario.

Finally, the influence of training sample size and age distribution on classification performance was assessed by evaluating the ability to distinguish AD and HC groups based on deviation scores from the normative models (i.e. Z-scores). ROC-AUC increased when increasing sample size for all sampling strategies, reaching a performance comparable to the models trained with the full training sample size (AUC = 0.86) at around 15 samples, and stabilized from that point onwards (5C). This suggests that small training sets may already capture relevant group-level patterns.

Importantly, all findings from the clinical validation analyses were replicated in the independent AIBL dataset (Figures S22, Tables S7–S8).

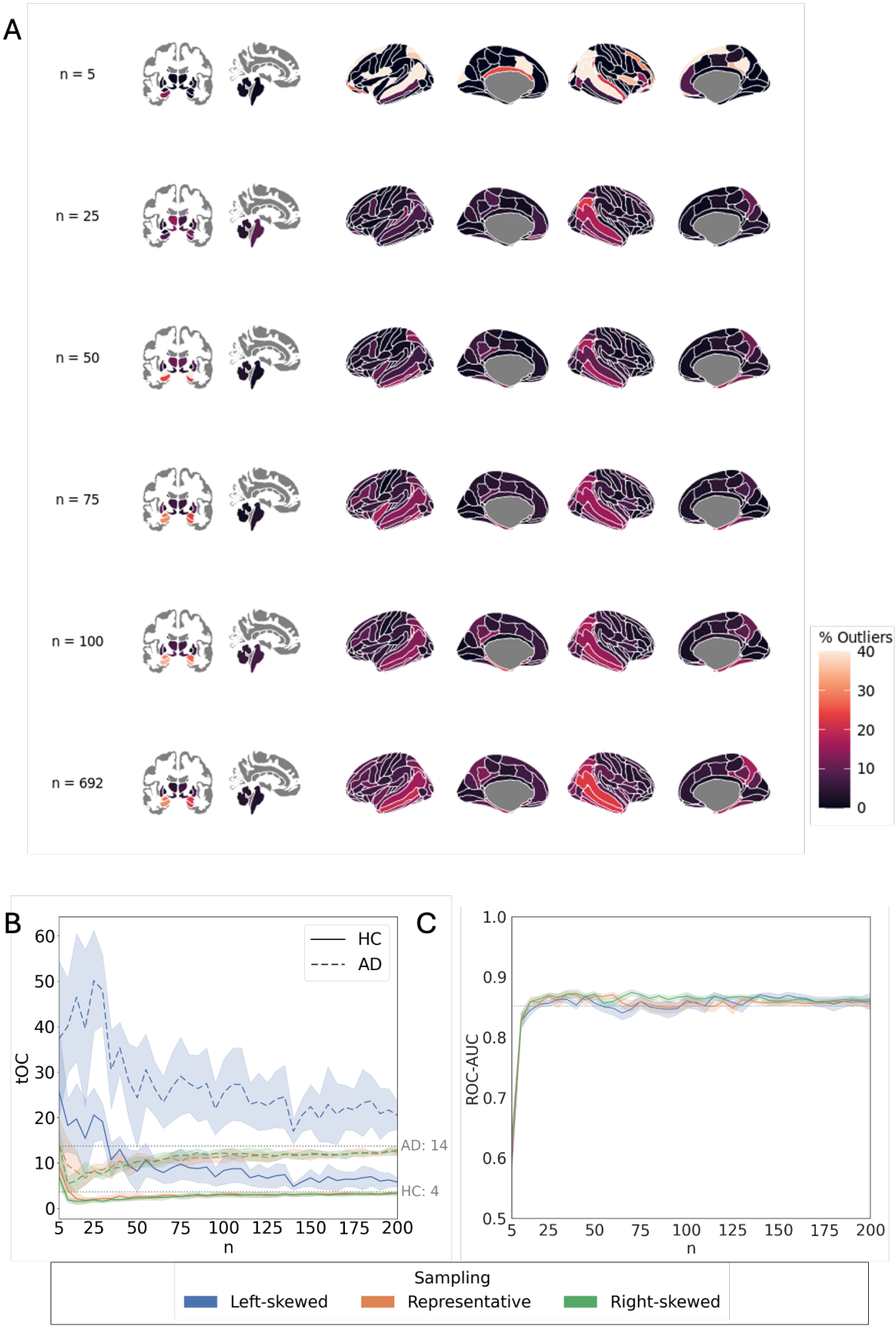
A
n = 5
n = 25
n = 50
n = 75
n = 100
n = 692
% Outliers
40
30
20
10
0
B
C
HC
AD
tOC
AD: 14
HC: 4
n
ROC-AUC
Sampling
Left-skewed
Representative
Right-skewed

Figure 5. Clinical validation: effect of sample size and age distributions in outlier detection and classification performance. A: Percentage of participants with extreme negative deviation (Z < -1.96) in each brain region in the AD group, shown for independent example iterations at different sample sizes with representative age distribution. B: Average total Outlier Count (tOC) in the AD and HC are represented with dashed and solid lines respectively as a function of sample size (n). The shaded areas indicate the standard deviation across iterations. Dotted grey lines correspond to the estimations of tOC for HC and AD groups obtained with full sample size. C: ROC-AUC of Support Vector Classifier with 10-fold cross-validation as a function of sample size in the training set. The solid lines represent the average AUC across iterations for each sampling strategy. The dotted line shows the performance of the models trained with full sample size.

2.4. Adaptation from large dataset

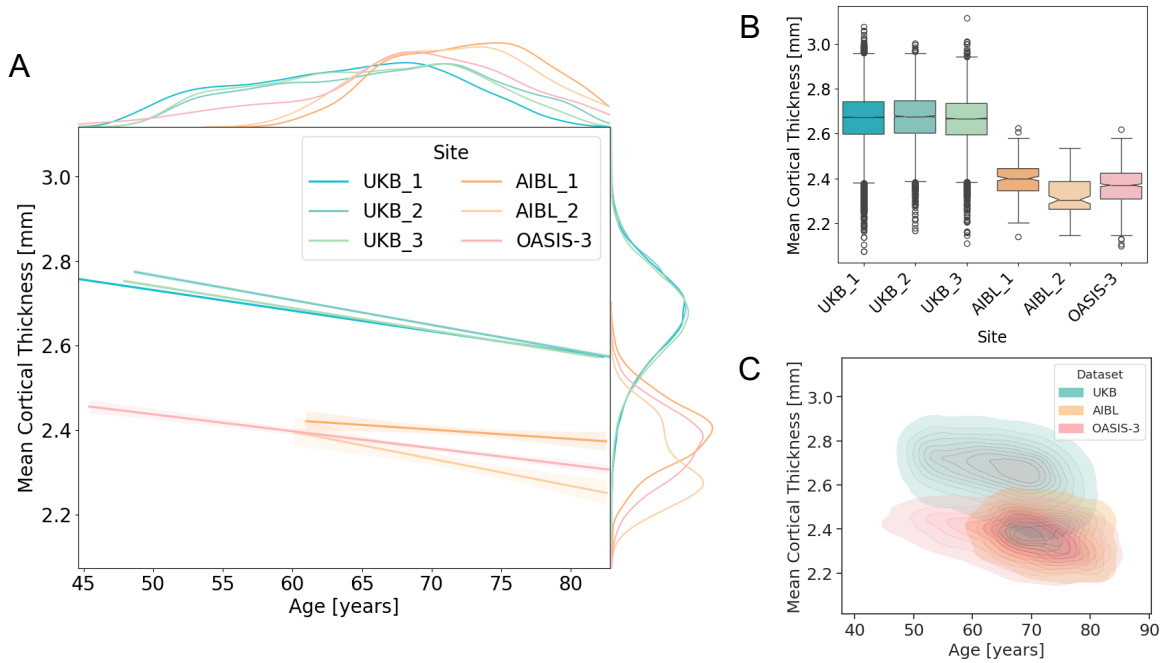

Figure 6. Site effects on mean cortical thickness (whole-brain average) in the UK Biobank (UKB), AIBL, and OASIS-3 datasets for HC. A: Regression plots showing the relationship between age and mean cortical thickness across different datasets and imaging sites (UKB: 3 sites, AIBL: 2 sites, OASIS-3: 1 site). Marginal density plots are displayed on the sides of each axis, illustrating the distribution of mean cortical thickness and age for each dataset's imaging sites. The UKB exhibits notable differences in mean cortical thickness estimations compared to AIBL and OASIS-3, unrelated to the age distribution of participants. B: Boxplots representing the mean cortical thickness for participants in each dataset and imaging site. C: Bivariate kernel density estimates plots showing the joint distribution of age and mean cortical thickness for each dataset. The contour plots represent the density of data points, with filled areas reflecting higher concentrations of data.

