Discover Data

RESEARCH Open Access

Data harmonizing via interpolation applied to brain age prediction

Nicolás Nieto[1,2*], Aditi Asati[3], Kshitij Jadhav[4†] and Kaustubh R. Patil[1†]

†Kshitij Jadhav and Kaustubh R. Patil have contributed equally to this study.

*Correspondence:
Nicolás Nieto
n.nieto@fz-juelich.de
[1]Forschungszentrum Jülich, Brain and Behaviour (INM-7), Institute of Neurosciences and Medicine, Jülich, Germany
[2]Institute of Systems Neuroscience, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[3]Mathematics Department, Indian Institute of Science Education and Research, Tirupati, India
[4]Koita Centre for Digital Health, Indian Institute of Technology, Bombay, India

Abstract

Brain age estimation using magnetic resonance imaging is a promising biomarker for detecting accelerated aging and neurodegenerative disorders. However, the development of robust clinical models is severely hampered by the "Effects of Site", where scanner-specific biases obscure biological signals in multi-center datasets. In this study, we propose a novel harmonization strategy, Inter-Site SMOTE, which generates synthetic training data by interpolating between age- and gender-matched participants from different sites. We hypothesize that these synthetic samples populate the sparse regions between site distributions, effectively bridging domain gaps while preserving biological integrity. We systematically evaluated this approach using four large neuroimaging datasets ($N = 2031$) in a leave-one-site-out regression task. Our results demonstrate that Inter-Site SMOTE significantly improves generalization to unseen scanners compared to standard data pooling. Crucially, we show that standard statistical harmonization (ComBat) fails to improve predictive performance in this setting due to inference-time assumptions, whereas our data-centric approach enhances robustness. Furthermore, we provide empirical evidence that the improvement is driven by the specific geometry of cross-site interpolation, as intra-site augmentation failed to yield comparable gains. This work presents a simple, effective solution for multi-site harmonization that circumvents the limitations of statistical adjustment methods, paving the way for more generalizable prediction models.

Keywords Brain age prediction, Synthetic minority over-sampling technique, Data harmonization, Data augmentation

1 Introduction

Neurodegenerative diseases represent a significant and growing global public health challenge [1, 2]. Early identification and intervention are crucial for effective disease management and the exploration of potential therapeutic strategies. In recent years, advances in machine learning (ML) and the availability of large-scale magnetic resonance imaging (MRI) datasets have driven the development of several biomarkers. Among these, brain age gap or age delta, the difference between ML-estimated brain age and chronological age, has emerged as a promising biomarker for detecting accelerated brain aging. This marker has shown utility in the early detection of neurodegenerative

diseases, such as Alzheimer's disease (AD) and Parkinson's disease (PD), and has provided insights into changes in brain structure and function [3–5]. Older appearing brains, with a higher age delta, are associated with cognitive impairment and AD, underscoring the potential of brain age models to predict future cognitive decline and identify early signs of AD, potentially enhancing individual risk assessment for neurodegenerative diseases and guiding personalized interventions [6]. Overall, brain age models hold promise for developing early detection and intervention strategies by identifying individuals at risk of accelerated brain aging.

Data collected from different MRI scanners, even the same model from the same manufacturer, show systematic differences. This site-specific variability is enhanced due to differences in scanner types, imaging protocols, magnetic field strengths, acquisition parameters, and reconstruction algorithms [7–10]. Known as the Effects of Site (EoS), this variability can manifest as discrepancies in image intensities, tissue volume measures, and connectivity, complicating data analysis and interpretation if not properly addressed [11, 12]. Since EoS are unrelated to biological information and stem solely from acquisition site differences, their removal can significantly enhance subsequent data analyses [13–17]. To address this challenge, numerous Methods Aiming to Remove the Effects of Site (MAREoS) have been proposed and developed [18, 19]. Consequently, to develop generalizable ML models that are accurate for different scanners, large multisite datasets are needed while removing the EoS.

Various harmonization strategies have been proposed to mitigate EoS, which can be broadly categorized into two main approaches: statistical methods and Deep Learning (DL) methods. Among statistical methods, ComBat is the most widely used and extensively developed [20, 21]. Originally designed for genomic data, ComBat performs harmonization by estimating location (additive) and scale (multiplicative) parameters for each feature and site. Several variants of ComBat have been introduced to address specific challenges, such as ComBat-GAM [22], which preserves nonlinear covariate effects using generalized additive models (GAM), and GMM-ComBat [23], which incorporates Gaussian mixture models to handle multimodal distributions, among others [19, 24]. Despite the availability of numerous ComBat variations, these methods often face limitations due to the limited and potentially imbalanced data from individual sites, where site and target variables may be correlated. This poses challenges for statistical methods like ComBat, as their leakage-free application may not always yield the expected benefits [25, 26]. Lastly, ComBat cannot be used for data from an unseen site.

On the other hand, several deep learning (DL) methods have also been developed for harmonization. For instance, a conditional variational autoencoder (CVAE) was used to generate data conditioned on additional covariates [27]. At the feature level, methodologies have been proposed that leveraged CVAE concepts to learn latent-space representations that are independent of imaging sites, utilizing batch-conditioned encoder-decoder pairs [28]. Other architectures, including generative adversarial networks (GANs), have also been explored. Examples include Cycle-consistent GANs (image-level) [29] and surface-to-surface GANs (S2SGAN) [30]. However, the training of deep learning harmonization models is often hindered by small sample sizes, as these models typically involve a large number of parameters and require significantly more data than is usually available.

