Calibration of MRI-based reference intervals to new samples

Andrew A. Chen[a,*], Jakob Seidlitz[b,c,d,e], Margaret Gardner[b,f], Richard A. I. Bethlehem[g], Lena Dorfschmidt[b,c,d], Eren Kafadar[b,f], Lifespan Brain Chart Consortium, Andreana Benitez[h], Jens H. Jensen[i], Simon Vandekar[j], Theodore D. Satterthwaite[c,k,l], and Aaron F. Alexander-Bloch[b,c,d,e]

[a] *Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, USA*
[b] *Brain-Gene-Development Lab, The Children's Hospital of Philadelphia and Penn Medicine, Philadelphia, PA, USA*
[c] *Lifespan Brain Institute, The Children's Hospital of Philadelphia and Penn Medicine, Philadelphia, PA, USA*
[d] *Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA*
[e] *Department of Child and Adolescent Psychiatry and Behavioral Science, The Children's Hospital of Philadelphia, Philadelphia, PA, USA*
[f] *Neuroscience Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[g] *Department of Psychology, University of Cambridge, Cambridge, UK*
[h] *Department of Neurology, Medical University of South Carolina, Charleston, SC, USA*
[i] *Department of Neuroscience, Medical University of South Carolina, Charleston, SC, USA*
[j] *Department of Biostatistics, Vanderbilt University, Nashville, TN, USA*
[k] *Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[l] *The Penn Lifespan Informatics and Neuroimaging Center, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*

***Correspondence: Andrew A. Chen**
chenandr@musc.edu

Abstract

Reference intervals, defined as intervals containing a new observation with a specified probability relative to reference data, would be clinically useful in assessing brain magnetic resonance imaging (MRI). Brain charts, which are estimates of MRI phenotypes across covariates such as age and sex, can be used to construct reference intervals. However, the reference data used to fit intervals often differs from a new sample in terms of study design, MRI acquisition, and image preprocessing. Application of MRI reference intervals to new samples remains a challenging problem. Here, we propose a new method called Reference interval calibration via conFormal prediction (ReForm) that adjusts reference intervals for a new sample. Our method builds on recent work in conformal prediction, which yields intervals with guaranteed coverage for new observations. Through resampling experiments in Lifespan Brain Chart Consortium cortical thickness data, we compare ReFormed reference intervals to refitting intervals, statistical harmonization methods, and model-based adjustment of intervals. Notably for patient privacy concerns, ReForm does not require sharing of reference data. Yet, our empirical results demonstrate that ReForm controls FPR similarly or better than alternative methods which require sharing reference data. Finally, we provide recommendations for practical applications of ReForm and an R package (`https://github.com/andy1764/ReForm`) for calibrating reference intervals using ReForm.

1 Introduction

Imaging consortia increasingly aggregate massive databases of magnetic resonance imaging (MRI) scans for precise mapping of the brain across the lifespan. Their goal is to estimate *brain charts* for a given MRI phenotype, which are models that capture nonlinear trajectories across neurodevelopment and aging while accounting for known sex differences and other covariates (Bethlehem et al., 2022; Dima et al., 2022; Frangou et al., 2022; Rutherford et al., 2022). Numerous consortia have recently published brain charts across additional types of MRI scans or open-access resources for other researchers to fit their own brain charts (Kim et al., 2025; Shafiei et al., 2025; Zhu et al., 2025). The promise is that brain charts can enable scientists and clinicians to compare new scans to a well-characterized healthy reference population.

However, the promise of brain charts is limited due to differences in MRI acquisition and preprocessing. Differences across MRI acquisition have been recognized to introduce bias and reduce power (Han et al., 2006) and several recent studies highlight this issue across multiple MRI modalities (Cetin Karayumak et al., 2019; Fortin et al., 2017, 2018; Shinohara et al., 2017; Yu et al., 2018). Simultaneously, several groups have acknowledged that different preprocessing steps can lead to substantially different results across a wide range of MRI modalities and derived phenotypes (Antonopoulos et al., 2023; Bhagwat et al., 2021; Botvinik-Nezer et al., 2020; Haddad et al., 2019).

Recent work in brain charts has led to several statistical approaches to dealing with technical biases resulting from heterogeneity in MRI acquisition and preprocessing. One category is *harmonization* methods, which aim to transform data and eliminate systematic biases attributable to acquisition and preprocessing differences (Fortin et al., 2017, 2018; Gardner et al., 2025; Hu et al., 2023; Johnson et al., 2007; Ling et al., 2022). Another category is *model adjustment* which aims to fit model-specific adjustment parameters in new data to account for differences across study sites. Model adjustments are tailored to a specific choice of model used to fit the brain chart, typically a generalized additive model for location, scale, and shape (GAMLSS; Rigby and Stasinopoulos, 2005). Both categories of methods have found widespread use across a number of recent brain chart studies (Bethlehem et al., 2022; Gardner et al., 2025; Ge et al., 2024; Kim et al., 2025; Shafiei et al., 2025; Zhu et al., 2025).

In clinical chemistry, healthy populations are typically used to construct reference intervals that describe the range of normal values for a given phenotype (Gräsbeck and Fellman, 1968; Ozarda, 2016). The definition of reference intervals requires that an observation from the reference data falls inside the interval with a specified probability. Practically, the Clinical and Laboratory Standards Institute (CLSI) EP28-A3c guidelines adopted in 2008 recommend verifying reference intervals by taking a new sample with a specified tolerance for the proportion of samples falling outside the interval (Clinical and Laboratory Standards Institute, 2008), which we refer to as the false positive rate (FPR). Inspired by these guidelines, we propose a stricter definition for a valid reference interval, requiring a new observation to fall above and below the interval with a specified probability.

Brain charts, which are an estimate of the distribution of MRI phenotypes conditional on age and sex, can be used to construct reference intervals. Several studies publish their brain charts or reference intervals for other researchers to utilize (Bethlehem et al., 2022; Kim et al., 2025; Zhu et al., 2025). However, MRI reference intervals currently offer no guarantees of FPR control for a new sample. Neither harmonization nor model adjustment methods have been tested for constructing reference intervals that control FPR.

Here, we draw from recent work in conformal prediction (Lei et al., 2018; Romano et al., 2019b) to propose a method for adjusting reference intervals called Reference interval calibration via conFormal prediction (ReForm). ReForm differs from harmonization and model adjustment methods in that it guarantees control of the FPR in a new observation and thus produces valid reference intervals. Through resampling studies in the Lifespan Brain Chart Consortium (LBCC) cortical thickness data, we demonstrate the performance of ReForm compared to several alternative methods. We find that ReForm controls FPR across a number of different experiments while requiring reasonable sample sizes to achieve adequate FPR empirically. Compared to harmonization methods, we find that ReForm substantially improves FPR control while eliminating dependency on harmonization model assumptions and reference set sample size. Versus model adjustment, we find that ReForm performs similarly even without access to reference data, so long as an adequately sized calibration set is available. In Alzheimer's disease (AD) patients, we demonstrate that our ReFormed reference intervals are sensitive to deviations in entorhinal thickness values. Finally, we offer practical recommendations for using ReForm as a way to calibrate reference intervals from any MRI phenotype for new samples.

2 Methods

2.1 Problem setting

Denote our full data (X_i, Y_i) where X_i are covariates, Y_i are outcomes, and $i = 1, 2, \ldots, n$. Our full data consists of a reference set N_R and *calibration set* N_C with no overlap between the sets, such that $N_R \cup N_C = \{1, 2, \ldots, n\}$ and $N_R \cap N_C = \varnothing$. In neuroimaging, the reference and calibration sets often come from separate datasets with differences in study protocols, scanners, and acquisition parameters. Thus, the distribution of (X_i, Y_i) may differ between the reference and calibration sets.

Our goal is to construct *reference intervals* denoted $C(X) = [L(X), U(X)]$ that are *valid* at a chosen level α. That is, for a new sample (X_{n+1}, Y_{n+1}) drawn from the same dataset as the calibration set, a valid $100 \times (1-\alpha)\%$ reference interval $C(X)$ has the property that a new observation falls above the reference interval with probability $\alpha/2$ and below with probability $\alpha/2$, or equivalently

$$P\left(Y_{n+1} \leq L(X_{n+1}) \middle| (X_i, Y_i) : i \in N_R\right) \leq \alpha/2 \tag{1}$$

and

$$P\left(Y_{n+1} \geq U(X_{n+1}) \middle| (X_i, Y_i) : i \in N_R\right) \leq \alpha/2. \tag{2}$$

Note that this definition of validity is stricter than many standard definitions, which only consider observations within the reference dataset. Additionally, our definition is more general than common definitions of reference intervals, which often refer only to the 95% reference interval (Ozarda, 2016). Validity may not be feasible in the general setting; however, we will show that validity is achievable when the probability is taken over the calibration set and the new observation.

2.2 Reference interval calibration via conformal prediction (ReForm)

Here, we propose a new method that can adjust a fitted reference interval using the calibration set and produce a valid reference interval. We will begin by reviewing conformalized quantile regression, which is a type of conformal prediction method (Lei et al., 2018; Romano et al., 2019b). Suppose our full data are exchangeable and drawn from the same joint distribution such that $(X_i, Y_i) \sim P$, $i = 1, 2, \ldots, n$. This assumption is reasonable when the reference and calibration sets are taken from the same dataset. In this setting, conformalized quantile regression (CQR) yields a valid $100 \times (1-\alpha)\%$ reference interval. CQR first fits a quantile regression model in the reference set to construct a *naïve* reference interval $\tilde{C}(X) = [\hat{q}_{\alpha/2}(X), \hat{q}_{1-\alpha/2}(X)]$ where $\hat{q}_\tau(X)$ is the estimated conditional τ-th quantile function of Y given X. Here,

we focus on the version of CQR which controls the upper and lower coverage of the interval. Define the *lower residuals* as $E_i = \hat{q}_{\alpha/2}(Xi) - Y_i$ and *upper residuals* $R_i = Y_i - \hat{q}_{1-\alpha/2}(X_i)$ where $i \in N_C$. Denote the size of the calibration set as n_C where $n_C = |N_C|$. We estimate calibration constants $\hat{c}_{\alpha/2}$ as the $(1-\alpha/2)(1+1/n_C)$-th empirical quantile of the E_i and $\hat{c}_{1-\alpha/2}$ and $(1-\alpha/2)(1+1/n_C)$-th empirical quantile of the R_i, $i \in N_C$. These empirical quantiles are computed as the $\lceil (1-\alpha/2)(n_C+1) \rceil$ smallest of E_i and R_i respectively, $i \in N_C$. Under exchangeability of the full data and new observation and continuous-valued residuals, Romano et al. (2019b) prove that the reference interval $\hat{C}(X) = [\hat{q}_{\alpha/2}(X) - \hat{c}_{\alpha/2}, \hat{q}_{1-\alpha/2}(X) + \hat{c}_{1-\alpha/2}]$ has the properties

$$P\left(Y_{n+1} \leq \hat{q}_{\alpha/2}(X) - \hat{c}_{\alpha/2} \middle| (X_i, Y_i) : i \in N_R\right) \leq \alpha/2, \tag{3}$$

and

$$P\left(Y_{n+1} \geq \hat{q}_{1-\alpha/2}(X) + \hat{c}_{1-\alpha/2} \middle| (X_i, Y_i) : i \in N_R\right) \leq \alpha/2, \tag{4}$$

where the probability is taken over the calibration set and the new observation. That is, the reference interval $\hat{C}(X)$ satisfies Properties 1 and 2 and is thus a valid reference interval.

In our setting however, the reference set and calibration set may have different distributions. In neuroimaging, this situation occurs when a reference interval is fit in reference data and applied to a new dataset, which differs in study design, MRI acquisition, or image preprocessing. Now, we assume that the reference and calibration sets differ in distribution such that $(X_i, Y_i) \sim P$, $i \in N_R$ and $(X_i, Y_i) \sim Q$, $i \in N_C$. This setting violates the original assumptions of both conformal prediction and CQR since the full data is no longer exchangeable (Lei et al., 2018; Romano et al., 2019b). However, the proof of Equations 3 and 4 only depends on exchangability of the calibration set and new observation.

Here, we assume the calibration set (X_i, Y_i), $i \in N_C$ and new observation (X_{n+1}, Y_{n+1}) are exchangeable and distributed as Q. We first use the reference set $(X_i, Y_i), i \in N_R$ to estimate a naïve reference interval $\tilde{C}(X) = [\hat{q}_{\alpha/2}(X), \hat{q}_{1-\alpha/2}(X)]$. Next, we aim to *calibrate* $\tilde{C}(X)$ to obtain a valid reference interval for the new observation (X_{n+1}, Y_{n+1}). Using the estimation procedure from CQR, we estimate calibration constants $\hat{c}_{\alpha/2}$ as the $(1-\alpha)(1+1/n_C)$-th empirical quantile of the E_i and $\hat{c}_{1-\alpha/2}$ and $(1-\alpha)(1+1/n_C)$-th empirical quantile of the R_i, $i \in N_C$. We call this method Reference interval calibration via conFormal prediction (ReForm) and our derived interval a *ReFormed* reference interval given by

$$\hat{C}(X) = [\hat{q}_{\alpha/2}(X) - \hat{c}_{\alpha/2}, \hat{q}_{1-\alpha/2}(X) + \hat{c}_{1-\alpha/2}]. \tag{5}$$

Using the same definition of lower residuals E_i and upper residuals R_i from CQR, we know that E_i and R_i are both continuous-valued and exchangeable among $i \in N_C$ and $i = n+1$. Thus, from Romano et al. (2019b), our ReFormed reference interval satisfies Equations 3 and 4 where the probability is taken over the calibration set and the new observation. Therefore, our Reformed reference interval is a valid reference interval satisfying Properties 1 and 2. Note that first, we make no assumptions on the reference set distribution P or the calibration set distribution Q. Second, if the naïve reference interval $\tilde{C}(X)$ is available, the reference set is not required to obtain the ReFormed reference interval.