To motivate the need for model adaptation across cohorts (Figure 1B), we examined differences in mean cortical thickness between UKB and the clinical datasets. Systematic differences were evident, with UKB consistently showing higher mean cortical thickness compared to AIBL and OASIS-3, regardless of age (Figure 6). These differences most likely reflect

acquisition-related variability and highlight the necessity of adapting models when applying the method to new datasets.

Adapted models based on pre-trained UKB data and transferred to OASIS-3 rapidly achieved the performance level of models trained directly within the cohort (Figure 7). Across all evaluation metrics (*MSLL*, *SMSE*, *EV*, *Rho*, and *MSE*) adapted models plateaued around n = 50, whereas models trained directly within cohort required approximately n = 200 to reach comparable performance. Explained variance (EV) and correlation (Rho) remained stable across all adaptation sample sizes, as adaptation modifies only the mean of the predictions without changing their variance or rank structure.

Despite the convergence in evaluation metrics, outlier detection at full adaptation sample size (n = 692) revealed differences between the two modeling strategies, with and without adaptive transfer learning. In the left amygdala, the adapted model detected 15.0% outliers compared to 28.1% for the models trained directly without adaptation (–13.1%); in the left hippocampus, 17.4% versus 30.5% (–13.1%); in the right amygdala, 11.4% versus 23.4% (–12.0%); and in the right hippocampus, 15.0% versus 22.2% (–7.2%). Cortical regions also showed discrepancies, such as the right middle temporal gyrus (17.4% adapted vs. 22.8% within-cohort, –5.4%) and the left parahippocampal gyrus (14.4% vs. 16.2%, –1.8%).

Adapted models, while converging faster to the full adaptation set model estimation compared to within-cohort trainings, still showed instability in outlier detection at small adaptation sample sizes. In the example iterations in Figure 7B, at n = 25, the percentage of outliers was 7.2% in the left amygdala and 1.8% in the left hippocampus (–7.8 and –15.6%, respectively); at n = 50, values reached 19.2% and 9.6% (+4.2 and –7.8%). At n = 100, percentages were 15.0% in the left amygdala and 18.0% in the left hippocampus, closely aligning with the full model (differences under ±1%).

Similar to the direct training case, these values correspond to individual example iterations and are shown for illustration. The full replication of the previous analysis using UKB-adapted models to OASIS-3 and AIBL is provided in the Supplementary Materials (Figures S9-16 and S26-33 respectively).

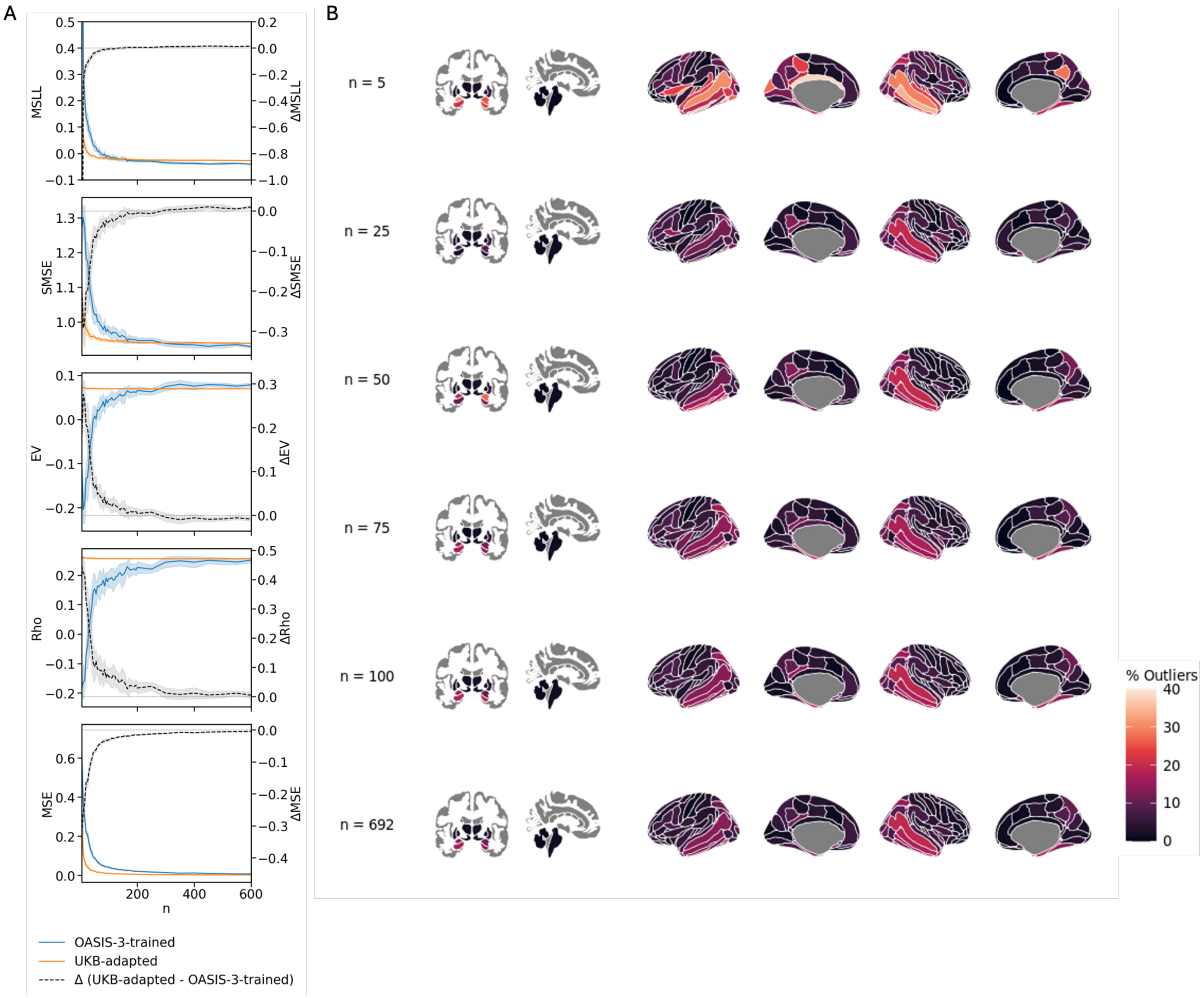

Figure 7. Adaptive transfer learning evaluation. A. Comparison of model performance between direct training models (OASIS-3-trained) and models pretrained on UKB then adapted to OASIS-3 (UKB-adapted). The grey curve shows the difference in performance (within-cohort minus adapted), plotted on a separate y-axis (right). For MSLL, SMSE, and MSE, lower values indicate better performance; therefore, positive differences indicate better performance of the within-cohort models. For EV and Rho, higher values indicate better performance; hence, positive differences reflect better performance of the adapted models. EV and Rho values remain constant across models, as the adaptation procedure only modifies mean of models and does not affect the shape. B. Example adaptation iteration showing the percentage of individuals identified as outliers per region. Last row shows the outlier detection obtained using the full adaptation sample (n = 692).

3. Discussion

In this study, we systematically examined how sample size and covariate distributions impact the fit and clinical utility of normative models to AD datasets. To ensure the robustness and generalizability, we leveraged a large-scale discovery sample (OASIS-3) and a replication sample (AIBL). Models were first fitted on 80% of the HC participants from the OASIS-3, representing the optimal scenario in which the largest available dataset is leveraged to maximize accuracy and reliability. These full-sample training models served as a benchmark against which we evaluated the impact of reduced sample sizes and covariate misalignment. We systematically sub-sampled the full training set to vary sample sizes and manipulate covariate distributions (i.e. age and sex) and then assessed their effects on model fit in the HC test set as well as on deviation scores and outlier detection in both the HC test set and the AD cohorts. Finally, we applied the same sampling and evaluation framework to an adaptive transfer learning setting. In which pre-trained UK Biobank models were adapted using the same clinical training subsets method previously used for direct training.

Across all analyses, increasing sample size yielded to more stable model fits and more consistent results. In contrast, misaligned covariate distributions, particularly with respect to age, introduced systematic distortions in model predictions and outlier detection. In the following sections, we provide a detailed analysis of how sample size and covariate distributions affected model evaluation and practical application.