Another approach, known as "prospective harmonization", involves recruiting traveling participants who visit every scanner used in a study. This allows for the effective

estimation of scanner-related biases, which can then be removed from the data of other participants [31]. While promising, this approach is challenging to implement in practice due to the challenges and cost of data collection. Taken together, there is a need for novel approaches to leverage existing data more effectively, enabling the development of generalizable and accurate models.

To address these limitations, we propose a novel data-centric approach utilizing the synthetic minority over-sampling technique (SMOTE) [32] to generate synthetic data that bridges the domain gap between sites. Unlike standard augmentation, which generates data within an existing distribution, we create synthetic samples by interpolating between age- and gender-matched individuals across different sites. The rationale is that while matched pairs share similar biological signals, their site-specific noise profiles differ. Interpolating between them preserves the biological information while averaging out site-specific artifacts, effectively generating training samples in the domain-invariant space between scanners. We hypothesize that incorporating these "inter-site" synthetic samples forces the model to ignore site-specific noise, thereby learning robust, site-invariant features. To validate our approach, we conducted systematic analyses using three large-scale datasets (N = 2031) and evaluated performance in a leave-one-site-out scheme. Our results demonstrate that the inclusion of synthetic data enhances model performance, although the precise mechanisms underlying these improvements remain challenging to disentangle due to limitations in data size. To facilitate reproducibility and further research, we have made the corresponding Python code publicly available: https://github.com/Aditi-Asati/Interpolation-And-Brain-Age-Prediction.

2 Methodology

2.1 Creation of synthetic samples via inter-site SMOTE

To mitigate site effects, we propose a novel adaptation of the synthetic minority oversampling technique (SMOTE). While standard SMOTE generates samples by interpolating between neighbors within the same class, our approach (Inter-Site SMOTE) generates synthetic samples by interpolating between matched pairs of participants from two *different* sites.

We designate one site as the *base* site (S_{base}) and the other as the *target* site (S_{target}). To maintain biological plausibility, interpolation is strictly constrained to pairs of participants matched by gender and age. We hypothesize that by linearly interpolating between these matched cross-site pairs, the resulting synthetic data will average out site-specific noise vectors while preserving the shared biological signal.

2.1.1 Formal definition

Let $\mathrm{b} \in \mathbb{R}^D$ represent a feature vector from the *base* site and $\mathrm{t} \in \mathbb{R}^D$ represent a feature vector from the *target* site. A synthetic sample s is generated via linear interpolation:

$$s = b + SMOTE_{\alpha}(t - b) \tag{1}$$

where $SMOTE_{\alpha} \in [0, 0.5]$ is a uniform random weighting factor. By restricting $SMOTE_{\alpha}$ to this range, the synthetic sample s remains in the local neighborhood of the base sample b, effectively acting as a "site-regularized" augmentation of the base site.

2.1.2 Metadata propagation

The metadata for the synthetic sample s is derived to ensure consistency with the feature interpolation:

- *Site Label:* Assigned as the base site. Due to the restriction in $SMOTE_\alpha$, the generated sample s will always be more similar to the *base* site.
- *Gender:* Inherited directly from b and t, since $\text{Gender}(b) = \text{Gender}(t)$ by design.
- *Age:* Computed using the same $SMOTE_\alpha$ used in Eq. (1) and the $\Delta age = \text{Age}(t) - \text{Age}(b)$:

$$\text{Age}(s) = \text{Age}(b) + SMOTE_\alpha * \Delta age \tag{2}$$

2.1.3 Algorithm

The data generation process is detailed in Algorithm 1.

Input: Datasets D_A (Site A) and D_B (Site B); Oversampling rate k.
Output: Augmented dataset including synthetic samples.
Procedure: Augment Site A (Base) using Site B (Target)
for each sample $\mathbf{b}_i \in D_A$ **do**
    Match: Select subset $M_i \subset D_B$ where $\text{Gender}(\mathbf{t}) = \text{Gender}(\mathbf{b}_i)$ for all $\mathbf{t} \in M_i$.
    Filter: Sort M_i by $|\text{Age}(\mathbf{b}_i) - \text{Age}(\mathbf{t})|$ and select the top k nearest neighbors $\{\mathbf{t}_1, \ldots, \mathbf{t}_k\}$.
    for $j = 1$ to k **do**
        Sample $SMOTE_\alpha \sim U(0, 0.5)$
        Generate $\mathbf{s}_{ij} = \mathbf{b}_i + SMOTE_\alpha(\mathbf{t}_j - \mathbf{b}_i)$
        Calculate $\text{Age}(\mathbf{s}_{ij}) = \text{Age}(\mathbf{b}_i) + SMOTE_\alpha * \Delta age_{ij}$
        Add $\mathbf{s}_{ij}$ to training set with label Site A.
    end for
end for
Repeat procedure swapping Base (Site B) and Target (Site A).