We implement our method in the *ReForm* R package available via GitHub (`https://github.com/andy1764/ReForm`), which includes functions tailored to widely-used regression models.

2.3 Alternative methods

Several alternative methods have been proposed in the literature for either adjusting data or adjusting models in order to mitigate differences between datasets. Here, we compare ReForm to methods from three categories: refitting, harmonization, and model adjustment.

Refitting is the simplest solution where, instead of using the reference interval, we fit new intervals entirely in the calibration set. Refitting is recommended by the CLSI when a reference interval fails to be verified in a new sample (Clinical and Laboratory Standards Institute, 2008). Here, this corresponds to a reference interval $\hat{C}'(X) = [\hat{q}'_{\alpha/2}(X), \hat{q}'_{1-\alpha/2}(X)]$ where $\hat{q}'_{\alpha/2}(X)$ and $\hat{q}'_{1-\alpha/2}(X)$ are estimated conditional quantiles fit in the calibration set, $i \in N_C$. Refitting is not guaranteed to yield a valid reference interval and may provide inaccurate fits particularly in small calibration sets.

Harmonization methods aim to transform the new observation to have the same conditional distribution as the reference data. In our setting where the reference set $(X_i, Y_i) \sim P = P_X P_{Y|X}$, $i \in N_R$, this corresponds

to estimating a function h such that $h(Y_{n+1}, X_{n+1}) \sim P_{Y|X}$. For harmonization, we can obtain an estimate $\hat{h}$ from the full data $i = 1, 2, \ldots, n$ and apply to the $(n+1)$-th observation. The form of $\hat{h}$ depends on the chosen harmonization method, which can employ: linear models (Johnson et al., 2007); generalized additive models (Pomponio et al., 2020); generalized additive models for location, scale, and shape (GAMLSS; Gardner et al., 2025); deep learning (Hu et al., 2024); and other models (Beer et al., 2020; Hu et al., 2023; Ling et al., 2022). Recent work suggests that fitted harmonization parameters from ComBat and several of its extensions can be applied to new sites without using the reference set (Xin et al., 2025).

Model adjustment methods fit a reference interval in the full data and apply that interval directly to the new observation. The specific model used to fit the interval varies, but typically includes a term for *batch*. In neuroimaging, batch is typically used to indicate the study site or MRI scanner for each observation. Assuming that the new observation is from a batch in the full data, we can directly estimate $\tilde{C}(X)$ as described previously and apply to the new observation (X_{n+1}, Y_{n+1}), where here $X_i, i = 1, 2, \ldots, n+1$ includes indicators for batch. In a recent study, Bethlehem et al. (2022) develop a new method based on GAMLSS with a random effect for batch to fit the random effect in a new batch, without access to the reference set. However, model adjustment methods do not guarantee validity of reference intervals.

Figure 1 illustrates unadjusted reference intervals (Unadjusted, 1a), harmonization via ComBatLS (1b), refitting in the calibration set (Refit, 1c), and ReForm (1d), all using empirical data from our study.

2.4 Considerations for ReForm

ReForm only requires exchangeability of the calibration set and new observation to yield valid reference intervals. However, ReForm does not guarantee that the ReFormed reference interval accurately reflects the underlying distribution of the new observation. That is, for a new observation $(X_{n+1}, Y_{n+1}) \sim Q$ where $Q = Q_X Q_{Y|X}$, the ReFormed interval $\hat{C}(X)$ from Equation 5 is not guaranteed to be close to the actual population $\alpha/2$ or $1 - \alpha/2$ quantiles of $Q_{Y|X}$. ReForm adjusts the upper bound and lower bound of the naïve reference interval $\tilde{C}(X)$ by separate constants. This adjustment would be ideal if the $\alpha/2$ or $1 - \alpha/2$ quantiles of the conditional distributions of the reference set $P_{Y|X}$, where $P = P_X P_{Y|X}$, and calibration set $Q_{Y|X}$ differ by separate constants. If that relationship is true and the naïve reference interval is close to the true $\alpha/2$ or $1 - \alpha/2$ quantiles of $P_{Y|X}$, then the ReFormed reference interval should be close to the quantiles of $Q_{Y|X}$. This situation is similar to a common assumption in harmonization methods, which is that batches differ in mean and variance (Beer et al., 2020; Gardner et al., 2025; Johnson et al., 2007; Pomponio et al., 2020). This situation is also comparable to the assumptions of ConQuR, which adjusts the conditional quantiles across batches to perform harmonization (Ling et al., 2022), and GAMLSS with random effects, which are typically specified for both the mean and variance (Bethlehem et al., 2022). In our ReForm R package, we provide diagnostic plots based on the upper and lower residuals for qualitative evaluation.

Since our primary goal is to provide a reference interval calibration method that controls FPR in a new individual, we do not evaluate our method based on recovery of the true quantiles of $Q_{Y|X}$. However, we compare ReForm to refitting (Refit) throughout all experiments, where refitting uses the calibration set to obtain an estimate of the population quantiles of $Q_{Y|X}$.

Our setting is a special case of covariate shift and distribution shift, where the new observation is drawn from a different distribution than the full data (Barber et al., 2023; Tibshirani et al., 2019). In our setting however, we have a calibration set available that is drawn from the same distribution as the new observation. Therefore, the results of CQR hold when using that calibration set. However, distribution shift relative to the calibration set could occur in situations where the sampling scheme or data acquisition changes over the course of an MRI study.

2.5 Image acquisition and preprocessing

All data in this study are from the Lifespan Brain Chart Consortium (LBCC) with preprocessing details previously published (Bethlehem et al., 2022; Dorfschmidt et al., 2025). In summary, all studies collected T1-weighted images and certain studies collected T2-FLAIR images as well. When both T1-weighted and T2-FLAIR images were available, FreeSurfer's combined T1-T2 *recon-all* was used to extract volumetric and cortical measures. When only T1-weighted images were available, FreeSurfer's standard *recon-all* was used

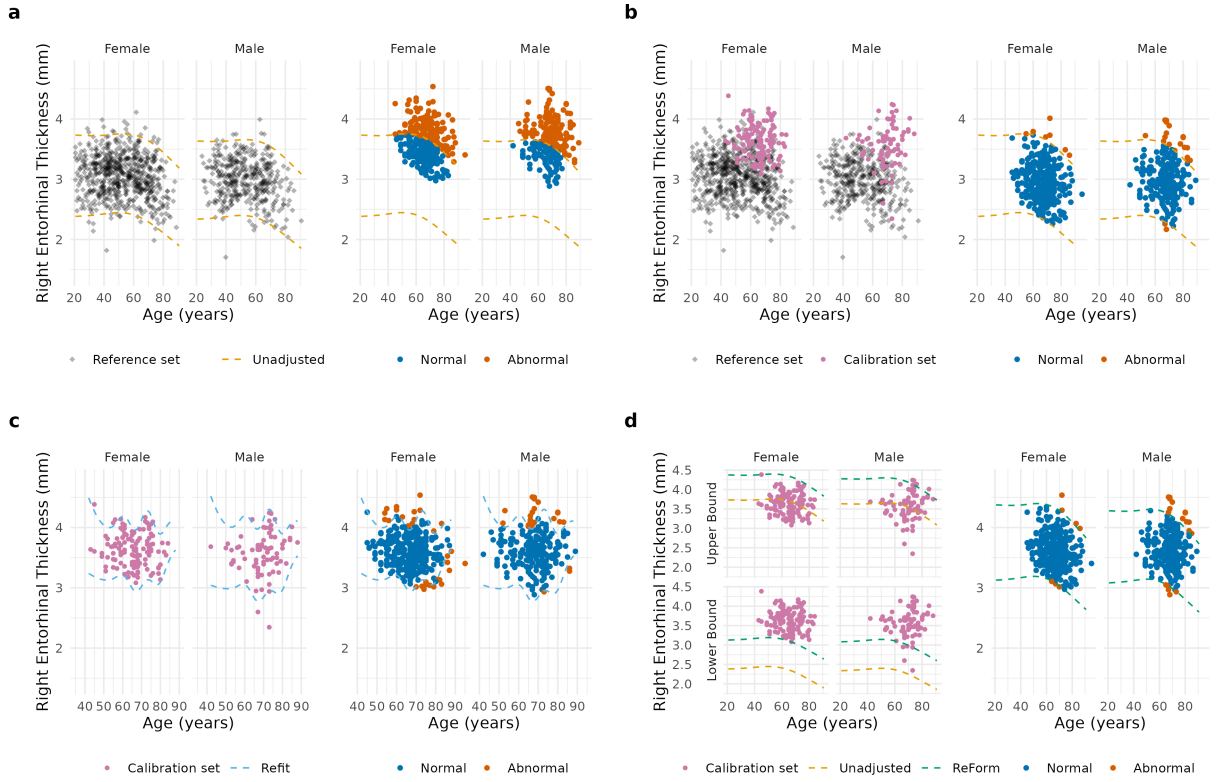

Figure 1: Illustration of methods employed in Experiment 1, which fitted reference intervals in ARWiBO 1.0 T and applied them to OASIS 3.0 T scans. All reference intervals were applied to the same validation set consisting of 588 OASIS 3.0 T scans, represented as colored points where blue is inside the reference interval (normal) and orange is outside (abnormal). Data shown is from a single run from Experiment 1. (a) 1032 ARWiBO scans (diamond points) were used to fit 95% reference intervals using GAMLSS (left, light orange dashed line), which were directly applied to OASIS 3.0 T scans yielding 314 (53.4%) outside the interval. (b) 1032 ARWiBo scans were combined with 200 OASIS 3.0 T scans (pink) to harmonize across sites (left), which shifted the validation set (right) yielding 27 outside (4.6%) (c) A 95% reference interval was fit in 200 OASIS 3.0 T scans (pink) using GAMLSS (left, light blue dashed line) and applied to the validation set yielding 57 outside (9.7%). (d) The upper and lower bounds of the reference interval from (a) were calibrated using Reference interval calibration via conFormal prediction (ReForm, left, green dashed line) in 200 OASIS 3.0 T scans and applied to the validation set yielding 26 outside (4.4%).

instead. Briefly, *recon-all* involved non-uniformity correction, projection to Talairach space, intensity normalisation, skull-stripping, automatic tissue and subcortical segmentation, surface interpolation, tessellation and registration. Different FreeSurfer versions were used across the LBCC studies with additional details available elsewhere (Bethlehem et al., 2022; Dorfschmidt et al., 2025). All data analyzed in this work used FreeSurfer version 6.0.142.

Subject details were provided by the individual studies included in the LBCC. We used a subset of baseline scans from AD studies for our experiments. Our inclusion and study-specific diagnostic criteria are detailed under the descriptions of each experiment in this study.

2.6 Resampling experiments in LBCC data

We chose to use right entorhinal cortical thickness as our primary outcome. Our rationale is that atrophy in this region has been identified in mild cognitive impairment preceding AD with further reductions across stages of AD (Juottonen et al., 1999; Tapiola et al., 2008). We chose to examine cortical thickness owing to its high sensitivity to MRI acquisition and preprocessing differences (Chen et al., 2022; Fortin et al., 2018; Han et al., 2006).

Across all experiments, we evaluated methods via FPR. For a valid reference interval, FPR should be controlled at α across randomly sampled new observations for a given calibration set. However, empirical FPR may vary depending on the demographics and outcomes included in the calibration set. FPR should also depend on the size of the calibration set, following results from conformal prediction that size impacts the upper bound on coverage (Lei et al., 2018; Romano et al., 2019b). Furthermore, FPR is computed from the data, so the theoretical guarantees of ReForm may not correspond to exact control of FPR. For comparison against the target FPR, we applied one-sample Wilcoxon signed-rank tests. For comparison between methods, we used Wilcoxon signed-rank tests across pairs of methods applied to the same data splits. All tests were performed with a significance level of 0.05.

Our experiments estimated 90%, 95%, and 99% reference intervals which have target FPRs of 0.1, 0.05, and 0.01 respectively. In all experiments, we designated separate reference dataset and test datasets as different sites from the LBCC data. Then, we randomly drew reference sets from the reference dataset and calibration sets from the test dataset, both without replacement. Across experiments, reference set sizes varied across $\{50, 200, n_T\}$ where n_T refers to the size of the entire reference dataset. Calibration sets varied across $\{20, 40, 80, 120, 160, 200, 300, 400, 500, 600\}$ throughout all experiments. Once the calibration set was drawn, the validation set consisted of all remaining observations in the test dataset. We primarily used GAMLSS to fit reference intervals assuming a Box-Cox t-distribution (BCT) with a fixed effect for biological sex and a smooth age term with default settings in both the mean and variance. The reference interval from GAMLSS was estimated from the population quantiles of the estimated BCT distributions. In Supplementary Materials, we report results using additive quantile regression instead, including a fixed effect for biological sex and a smooth age term with default settings. Each experiment was repeated 1,000 times for each unique combination of reference set size, calibration set size, and model choice.