Across all analyses, models trained on the full training set of available HC yielded the best overall fit, demonstrating superior performance when all data are utilized. While we acknowledge that this model does not represent an absolute ground truth, it served as a robust reference for evaluation how variations in sample size and covariate distribution affect model fit. Models trained directly on OASIS-3 achieved model fit comparable to UKB-pretrained models, indicating that training solely on the target cohort was sufficient for reliable estimation. The models fitted with full training set also provided a strong reference for biological plausibility and interpretability, producing results consistent with established AD pathology (Igarashi, 2023; Planche et al., 2022). Specifically, in the HC test set, full training set models performed as expected, with the average *tOC* per individual around 4 in both datasets, aligning with expectations given the defined threshold for outlier detection (bottom 2.5% among 167 ROIs). In contrast, individuals with AD exhibited elevated *tOC* values, averaging at 14 in OASIS-3 and 20 in AIBL. Importantly, regions surrounding the entorhinal cortex, hippocampus, and amygdala exhibited the highest outlier percentage, highlighting their well-established involvement in AD pathology (Igarashi, 2023).

Subsampling analyses revealed that the largest improvements in model performance occurred up to approximately 50 subjects, at which point metrics such as the intraclass correlation coefficient (ICC) reached excellent reliability and 60–90% of total performance gains were achieved. Beyond 200 subjects, performance began to plateau with diminishing returns up to 300 and with only marginal improvements up to 600 subjects.

Sample-size requirements depend strongly on the intended use of a normative model. In the present study we exemplified two applications commonly used in recent normative-modeling work: (1) outlier detection through regional deviation mapping across the brain and the tOC per individual (Bhome et al., 2024; Floris et al., 2021; Loreto et al., 2024; Rutherford et al., 2023; Verdi et al., 2023, 2024; Zabihi et al., 2020); (2) case-control classification to assess whether the relative population ranks were preserved. Subsampling of real data showed that approximately 100-200 well-characterized HC were sufficient to generate clinically robust deviation maps and *tOC* values, while adding additional participants primarily reduced residual variance at a high acquisition cost. For the classification task, performance in OASIS-3 was indistinguishable from the full model with only fifteen training samples.

Sample size had no significant effect on *tOC* in the HC group, likely due to the low number of deviations expected in healthy aging. In contrast, *tOC* estimates in the AD group stabilized with increasing sample size. This was particularly evident in regions such as the hippocampus and amygdala, where outlier detection was inconsistent at lower sample sizes despite their known vulnerability in AD (Barnes et al., 2009; Qu et al., 2023). Previous work using simulated hippocampal-volume data in order to suggested that several thousand subjects are needed to stabilize the extreme lower percentiles (Bozek et al., 2023). In practice, achieving cohorts of that scale is hindered by high data acquisition costs, privacy constraints, and the logistical challenges of harmonizing data from multiple sites. Moreover, our clinical-validation tasks illustrated that much lower sample sizes saturates clinical readouts estimation. Our results therefore do not prescribe a universal threshold; rather, they provide a reference for the precision that can reasonably be expected from sample sizes commonly attainable in clinical studies. Roughly 200–300 participants emerge as a pragmatic baseline for deviation mapping in ageing cohorts, while larger data sets are strongly recommended for accurate percentile screening. Notably, broader lifespan applications may require larger samples to ensure coverage across the full age range. This is further supported by our analysis of age-skewed distributions, which highlights the importance of aligning the training cohort with the target population.

Indeed, beyond sample size, skewed age distributions affected both models fit and the accuracy of deviation scores. Deviation scores were systematically overestimated in older individuals when models were trained on younger (left-skewed) samples and underestimated in younger individuals when trained on older (right-skewed) samples. These patterns align well with well-documented age-related brain changes, including cortical thinning and subcortical atrophy with increasing age (Lemaitre et al., 2012). Left-skewed samplings inflated normative means, leading to more pronounced negative deviations in older individuals, whereas right-skewed samplings lowered references, underestimating deviations in the young. As also observed in the simulation study by Bozek et al. (2023), errors were most pronounced at the extremes of the age distribution, where data were sparse. Centile analyses confirmed that models performed reliably across well-sampled age ranges but produced unstable estimates at underrepresented ages, even under representative sampling. These edge effects persisted despite covariate alignment and were attenuated with larger sample sizes.

Our findings highlight the critical importance of age distribution in clinical applications. When the age range of the training data failed to align with that of the target population, deviation estimates became systematically biased, affecting both global outlier counts and regional detection patterns. This was most notably under the left-skewed sampling, where younger-biased training sets inflated deviation scores across both HC and AD groups, with pronounced effects in regions such as the hippocampus and amygdala. Importantly, these biases were not fully resolved until relatively large sample sizes were reached. While error under right-skewed sampling gradually diminished with increasing sample size, overestimation persisted even at large sample sizes in left-skewed distributions.

The left-skewed sampling had overall a greater effect than right-skewed sampling in both model evaluation and clinical validation, likely due to (1) the dataset's original bias toward older individuals, making younger-skewed samples less representative, and (2) the older age structure of the AD population, which exacerbates mismatch when younger HC are used to calibrate models in the clinical population. Although we generated skewed distributions artificially, they mirror real-world imbalances frequently observed in neuroimaging studies, including in our own dataset, where recruitment constraints often result in age-biased cohorts (Bethlehem et al., 2022; LeWinn et al., 2017).

Together, these findings underscore the need for careful covariate matching, when developing normative models for clinical application, not only in terms of sample size but also distributional structure. This is particularly important in disorders such as AD, where age is tightly

linked to disease progression and anatomical change. Deviations from a representative age distribution can propagate errors that directly affect individual-level interpretation. However, recruiting very old healthy individuals is inherently challenging. With increasing age, clinical and biological heterogeneity also increases, due in part to the greater inter-individual variability in brain structure (Fjell & Walhovd, 2010; Lemaitre et al., 2012), accumulation of comorbidities, and the difficulty of confidently excluding preclinical or undiagnosed conditions. Additionally, selective survival effects may introduce sampling biases, further complicating the construction of representative reference cohorts in the oldest age ranges.

Given the higher prevalence of AD in women and documented sex-related differences in brain anatomy (Ferretti et al., 2018; Küchenhoff et al., 2024; Riedel et al., 2016; Ruigrok et al., 2014), we examined whether imbalances in sex distribution influenced model fit or deviation scores. Overall, sex imbalances had minimal effects in both model evaluation and clinical validation. Model fit remained stable, and only small improvements in deviation scores accuracy were observed. Deviation score accuracy was higher for individuals whose sex was more represented in the training data, and the strength of this association increased with the degree of sex imbalance. While sex imbalances had little impact on global model performance, some interaction effects with diagnosis were significant in deviation scores and outlier detection, particularly under extreme sex ratios. However, these effects were small and inconsistent, indicating limited practical impact overall. This limited sensitivity suggests that sex-related anatomical variation may generalize more robustly than age-related changes, especially in ageing populations. Moreover, unlike age, which introduces continuous, nonlinear variability, sex is a discrete covariate, which makes it easier to model. These findings indicate that moderate sex imbalances are unlikely to affect overall model performance, although sex-matched training data may still be beneficial for improving individual-level assessments.

Our findings highlight the potential of adaptive transfer for efficient model fitting in clinical cohorts if large datasets are available. Normative models pre-trained on the external large-scale UKB dataset rapidly achieved performance levels comparable to models trained directly within the cohort, requiring substantially fewer individuals. Evaluation metrics plateaued around 50 participants for adapted models, whereas within-cohort models required approximately 200 to achieve similar fit. This suggests that pre-trained models can be effectively recalibrated with modest sample sizes, offering practical advantages in clinical or low-resource contexts where large datasets are often unavailable. Notably, this stabilization threshold of 50 participants was higher than the 25 participants reported by Gaiser et al. (2024), who used Hierarchical Bayesian

Regression (HBR). HBR's hierarchical structure allows information sharing across sites, which may facilitate convergence with fewer samples. In contrast, our study used Bayesian Linear Regression (BLR), a method that does not account for cross-site dependencies and may require more data to accommodate population heterogeneity. Nevertheless, BLR remains widely used in recent normative modeling applications (Bhome et al., 2024; Corrigan et al., 2024; Holz et al., 2023; Meijer et al., 2024; Rutherford, Kia, et al., 2022; Savage et al., 2024) because of its computational efficiency, ease of implementation, and flexibility. Another factor likely contributing to this difference in stabilization thresholds is the demographic composition of the studied populations. Gaiser et al. (2024) examined on a younger, developmentally homogeneous cohort (ages 6–17), whereas our sample spanned older adults (44–82 years), a range characterized by greater inter-individual variability in brain structure (Fjell & Walhovd, 2010; Lemaitre et al., 2012). This increased heterogeneity may necessitate larger adaptation sets to achieve stable and reliable deviation estimates.