Algorithm 1 Inter-Site Interpolation Algorithm

2.2 Datasets

We utilized T1-weighted (T1w) magnetic resonance imaging (MRI) data from healthy subjects spanning the adult age range (18–88 years), sourced from three large neuroimaging datasets: the Cambridge Center for Ageing and Neuroscience (CamCAN) [33], the Information eXtraction from Images (IXI) dataset (https://brain-development.org/ixi-dataset/), and the enhanced Nathan Kline Institute-Rockland sample (eNKI) [34]. The IXI dataset comprises data acquired from three distinct imaging centers–Guy's Hospital, Hammersmith Hospital, and the Institute of Psychiatry–each treated as a separate site due to the use of different scanners (see Table 1 for detailed data characteristics). Due to its limited sample size, the IXI-IOP dataset was only used as a test site and never included in the training.

For all MRI images, Voxel-Based Morphometry (VBM) was performed using the Computational Anatomy Toolbox (CAT) version 12.8 [35] to extract modulated gray matter volume (GMV). The GMV maps were linearly resampled to a voxel size of $8 \times 8 \times 8\, mm^3$, resulting in 3747 features per image.

Table 1 Characteristics of the original MRI datasets used in the study

Dataset name	N images	Mean age	SD age	% Female
eNKI	818	46.90	17.73	65
CamCAN	651	54.27	18.59	50
IXI/Guys	313	50.84	15.88	56
IXI/HH	181	47.36	16.71	52
IXI/IOP	68	42.37	16.60	65

SD: standard deviation

Table 2 Hyperparameters for Kernel Ridge Regressor optimization

Hyperparameter	Values
λ	1×10^{-7} to 1 (logarithmic scale)
γ	1×10^{-6} to 1 (logarithmic scale)
Kernel	[linear, poly, rbf]
Degree	[2, 3, 4]

2.3 Experiments

2.3.1 Site-specific effects and age prediction on synthetic dataset

The first experiment was designed to validate our core hypothesis: that cross-site synthetic data, generated by matching biological variables, reduces site-specific signatures while preserving age-relevant biological signal.

Data from four sites were pooled and split into training (80%) and testing (20%) sets using a stratified cross-validation scheme (repeated 10 times). All features were scaled using a standard scaler fitted within the cross-validation loop.

Two distinct models were trained on the original training data (without synthetic augmentation):

(1) *Site Prediction:* A Support Vector Classifier (SVC) with a linear kernel ($C = 1$).
(2) *Age Prediction:* A Kernel Ridge Regressor (KRR) following [36].

The KRR hyperparameters were optimized via grid search using nested 10-fold cross-validation (Table 2), minimizing the negative Mean Absolute Error (MAE). The tuning parameters included:

- *Regularization strength* (λ)**:** Controls the penalty on the model weights (L2 norm) to prevent overfitting.
- *Kernel coefficient* (γ)**:** Determines the inverse of the radius of influence of samples for RBF/polynomial kernels.

To evaluate the properties of the interpolation method, we generated synthetic test sets derived from the 20% test partition. For each interpolation weight $\mathrm{SMOTE}_\alpha \in \{0.1, 0.15, \ldots, 0.5\}$, we created a corresponding synthetic dataset. We hypothesized that synthetic samples generated with a SMOTE_α near 0 would retain the characteristics of the *base* site, whereas samples generated with SMOTE_α approaching 0.5 would occupy the "site-neutral" space, showing minimal Effects of Site (EoS).

The trained SVC and KRR models were then evaluated on these synthetic test sets. We used Balanced Accuracy (bACC) for site classification (lower is better for harmonization) and MAE for age prediction (higher is better for biological retention).

2.3.2 Impact of synthetic training data on age prediction

The impact of including synthetic data on age prediction performance was evaluated within a leave-one-site-out scheme. Specifically, data from three of the four sites was used to train a KRR model whose performance was then evaluated on the held-out site. This process was repeated so that each site served as the test site once, ensuring a comprehensive evaluation.

The age prediction models were trained under three distinct scenarios:

- Baseline: We pooled original data from the three training sites to train a model. To ensure consistency the same hyperparameter ranges and search procedure were used as described in previous experiments.
- Synthetic Data: A KRR model was trained exclusively on the synthetic generated data using the three training sites. To generate synthetic data, our method requires a pair of sites: one designated as the *base* and the other as the *target*. Since three training sites were available, synthetic data was generated for all possible *base-target* pairs, ensuring a diverse and representative synthetic dataset. For each new synthetic data point, the interpolation parameter α was randomly selected between 0.3 and 0.5.
- Mixed Data: A KRR model was trained on a combination of original training data and synthetic data generated as described in the previous scenario.
- ComBat. We used a neuroHarmonize model (ComBat-GAM) to adjust the pooled data. As the model requires the covariates at train and test time, age was not used as a covariate.

To investigate the effect of the amount of synthetic data on the model's ability to generalize to unseen test sites, we varied the number of synthetic data points generated. Specifically, we used synthetic datasets of increasing size, where each increase was cumulative.