We compared ReForm to: the unadjusted naïve reference interval (Unadjusted), refitting reference intervals in the calibration set (Refit); two harmonization methods (ComBat-GAM and ComBatLS); and GAMLSS-based model adjustment where appropriate. ComBat-GAM and ComBatLS were chosen since both account for nonlinear age effects, but ComBatLS additionally models potential covariate effects in the scale of measurements (Gardner et al., 2025). GAMLSS-based model adjustment only applied to settings where the reference interval was constructed using a GAMLSS model and the reference dataset had multiple sites. For model adjustment, our chosen GAMLSS model (GAMLSS-RE) assumed a BCT distribution and specified fixed effects for biological sex, a smooth age term with default settings, and a random effect for batch, all in both the mean and variance. Both the harmonization and model adjustment methods employed can be fit without the reference set; however, omitting the reference introduces error (Bethlehem et al., 2022; Xin et al., 2025). Here, we compared ReForm to methods that use the reference set in order to assess performance against the optimal implementation of existing methods.

GAMLSS was implemented via the *gamlss2* R package v0.1-0 (`https://github.com/gamlss-dev/gamlss2`). Additive quantile regression was fit using the *qgam* R package v2.0.0 on CRAN (Fasiolo et al., 2021). ComBat-GAM and ComBatLS were both implemented using the *CombatFamily* R package v0.2.1 (`https://github.com/andy1764/ComBatFamily`), which allows for estimated harmonization parameters

to be applied to the validation set. ComBat-GAM was fit using a smooth term for age and fixed effect for sex, both specified using the *gam* function from the *mgcv* R package v.1.9-1. ComBatLS was fit using smooth terms for age (*ps* function) and fixed effects for sex in both the mean and variance using the *gamlss* R package v.5.4-22. Since our outcome involved a single outcome, ComBat-GAM and ComBatLS are implemented without the empirical Bayes shrinkage across outcomes (Johnson et al., 2007).

2.7 Experiment 1: Different study, different acquisition

Our first experiment used 1.0 T scans from the Alzheimer's Disease Repository Without Borders (ARWiBo; Frisoni et al., 2009) study to fit reference intervals and applied them to 3.0 T scans from the Open Access Series of Imaging Studies (OASIS 3.0 T; LaMontagne et al., 2019; Marcus et al., 2007, 2010). Our goal for this experiment was to assess how severely FPR could be impacted by study differences, including a large magnetic field strength difference (1.0 T vs 3.0 T) in MRI acquisition. The ARWiBo study defined cognitively normal (CN) as the absence of subjective memory complaints, objective memory impairment based on several standard exams, and AD assessed via the NINCDS-ADRDA criteria Frisoni et al. (2009); McKhann et al. (1984). The OASIS study defined CN as subjects assessed via the Clinical Dementia Rating (CDR) scale (LaMontagne et al., 2019; Morris, 1993) to have a CDR of less than 0.5. Here, the aforementioned GAMLSS model adjustment method could not be applied since the reference dataset only contains a single batch.

2.8 Experiment 2: Same study, different acquisition

Our second experiment used 1.5 T scans from OASIS study (OASIS 1.5 T; Marcus et al., 2007, 2010) to fit reference intervals and applied them to the OASIS 3.0 T. Unlike our first experiment, the reference and test dataset here were both from the OASIS study with identical inclusion criteria, the same definition of CN, and comparable acquisition parameters (LaMontagne et al., 2019). However, we anticipated that there would still notable differences due to the difference in magnetic field strength (1.5 T vs 3.0 T). In this experiment, we compared ReForm to refitting (Refit) and ComBatLS. Again, the GAMLSS model adjustment method could not be used in this setting.

2.9 Experiment 3: Multi-site reference dataset

Our third experiment used multi-site reference data from the National Alzheimer's Coordinating Center (NACC), which used several 3.0 T scanners across multiple study sites with details reported previously (Morris et al., 2006). We fitted reference intervals in NACC and applied these reference intervals to the OASIS 3.0 T. The NACC database used its own diagnostic criteria based on several screening tools, including the CDR, to categorize subjects as either CN, mild cognitive impairment, or AD patients (Morris et al., 2006). In this study, we included CN from both studies and AD subjects from the OASIS 3.0 T, where the OASIS study defined AD as subjects having a CDR of 0.5 or greater (LaMontagne et al., 2019).

Here, we first compared in CN subjects the following methods: ReForm, the unadjusted naïve reference interval (Unadjusted), refitting (Refit), the same two harmonization methods (ComBat-GAM and ComBatLS), and GAMLSS model adjustment treating batch as a random effect (GAMLSS-RE). Batches included the 22 NACC study sites and OASIS 3.0 T. Then, we drew calibration sets from OASIS 3.0T CN subjects and compared all methods in the OASIS 3.0 T AD patients. Our goal was to measure the *positive rate* (PR) estimated as the proportion of OASIS 3.0 T AD patients with observations outside a given reference interval. We aimed to demonstrate similar or increased PR while controlling FPR at our designated level. As a supplementary analysis, we also examine PR subsetting to trials where the FPR close to our target FPR (α), within $\alpha \pm 0.2 \times \alpha$. Finally, we applied our NACC reference intervals to the ARWiBo 1.0 T study described previously.

Table 1: Subject characteristics by study and site.

	ARWiBo	OASIS		NACC
	1.0 T (Exp. 1 Reference)	1.5 T (Exp. 2 Reference)	3.0 T (Exp. 1-3 Test)	All 22 Sites (Exp. 3 Reference)
Sample size	1155	210	1107	4530
Age (years)				
Mean (SD)	55.0 (16.1)	66.9 (11.4)	69.4 (8.99)	71.8 (10.6)
Median [Min, Max]	54.8 [20.8, 91.8]	67.0 [45.0, 95.0]	70.0 [42.0, 95.0]	72.5 [19.7, 101]
Sex				
Female	735 (63.6%)	145 (69.0%)	625 (56.5%)	2800 (61.8%)
Male	420 (36.4%)	65 (31.0%)	482 (43.5%)	1730 (38.2%)
Diagnosis				
CN	1032 (89.4%)	172 (81.9%)	788 (71.2%)	2823 (62.3%)
AD	123 (10.6%)	38 (18.1%)	319 (28.8%)	1707 (37.7%)

Abbreviations: Exp., Experiment; ARWiBo, Alzheimer's Disease Repository Without Borders; OASIS, Open Access Series of Imaging Studies; NACC, National Alzheimer's Coordinating Center; SD, Standard deviation; CN, Cognitively normal; AD, Alzheimer's disease

3 Results

3.1 Data subset from the LBCC

Table 1 reports the sample used across all experiments labeled by the experiment (e.g. Exp. 1 Reference refers to Experiment 1 reference set). For this investigation, we only included subjects that are labeled as either cognitively normal (CN) or Alzheimer's disease (AD) based on the individual study's criteria detailed in Section 2. Experiment 1 included CN subjects from the ARWiBo and OASIS 3.0 T studies ($n = 1820$). Experiment 2 included CN subjects from the ARWiBo and OASIS 3.0 T studies ($n = 960$). Experiment 3 included CN subjects from the NACC and OASIS 3.0 T studies ($n = 3611$ across 23 study sites) and AD subjects from the OASIS 3.0 T study ($n = 319$).

3.2 ReForm versus refitting intervals

In general, we found that ReForm outperformed the Unadjusted and Refit methods for controlling FPR. Figure 2 shows FPR as medians and interquartile ranges (IQR) from Experiment 1. For 95% reference intervals with the full reference dataset and calibration set size of 40, the Unadjusted method resulted in median FPR of 0.527, far exceeding the target FPR of 0.05. ReForm demonstrated better control of FPR (median FPR 0.040, IQR 0.022-0.066) than Refit (median FPR 0.107, IQR 0.074-0.148; Refit vs. Reform Wilcoxon signed-rank test $p < 0.001$). In a larger calibration set of size 400, ReFit (median FPR 0.052, IQR 0.041-0.062) and ReForm (median FPR 0.049, IQR 0.039-0.059) showed more comparable performance. For 99% reference intervals and a small calibration set of size 40, both Refit (median FPR 0.043, IQR 0.023-0.071) and ReForm (median FPR 0.041, IQR 0.023-0.068) failed to control FPR at $\alpha = 0.01$. With 160 calibration samples, Refit still showed poor control (median FPR 0.014, IQR 0.010-0.021) while ReForm significantly outperformed (median FPR 0.010, IQR 0.005-0.018; Refit vs ReForm $p < 0.001$). ReForm was uniquely able to achieve FPR across trials not significantly different from 0.01 across all reference set sizes at calibration set sizes of 400 and 600. Using additive quantile regression instead of GAMLSS, we found that ReForm controlled the FPR of 95% and 90% reference intervals better than Refit for small calibration sets and was not significantly different from the target FPR in large calibration sets (Supplementary Figure 1).

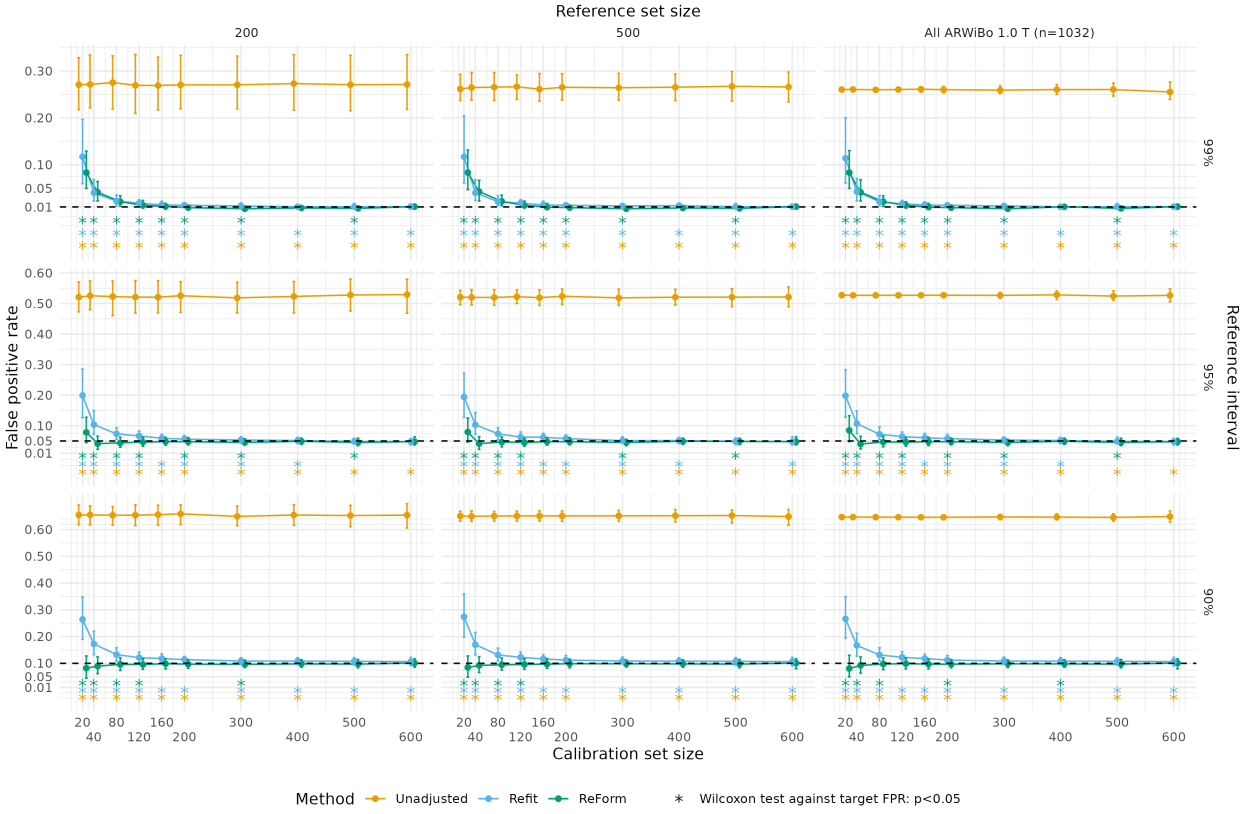

Figure 2: False positive rates across unadjusted, refit, and ReFormed reference intervals from Experiment 1, where we fitted reference intervals in ARWiBO 1.0 T scans and applied them to OASIS 3.0 T scans. Median and interquartile range are plotted across 1,000 trials. Plots assess the validity of estimated 99% reference intervals (top panel, target FPR of 0.01); 95% reference intervals (middle panel, target FPR of 0.05; and 90% reference interval (bottom panel, target FPR of 0.1). Asterisks used to show significant differences from the target FPR (dashed line) using a one-sample Wilcoxon signed-rank test. Unadjusted represents direct application of reference intervals fit using GAMLSS. Refit refers to refitting using GAMLSS in the calibration set of a designated size. ReForm is our proposed method which calibrates the unadjusted reference interval. All FPR values are computed in the same OASIS 3.0 T scans with varying sample size left after holding out the calibration set. In Refit, some trials yielded errors during fitting and are not shown.