While evaluation metrics indicated convergence in overall model performance, differences remained in outlier detection between adapted and within-cohort models. In several regions commonly affected in AD, adapted models consistently identified fewer outliers, even at the largest adaptation sample size. These discrepancies suggest that, although adaptation effectively aligns prediction accuracy, subtle differences in the underlying normative distributions may persist and influence deviation-based measures. Furthermore, the age distribution of the adaptation set continued to influence both model fit and outlier detection, despite the pre-trained UKB model already covering the same age range extensively. In contrast, sex distribution had minimal influence on the adaptation outcomes.

Overall, these results support the use of adaptive transfer as a resource-efficient alternative to full retraining, particularly when pre-trained models are available for the targeted demographics.

Some limitations should be considered. This study relied on datasets in which adults of European ancestry were overrepresented. Although, the UK Biobank, which was used to pre-train the model, provided a robust reference distribution for the targeted populations of OASIS-3 and AIBL in this study, expanding reference datasets to include populations with more diverse socioeconomic and ethnic backgrounds would further improve representativeness and ensure broader applicability of normative models (Harnett et al., 2024). Moreover, we did not simultaneously manipulate age and sex distributions in the training and adaptation sets. Exploring these factors jointly would have required a substantially more complex experimental

design and statistical modeling, as their interactions could introduce confounding effects that are difficult to disentangle within the current framework. Finally, our analysis was restricted to cortical thickness for cortical regions and volumes for subcortical structures. Thickness was preferred in the cortex as it is more sensitive to age-related cellular alterations (Lemaitre et al., 2012), while volumetric measures were used for subcortical regions, where the absence of a clear laminar organization and current MRI resolution limit reliable thickness estimation (Dima et al., 2021; Lemaitre et al., 2012). Future work could incorporate additional structural or microstructural metrics to enhance the characterization of brain changes and improve the applicability of normative models across conditions.

Collectively, our results indicate that the benefits of enlarging a reference cohort depend critically on how closely its demographic profile mirrors that of the clinical sample. Gains from additional participants plateau when age alignment is poor, whereas even moderate-sized cohorts perform well when demographic concordance is preserved. This nuance underscores the importance of balancing sample size with representativeness when assembling reference datasets for normative modeling in ageing and neurodegeneration research. Our findings provide practical guidelines for maximizing the value of moderately sized but deeply phenotyped datasets, thereby reducing costs and limiting the need for additional large-scale data collection without compromising accuracy.

1. Methods

3.1. Data

3.1.1. Cohorts

We used data from three publicly available datasets: two clinical datasets including individuals diagnosed with AD and HC, namely the Open Access Series of Imaging Studies-3 (OASIS-3) and the Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL) (Fowler et al., 2021) for replication. In addition the UKB (Sudlow et al., 2015) was used for pre-training of normative models for the adaptive transfer learning experiments. In the following, we provide more details on each dataset (see also Table 1):

OASIS-3

The OASIS-3 sample included 1,098 participants from the United States. For the present project, we included participants from the OASIS-3 whose ages overlapped with the UKB age range, had available T1-weighted MRI scans, passed QC checks, and were diagnosed with either AD or identified as HC. Only baseline scans were included to ensure independence across observations, and individuals with mild cognitive impairment or other dementia-related diagnoses were excluded. The final OASIS-3 sample consisted of 1,032 individuals aged from 45 to 82 years (58% female), including 865 HC participants and 167 diagnosed with AD (LaMontagne et al., 2019). Image acquisitions were performed at a single site. Detailed acquisition protocols are provided in https://bpb-us-w2.wpmucdn.com/sites.wustl.edu/dist/6/4383/files/2024/04/OASIS-3_Imaging_Data_Dictionary_v2.3-a93c947a586e7367.pdf. Ethics approval was obtained from the institutional human research ethics committees of Austin Health, St Vincent's Health, Hollywood Private Hospital, and Edith Cowan University, and all participants provided written informed consent.

AIBL

AIBL is a longitudinal Australian cohort study comprising 2,359 participants. Data was collected by the AIBL study group. AIBL study methodology has been reported previously (Ellis et al., 2009). Participants were included if they had available T1-weighted MRI scans that passed QC checks (see section 3.1.2) were aged were aged 44–82 years (overlapping with the UKB range), and were classified as HC or diagnosed with AD. Individuals with other dementia-related

diagnoses were excluded, and only baseline time points were retained, consistent with OASIS-3. The final AIBL sample consisted of 462 individuals aged 58–82 years (58% female), including 403 HC and 60 AD participants. MRI data were acquired at two sites; acquisition protocols are described in detail elsewhere (Ayton et al., 2017; Dima et al., 2021). Ethics approval was obtained from the institutional human research ethics committees of Austin Health, St Vincent's Health, Hollywood Private Hospital, and Edith Cowan University, and all participants provided written informed consent.

UKB

The UKB is a large-scale health and imaging databank of 500,000 participants, from which 42,747 healthy individuals aged 44 to 82 years (53% female) were selected based on the availability of T1-weighted MRI scans and their quality control (QC) assessment (see section 3.1.2. for more details on quality control and preprocessing steps of all datasets). Only baseline time points were included when longitudinal data were available. MRI data were collected at three different sites using harmonized protocols. The detailed acquisition protocols are provided in https://www.fmrib.ox.ac.uk/ukbiobank/protocol/. All participants provided full informed consent, and the study received ethics approval from the National Health Service National Research Ethics Service (Ref 11/NW/0382). This research was conducted using the UKB resource under application number 96841.

3.1.2. MRI pre-processing and ROI extraction

All datasets underwent processing using FreeSurfer software, with quality control (QC) performed primarily after segmentation (Fischl, 2012). The standard FreeSurfer pipeline included intensity normalization, skull stripping, and segmentation of grey matter (GM) and white matter (WM) surfaces, ultimately producing a tessellated representation of the GM/WM boundary. These surfaces were then corrected for topological errors and aligned to a spherical atlas based on individual cortical folding patterns. Cortical thickness was defined as the shortest distance between the vertices of the GM/WM boundary and the pial surface, and cortical thickness was measured for each individual using the Destrieux parcellation scheme, which included 74 regions per hemisphere (Destrieux et al., 2010). For a comprehensive overview of the pipeline and its documentation, refer to the official FreeSurfer website (https://surfer.nmr.mgh.harvard.edu/).

In this study, we focused on cortical thickness and on subcortical volumes. Cortical thickness was included as it is a sensitive biomarker of age-related atrophy, particularly in the parietal cortex, and provides valuable insights into neurodegeneration in AD (Lemaitre et al.,

2012; McGinnis et al., 2011). Cortical volumes, which account for both cortical folding and thickness, may be less sensitive to specific cellular changes such as neuronal, dendritic, and synaptic changes (Lemaitre et al., 2012). Subcortical volumes were also considered, as AD is known to particularly affect the entorhinal cortex and surrounding regions (Qu et al., 2023; Yan et al., 2019; Zhao et al., 2019). These structures lack the laminar organization of the cortex, making thickness estimates unreliable, as a result, volumetric measures are typically used for subcortical regions. (Dima et al., 2021; Lemaitre et al., 2012).

For labeling subcortical tissue classes, the volume-based stream was utilized. After affine registration to MNI305 space, initial volumetric labeling, and B1 bias field intensity correction, a high-dimensional nonlinear volumetric alignment to the MNI305 atlas was performed, with the volume labeled according to the ASEG atlas (Fischl et al., 2002). From this atlas, we retained 19 subcortical regions.

3.1.3. Quality Control

QC procedures were implemented across all datasets. For the OASIS-3 and AIBL datasets, the QC process was performed internally by our research team (BHV and CE). We used the quality assessment tools bundled with FreeSurfer which included several automated QC measures, such as signal-to-noise ratio, topological defects, Euler number, and rotation parameters, to identify potential issues with MRI data. For each scan flagged by these measures, we conducted a thorough visual inspection. Scans with poor contrast, orientation, or significant artifacts were excluded. Consequently, 39 scans were excluded from the OASIS-3 dataset, while no exclusion was necessary for the AIBL dataset. For the UKB, QC was conducted as part of the original data release using a semi-automated pipeline. T1-weighted images were scored using an automated classifier, with manual review performed for cases close to the 'bad data' threshold. Data deemed seriously problematic were excluded. FreeSurfer outputs were also QC checked using the Qoala-T approach (Klapwijk et al., 2019), supplemented by manual review, ensuring only high-quality data were used. For more details, refer to the documentation available at https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf.