2.3.3 Disentangling the impact of harmonization and information generation

As the proposed method is a data augmentation process, we aimed to disentangle the relative contributions of two factors, the impact of harmonization and the generation of additional information through synthetic data. To this end, a systematic study was designed where the largest site, eNKI, was used to create two subsets, called A_1 and A_2, with similar age distributions and sizes. The second-largest site, CamCAN, was used as an extra site, called *B*, which was matched in age distribution and size to A_1. From these three sites (A_1, A_2 and *B*), two synthetic datasets were generated. A within-site synthetic dataset (S_1) was created by interpolating between A_1 and A_2. A cross-site synthetic dataset (S_2) was created by interpolating between A_1 (eNKI) and *B* (CamCAN). Additionally, we increased the number of samples in *S* created by augmenting the upsampling number *k*. As before, the interpolation parameter α was randomly sampled between 0.3 and 0.5 for each new synthetic sample. Two KRR models were trained on the generated combinations, $A_1 + A_2 + B + S_1$ and $A_1 + A_2 + B + S_2$. These models were evaluated on the unseen IXI sites (IXI/Guys, IXI/HH, and IXI/IOP), with performance measured using MAE and R^2. To account for randomness in subset generation and synthetic data creation the process was repeated 20 times.

3 Results

3.1 Site-specific effects and age prediction on synthetic dataset

To assess whether the synthetic data retains EoS, we examined how the bACC score for site classification varied with increasing α values. Increasing α led to a constant reduction in bACC (Fig. 1a). At $\alpha = 0.5$, the bACC score was at its lowest (0.4), indicating minimal EoS in the synthetic data. However, the bACC did not reach the chance level (0.25), indicating the presence of some site-specific information, albeit minimal.

We evaluated how well the age information was preserved in the synthetic dataset by analyzing MAE of age prediction (Fig. 1b). The MAE decreased with increasing α, suggesting that the synthetic data better retained age-related signal despite stronger interpolation between two sites. This suggests that interpolating did not remove the meaningful age-related signal.

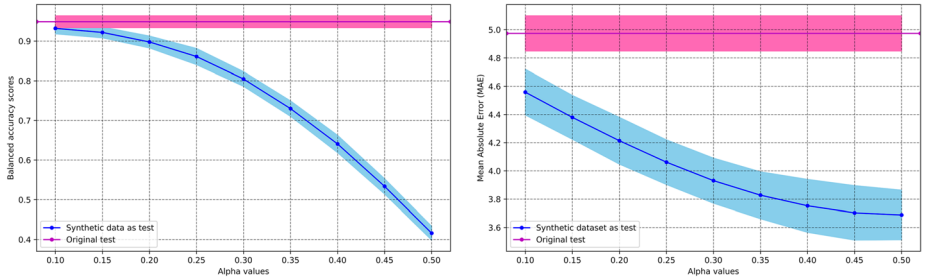
Overall, our findings support our hypothesis that interpolating between demographically matched samples in two datasets reduces site-specific effects while maintaining biological relevance.

3.1.1 Impact of synthetic training data on age prediction

We evaluated the effect on age prediction performance when incorporating synthetic data in the training set in a leave-one-site-out scheme. Regardless of the test site, models trained exclusively on synthetic data initially exhibited high MAE and drastically improved as more synthetic data was used. This result suggests that large synthetic data are needed to cover the original data variance.

The best performance was observed when synthetic data was combined with the original training data. This hybrid approach led to a significant reduction in MAE for the test sites IXI/HH and IXI/Guys (Fig. 2a, b) and yielded comparable performance for eNKI and CamCAN (Fig. 2c, d). These results indicate that combining synthetic data and original data improves the robustness of the models, leading to better generalization on data from a new unseen site.

In addition, we observed that the *control* performance (pink line in Fig. 2a, b) improved with the size of training data, e.g. when using the smallest sites as test sites (IXI/HH and IXI/Guys). Even in the scenarios where the models achieved the best *control* performance, including synthetic data led to improvement. This suggests that a larger training



(a) **Site-effects in the synthetic dataset as alpha varies.** The plot depicts balanced accuracy scores evaluated on the synthetic dataset (blue curve) and original test set (magenta line). The X-axis represents the alpha value, which governs the location of synthetic data generation between each pair of sites.

(b) **Age signal in the synthetic dataset as alpha varies.** The plot depicts mean absolute error (MAE) evaluated on the synthetic dataset (blue curve) and original test set (magenta line). The X-axis represents the alpha value, which governs the location of synthetic data generation between each pair of sites.

Fig. 1 Analysis of site-effects and age signal in the synthetic dataset as alpha varies. **a** Balanced accuracy scores for site prediction, and **b** mean absolute error (MAE) for age prediction. In both plots, the blue curve represents results for the synthetic dataset, while the magenta line represents results for the original test set. The X-axis in both plots represents the alpha value, which controls the interpolation between pairs of sites