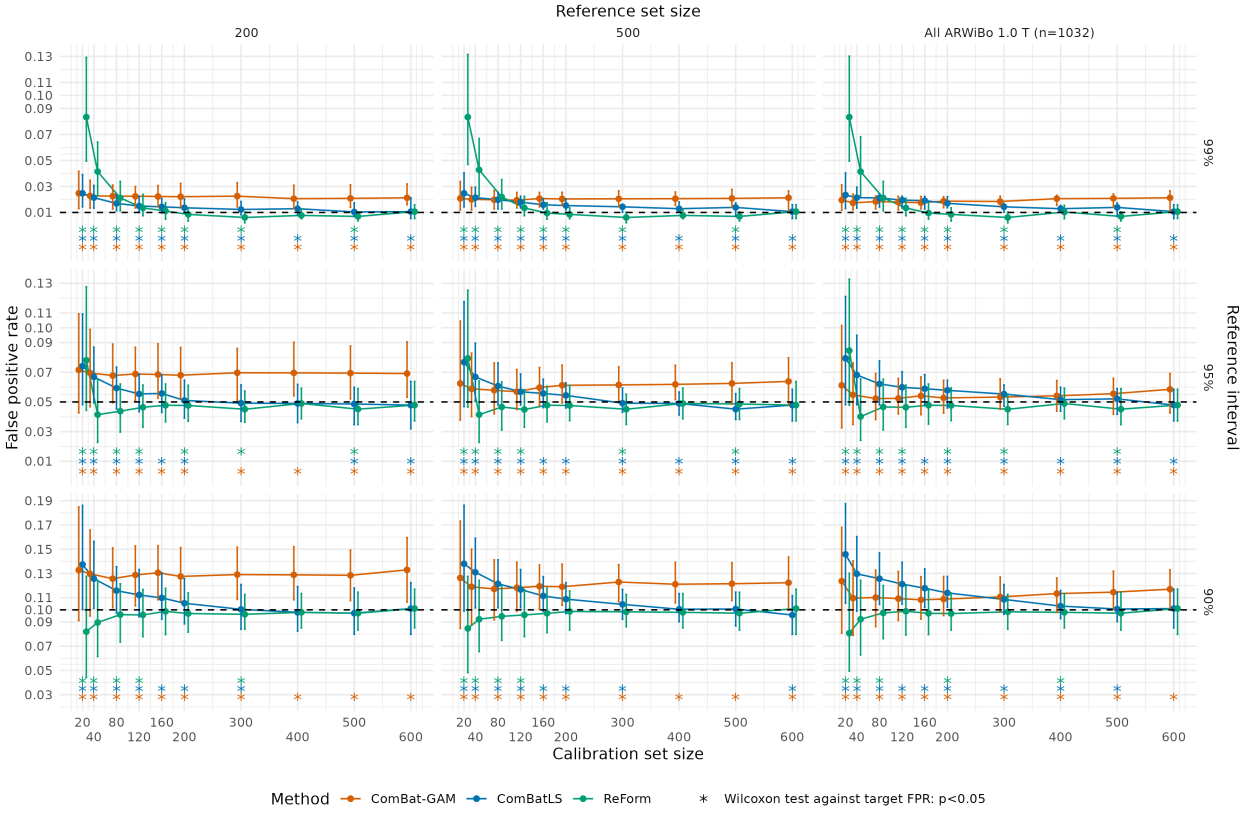

Figure 3: False positive rates across harmonization methods and ReForm from Experiment 1, where we fitted reference intervals in ARWiBO 1.0 T scans and applied them to OASIS 3.0 T scans. Median and interquartile range are plotted across 1,000 trials. Plots assess the validity of estimated 99% reference intervals (top panel, target FPR of 0.01); 95% reference intervals (middle panel, target FPR of 0.05; and 90% reference interval (bottom panel, target FPR of 0.1). Asterisks used to show significant differences from the target FPR (dashed line) using a one-sample Wilcoxon signed-rank test. ComBat-GAM and ComBatLS are fit on the entire reference set combined with the calibration set. ReForm is our proposed method which calibrates the unadjusted reference interval. All FPR values are computed in the same OASIS 3.0 T scans with varying sample size left after holding out the calibration set.

3.3 ReForm versus statistical harmonization

We first observed that ComBatLS notably outperformed ComBat-GAM for FPR control, but both harmonization methods failed to control FPR in certain settings. Figure 3 shows FPR of ComBat-GAM, ComBatLS, and ReForm across trials in Experiment 1, where intervals are fit in the ARWiBo 1.0 T and applied to OASIS 3.0 T scans. For 90% reference intervals, ComBat-GAM exhibited dependence on the reference set size. Fixing the calibration set size at 200 and comparing reference set sizes of 200 and 1033, ComBat-GAM median FPR goes from 0.128 (IQR 0.107-0.151) to 0.109 (IQR 0.095-0.128). Comparing ComBat-GAM and ComBatLS in an ideal setting where both methods have access to the full reference set and a large calibration set of size 500, ComBatLS showed significantly lower FPR (median FPR 0.101, IQR 0.090-0.115) than ComBat-GAM (median FPR 0.115, IQR 0.101-0.132; one-sided Wilcoxon signed-rank test $p < 0.001$). Comparing ComBatLS versus ReForm for 90% reference intervals, ComBatLS demonstrated poor control of FPR in calibration sets of size 120 (median FPR 0.121, IQR 0.105-0.139) while ReForm exhibited FPR not significantly different than the target FPR of 0.1 (median FPR 0.099, IQR 0.079-0.120; one-sample Wilcoxon signed-rank test against 0.1, $p = 0.67$).

The superior performance of ReForm over ComBatLS was more stark in Experiment 2, which fits reference intervals in the OASIS 1.5 T for application to OASIS 3.0 T. As hypothesized, we observed that the FPRs

of unadjusted intervals in Experiment 2 were closer to the target FPRs compared to Experiment 1 (Figure 3 vs. Figure 4). However, we found that ComBatLS showed poor FPR control even in ideal settings. For 95% reference intervals with access to the full reference set and calibration set size of 400, ComBatLS exhibits FPR significantly different than the target FPR of 0.05 (median 0.075, IQR 0.070-0.082; one-sample Wilcoxon signed-rank test against 0.05, $p < 0.001$). In this setting, ReForm achieves FPR close to $\alpha = 0.05$ (median FPR 0.049, IQR 0.039-0.059; $p = 0.44$). For 99% reference intervals in the same setting, performance of ComBatLS (median FPR 0.026, IQR 0.021-0.028) was close to Unadjusted (median FPR 0.026, IQR 0.023-0.031). In contrast, ReForm demonstrated effective control of FPR at close to, but significantly less than, the target FPR of 0.01 (median FPR 0.008, IQR 0.005-0.013; one-sample Wilcoxon signed-rank test against 0.01, $p = 0.01$). Moreover, we observed that ComBatLS depended on the reference set size while ReForm did not. Moving from 50 reference samples to the full reference set to fit 99% reference intervals, ComBatLS median FPR decreases substantially from 0.054 (IQR 0.046-0.062) to 0.010 (IQR 0.008-0.015) whereas ReForm is stable at 0.008 (IQR 0.005-0.013) for both reference set sizes.

In Experiment 3, Figure 5 shows that ComBatLS fit across NACC study sites and OASIS 3.0 T failed to control FPR across most settings. For instance, ComBatLS for 95% reference intervals showed poor control in both small calibration sets of size 40 (median FPR 0.066, IQR 0.048-0.090; one-sample Wilcoxon signed-rank test against 0.05, $p < 0.001$) and large calibration sets of size 400 (median FPR 0.054, IQR 0.046-0.062; $p < 0.001$).

3.4 ReForm versus GAMLSS in pooled multi-site data

Compared to fitting GAMLSS across all sites with a random effect for site (GAMLSS-RE), we found that ReForm showed less stable FPR control in small calibration sets but more precise control in larger calibration sets. Figure 5a illustrates these methods in a single trial from Experiment 3. Figure 5b compares FPR across Unadjusted, ComBatLS, and fitting GAMLSS across all sites with a random effect for site (GAMLSS-RE), and ReForm. In Unadjusted, we observed in 99% reference intervals that the FPR is already close to the target of 0.01 (Unadjusted median FPR 0.008). GAMLSS-RE performed consistently even across small calibration samples taken from OASIS 3.0 T. For example, 99% reference intervals constructed via GAMLSS-RE controlled FPR consistently even with 20 calibration samples (median FPR 0.010, IQR 0.008-0.018), whereas ReForm did not (median FPR 0.081, IQR 0.048-0.124, Wilcoxon rank-sum test $p < 0.001$). In 90% reference intervals, we observed GAMLSS-RE was slightly conservative even with 600 calibration samples (median FPR 0.096, IQR 0.080-0.112) compared to ReForm (median FPR 0.101, IQR 0.085-0.117, two-sided Wilcoxon signed-rank test, $p < 0.001$). Furthermore, we found that ReForm able to achieve FPR control not significantly different from the target FPRs for 90%, 95% and 99% reference intervals with 600 calibration samples, while GAMLSS-RE did not (Figure 5). ReForm was not significantly different than the target FPR for 80 calibration samples in 90% reference intervals and 120 calibration samples in 95% reference intervals (Figure 5). In Supplementary Materials, we observe similar results in application to the ARWiBo study where ReForm achieves comparable performance to GAMLSS-RE at 80 samples for 90% reference intervals, 120 for 95% reference intervals, and 160 samples for 99% reference intervals (Supplementary Figure 3a).

3.5 Reference intervals applied to Alzheimer's disease patients

In applying reference intervals to AD patients, ReForm and GAMLSS-RE both considerably outperformed Unadjusted in terms of positive rate (PR), the proportion of AD patients with right entorhinal thickness outside of a given reference interval. Figure 6a illustrates a ReFormed reference interval compared to GAMLSS-RE and direct application of unadjusted intervals (Unadjusted), all from a single run of Experiment 3. Figure 6b compares PR across Unadjusted, GAMLSS-RE, and ReForm across all trials. Results for $\alpha = 0.01$ are omitted since we observed that neither method controlled FPR adequately; less than 15% of trials had both methods yield FPR between 0.008 and 0.012 (Supplementary Figure 4a). First, we observed that the unadjusted reference intervals yield a median PR below 0.12 across 90% (median FPR 0.113), 95% (median FPR 0.066), and 99% reference intervals (median FPR 0.013). Next, we compared GAMLSS-RE and ReForm at a large calibration set size of 400, where both methods showed good control of FPR. Comparing GAMLSS-RE and ReForm for 95% intervals, both showed similar performance (GAMLSS-RE median PR 0.176, IQR 0.169-0.182; ReForm median PR 0.182, IQR 0.176-0.188). For 90% reference intervals, Re-

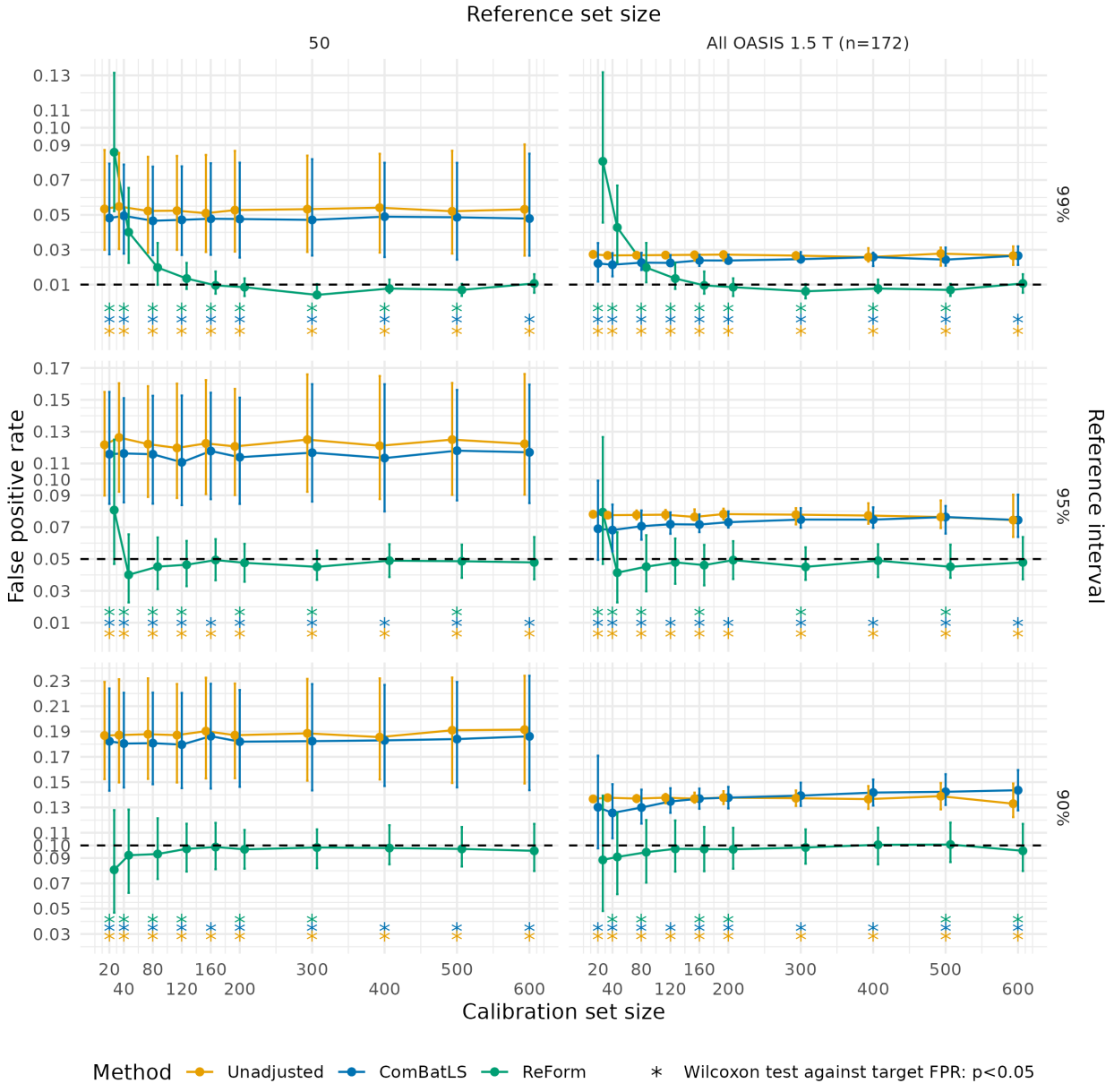

Figure 4: False positive rates across unadjusted, ComBatLS, and ReForm for fitting reference intervals from Experiment 2, where we fitted reference intervals in OASIS 1.5 T scans and applied them to OASIS 3.0 T scans. Median and interquartile range are plotted across 1,000 trials. Plots assess the validity of estimated 99% reference intervals (top panel, target FPR of 0.01); 95% reference intervals (middle panel, target FPR of 0.05; and 90% reference interval (bottom panel, target FPR of 0.1). Asterisks used to show significant differences from the target FPR (dashed line) using a one-sample Wilcoxon signed-rank test. Unadjusted represents direct application of reference intervals fit using GAMLSS. ComBatLS is fit on the entire reference set combined with the calibration set. ReForm is our proposed method which calibrates the unadjusted reference interval. All FPR values are computed in the same OASIS 3.0 T scans with varying sample size left after holding out the calibration set.