3.2. Normative modeling

3.2.1. Estimation of normative models

To capture individual deviations from typical ageing patterns, normative models were fitted independently within each dataset (OASIS-3 and AIBL) for each ROI (Figure 1A). Each

dataset was stratified by site, sex, and age (5-year bins), and split into an training set (80% of HC; n=691 for OASIS-3, n=322 for AIBL) and a test set (20% of HC plus all AD participants; n= 174 HC, n = 167 AD for OASIS-3, and n = 81 HC, n = 60 AD for AIBL) (Table 1).

We employed Warped Bayesian Linear Regression (BLR) with a B-spline basis expansion, using Python 3.9.16 and PCNtoolkit version 0.29 (Fraza et al., 2021). All procedures followed the recommendations for BLR models, as detailed in (Rutherford, Kia, et al., 2022). The covariates (age, sex, and a dummy-coded site variable) were incorporated into the model. Age and sex were modeled using a cubic B-spline basis with three evenly spaced knots, allowing for the creation of a smooth curve that can flexibly represent the changes in brain structure across different ages and sexes. The Sinh-Arcsinh function and the respective warped likelihood were used to handle non-Gaussian distributions in the data (Fraza et al., 2021; Jones & Pewsey, 2009) and optimized the model by minimizing the negative log-likelihood using Powell's conjugate direction method.

From normative models, deviation scores (Z-scores) were derived to quantify how much an individual's measurement deviates from the predicted normative range. These scores were computed as the difference between the observed value and the model's predicted conditional mean, scaled by the estimated variance. Unlike traditional Z-scores, which assume a fixed population variance, normative modeling Z-scores account for two sources of uncertainty commonly defined in machine learning: (1) aleatoric uncertainty (irreducible uncertainty), which captures true underlying variability across individuals that cannot be reduced with more data; and (2) epistemic uncertainty (reducible uncertainty), which reflects parameter uncertainty due to limited data and can be minimized as more data becomes available (Fraza et al., 2021; Marquand et al., 2016). By integrating these components, Z-scores provide a standardized measure of deviation, indicating whether an individual's measurement falls within or outside the normative range and how atypical it is relative to the reference population.

3.3. Subsampling

We first fitted models for each ROI using 80% of the HC participants from OASIS-3 or AIBL for the replication analysis. These models represent the optimal scenario, leveraging the maximum available data to achieve the highest accuracy and reliability. These models trained on the full sample then served as a reference benchmark for evaluating the impact of reduced sample sizes and covariate misalignment. To assess how systematic variations in (1) sample size and (2) covariate distributions influenced model fit performance, we applied a series of subsampling strategies to generate training sets with varying sizes and age or sex distributions.

For each subsampling condition, a set of individuals was selected to form the training set, which was then used to fit separate models for each ROI. This approach allowed us to systematically assess how variations in sample composition affected model fit across brain regions. Across all subsampling strategies, the test set remained identical (described in 2.2.1), ensuring comparability. All primary results are presented for the OASIS-3 dataset, with replication in the AIBL dataset detailed in the supplementary material.

3.3.1. Sample Size Variations

Subsamples were drawn in increments of 5 from size 5 to 200, and in increments of 50 beyond 200 up to 600. Finer steps at smaller sample sizes allowed closer track of changes in model performance at smaller sample sizes, while larger steps were applied beyond 200 to reduce computational demands, after convergence was expected. For each sample size, 10 random iterations were performed to estimate variability and reduce the influence of individual draws. For replication in AIBL, samples were drawn from 5 to 100 per site, due to more limited available data.

3.3.2. Sub-sampling with different covariates distributions

We assessed the influence of age and sex distributions in the training set by using a representative age distribution with balanced sex ratios and then altering either age distribution (using skewed subsampling) or sex balance (using uneven ratios of females and males) while also systematically varying sample size as described above.

To generate subsamples that mimicked the original age distributions of the data, we partitioned the age distribution into 10 bins using a quantile discretizer, which ensured equal numbers of participants in each bin. An equal number of males and females (±1) was then randomly selected across bins. This approach, referred to as *representative sampling*, ensured that the subsampled training sets closely matched the original age distribution of the dataset while maintaining a balanced sex ratio.

To create training sets with skewed (non-representative) age distributions, we applied two strategies that preferentially selected either younger (left-skewed) or older (right-skewed) individuals. More specifically, the age distribution was divided into 10 equally spaced age bins, using a uniform discretizer (which divides the age range into equal-width intervals but does not ensure an equal number of participants per bin as in the representative subsampling). Samples were randomly drawn from these bins, and a beta distribution was applied to weight the

probability of drawing samples from each bin. For the left-skewed sampling, we used parameters α=2 and β=5, increasing the likelihood of selecting younger samples. Conversely, for the right-skewed sampling, parameters α=5 and β=2 were used, concentrating more samples in the older age bins.

To examine the influence of sex imbalance, we adapted the representative age sampling strategy but altered the male-to-female ratio according to predefined levels (1:4; 1:10; 4:1, 10:1). This enables assessment how sex distribution imbalances affected model performance and clinical interpretation of deviations.

3.4. Model evaluation

Subsequently, models trained on subsampled data were evaluated on the same fixed test set of participants, assessing model fit in the HC test set and Z-score errors in both HC and AD test sets (Figure 1C). Performance of the subsampled models was then compared with that of the models trained on the full training set across three subsampling scenarios: 1) representative age distribution with balanced sex ratios, 2) skewed age subsampling with balanced sex ratios, and 3) representative age sampling with unbalanced sex ratios.

3.4.1. Evaluation of normative models fitting

Every derived model was evaluated using state-of-the-art model evaluation metrics including: MSLL, SMSE, EV and Rho (Fraza et al., 2021; Rutherford, Kia, et al., 2022) and *ICC* (Koo & Li, 2016). Evaluation was restricted to HC individuals, as model fit is intended to assess performance relative to the reference population. The AD cohort was instead considered for Z-scores errors and clinical validation. *MSLL*, standardized by the mean loss of the training dataset (Bayer et al., 2022; Rasmussen & Williams, 2008), indicates the predictive performance of a model compared to the mean of the training set, with more negative values being preferable as they indicate better predictive accuracy. *SMSE* assesses the average squared differences between the predicted and actual values, standardized by the variance of the test data. Lower *SMSE* values indicate better predictive performance, as they signify smaller errors. *EV* measures the proportion of variance in the test data that is explained by the model. Higher values of EV indicate that the model accounts for a greater portion of the variability in the data, reflecting better model performance. *Rho* measures the linear correlation between the predicted and actual values. A Pearson correlation close to 1 indicates a strong positive correlation, implying that the model's predictions are closely aligned with the actual values. The *ICC* quantifies the reliability of the models' outputs across iterations. The *ICC* was computed on the Z-scores of HC

test set for each model using a two-way random-effects model for single measurements, with absolute agreement, to assess the consistency of Z-scores across iterations (where iterations refer to repeated random sampling of the same sample size within a subsampling strategy).

3.4.2. Evaluation of Z-scores estimations

Z-score errors were computed for each sampling strategy and sample size by comparing models trained on subsampled data to the ones trained on the full training set. We used *MSE* to summarize overall model accuracy of estimates by averaging the squared deviations across ROIs for each individual. *MSE* captures the overall magnitude of errors, penalizing larger deviations more strongly due to the squaring operation, which makes it well suited for assessing general model accuracy at the individual level. To reduce the influence of rare iterations producing atypical very large errors, outliers were excluded using a conservative interquartile range threshold (3×IQR) to exclude only extreme values while preserving most of the data. Age effects were evaluated using cubic regressions across the age range; sex effects were assessed by comparing *MSEs* between females and males under imbalanced sampling.

We additionally computed *MBE* to assess the direction of Z-score differences, indicating whether estimates tend to be systematically over- or underestimated. Because opposing errors can cancel each other out, *MBE* may underestimate the overall level of error. *MSE* and *MBE* together provide complementary information: *MSE* captures the total error magnitude, while *MBE* reveals directional biases.