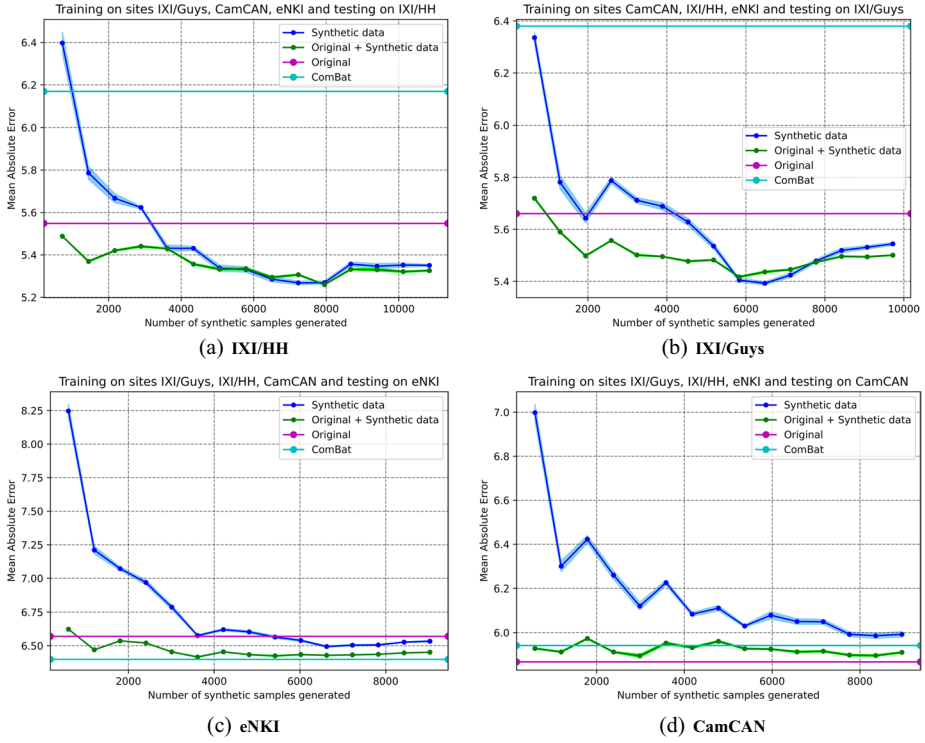


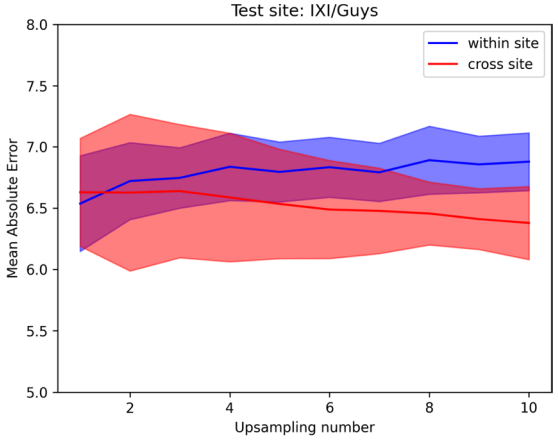
Fig. 2 Impact of synthetic data size and harmonization method on KRR model generalization to unseen test sites. The Y-axis represents the mean absolute error (MAE) in years, and the X-axis represents the number of synthetic samples generated. The Blue curves represent models trained exclusively on synthetic data. The Green curves represent models trained on a combination of original observed data and synthetic data. The Magenta line represents the baseline performance trained only on original observed data (no augmentation). The Cyan line represents the performance when the training data is harmonized using ComBat prior to model training. Each panel displays the results for a specific leave-one-site-out fold: **a** Testing on IXI/HH, **b** Testing on IXI/Guys, **c** Testing on eNKI, and **d** Testing on CamCAN. Shaded regions indicate the standard error across repetitions

set not only benefits model learning but also improves the effectiveness of synthetic data generation by better capturing inter-subject variability.

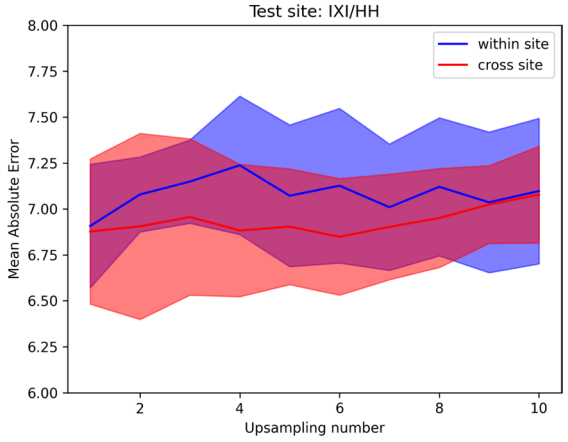
To benchmark the proposed method against standard practices, we compared our approach with ComBat harmonization (Cyan line, Fig. 2). The results reveal that ComBat does not consistently improve generalization performance in this leave-one-site-out regression task. In the IXI/HH (Fig. 2a) and IXI/Guys (Fig. 2b) test sets, ComBat harmonization resulted in a significant increase in Mean Absolute Error (MAE ≈ 6.18 and 6.4 respectively) compared to the un-harmonized "Original" baseline (MAE ≈ 5.55 and 5.65). A similar, though smaller, degradation was observed in the CamCAN cohort (Fig. 2d). The only scenario where ComBat provided a performance benefit over the baseline was the eNKI site (Fig. 2c), where it achieved an MAE of ≈ 6.4, surpassing the original baseline (6.55) and matching the performance of our proposed method.