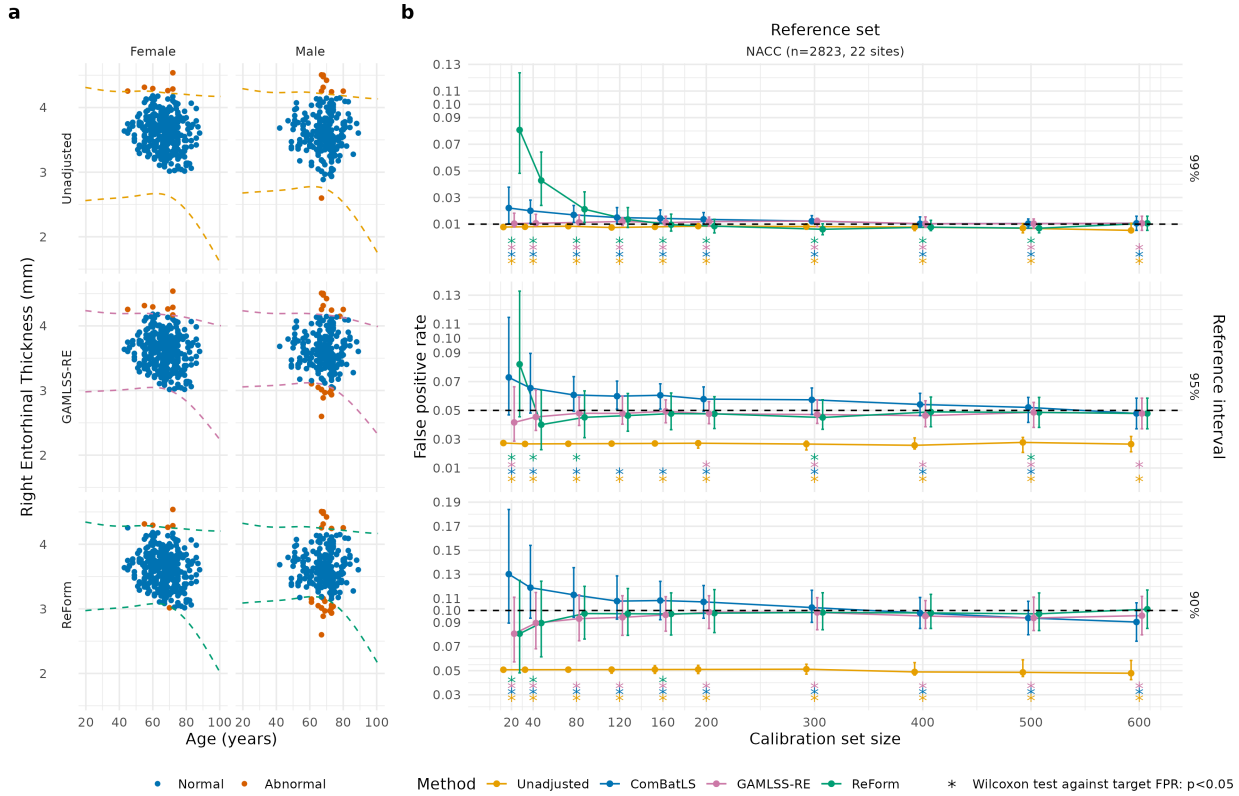

Figure 5: Application of reference intervals fitted in the NACC reference dataset to OASIS 3.0 T healthy subjects in Experiment 3. (a) Illustration of methods using a single trial from Experiment 3, where we fitted reference intervals in the multi-site NACC study and applied them to OASIS 3.0 T scans. 588 OASIS 3.0 T healthy scans are shown. Direct application of 95% reference intervals (Unadjusted) yielded 15 outside (2.5%). A 95% reference interval constructed using GAMLSS with a random effect across the 2928 NACC scans and 200 OASIS 3.0 T healthy scans yielded 27 outside (4.6%). A ReFormed reference interval yielded 30 outside (5.1%) (b) False positive rates across unadjusted, ComBatLS, GAMLSS-RE and ReForm for Experiment 3. Median and interquartile range are plotted across 1,000 trials. Plots assess the validity of estimated 99% reference intervals (top panel, target FPR of 0.01); 95% reference intervals (middle panel, target FPR of 0.05; and 90% reference interval (bottom panel, target FPR of 0.1). Asterisks used to show significant differences from the target FPR (dashed line) using a one-sample Wilcoxon signed-rank test. Unadjusted represents direct application of reference intervals fit using GAMLSS. ComBatLS is fit on the entire reference set combined with the calibration set. GAMLSS-RE refers to fitting a GAMLSS with random effect for study site across the combined reference and calibration sets. ReForm is our proposed method which calibrates the unadjusted reference interval. All FPR values are computed in the same OASIS 3.0 T scans with varying sample size left after holding out the calibration set.

Form achieved a higher median PR (0.260, IQR 0.251-0.283) than GAMLSS-RE (0.248, IQR 0.229-0.254; one-sided Wilcoxon signed-rank test $p < 0.001$). This result may have been related to GAMLSS-RE being slightly slightly overconservative in terms of FPR (Section 3.4). This PR difference corresponded to median 6 (IQR 2-10) more AD subjects being labelled as abnormal based on the ReFormed reference interval versus the GAMLSS-RE interval. Supplementary Figure 4b shows performance across trials where FPR for both methods is controlled within 20% of the target FPR, where we made similar observations. Supplementary Figures 3b and 5 show results for PR when applying reference intervals instead to the ARWiBo AD patients ($n = 123$), generally showing similar results despite the severely inflated PR in Unadjusted intervals.

4 Discussion

We have introduced ReForm, a new method for calibration of reference intervals, and demonstrated its utility for MRI studies. As an illustrative example, we have focused on right entorhinal cortical thickness which is known to be lower in AD patients, and showed that ReForm considerably outperforms unadjusted intervals, refitting, and harmonization. When multi-site reference data was available, ReForm performed similarly to GAMLSS-based model adjustment for sufficiently sized samples. In application to AD patients, we show that ReForm and GAMLSS model adjustment increased the positive rate, defined as the proportion of AD patients identified as abnormal. ReForm achieved slightly higher positive rates depending on the chosen reference interval. While we compared ReForm to methods that utilize the reference dataset, ReForm does not require any reference data. Not requiring reference data is of high importance in practice due to privacy concerns related to sharing patient data as well as computational concerns if models must be refit on reference data. Moreoever, across our experiments, ReForm performance did not depend on the reference set size, model choice, or underlying data distribution——unlike other strategies. Yet, ReForm outperformed harmonization methods and performs comparably to a GAMLSS model adjustment method, both of which require using the entire reference dataset for estimation.

Across experiments, ReForm performance depended on the calibration set size. In general, large calibration sets (≥ 120) were necessary to achieve no significant difference from the target level α for a given reference interval. For adequate performance, we found that about 40 samples were necessary for 90% reference intervals and 80 samples for 95% reference intervals. For 95% and 99% reference intervals across all experiments, we observed that ReForm was significantly overconservative in calibration set sizes of 300 and 500. This phenomenon may be related to rounding in the computation of empirical quantiles used in ReForm. In 95% reference intervals for example, the calibration set size of 300 is not divisible by 40, which leads to rounding in the 97.5% empirical quantile of the lower and upper residuals used in ReForm. Future studies should consider interpolation as a potential solution to this issue. In Experiment 3, we observed in 99% reference intervals that ReForm exhibited overconservative FPR and lowered PR in AD subjects. For a calibration set size of 200 or less, ReForm adjusts the upper bound to the largest observation and lower bound to the smallest observation. In this setting, ReForm can thus exhibit dramatically different behavior depending on outlier observations in the calibration set. For 99% reference intervals, we found that ReForm showed good control of FPR in calibration set sizes of 120 or greater, but advise caution due to the large variability. These findings are consistent with previous studies of GAMLSS and additive quantile regression, which suggest that large samples are necessary to estimate more extreme quantiles (Bozek et al., 2023; Jennen-Steinmetz, 2014). Overall, we suggest that the calibration set size and observations are both carefully chosen to minimize issues with FPR using ReForm. Future work could also consider other methods for conformal prediction under distribution shift, which offer solutions when less or no observations are available from the new sample (Barber et al., 2023; Ai and Ren, 2025).

While refitting is recommended by the CLSI when reference intervals fail to be verified in a new sample (Clinical and Laboratory Standards Institute, 2008), our results suggest that refitting is rarely the best option. Refitting was particularly prone to error in smaller sample sizes, and led to inflated FPR in several settings. Furthermore, refitting was dependent on the choice of model, which we demonstrated by comparing additive quantile regression versus GAMLSS. Refitting can also lead to reference intervals with unreasonable age trends as illustrated in Figure 1c. Nonetheless, in cases where no reference data or reference intervals are available, refitting shows good control of FPR with sample sizes of 160 or above. These findings are consistent with the CLSI guidelines for fitting nonparametric 95% reference intervals, which recommend at

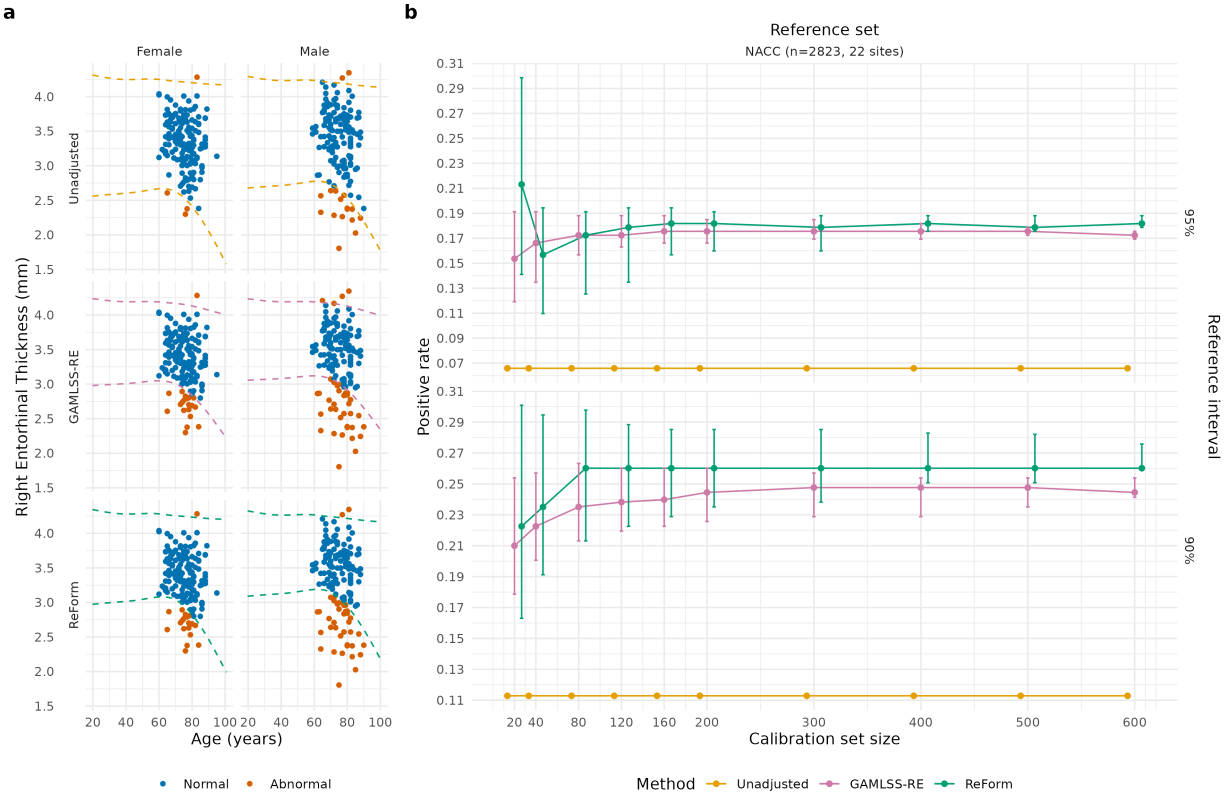

Figure 6: Application of reference intervals fitted in the NACC reference dataset to OASIS 3.0 T Alzheimer's disease (AD) subjects in Experiment 3, where we fitted reference intervals in multi-site NACC healthy individuals and applied them to OASIS 3.0 T Alzheimer's disease patients. (a) Illustration of methods using a single trial from Experiment 3. 319 OASIS 3.0 T AD scans are shown. Direct application of 95% reference intervals (Unadjusted) yielded 21 outside (6.5%). A 95% reference interval constructed using GAMLSS with a random effect across the 2928 NACC scans and 200 OASIS 3.0 T healthy scans yielded 58 outside (18.2%). A ReFormed reference interval yielded 60 outside (18.8%) (b) Positive rates across unadjusted, ComBatLS, GAMLSS-RE and ReForm for Experiment 3. Median and interquartile range are plotted across 1,000 trials. Unadjusted represents direct application of reference intervals fit using GAMLSS. Here, calibration sets are drawn from the OASIS 3.0 T cognitively normal individuals. ComBatLS is a harmonization method that uses the reference data and calibration set to adjust the data. GAMLSS-RE refers to fitting a GAMLSS with random effect for study site across the combined reference and calibration sets. ReForm is our proposed method which calibrates the unadjusted reference interval. All positive rate values are computed in the same 319 OASIS 3.0 T AD scans.

least 120 samples (Clinical and Laboratory Standards Institute, 2008).