3.5. Clinical validation

To evaluate clinical validity, we assessed whether deviation patterns aligned with known AD-related neurobiological evidence and whether they could reliably distinguish HC from AD. Specifically, we examined whether the most deviating regions (i.e., outliers) corresponded to established patterns of AD pathology (i.e., biological plausibility) and whether classification based on deviation scores could distinguish HC from AD. For both analyses, we compared the performance of the model fitted with the full training set with subsampled training sets models to determine whether deviation patterns and classification performance remained consistent across sampling strategies and sample sizes.

3.5.1. Outlier detection and biological plausibility

To assess biological plausibility, we computed the percentage of outliers per ROI in the test set to assess the consistency and group-specific patterns of deviations in HC and AD, across the model trained on the full set and those trained on each subsampled set. A data point was defined as an outlier if its deviation score was $Z < -1.96$. This threshold, representing the bottom 2.5% of the normative range, has been commonly used to report extreme deviations in previous works (Bhome et al., 2024; F et al., 2024; Verdi et al., 2023, 2024). Only negative outliers are considered here, as AD is mainly characterized by diffuse, progressive atrophy of brain regions (Pini et al., 2016; Sabuncu et al., 2011; Ten Kate et al., 2018). Using this framework, a properly fitted model is expected to show outlier distributions that align with known neuroanatomical patterns of AD, with regions commonly affected by the disease exhibiting higher outlier counts.

To further capture the overall trends in outlier detection, we analyzed the mean *tOC* per individual in both HC and AD groups. The *tOC*, defined as the sum of ROI with negative outliers for each participant (Loreto et al., 2024; Verdi et al., 2021, 2023, 2024), was computed within each group (i.e. HC and AD) for every sample size and subsampling iteration.

3.5.2. Classification performance

We evaluated classification performance to determine whether deviations from normative models (i.e., Z-scores) retained their clinical validity and preserved the relative ranking of individuals for each sampling configuration. Support Vector Classifiers (SVCs) were trained using the ROI-level deviation scores from the test set. Classification was performed using 10-fold cross-validation and the *Receiver Operating Characteristic - Area Under the Curve* (*ROC-AUC*) was computed for each fold. For each sampling configuration, *ROC-AUC* values were averaged across folds and iterations to obtain a robust measure of classification performance.

3.6. Statistical analysis

We used linear mixed-effects models (LMMs) to evaluate the effects of sample size and subsampling strategy on model fit metrics (*MSLL*, *EV*, *SMSE*, *Rho*, and *ICC*) and deviation scores (*MSE* and *tOC*) (i.e., outcome). Subsampling strategies included age-skewed (left-skewed, right-skewed) and sex-imbalanced (1:4, 1:10, 4:1, 10:1 female-to-male ratios). The representative sampling strategy, which preserved the original age and sex distribution served as the reference.

For model fit, LMMs included sample size (n), sampling strategy, and their interaction as fixed effects, with a random intercept for ROI to account for variability across regions:

$$\text{Outcome} \sim \log(n) \times \text{Sampling Strategy} + (1 \mid \text{ROI})$$

The relationship between sample size and model performance is characterized by steep improvements at small n that gradually level off with larger samples (Figures 2-3). To capture this non-linear pattern and avoid overemphasizing differences at high n, we applied a logarithmic transformation to sample size. This transformation linearizes the effect, resulting in more appropriate estimates of the influence of n and its interactions with sampling strategies. Both performance metrics and sample size were standardized prior to modeling to facilitate comparability.

For deviation scores, LMMs additionally included group (HC or AD) and its interactions with sample size and sampling strategy as fixed effects, with a random intercept for individuals to account for subject-level variability. As before, sample size was log-transformed, and both outcomes and sample size were standardized prior to modeling:

$$\text{Outcome} \sim \log(n) \times \text{Sampling Strategy} \times \text{Group} + (1 \mid \text{Individual})$$

Again, the representative sampling strategy and HC group served as baseline levels for comparison. Fixed effects and interaction terms captured the influence of sample size, sampling strategy, and group on the outcomes.

To illustrate the effect of sampling strategies and sample size for specific brain regions, additional LMMs were computed independently for each ROI. These results are presented in the supplementary materials and were corrected using the false discovery rate (FDR) to account for multiple comparisons.

3.7. Adaptive transfer learning

To evaluate whether models trained on large external datasets can generalize to clinical cohorts, we additionally compared our models trained within cohorts (i.e. OASIS-3 or AIBL) with models adapted from a large independent cohort using adaptive transfer learning. To motivate the need for model adaptation across cohorts, we first examined differences in mean cortical thickness between UKB and the clinical datasets. We compared mean cortical thickness and mean subcortical grey volumes for HC from the UKB, OASIS-3 and AIBL datasets. Differences in mean cortical thickness across datasets were assessed using boxplots and bivariate density plots (mean cortical thickness vs. age). Separate linear regression models were fitted for each dataset, with mean cortical thickness as the dependent variable and age as the independent variable.

Adaptive transfer learning was applied to adjust the UKB-based normative models to the clinical cohorts (Figure 1B). Models were pre-fitted using 80% of UKB HC, with 20% held out for internal evaluation. Adaptation sets and within-cohort training sets were sampled from the same training cohort and tested on the same evaluation set, ensuring direct comparability between models. This approach procedure assumes that biological effects, such as age-related trajectories, are shared across datasets, and only cohort-specific shifts require adjustment. The adaptation assumes that normative patterns learned from a large reference cohort remain valid across datasets, and that residual differences, such as scanner or cohort biases, can be captured through distributional adjustments at the mean level.

To evaluate the two modeling strategies, either directly trained on the target dataset (i.e. OASIS-3 or AIBL) or adapted from a pre-trained UKB model, we compared model fit using the same metrics as previously (*MSLL*, *SMSE*, *EV*, *Rho*, and *ICC*), assessed Z-score errors relative to the full adaptation model, and illustrated the percentage of outliers per brain region across sample sizes. To facilitate direct comparison, we also computed the difference in performance between the two approaches (direct training minus adapted) for each metric. Additionally, we replicated all analyses conducted in the direct training setting, including covariate skewing, and present the adapted versions in the Supplementary Materials.

Acknowledgments

This research was supported by the Dementia Research Switzerland - Synapsis Foundation [No. 2021-PI02] and the Swiss National Science Foundation (SNSF) under the following grants [10001C 197480 & IC00I0 227750].

Data were provided in part by OASIS-3: Longitudinal Multimodal Neuroimaging: Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P30 AG066444, P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

References

Ayton, S., Fazlollahi, A., Bourgeat, P., Raniga, P., Ng, A., Lim, Y. Y., Diouf, I., Farquharson, S., Fripp, J., Ames, D., Doecke, J., Desmond, P., Ordidge, R., Masters, C. L., Rowe, C. C., Maruff, P., Villemagne, V. L., the Australian Imaging Biomarkers and Lifestyle (AIBL) Research Group, Salvado, O., & Bush, A. I. (2017). Cerebral quantitative susceptibility mapping predicts amyloid-β-related cognitive decline. *Brain*, *140*(8), 2112–2119. https://doi.org/10.1093/brain/awx137

Barnes, J., Bartlett, J. W., van de Pol, L. A., Loy, C. T., Scahill, R. I., Frost, C., Thompson, P., & Fox, N. C. (2009). A meta-analysis of hippocampal atrophy rates in Alzheimer's disease. *Neurobiology of Aging*, *30*(11), 1711–1723. https://doi.org/10.1016/j.neurobiolaging.2008.01.010

Bayer, J. M. M., Dinga, R., Kia, S. M., Kottaram, A. R., Wolfers, T., Lv, J., Zalesky, A., Schmaal, L., & Marquand, A. (2022). Accommodating site variation in neuroimaging data using normative and hierarchical Bayesian models. *NeuroImage*, *264*, 119699. https://doi.org/10.1016/j.neuroimage.2022.119699

Bethlehem, R. a. I., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A., Astle, D. E., Auyeung, B., Ayub, M., Bae, J., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S. A., Benegal, V., … Alexander-Bloch, A. F. (2022). Brain charts for the human lifespan. *Nature*, *604*(7906), 525–533. https://doi.org/10.1038/s41586-022-04554-y

Bhome, R., Verdi, S., Martin, S. A., Hannaway, N., Dobreva, I., Oxtoby, N. P., Castro Leal, G., Rutherford, S., Marquand, A. F., Weil, R. S., & Cole, J. H. (2024). A neuroimaging measure to capture heterogeneous patterns of atrophy in Parkinson's disease and dementia with Lewy bodies. *NeuroImage. Clinical*, *42*, 103596. https://doi.org/10.1016/j.nicl.2024.103596