3.2 Disentangling the impact of harmonization and information generation

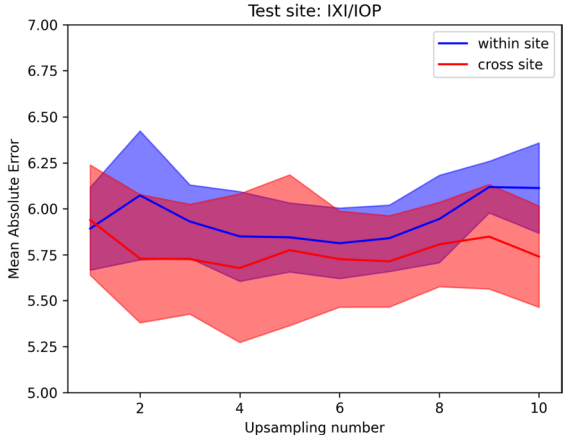
On average, the inclusion of cross-site synthetic data led to a better MAE, compared to the inclusion of within-site synthetic data for all upsampling levels (Fig. 3a–c). For IXI/Guys and IXI/IOP as test datasets, cross-site data generation resulted in superior generalization at the highest up-sampling level of 10 (Fig. 3a, c). For IXI/HH dataset, the difference was negligible (Fig. 3b). These results support the role of the proposed



(a) Test site: IXI/Guys



(b) Test site: IXI/HH



(c) Test site: IXI/IOP

Fig. 3 Impact of synthetic data size on KRR model's ability to generalize to unseen test sites. The Y-axis represents the mean absolute error (MAE), and the X-axis represents the number of synthetic samples generated. The blue curves correspond to models trained exclusively on synthetic data, while the green curves represent training with a combination of original data from IXI/Guys, IXI/HH, and eNKI, along with synthetic data generated from these three sites

harmonization, as the introduction of more diverse and less site-specific data allows the model to develop a broader representation of age-related patterns, reducing overfitting to the characteristics of a particular site. Additionally, cross-site data did not degrade performance in any case, reaffirming its robustness as a strategy for improving model generalization.

4 Discussion

Brain age estimation models hold significant potential for clinical practice, particularly in neurodegenerative disease management, where the "brain age gap" serves as a critical biomarker for early detection. However, the deployment of such models is currently hindered by the lack of cross-site generalizability. MRI data is inherently sensitive to acquisition parameters, resulting in systematic "Effects of Site" (EoS) that obscure biological signals and degrade model performance on unseen scanners. In this work, we proposed a novel harmonization framework that leverages Inter-Site SMOTE to generate synthetic training data. Our core hypothesis posits that interpolating between biologically matched samples (age and gender) from distinct sites creates a synthetic manifold that preserves biological information while averaging out site-specific noise vectors. By training on this "site-neutral" data, models are encouraged to learn robust, invariant features that generalize better to unseen domains.

Through systematic validation, we confirmed that this interpolation strategy effectively decouples site identity from biological signal. Our experiments demonstrated that as the interpolation weight increases, the ability of a classifier to detect the site of origin diminishes, while age prediction accuracy on synthetic data improves. Crucially, this creates a training regime where the model is penalized for relying on site-specific correlations. Consequently, models trained with our augmented dataset consistently outperformed the baseline of simply pooling raw data, achieving lower mean absolute error (MAE) across unseen test sites. This suggests that the synthetic data effectively bridges the domain gap, populating the sparse regions of the feature space between site distributions.

A key finding of this study is the superior performance of our data-centric approach compared to standard statistical harmonization methods like ComBat. While ComBat is the gold standard for retrospective group-level analyses, our results showed that it frequently degraded performance in this predictive machine learning (ML) setting. This failure highlights a fundamental conflict between the assumptions of statistical harmonization and prospective prediction. First, ComBat requires the biological covariate (in our case, age) to be explicitly modeled during the harmonization of the test set to preserve biological variance. However, in a prediction task, age is the unknown target variable; using it for test-set harmonization constitutes data leakage, while omitting it risks removing the very signal of interest [26]. Finally, statistical harmonization is sensitive to sample size imbalances, potentially biasing feature transformations toward the distribution of larger sites (e.g., eNKI) at the expense of smaller cohorts.

Furthermore, we provided empirical evidence that the geometry of the data generation matters more than the mere quantity of samples. By comparing our "Inter-Site" approach against an "Intra-Site" control (where synthetic samples were generated within the same site), we showed that the performance gains are not simply attributable to data augmentation. While Intra-Site SMOTE provided a marginal benefit by densifying

the training distribution, it failed to remove site effects. The significant boost in generalization observed only with Inter-Site interpolation confirms that the mechanism of improvement is indeed the "bridging" of site domains, rather than simple regularization via sample size increase.

A limitation of the current approach lies in the linear nature of the SMOTE interpolation. Brain aging trajectories and scanner-induced variations likely reside on complex, non-linear manifolds in the high-dimensional feature space. While we mitigate this by strictly matching subjects on age and gender–relying on the assumption that the data manifold is locally linear within small neighborhoods–there is a theoretical risk that linearly interpolated samples may represent biologically implausible combinations of brain regional volumes. The observed performance gains may stem partly from the synthetic data acting as a form of "site-aware regularization", smoothing the decision boundary across the domain gap, rather than perfectly reconstructing biological anatomy.

In the present study, synthetic samples were generated by interpolating between subjects matched specifically for age and gender. However, the flexibility of our framework allows for the inclusion of additional covariates, such as ethnicity, race, education level, or clinical status. Including these factors in the matching process could further refine the harmonization by ensuring that the interpolated path between sites remains strictly within specific demographic or biological sub-populations. This would be particularly beneficial for reducing demographic bias and ensuring that the learned features are truly representative of aging rather than confounding demographic variables. It is important to note, however, that adding matching criteria imposes stricter constraints on the data generation process; as the number of covariates increases, the likelihood of finding exact matches across sites decreases (sparsity), which may require larger datasets to implement effectively. Future work will explore high-dimensional matching strategies to accommodate these richer covariate profiles.