When a multi-site reference dataset is available, model adjustment provides a solution with reasonable control of FPR. In fact, model adjustment using the entire reference dataset outperformed ReForm in smaller samples. Our result is consistent with recent studies which release GAMLSS models accompanied by out-of-sample fitting of random effects, showing generally good performance with at least 100 samples (Bethlehem et al., 2022). Our current work only considers GAMLSS model adjustment using the full reference dataset, which represents the best possible scenario for model adjustment. Future studies should evaluate how limited access to the reference dataset impacts performance of model adjustment for controlling FPR in reference intervals.

Compared to harmonization, ReForm outperformed both ComBat-GAM and ComBatLS for control of FPR across experiments with adequate calibration set size. Both ComBat-GAM and ComBatLS adjust for batch effects in mean and variance, while accounting for covariate effects. As a result, neither method explicitly models the tails of the distribution, which drive the constructed reference intervals. Comparing ComBat-GAM versus ComBatLS, we observed that ComBatLS performs substantially better, likely due to its modeling of covariate effects in the variance. We also observed that these harmonization methods can be either overly conservative or liberal in terms of FPR depending on the experimental setting. These harmonization methods were also highly dependent on the size of the reference dataset provided, which is important to consider for studies that share reference data (Zhu et al., 2025).

Comparing Experiments 1 and 2, where Experiment 2 has a lesser difference between the reference and test datasets, we observed a counterintuitive result in our harmonization methods considered. As hypothesized, Experiment 1 demonstrated poorer FPR control in unadjusted intervals compared to Experiment 2. ComBatLS offered notable benefit over unadjusted intervals in Experiment 1, but almost no benefit in Experiment 2. Several potential reasons should be considered. First, as pointed out previously, ComBatLS adjusts for batch effects in mean and variance and does not necessarily adjust the tails of the distribution used in reference interval construction. Second, minor batch effects in the mean and variance lead to minor harmonization adjustments overall. Thus, it is possible that the batch effects observed in Experiment 2 do not substantially impact the mean and variance of measurements, leading to insufficient adjustment by ComBatLS.

Several limitations of the current work limit the generalizability of our findings. First, we only considered a single MRI phenotype for its relevance to AD research. Second, our experiments involved a small number of MRI studies and cannot directly generalize to other MRI studies. While ReForm makes no assumptions about the data distribution, the demographic and phenotypical variances may impact empirical performance. Further studies should assess how ReForm across other MRI phenotypes and studies, including volumetric phenotypes and MRI studies collecting functional and diffusion MRI.

We compared ReForm to alternative methods on the basis of FPR, not on the basis of recovering the true underlying population quantiles. Our reasoning was that ReForm is a way to calibrate existing reference intervals, not construct new ones from the reference dataset. As such, we did not aim to perform a comprehensive comparison across modeling methods. Prior work has provided comparisons across methods for building brain charts in both collected MRI data (Ge et al., 2024) and simulated data (Bozek et al., 2023), as well as demonstrations of conformal prediction for controlling FPR in simulation (Lei et al., 2018; Romano et al., 2019b). Future studies could examine how well ReForm recovers true population quantiles by simulating different reference and test dataset distributions.

Our alternative methods were chosen on the basis of potential feasibility in the same setting as ReForm. Through recent papers, approximations of the statistical harmonization and GAMLSS model adjustment methods in our study can be implemented without access to the reference dataset (Bethlehem et al., 2022; Xin et al., 2025). However, our study applied those methods with access to the reference to represent the best possible performance of the alternative methods. Other harmonization methods including deep learning approaches were not implemented since our data only includes derived MRI phenotypes rather than the full MRI image (Hu et al., 2023). We chose not to compare to the ConQuR harmonization method since it does not have an implementation for out-of-sample scans (Ling et al., 2022). Our goal was to compare to a subset of prominent model frameworks while acknowledging that there are likely other approaches that could be utilized in this setting. Future studies could compare ReForm to additional methods.

Throughout this study, we evaluated FPR across all ages and sexes in the validation sample. This FPR measure has limitations though, since it does not guarantee that reference intervals control FPR within age

and sex bins. An extension of conformal prediction suggests that sampling calibration sets across groups of interest is a viable solution to this problem (Romano et al., 2019a). Other recent works also consider this problem and propose extensions of conformal prediction that guarantee coverage conditional on covariates (Gibbs et al., 2023; Guan, 2023; Hore and Barber, 2024). Further investigation is needed to determine if their methodologies can be applied in our setting and to assess the sample size requirements and empirical performance. Follow-up studies in MRI reference intervals should consider defining appropriate age and sex bins for evaluation of FPR within demographic groups.

Finally, we advise caution in interpreting MRI reference intervals. To quote Leo Tolstoy's *Anna Karenina*, "[a]ll happy families are alike; each unhappy family is unhappy in its own way". In other words, an observation that falls outside an MRI reference interval may be abnormal owing to a multitude of reasons. In this work, we do not investigate potential causes. In practice, it is important to treat an abnormal observation with particular attention and closely examine other measurements to understand potential biological or technical causes.

5 Conclusion

We propose a novel method called ReForm which calibrates reference intervals to a new sample. Unlike other methods, we show that the performance of ReForm does not depend on the size of the reference dataset or nature of the modeling approach used to fit reference intervals. If a large multi-site reference dataset is available, then model-based batch effect adjustment offers a similarly viable solution that performs well even in small sample sizes. Our recommendations are to consider ReForm across a wide range of settings, so long as the sample size is sufficient for type of reference interval considered. We find that 40 samples is adequate for 90% reference intervals and 80 samples for 95% reference intervals. For the challenging problem of 99% reference intervals, we find that ReForm requires at least 120 samples to ensure acceptable control of FPR, but advise caution due to sensitivity to outliers. These recommendations are based on a limited set of MRI studies and benchmarks, which should be expanded in future work. While we have focused on a single MRI phenotype in our study, our ReForm method is broadly applicable to any phenotype that differs across study sites.

Data and Code Availability

An R implementation of the proposed method, ReForm, is available via GitHub (`https://github.com/andy1764/ReForm`). Data used in this work were from the Lifespan Brain Chart Consortium (LBCC), which includes multiple publicly-available MRI datasets. Our work included three datasets, each with their own data access and sharing policies, which can be accessed via the following websites: The Alzheimer's Disease Repository Without Borders (ARWiBo, `https://www.arwibo.it`); Open Access Series of Imaging Studies (OASIS, `https://sites.wustl.edu/oasisbrains/`); and the National Alzheimer's Coordinating Center (NACC, `https://naccdata.org/`).

Author Contributions

A.A.C.: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing—Original Draft, Writing—Review and Editing, Visualization, Funding Acquisition. J.S.: Resources, Data Curation, Writing—Review and Editing. Resources, Data Curation, Writing—Review and Editing. M.G.: Resources, Data Curation, Writing—Review and Editing. I.G.: Methodology, Writing—Review and Editing. R.A.I.B: Resources, Data Curation, Writing—Review and Editing. L.D.: Resources, Data Curation, Writing—Review and Editing. E.K.: Resources, Data Curation, Writing—Review and Editing. A.B.: Writing—Review and Editing. J.J.: Writing—Review and Editing. S.V.: Methodology, Writing—Review and Editing. T.D.S.: Writing—Review and Editing. A.A.B.: Conceptualization, Resources, Data Curation, Writing—Original Draft, Writing—Review and Editing, Visualization, Supervision, Funding Acquisition.

Declaration of Competing Interests

J.M.S. and R.A.I.B. are directors and J.M.S., R.A.I.B., and A.A.B. have equity in Centile Bioscience.

Ethics

All subjects gave written informed consent prior to participation and all studies were approved by the regional ethics committees.

Acknowledgments

This work was supported by the National Institute of Mental Health (award numbers R01MH133843, R01MH123550, R01MH112847, R01MH113550, R37MH125829, and R01MH120482). A.B. and J.J. were supported by the National Institutes of Health (R01AG054159); this grant provided effort support and did not fund the data collection or analyses reported in this manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Ai, J. and Ren, Z. (2025). Not all distributional shifts are equal: Fine-grained robust conformal inference.

Antonopoulos, G., More, S., Raimondo, F., Eickhoff, S. B., Hoffstaedter, F., and Patil, K. R. (2023). A systematic comparison of VBM pipelines and their application to age prediction. *NeuroImage*, 279:120292.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845.

Beer, J. C., Tustison, N. J., Cook, P. A., Davatzikos, C., Sheline, Y. I., Shinohara, R. T., and Linn, K. A. (2020). Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. *NeuroImage*, 220:117129.

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., Beyer, F., Blangero, J., Blesa Cábez, M., Boardman, J. P., Borzage, M., Bosch-Bayard, J. F., Bourke, N., Calhoun, V. D., Chakravarty, M. M., Chen, C., Chertavian, C., Chetelat, G., Chong, Y. S., Cole, J. H., Corvin, A., Costantino, M., Courchesne, E., Crivello, F., Cropley, V. L., Crosbie, J., Crossley, N., Delarue, M., Delorme, R., Desrivieres, S., Devenyi, G. A., Di Biase, M. A., Dolan, R., Donald, K. A., Donohoe, G., Dunlop, K., Edwards, A. D., Elison, J. T., Ellis, C. T., Elman, J. A., Eyler, L., Fair, D. A., Feczko, E., Fletcher, P. C., Fonagy, P., Franz, C. E., Galan-Garcia, L., Gholipour, A., Giedd, J., Gilmore, J. H., Glahn, D. C., Goodyer, I. M., Grant, P. E., Groenewold, N. A., Gunning, F. M., Gur, R. E., Gur, R. C., Hammill, C. F., Hansson, O., Hedden, T., Heinz, A., Henson, R. N., Heuer, K., Hoare, J., Holla, B., Holmes, A. J., Holt, R., Huang, H., Im, K., Ipser, J., Jack, C. R., Jackowski, A. P., Jia, T., Johnson, K. A., Jones, P. B., Jones, D. T., Kahn, R. S., Karlsson, H., Karlsson, L., Kawashima, R., Kelley, E. A., Kern, S., Kim, K. W., Kitzbichler, M. G., Kremen, W. S., Lalonde, F., Landeau, B., Lee, S., Lerch, J., Lewis, J. D., Li, J., Liao, W., Liston, C., Lombardo, M. V., Lv, J., Lynch, C., Mallard, T. T., Marcelis, M., Markello, R. D., Mathias, S. R., Mazoyer, B., McGuire, P., Meaney, M. J., Mechelli, A., Medic, N., Misic, B., Morgan, S. E., Mothersill, D., Nigg, J., Ong, M. Q. W., Ortinau, C., Ossenkoppele, R., Ouyang, M., Palaniyappan, L., Paly, L., Pan, P. M., Pantelis, C., Park, M. M., Paus, T., Pausova, Z., Paz-Linares, D., Pichet Binette, A., Pierce, K., Qian, X., Qiu, J., Qiu, A., Raznahan, A., Rittman, T., Rodrigue, A., Rollins, C. K., Romero-Garcia, R., Ronan, L., Rosenberg, M. D., Rowitch, D. H., Salum, G. A., Satterthwaite, T. D., Schaare, H. L., Schachar, R. J., Schultz, A. P., Schumann, G., Schöll, M., Sharp, D., Shinohara, R. T., Skoog, I., Smyser, C. D., Sperling, R. A., Stein, D. J., Stolicyn, A., Suckling, J., Sullivan, G., Taki, Y., Thyreau, B., Toro, R., Traut, N., Tsvetanov, K. A., Turk-Browne, N. B., Tuulari, J. J., Tzourio, C., Vachon-Presseau, É., Valdes-Sosa, M. J., Valdes-Sosa, P. A., Valk, S. L., van Amelsvoort, T., Vandekar, S. N., Vasung, L., Victoria, L. W., Villeneuve, S., Villringer, A., Vértes, P. E., Wagstyl, K., Wang, Y. S., Warfield, S. K., Warrier, V., Westman, E., Westwater, M. L., Whalley, H. C., Witte, A. V., Yang, N., Yeo, B., Yun, H., Zalesky, A., Zar, H. J., Zettergren, A., Zhou, J. H., Ziauddeen, H., Zugman, A., Zuo, X. N., Bullmore, E. T., and Alexander-Bloch, A. F. (2022). Brain charts for the human lifespan. *Nature*, 604(7906):525–533.