Bozek, J., Griffanti, L., Lau, S., & Jenkinson, M. (2023). Normative models for neuroimaging markers: Impact of model selection, sample size and evaluation criteria. *NeuroImage*, *268*, 119864. https://doi.org/10.1016/j.neuroimage.2023.119864

Corrigan, N. M., Rokem, A., & Kuhl, P. K. (2024). COVID-19 lockdown effects on adolescent brain structure suggest accelerated maturation that is more pronounced in females than in males. *Proceedings of the National Academy of Sciences*, *121*(38), e2403200121. https://doi.org/10.1073/pnas.2403200121

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15. https://doi.org/10.1016/j.neuroimage.2010.06.010

Dima, D., Modabbernia, A., Papachristou, E., Doucet, G. E., Agartz, I., Aghajani, M., Akudjedu, T. N., Albajes-Eizagirre, A., Alnæs, D., Alpert, K. I., Andersson, M., Andreasen, N. C., Andreassen, O. A., Asherson, P., Banaschewski, T., Bargallo, N., Baumeister, S., Baur-Streubel, R., Bertolino, A., … Frangou, S. (2021). Subcortical volumes across the lifespan: Data from 18,605 healthy individuals aged 3–90 years. *Human Brain Mapping*, *43*(1), 452–469. https://doi.org/10.1002/hbm.25320

Duara, R., & Barker, W. (2022). Heterogeneity in Alzheimer's Disease Diagnosis and Progression Rates: Implications for Therapeutic Trials. *Neurotherapeutics*, *19*(1), 8–25. https://doi.org/10.1007/s13311-022-01185-z

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

F, L., S, V., Sm, K., A, D., H, H., A, F., N, P., J, L., Z, W., R, P., Af, M., Jh, C., & P, M. (2024). Alzheimer's disease heterogeneity revealed by neuroanatomical normative modeling. *Alzheimer's & Dementia (Amsterdam, Netherlands)*, *16*(1). https://doi.org/10.1002/dad2.12559

Ferretti, M. T., Iulita, M. F., Cavedo, E., Chiesa, P. A., Schumacher Dimech, A., Santuccione Chadha, A., Baracchi, F., Girouard, H., Misoch, S., Giacobini, E., Depypere, H., & Hampel, H. (2018). Sex differences in Alzheimer disease—The gateway to precision medicine. *Nature Reviews Neurology*, *14*(8), 457–469. https://doi.org/10.1038/s41582-018-0032-9

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

Fjell, A. M., & Walhovd, K. B. (2010). Structural brain changes in aging: Courses, causes and cognitive consequences. *Reviews in the Neurosciences*, *21*(3), 187–221. https://doi.org/10.1515/revneuro.2010.21.3.187

Floris, D. L., Wolfers, T., Zabihi, M., Holz, N. E., Zwiers, M. P., Charman, T., Tillmann, J., Ecker, C., Dell'Acqua, F., Banaschewski, T., Moessnang, C., Baron-Cohen, S., Holt, R., Durston, S., Loth, E., Murphy, D. G. M., Marquand, A., Buitelaar, J. K., Beckmann, C. F., … Zwiers, M. P. (2021). Atypical Brain Asymmetry in Autism—A Candidate for Clinically Meaningful Stratification. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *6*(8), 802–812. https://doi.org/10.1016/j.bpsc.2020.08.008

Fowler, C., Rainey-Smith, S. R., Bird, S., Bomke, J., Bourgeat, P., Brown, B. M., Burnham, S. C., Bush, A. I., Chadunow, C., Collins, S., Doecke, J., Doré, V., Ellis, K. A., Evered, L., Fazlollahi, A., Fripp, J., Gardener, S. L., Gibson, S., Grenfell, R., … Ames, D. (2021). Fifteen Years of the Australian Imaging, Biomarkers and Lifestyle (AIBL) Study: Progress and Observations from 2,359 Older Adults Spanning the Spectrum from Cognitive Normality to Alzheimer's Disease. *Journal of Alzheimer's Disease Reports*, *5*(1), 443–468. https://doi.org/10.3233/ADR-210005

Fraza, C. J., Dinga, R., Beckmann, C. F., & Marquand, A. F. (2021). Warped Bayesian linear

regression for normative modelling of big data. *NeuroImage*, *245*, 118715. https://doi.org/10.1016/j.neuroimage.2021.118715

Gaiser, C., Berthet, P., Kia, S. M., Frens, M. A., Beckmann, C. F., Muetzel, R. L., & Marquand, A. (2023). *Estimating cortical thickness trajectories in children across different scanners using transfer learning from normative models* (p. 2023.03.02.530742). bioRxiv. https://doi.org/10.1101/2023.03.02.530742

Goodkin, O., Pemberton, H., Vos, S. B., Prados, F., Sudre, C. H., Moggridge, J., Cardoso, M. J., Ourselin, S., Bisdas, S., White, M., Yousry, T., Thornton, J., & Barkhof, F. (2019). The quantitative neuroradiology initiative framework: Application to dementia. *British Journal of Radiology*, *92*(1101). https://doi.org/10.1259/bjr.20190365

Harnett, N. G., Fani, N., Rowland, G., Kumar, P., Rutherford, S., & Nickerson, L. D. (2024). Population-level normative models reveal race- and socioeconomic-related variability in cortical thickness of threat neurocircuitry. *Communications Biology*, *7*(1), 1–11. https://doi.org/10.1038/s42003-024-06436-7

Holz, N. E., Zabihi, M., Kia, S. M., Monninger, M., Aggensteiner, P.-M., Siehl, S., Floris, D. L., Bokde, A. L. W., Desrivières, S., Flor, H., Grigis, A., Garavan, H., Gowland, P., Heinz, A., Brühl, R., Martinot, J.-L., Martinot, M.-L. P., Orfanos, D. P., Paus, T., ... Marquand, A. F. (2023). A stable and replicable neural signature of lifespan adversity in the adult brain. *Nature Neuroscience*, *26*(9), 1603–1612. https://doi.org/10.1038/s41593-023-01410-8

Igarashi, K. M. (2023). Entorhinal cortex dysfunction in Alzheimer's disease. *Trends in Neurosciences*, *46*(2), 124–136. https://doi.org/10.1016/j.tins.2022.11.006

Jones, M. C., & Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, *96*(4), 761–780. https://doi.org/10.1093/biomet/asp053

Kia, S. M., Huijsdens, H., Rutherford, S., Dinga, R., Wolfers, T., Mennes, M., Andreassen, O. A., Westlye, L. T., Beckmann, C. F., & Marquand, A. F. (2021). *Federated Multi-Site Normative Modeling using Hierarchical Bayesian Regression* (p. 2021.05.28.446120). bioRxiv. https://doi.org/10.1101/2021.05.28.446120

Klapwijk, E. T., van de Kamp, F., van der Meulen, M., Peters, S., & Wierenga, L. M. (2019). Qoala-T: A supervised-learning tool for quality control of FreeSurfer segmented MRI data. *NeuroImage*, *189*, 116–129. https://doi.org/10.1016/j.neuroimage.2019.01.014

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Küchenhoff, S., Bayrak, Ş., Zsido, R. G., Saberi, A., Bernhardt, B. C., Weis, S., Schaare, H. L., Sacher, J., Eickhoff, S., & Valk, S. L. (2024). Relating sex-bias in human cortical and hippocampal microstructure to sex hormones. *Nature Communications*, *15*(1), 7279. https://doi.org/10.1038/s41467-024-51459-7

Lam, B., Masellis, M., Freedman, M., Stuss, D. T., & Black, S. E. (2013). Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome. *Alzheimer's Research & Therapy*, *5*(1), 1. https://doi.org/10.1186/alzrt155

LaMontagne, P. J., Benzinger, T. L., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E.,

Hassenstab, J., Moulder, K., Vlassenko, A. G., Raichle, M. E., Cruchaga, C., & Marcus, D. (2019). *OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease* (p. 2019.12.13.19014902). medRxiv. https://doi.org/10.1101/2019.12.13.19014902

Lemaitre, H., Goldman, A. L., Sambataro, F., Verchinski, B. A., Meyer-Lindenberg, A., Weinberger, D. R., & Mattay, V. S. (2012). Normal age-related brain morphometric changes: Nonuniformity across cortical thickness, surface area and gray matter volume? *Neurobiology of Aging*, *33*(3), 617.e1-9. https://doi.org/10.1016/j.neurobiolaging.2010.07.013