While our study demonstrated efficacy using a Kernel Ridge Regression (KRR) model, the proposed framework is model-agnostic. Finally, while the proposed augmentation strategy enhances model robustness, the resulting increase in sample size may impose computational constraints on models with high complexity (like kernel-based methods like KRR due to their cubic complexity ($O(n^3)$), thereby favoring the use of more scalable algorithms (e.g., deep learning or linear models) in extremely large-scale settings. Future work should explore the application of this method to more complex deep learning architectures and investigate the sensitivity of the hyperparameters (interpolation range SMOTE_α and upsampling factor k) across different neuroimaging modalities.

In conclusion, these findings underscore the importance of cross-site synthetic data generation as a powerful tool for developing generalizable biomarkers. By actively reducing site-specific dependencies through targeted interpolation, this approach enhances model robustness without the strict assumptions required by statistical harmonization, offering a promising pathway for the deployment of reliable AI tools in multi-center clinical environments.

5 Conclusion

In summary, this study highlights the potential of Inter-Site synthetic data generated by interpolation to enhance the generalizability of age prediction models in multi-site neuroimaging studies, with implications for early detection and intervention in

neurodegenerative diseases. By leveraging a novel adaptation of SMOTE to generate synthetic data, we showed that interpolating between age- and gender-matched samples from different sites can reduce site-specific biases while preserving biologically relevant information. Crucially, our experiments demonstrated that this data-centered approach outperforms both standard data pooling and traditional statistical harmonization (ComBat), specifically avoiding the pitfalls of test-time data leakage and assumption violations inherent to statistical methods in predictive tasks. Furthermore, we revealed that interpolating *between* sites effectively bridges domain gaps to learn invariant features, whereas generating samples *within* sites yields significantly lower performance gains. Importantly, the proposed method provides an easy and effective way to generate synthetic samples that acts as a site-aware regularization tool, which can be applied to other task domains. By addressing challenges related to data scarcity and site-specific variability, our approach paves the way for more robust and reliable prediction models without relying on complex generative architectures.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s44248-026-00100-7.

Supplementary file 1 (pdf 137 KB)

Author contributions

N.N. and A.A. wrote the first manuscript draft, developed the code, conducted the experiments, and prepared all figures. K.J. and K.P. were responsible for funding acquisition and project administration. All authors contributed to the conceptualization and reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was partly supported by the MODS project funded from the program "Profilbildung 2020" (Grant No. PROFILNRW-2020-107-A), an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia, Helmholtz-AI project BrainAge4AD (ZT-I-PF-5-163), and Helmholtz Portfolio Theme Supercomputing and Modeling for the Human Brain.

Availability of data and materials

Information eXtraction from Images (IXI) dataset is publicly available and can be found in the following URL: https://brain-development.org/ixi-dataset/. The Cambridge Center for Ageing and Neuroscience can be accessed following the instructions provided at: https://opendata.mrc-cbu.cam.ac.uk/access/. Finally, the enhanced Nathan Kline Institute-Rockl and sample (eNKI) can be accessed following the instructions provided at: https://fcon-1000.projects.nitrc.org/indi/enhanced/neurodata.html.

Declarations

Ethics approval and consent to participate

A re-analysis of the anonymized data was approved by the ethics committee of the Heinrich Heine University Düsseldorf (2018-317-RetroDEuA). Participants' consent was acquired in the respective data collection efforts. No new data were acquired in this research; thus, no new participants' consents were collected.

Consent for publication

All authors agreed to submit this manuscript for publication.

Competing interests

The authors declare no competing interests.

Received: 18 October 2025 / Accepted: 23 January 2026

Published online: 10 March 2026

References

1. Feigin VL, Nichols E, Alam T, Bannick MS, Beghi E, Blake N, et al. Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the global burden of disease study 2016. Lancet Neurol. 2019;18(5):459–80.
2. Reitz C, Brayne C, Mayeux R. Epidemiology of Alzheimer disease. Nat Rev Neurol. 2011;7(3):137–52.
3. Eickhoff CR, Hoffstaedter F, Caspers J, Reetz K, Mathys C, Dogan I, et al. Advanced brain ageing in Parkinson's disease is related to disease duration and individual impairment. Brain Commun. 2021;3(3):fcab191.