Bhagwat, N., Barry, A., Dickie, E. W., Brown, S. T., Devenyi, G. A., Hatano, K., DuPre, E., Dagher, A., Chakravarty, M., Greenwood, C. M. T., Misic, B., Kennedy, D. N., and Poline, J.-B. (2021). Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses. *GigaScience*, 10(1):giaa155.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., Benoit, R. G., Berkers, R. M., Bhanji, J. P., Biswal, B. B., Bobadilla-Suarez, S., Bortolini, T., Bottenhorn, K. L., Bowring, A., Braem, S., Brooks, H. R., Brudner, E. G., Calderon, C. B., Camilleri, J. A., Castrellon, J. J., Cecchetti, L., Cieslik, E. C., Cole, Z. J., Collignon, O., Cox, R. W., Cunningham, W. A., Czoschke, S., Dadi, K., Davis, C. P., De Luca, A., Delgado, M. R., Demetriou, L., Dennison, J. B., Di, X., Dickie, E. W., Dobryakova, E., Donnat, C. L., Dukart, J., Duncan, N. W., Durnez, J., Eed, A., Eickhoff, S. B., Erhart, A., Fontanesi, L., Fricke, G. M.,

Fu, S., Galván, A., Gau, R., Genon, S., Glatard, T., Glerean, E., Goeman, J. J., Golowin, S. A. E., González-García, C., Gorgolewski, K. J., Grady, C. L., Green, M. A., Guassi Moreira, J. F., Guest, O., Hakimi, S., Hamilton, J. P., Hancock, R., Handjaras, G., Harry, B. B., Hawco, C., Herholz, P., Herman, G., Heunis, S., Hoffstaedter, F., Hogeveen, J., Holmes, S., Hu, C.-P., Huettel, S. A., Hughes, M. E., Iacovella, V., Iordan, A. D., Isager, P. M., Isik, A. I., Jahn, A., Johnson, M. R., Johnstone, T., Joseph, M. J. E., Juliano, A. C., Kable, J. W., Kassinopoulos, M., Koba, C., Kong, X.-Z., Koscik, T. R., Kucukboyaci, N. E., Kuhl, B. A., Kupek, S., Laird, A. R., Lamm, C., Langner, R., Lauharatanahirun, N., Lee, H., Lee, S., Leemans, A., Leo, A., Lesage, E., Li, F., Li, M. Y., Lim, P. C., Lintz, E. N., Liphardt, S. W., Losecaat Vermeer, A. B., Love, B. C., Mack, M. L., Malpica, N., Marins, T., Maumet, C., McDonald, K., McGuire, J. T., Melero, H., Méndez Leal, A. S., Meyer, B., Meyer, K. N., Mihai, G., Mitsis, G. D., Moll, J., Nielson, D. M., Nilsonne, G., Notter, M. P., Olivetti, E., Onicas, A. I., Papale, P., Patil, K. R., Peelle, J. E., Pérez, A., Pischedda, D., Poline, J.-B., Prystauka, Y., Ray, S., Reuter-Lorenz, P. A., Reynolds, R. C., Ricciardi, E., Rieck, J. R., Rodriguez-Thompson, A. M., Romyn, A., Salo, T., Samanez-Larkin, G. R., Sanz-Morales, E., Schlichting, M. L., Schultz, D. H., Shen, Q., Sheridan, M. A., Silvers, J. A., Skagerlund, K., Smith, A., Smith, D. V., Sokol-Hessner, P., Steinkamp, S. R., Tashjian, S. M., Thirion, B., Thorp, J. N., Tinghög, G., Tisdall, L., Tompson, S. H., Toro-Serey, C., Tresols, J. J. T., Tozzi, L., Truong, V., Turella, L., van 't Veer, A. E., Verguts, T., Vettel, J. M., Vijayarajah, S., Vo, K., Wall, M. B., Weeda, W. D., Weis, S., White, D. J., Wisniewski, D., Xifra-Porxas, A., Yearling, E. A., Yoon, S., Yuan, R., Yuen, K. S., Zhang, L., Zhang, X., Zosky, J. E., Nichols, T. E., Poldrack, R. A., and Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88.

Bozek, J., Griffanti, L., Lau, S., and Jenkinson, M. (2023). Normative models for neuroimaging markers: Impact of model selection, sample size and evaluation criteria. *NeuroImage*, 268:119864.

Cetin Karayumak, S., Bouix, S., Ning, L., James, A., Crow, T., Shenton, M., Kubicki, M., and Rathi, Y. (2019). Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. *NeuroImage*, 184:180–200.

Chen, A. A., Srinivasan, D., Pomponio, R., Fan, Y., Nasrallah, I. M., Resnick, S. M., Beason-Held, L. L., Davatzikos, C., Satterthwaite, T. D., Bassett, D. S., Shinohara, R. T., and Shou, H. (2022). Harmonizing functional connectivity reduces scanner effects in community detection. *NeuroImage*, 256:119198.

Clinical and Laboratory Standards Institute (2008). Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline. *CLSI Document EP28_A3c*.

Dima, D., Modabbernia, A., Papachristou, E., Doucet, G. E., Agartz, I., Aghajani, M., Akudjedu, T. N., Albajes-Eizagirre, A., Alnaes, 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., Bonvino, A., Boomsma, D. I., Borgwardt, S., Bourque, J., Brandeis, D., Breier, A., Brodaty, H., Brouwer, R. M., Buitelaar, J. K., Busatto, G. F., Buckner, R. L., Calhoun, V., Canales-Rodríguez, E. J., Cannon, D. M., Caseras, X., Castellanos, F. X., Cervenka, S., Chaim-Avancini, T. M., Ching, C. R. K., Chubar, V., Clark, V. P., Conrod, P., Conzelmann, A., Crespo-Facorro, B., Crivello, F., Crone, E. A., Dannlowski, U., Dale, A. M., Davey, C., de Geus, E. J. C., de Haan, L., de Zubicaray, G. I., den Braber, A., Dickie, E. W., Di Giorgio, A., Doan, N. T., Dørum, E. S., Ehrlich, S., Erk, S., Espeseth, T., Fatouros-Bergman, H., Fisher, S. E., Fouche, J.-P., Franke, B., Frodl, T., Fuentes-Claramonte, P., Glahn, D. C., Gotlib, I. H., Grabe, H.-J., Grimm, O., Groenewold, N. A., Grotegerd, D., Gruber, O., Gruner, P., Gur, R. E., Gur, R. C., Hahn, T., Harrison, B. J., Hartman, C. A., Hatton, S. N., Heinz, A., Heslenfeld, D. J., Hibar, D. P., Hickie, I. B., Ho, B.-C., Hoekstra, P. J., Hohmann, S., Holmes, A. J., Hoogman, M., Hosten, N., Howells, F. M., Hulshoff Pol, H. E., Huyser, C., Jahanshad, N., James, A., Jernigan, T. L., Jiang, J., Jönsson, E. G., Joska, J. A., Kahn, R., Kalnin, A., Kanai, R., Klein, M., Klyushnik, T. P., Koenders, L., Koops, S., Krämer, B., Kuntsi, J., Lagopoulos, J., Lázaro, L., Lebedeva, I., Lee, W. H., Lesch, K.-P., Lochner, C., Machielsen, M. W. J., Maingault, S., Martin, N. G., Martínez-Zalacaín, I., Mataix-Cols, D., Mazoyer, B., McDonald, C., McDonald, B. C., McIntosh, A. M., McMahon, K. L., McPhilemy, G., Meinert, S., Menchón, J. M., Medland, S. E., Meyer-Lindenberg, A., Naaijen, J., Najt, P., Nakao, T., Nordvik, J. E., Nyberg, L., Oosterlaan, J., de la Foz, V. O.-G., Paloyelis, Y., Pauli, P., Pergola, G., Pomarol-Clotet, E., Portella, M. J., Potkin, S. G., Radua, J., Reif, A., Rinker, D. A., Roffman, J. L., Rosa, P. G. P., Sacchet,

M. D., Sachdev, P. S., Salvador, R., Sánchez-Juan, P., Sarró, S., Satterthwaite, T. D., Saykin, A. J., Serpa, M. H., Schmaal, L., Schnell, K., Schumann, G., Sim, K., Smoller, J. W., Sommer, I., Soriano-Mas, C., Stein, D. J., Strike, L. T., Swagerman, S. C., Tamnes, C. K., Temmingh, H. S., Thomopoulos, S. I., Tomyshev, A. S., Tordesillas-Gutiérrez, D., Trollor, J. N., Turner, J. A., Uhlmann, A., van den Heuvel, O. A., van den Meer, D., van der Wee, N. J. A., van Haren, N. E. M., Van't Ent, D., van Erp, T. G. M., Veer, I. M., Veltman, D. J., Voineskos, A., Völzke, H., Walter, H., Walton, E., Wang, L., Wang, Y., Wassink, T. H., Weber, B., Wen, W., West, J. D., Westlye, L. T., Whalley, H., Wierenga, L. M., Williams, S. C. R., Wittfeld, K., Wolf, D. H., Worker, A., Wright, M. J., Yang, K., Yoncheva, Y., Zanetti, M. V., Ziegler, G. C., Thompson, P. M., Frangou, S., and Karolinska Schizophrenia Project (KaSP) (2022). Subcortical volumes across the lifespan: Data from 18,605 healthy individuals aged 3-90 years. *Human Brain Mapping*, 43(1):452–469.

Dorfschmidt, L., White, S., Gardner, M., Bedford, S., Ball, G., Edwards, A. D., Gu, Y., He, Y., Huang, H., Karandikar, S., Kyriakopoulou, V., Ouyuang, M., Pulli, E. P., Raznahan, A., Robinson, E. C., Romero-García, R., Satterthwaite, T. D., Schabdach, J., Sebenius, I., Sha, Z., Shinohara, R. T., Tuulari, J., Vaghi, M. M., Vogel, J., Váša, F., Williams, L., Zimmerman, D., Warrier, V., Wagstyl, K., Bullmore, E. T., Bethlehem, R. A., Seidlitz, J., and Alexander-Bloch, A. (2025). Charting structural brain asymmetry across the human lifespan. *bioRxiv*, page 2025.07.21.665924.

Fasiolo, M., , Simon N., W., , Margaux, Z., , Raphaël, N., and and Goude, Y. (2021). Fast Calibrated Additive Quantile Regression. *Journal of the American Statistical Association*, 116(535):1402–1412.

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., McInnis, M., Phillips, M. L., Trivedi, M. H., Weissman, M. M., and Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120.

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., and Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, 161:149–170.

Frangou, S., Modabbernia, A., Williams, S. C. R., Papachristou, E., Doucet, G. E., Agartz, I., Aghajani, M., Akudjedu, T. N., Albajes-Eizagirre, A., Alnaes, 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., Bonvino, A., Boomsma, D. I., Borgwardt, S., Bourque, J., Brandeis, D., Breier, A., Brodaty, H., Brouwer, R. M., Buitelaar, J. K., Busatto, G. F., Buckner, R. L., Calhoun, V., Canales-Rodríguez, E. J., Cannon, D. M., Caseras, X., Castellanos, F. X., Cervenka, S., Chaim-Avancini, T. M., Ching, C. R. K., Chubar, V., Clark, V. P., Conrod, P., Conzelmann, A., Crespo-Facorro, B., Crivello, F., Crone, E. A., Dale, A. M., Dannlowski, U., Davey, C., de Geus, E. J. C., de Haan, L., de Zubicaray, G. I., den Braber, A., Dickie, E. W., Di Giorgio, A., Doan, N. T., Dørum, E. S., Ehrlich, S., Erk, S., Espeseth, T., Fatouros-Bergman, H., Fisher, S. E., Fouche, J.-P., Franke, B., Frodl, T., Fuentes-Claramonte, P., Glahn, D. C., Gotlib, I. H., Grabe, H.-J., Grimm, O., Groenewold, N. A., Grotegerd, D., Gruber, O., Gruner, P., Gur, R. E., Gur, R. C., Hahn, T., Harrison, B. J., Hartman, C. A., Hatton, S. N., Heinz, A., Heslenfeld, D. J., Hibar, D. P., Hickie, I. B., Ho, B.-C., Hoekstra, P. J., Hohmann, S., Holmes, A. J., Hoogman, M., Hosten, N., Howells, F. M., Hulshoff Pol, H. E., Huyser, C., Jahanshad, N., James, A., Jernigan, T. L., Jiang, J., Jönsson, E. G., Joska, J. A., Kahn, R., Kalnin, A., Kanai, R., Klein, M., Klyushnik, T. P., Koenders, L., Koops, S., Krämer, B., Kuntsi, J., Lagopoulos, J., Lázaro, L., Lebedeva, I., Lee, W. H., Lesch, K.-P., Lochner, C., Machielsen, M. W. J., Maingault, S., Martin, N. G., Martínez-Zalacaín, I., Mataix-Cols, D., Mazoyer, B., McDonald, C., McDonald, B. C., McIntosh, A. M., McMahon, K. L., McPhilemy, G., Meinert, S., Menchón, J. M., Medland, S. E., Meyer-Lindenberg, A., Naaijen, J., Najt, P., Nakao, T., Nordvik, J. E., Nyberg, L., Oosterlaan, J., de la Foz, V. O.-G., Paloyelis, Y., Pauli, P., Pergola, G., Pomarol-Clotet, E., Portella, M. J., Potkin, S. G., Radua, J., Reif, A., Rinker, D. A., Roffman, J. L., Rosa, P. G. P., Sacchet, M. D., Sachdev, P. S., Salvador, R., Sánchez-Juan, P., Sarró, S., Satterthwaite, T. D., Saykin, A. J., Serpa, M. H., Schmaal, L., Schnell, K., Schumann, G., Sim, K., Smoller, J. W., Sommer, I., Soriano-Mas, C., Stein, D. J., Strike, L. T., Swagerman, S. C., Tamnes, C. K., Temmingh, H. S., Thomopoulos, S. I., Tomyshev, A. S., Tordesillas-Gutiérrez, D., Trollor, J. N., Turner,

J. A., Uhlmann, A., van den Heuvel, O. A., van den Meer, D., van der Wee, N. J. A., van Haren, N. E. M., van 't Ent, D., van Erp, T. G. M., Veer, I. M., Veltman, D. J., Voineskos, A., Völzke, H., Walter, H., Walton, E., Wang, L., Wang, Y., Wassink, T. H., Weber, B., Wen, W., West, J. D., Westlye, L. T., Whalley, H., Wierenga, L. M., Wittfeld, K., Wolf, D. H., Worker, A., Wright, M. J., Yang, K., Yoncheva, Y., Zanetti, M. V., Ziegler, G. C., Karolinska Schizophrenia Project (KaSP), Thompson, P. M., and Dima, D. (2022). Cortical thickness across the lifespan: Data from 17,075 healthy individuals aged 3-90 years. *Human Brain Mapping*, 43(1):431–451.