LeWinn, K. Z., Sheridan, M. A., Keyes, K. M., Hamilton, A., & McLaughlin, K. A. (2017). Sample composition alters associations between age and brain structure. *Nature Communications*, *8*(1), 874. https://doi.org/10.1038/s41467-017-00908-7

Loreto, F., Verdi, S., Kia, S. M., Duvnjak, A., Hakeem, H., Fitzgerald, A., Patel, N., Lilja, J., Win, Z., Perry, R., Marquand, A. F., Cole, J. H., & Malhotra, P. (2024). Alzheimer's disease heterogeneity revealed by neuroanatomical normative modeling. *Alzheimer's & Dementia (Amsterdam, Netherlands)*, *16*(1), e12559. https://doi.org/10.1002/dad2.12559

Marquand, A. F., Rezek, I., Buitelaar, J., & Beckmann, C. F. (2016). Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biological Psychiatry*, *80*(7), 552–561. https://doi.org/10.1016/j.biopsych.2015.12.023

Meijer, J., Hebling Vieira, B., Elleaume, C., Baranczuk-Turska, Z., Langer, N., & Floris, D. L. (2024). Toward understanding autism heterogeneity: Identifying clinical subgroups and neuroanatomical deviations. *Journal of Psychopathology and Clinical Science*, *133*(8), 667–677. https://doi.org/10.1037/abn0000914

Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., Galluzzi, S., Marizzoni, M., & Frisoni, G. B. (2016). Brain atrophy in Alzheimer's Disease and aging. *Ageing Research Reviews*, *30*, 25–48. https://doi.org/10.1016/j.arr.2016.01.002

Planche, V., Manjon, J. V., Mansencal, B., Lanuza, E., Tourdias, T., Catheline, G., & Coupé, P. (2022). Structural progression of Alzheimer's disease over decades: The MRI staging scheme. *Brain Communications*, *4*(3), fcac109. https://doi.org/10.1093/braincomms/fcac109

Qu, H., Ge, H., Wang, L., Wang, W., & Hu, C. (2023). Volume changes of hippocampal and amygdala subfields in patients with mild cognitive impairment and Alzheimer's disease. *Acta Neurologica Belgica*, *123*(4), 1381–1393. https://doi.org/10.1007/s13760-023-02235-9

Rajan, K. B., Weuve, J., Barnes, L. L., McAninch, E. A., Wilson, R. S., & Evans, D. A. (2021). Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020-2060). *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *17*(12), 1966–1975. https://doi.org/10.1002/alz.12362

Rasmussen, C. E., & Williams, C. K. I. (2008). *Gaussian processes for machine learning* (3. print). MIT Press.

Riedel, B. C., Thompson, P. M., & Brinton, R. D. (2016). Age, APOE and sex: Triad of risk of Alzheimer's disease. *The Journal of Steroid Biochemistry and Molecular Biology*, *160*, 134–147. https://doi.org/10.1016/j.jsbmb.2016.03.012

Ruigrok, A. N. V., Salimi-Khorshidi, G., Lai, M.-C., Baron-Cohen, S., Lombardo, M. V., Tait, R. J., &

Suckling, J. (2014). A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews*, *39*, 34–50. https://doi.org/10.1016/j.neubiorev.2013.12.004

Rutherford, S., Barkema, P., Tso, I. F., Sripada, C., Beckmann, C. F., Ruhe, H. G., & Marquand, A. F. (2023). Evidence for embracing normative modeling. *eLife*, *12*, e85082. https://doi.org/10.7554/eLife.85082

Rutherford, S., Fraza, C., Dinga, R., Kia, S. M., Wolfers, T., Zabihi, M., Berthet, P., Worker, A., Verdi, S., Andrews, D., Han, L. K., Bayer, J. M., Dazzan, P., McGuire, P., Mocking, R. T., Schene, A., Sripada, C., Tso, I. F., Duval, E. R., … Marquand, A. F. (2022). Charting brain growth and aging at high spatial precision. *eLife*, *11*, e72904. https://doi.org/10.7554/eLife.72904

Rutherford, S., Kia, S. M., Wolfers, T., Fraza, C., Zabihi, M., Dinga, R., Berthet, P., Worker, A., Verdi, S., Ruhe, H. G., Beckmann, C. F., & Marquand, A. F. (2022). The normative modeling framework for computational psychiatry. *Nature Protocols*, *17*(7), 1711–1734. https://doi.org/10.1038/s41596-022-00696-5

Sabuncu, M. R., Desikan, R. S., Sepulcre, J., Yeo, B. T. T., Liu, H., Schmansky, N. J., Reuter, M., Weiner, M. W., Buckner, R. L., Sperling, R. A., Fischl, B., & Alzheimer's Disease Neuroimaging Initiative. (2011). The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Archives of Neurology*, *68*(8), 1040–1048. https://doi.org/10.1001/archneurol.2011.167

Savage, H. S., Mulders, P. C. R., van Eijndhoven, P. F. P., van Oort, J., Tendolkar, I., Vrijsen, J. N., Beckmann, C. F., & Marquand, A. F. (2024). Dissecting task-based fMRI activity using normative modelling: An application to the Emotional Face Matching Task. *Communications Biology*, *7*(1), 1–14. https://doi.org/10.1038/s42003-024-06573-z

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Ten Kate, M., Dicks, E., Visser, P. J., van der Flier, W. M., Teunissen, C. E., Barkhof, F., Scheltens, P., Tijms, B. M., & Alzheimer's Disease Neuroimaging Initiative. (2018). Atrophy subtypes in prodromal Alzheimer's disease are associated with cognitive decline. *Brain: A Journal of Neurology*, *141*(12), 3443–3456. https://doi.org/10.1093/brain/awy264

Verdi, S., Kia, S. M., Yong, K. X. X., Tosun, D., Schott, J. M., Marquand, A. F., & Cole, J. H. (2023). Revealing Individual Neuroanatomical Heterogeneity in Alzheimer Disease Using Neuroanatomical Normative Modeling. *Neurology*, *100*(24), e2442–e2453. https://doi.org/10.1212/WNL.0000000000207298

Verdi, S., Marquand, A. F., Schott, J. M., & Cole, J. H. (2021). Beyond the average patient: How neuroimaging models can address heterogeneity in dementia. *Brain: A Journal of Neurology*, *144*(10), 2946–2953. https://doi.org/10.1093/brain/awab165

Verdi, S., Rutherford, S., Fraza, C., Tosun, D., Altmann, A., Raket, L. L., Schott, J. M., Marquand, A. F., Cole, J. H., & Alzheimer's Disease Neuroimaging Initiative. (2024). Personalizing progressive changes to brain structure in Alzheimer's disease using normative modeling. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *20*(10), 6998–7012. https://doi.org/10.1002/alz.14174

Vernooij, M. W., Pizzini, F. B., Schmidt, R., Smits, M., Yousry, T. A., Bargallo, N., Frisoni, G. B., Haller, S., & Barkhof, F. (2019). Dementia imaging in clinical practice: A European-wide survey of 193 centres and conclusions by the ESNR working group. *Neuroradiology*, *61*(6), 633–642. https://doi.org/10.1007/s00234-019-02188-y

Yan, T., Wang, Y., Weng, Z., Du, W., Liu, T., Chen, D., Li, X., Wu, J., & Han, Y. (2019). Early-Stage Identification and Pathological Development of Alzheimer's Disease Using Multimodal MRI. *Journal of Alzheimer's Disease*, *68*(3), 1013–1027. https://doi.org/10.3233/JAD-181049

Zabihi, M., Floris, D. L., Kia, S. M., Wolfers, T., Tillmann, J., Arenas, A. L., Moessnang, C., Banaschewski, T., Holt, R., Baron-Cohen, S., Loth, E., Charman, T., Bourgeron, T., Murphy, D., Ecker, C., Buitelaar, J. K., Beckmann, C. F., & Marquand, A. (2020). Fractionating autism based on neuroanatomical normative modeling. *Translational Psychiatry*, *10*(1), Article 1. https://doi.org/10.1038/s41398-020-01057-0

Zhao, W., Luo, Y., Zhao, L., Mok, V., Su, L., Yin, C., Sun, Y., Lu, J., Shi, L., & Han, Y. (2019). Automated Brain MRI Volumetry Differentiates Early Stages of Alzheimer's Disease From Normal Aging. *Journal of Geriatric Psychiatry and Neurology*, *32*(6), 354–364. https://doi.org/10.1177/0891988719862637