4. Gao J, Liu J, Yuhang X, Peng D, Wang Z. Brain age prediction using the graph neural network based on resting-state functional MRI in Alzheimer's disease. Front Neurosci. 2023;17:1222751.
5. More S, Antonopoulos G, Hoffstaedter F, Caspers J, Eickhoff SB, Patil KR, et al. Brain-age prediction: a systematic comparison of machine learning workflows. Neuroimage. 2023;270:119947.
6. Karim HT, Aizenstein HJ, Mizuno A, Ly M, Andreescu C, Minjie W, et al. Independent replication of advanced brain age in mild cognitive impairment and dementia: detection of future cognitive dysfunction. Mol Psychiatry. 2022;27(12):5235–43.
7. Chen J, Liu J, Calhoun VD, Arias-Vasquez A, Zwiers MP, Gupta CN, et al. Exploration of scanning effects in multi-site structural MRI studies. J Neurosci Methods. 2014;230:37–50.
8. Huynh KM, Chen G, Ye W, Shen D, Yap P-T. Multi-site harmonization of diffusion MRI data via method of moments. IEEE Trans Med Imaging. 2019;38(7):1599–609.
9. Li H, Smith SM, Gruber S, Lukas SE, Silveri MM, Hill KP, et al. Denoising scanner effects from multimodal MRI data using linked independent component analysis. Neuroimage. 2020;208:116388.
10. Wachinger C, Rieckmann A, Pölsterl S, Initiative ADN. Detect and correct bias in multi-site neuroimaging datasets. Med Image Anal. 2021;67:101879.
11. Bento M, Fantini I, Park J, Rittner L, Frayne R. Deep learning in large and multi-site structural brain MR imaging datasets. Front Neuroinform. 2022;15:805669.
12. Solanes A, Gosling CJ, Fortea L, Ortuño M, Lopez-Soley E, Llufriu S, et al. Removing the effects of the site in brain imaging machine-learning-measurement and extendable benchmark. Neuroimage. 2023;265:119800.
13. Acquitter C, Piram L, Sabatini U, Gilhodes J, Cohen-Jonathan EM, Ken S, et al. Radiomics-based detection of radionecrosis using harmonized multiparametric MRI. Cancers. 2022;14(2):286.
14. Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, et al. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage. 2018;167:104–20.
15. Ingalhalikar M, Shinde S, Karmarkar A, Rajan A, Rangaprakash D, Deshpande G. Functional connectivity-based prediction of autism on site harmonized abide dataset. IEEE Trans Biomed Eng. 2021;68(12):3628–37.
16. Li Y, Ammari S, Balleyguier C, Lassau N, Chouzenoux E. Impact of preprocessing and harmonization methods on the removal of scanner effects in brain MRI radiomic features. Cancers. 2021;13(12):3000.
17. Maikusa N, Zhu Y, Uematsu A, Yamashita A, Saotome K, Okada N, et al. Comparison of traveling-subject and combat harmonization methods for assessing structural brain characteristics. Hum Brain Mapp. 2021;42(16):5278–87.
18. Da-Ano R, Visvikis D, Hatt M. Harmonization strategies for multicenter radiomics investigations. Phys Med Biol. 2020;65(24):24TR02.
19. Hu F, Chen AA, Horng H, Bashyam V, Davatzikos C, Alexander-Bloch A, et al. Image harmonization: a review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. Neuroimage. 2023;274:120125.
20. Fortin J-P, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, et al. Harmonization of multi-site diffusion tensor imaging data. Neuroimage. 2017;161:149–70.
21. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
22. Pomponio R, Erus G, Habes M, Doshi J, Srinivasan D, Mamourian E, et al. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. Neuroimage. 2020;208:116450.
23. Horng H, Singh A, Yousefi B, Cohen EA, Haghighi B, Katz S, et al. Generalized combat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects. Sci Rep. 2022;12(1):4493.
24. Chen AA, Beer JC, Tustison NJ, Cook PA, Shinohara RT, Shou H. Mitigating site effects in covariance for machine learning in neuroimaging data. Hum Brain Mapp. 2022;43(4):1179–95.
25. El-Gazzar A, Thomas RM, van Wingen G. Harmonization techniques for machine learning studies using multi-site functional MRI data. bioRxiv. 2023;1:2023–06.
26. Nieto N, Eickhoff SB, Jung C, Reuter M, Diers K, Kelm M, Lichtenberg A, Raimondo F, Patil KR. Impact of leakage on data harmonization in machine learning pipelines in class imbalance across sites. 2024. arXiv:2410.19643.
27. Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. Adv Neural Inf Process Syst. 2015;28.
28. An L, Chen J, Chen P, Zhang C, He T, Chen C, Zhou JH, Yeo BTT, Lifestyle Study of Aging, Alzheimer's Disease Neuroimaging Initiative, et al. Goal-specific brain MRI harmonization. Neuroimage. 2022;263:119570.
29. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, 2017;2223–2232.
30. Kieselmann JP, Fuller CD, Gurney-Champion OJ, Oelfke U. Cross-modality deep learning: contouring of MRI data from annotated CT data only. Med Phys. 2021;48(4):1673–84.
31. Hawco C, Dickie EW, Herman G, Turner JA, Argyelan M, Malhotra AK, et al. A longitudinal multi-scanner multimodal human neuroimaging dataset. Sci Data. 2022;9(1):332.
32. Chawla NV, Bowyer KW, Hall LO, Philip Kegelmeyer W. Smote: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
33. Taylor JR, Williams N, Cusack R, Auer T, Shafto MA, Dixon M, et al. The Cambridge Centre for ageing and neuroscience (cam-can) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. Neuroimage. 2017;144:262–9.
34. Nooner KB, Colcombe SJ, Tobe RH, Mennes M, Benedict MM, Moreno AL, et al. The NKI-Rockland sample: a model for accelerating the pace of discovery science in psychiatry. Front Neurosci. 2012;6:152.
35. Gaser C, Dahnke R, Thompson PM, Kurth F, Luders E. Alzheimer's disease neuroimaging initiative. Cat—a computational anatomy toolbox for the analysis of structural MRI data. biorxiv, 2022;2022–06.

36. He RT, Kong AJ, Holmes MN, Sabuncu MR, Eickhoff SB, Bzdok D, et al. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. Neuroimage. 2020;206:116276.

Publisher's Note