Frisoni, G. B., Prestia, A., Zanetti, O., Galluzzi, S., Romano, M., Cotelli, M., Gennarelli, M., Binetti, G., Bocchio, L., Paghera, B., Amicucci, G., Bonetti, M., Benussi, L., Ghidoni, R., and Geroldi, C. (2009). Markers of Alzheimer's disease in a population attending a memory clinic. *Alzheimer's & Dementia*, 5(4):307–317.

Gardner, M., Shinohara, R. T., Bethlehem, R. A. I., Romero-Garcia, R., Warrier, V., Dorfschmidt, L., Lifespan Brain Chart Consortium, Shanmugan, S., Thompson, P., Seidlitz, J., Alexander-Bloch, A. F., and Chen, A. A. (2025). ComBatLS: A Location- and Scale-Preserving Method for Multi-Site Image Harmonization. *Human Brain Mapping*, 46(8):e70197.

Ge, R., Yu, Y., Qi, Y. X., Fan, Y.-n., Chen, S., Gao, C., Haas, S. S., New, F., Boomsma, D. I., Brodaty, H., Brouwer, R. M., Buckner, R., Caseras, X., Crivello, F., Crone, E. A., Erk, S., Fisher, S. E., Franke, B., Glahn, D. C., Dannlowski, U., Grotegerd, D., Gruber, O., Pol, H. E. H., Schumann, G., Tamnes, C. K., Walter, H., Wierenga, L. M., Jahanshad, N., Thompson, P. M., Frangou, S., Agartz, I., Asherson, P., Ayesa-Arriola, R., Banaj, N., Banaschewski, T., Baumeister, S., Bertolino, A., Borgwardt, S., Bourque, J., Brandeis, D., Breier, A., Buitelaar, J. K., Cannon, D. M., Cervenka, S., Conrod, P. J., Crespo-Facorro, B., Davey, C. G., de Haan, L., de Zubicaray, G. I., Giorgio, A. D., Frodl, T., Gruner, P., Gur, R. E., Gur, R. C., Harrison, B. J., Hatton, S. N., Hickie, I., Howells, F. M., Huyser, C., Jernigan, T. L., Jiang, J., Joska, J. A., Kahn, R. S., Kalnin, A. J., Kochan, N. A., Koops, S., Kuntsi, J., Lagopoulos, J., Lazaro, L., Lebedeva, I. S., Lochner, C., Martin, N. G., Mazoyer, B., McDonald, B. C., McDonald, C., McMahon, K. L., Medland, S., Modabbernia, A., Mwangi, B., Nakao, T., Nyberg, L., Piras, F., Portella, M. J., Qiu, J., Roffman, J. L., Sachdev, P. S., Sanford, N., Satterthwaite, T. D., Saykin, A. J., Sellgren, C. M., Sim, K., Smoller, J. W., Soares, J. C., Sommer, I. E., Spalletta, G., Stein, D. J., Thomopoulos, S. I., Tomyshev, A. S., Tordesillas-Gutiérrez, D., Trollor, J. N., van 't Ent, D., van den Heuvel, O. A., van Erp, T. G., van Haren, N. E., Vecchio, D., Veltman, D. J., Wang, Y., Weber, B., Wei, D., Wen, W., Westlye, L. T., Williams, S. C., Wright, M. J., Wu, M.-J., and Yu, K. (2024). Normative modelling of brain morphometry across the lifespan with CentileBrain: Algorithm benchmarking and model optimisation. *The Lancet Digital Health*, 6(3):e211–e221.

Gibbs, I., Cherian, J. J., and Candès, E. J. (2023). Conformal Prediction With Conditional Guarantees. https://arxiv.org/abs/2305.12616v4.

Gräsbeck, R. and Fellman, J. (1968). Normal values and statistics. *Scandinavian Journal of Clinical and Laboratory Investigation*, 21(3):193–195.

Guan, L. (2023). Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50.

Haddad, S. M. H., Scott, C. J. M., Ozzoude, M., Holmes, M. F., Arnott, S. R., Nanayakkara, N. D., Ramirez, J., Black, S. E., Dowlatshahi, D., Strother, S. C., Swartz, R. H., Symons, S., Montero-Odasso, M., Investigators, O., and Bartha, R. (2019). Comparison of quality control methods for automated diffusion tensor imaging analysis pipelines. *PLOS ONE*, 14(12):e0226715.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., and Fischl, B. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, 32(1):180–194.

Hore, R. and Barber, R. F. (2024). Conformal prediction with local weights: Randomization enables local guarantees.

Hu, F., Chen, A. A., Horng, H., Bashyam, V., Davatzikos, C., Alexander-Bloch, A., Li, M., Shou, H., Satterthwaite, T. D., Yu, M., and Shinohara, R. T. (2023). Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. *NeuroImage*, 274:120125.

Hu, F., Lucas, A., Chen, A. A., Coleman, K., Horng, H., Ng, R. W. S., Tustison, N. J., Davis, K. A., Shou, H., Li, M., Shinohara, R. T., and Initiative, T. A. D. N. (2024). DeepComBat: A statistically motivated, hyperparameter-robust, deep learning approach to harmonization of neuroimaging data. *Human Brain Mapping*, 45(11):e26708.

Jennen-Steinmetz, C. (2014). Sample size determination for studies designed to estimate covariate-dependent reference quantile curves. *Statistics in Medicine*, 33(8):1336–1348.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)*, 8(1):118–127.

Juottonen, K., Laakso, M. P., Partanen, K., and Soininen, H. (1999). Comparative MR Analysis of the Entorhinal Cortex and Hippocampus in Diagnosing Alzheimer Disease. *American Journal of Neuroradiology*, 20(1):139–144.

Kim, M. E., Gao, C., Ramadass, K., Newlin, N. R., Kanakaraj, P., Bogdanov, S., Rudravaram, G., Archer, D., Hohman, T. J., Jefferson, A. L., Morgan, V. L., Roche, A., Englot, D. J., Resnick, S. M., Held, L. L. B., Cutting, L., Barquero, L. A., D'archangel, M. A., Nguyen, T. Q., Humphreys, K. L., Niu, Y., Vinci-Booher, S., Cascio, C. J., Team, T. H.-H. S., Initiative, A. D. N., Team, T. B. S., Li, Z., Vandekar, S. N., Zhang, P., Gore, J. C., Landman, B. A., and Schilling, K. G. (2025). White matter microstructure and macrostructure brain charts across the human lifespan.

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., and Marcus, D. (2019). OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Ling, W., Lu, J., Zhao, N., Lulla, A., Plantinga, A. M., Fu, W., Zhang, A., Liu, H., Song, H., Li, Z., Chen, J., Randolph, T. W., Koay, W. L. A., White, J. R., Launer, L. J., Fodor, A. A., Meyer, K. A., and Wu, M. C. (2022). Batch effects removal for microbiome data via conditional quantile regression. *Nature Communications*, 13(1):5418.

Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2010). Open Access Series of Imaging Studies (OASIS): Longitudinal MRI Data in Nondemented and Demented Older Adults. *Journal of cognitive neuroscience*, 22(12):2677–2684.

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, 34(7):939–944.

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*, 43(11):2412–2414.

Morris, J. C., Weintraub, S., Chui, H. C., Cummings, J., Decarli, C., Ferris, S., Foster, N. L., Galasko, D., Graff-Radford, N., Peskind, E. R., Beekly, D., Ramos, E. M., and Kukull, W. A. (2006). The Uniform Data Set (UDS): Clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. *Alzheimer Disease and Associated Disorders*, 20(4):210–216.

Ozarda, Y. (2016). Reference intervals: Current status, recent developments and future considerations. *Biochemia Medica*, 26(1):5–11.

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I. M., Satterthwaite, T. D., Fan, Y., Launer, L. J., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Wolf, D. H., Gur, R., Gur, R., Morris, J., Albert, M. S., Grabe, H. J., Resnick, S. M., Bryan, R. N., Wolk, D. A., Shinohara, R. T., Shou, H., and Davatzikos, C. (2020). Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208:116450.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 54(3):507–554.

Romano, Y., Barber, R. F., Sabatti, C., and Candès, E. J. (2019a). With Malice Towards None: Assessing Uncertainty via Equalized Coverage.

Romano, Y., Patterson, E., and Candes, E. (2019b). Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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., Chang, S.-E., Penninx, B. W., Heitzeg, M. M., Burt, S. A., Hyde, L. W., Amaral, D., Wu Nordahl, C., Andreasssen, O. A., Westlye, L. T., Zahn, R., Ruhe, H. G., Beckmann, C., and Marquand, A. F. (2022). Charting brain growth and aging at high spatial precision. *eLife*, 11:e72904.

Shafiei, G., Esper, N. B., Hoffmann, M. S., Ai, L., Chen, A. A., Cluce, J., Covitz, S., Giavasis, S., Lane, C., Mehta, K., Moore, T. M., Salo, T., Tapera, T. M., Calkins, M. E., Colcombe, S., Davatzikos, C., Gur, R. E., Gur, R. C., Pan, P. M., Jackowski, A. P., Rokem, A., Rohde, L. A., Shinohara, R. T., Tottenham, N., Zuo, X.-N., Cieslak, M., Franco, A. R., Kiar, G., Salum, G. A., Milham, M. P., and Satterthwaite, T. D. (2025). Reproducible Brain Charts: An open data resource for mapping brain development and its associations with mental health. *Neuron*.

Shinohara, R. T., Oh, J., Nair, G., Calabresi, P. A., Davatzikos, C., Doshi, J., Henry, R. G., Kim, G., Linn, K. A., Papinutto, N., Pelletier, D., Pham, D. L., Reich, D. S., Rooney, W., Roy, S., Stern, W., Tummala, S., Yousuf, F., Zhu, A., Sicotte, N. L., Bakshi, R., and Cooperative, t. N. (2017). Volumetric Analysis from a Harmonized Multisite Brain MRI Study of a Single Subject with Multiple Sclerosis. *American Journal of Neuroradiology*, 38(8):1501–1509.

Tapiola, T., Pennanen, C., Tapiola, M., Tervo, S., Kivipelto, M., Hänninen, T., Pihlajamäki, M., Laakso, M. P., Hallikainen, M., Hämäläinen, A., Vanhanen, M., Helkala, E.-L., Vanninen, R., Nissinen, A., Rossi, R., Frisoni, G. B., and Soininen, H. (2008). MRI of hippocampus and entorhinal cortex in mild cognitive impairment: A follow-up study. *Neurobiology of Aging*, 29(1):31–38.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal Prediction Under Covariate Shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xin, Y., Gardner, M., Tustison, N., Cook, P., Gee, J., Benitez, A., Jensen, J. H., Initiative, A. D. N., Consortium, L. B. C., Bethlehem, R., Seidlitz, J., Alexander-Bloch, A. F., and Chen, A. A. (2025). ComBat-Predict enhances generalizability of neuroimaging models to new sites.

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., Trivedi, M. H., Weissman, M. M., Shinohara, R. T., and Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human Brain Mapping*, 39(11):4213–4227.

Zhu, A. H., Nir, T. M., Javid, S., Villalón-Reina, J. E., Rodrigue, A. L., Strike, L. T., de Zubicaray, G. I., McMahon, K. L., Wright, M. J., Medland, S. E., Blangero, J., Glahn, D. C., Kochunov, P., Williamson, D. E., Håberg, A. K., Thompson, P. M., and Jahanshad, N. (2025). Lifespan reference curves for harmonizing multi-site regional brain white matter metrics from diffusion MRI. *Scientific Data*, 12(1):748.