Neurocomputing 680 (2026) 133146

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Impact of leakage on data harmonization in machine learning pipelines in class imbalance across sites

Nicolás Nieto[a,b,c,*], Simon B. Eickhoff[a,b], Christian Jung[c,d], Martin Reuter[e,f], Kersten Diers[e], for the Alzheimer's Disease Neuroimaging Initiative, Malte Kelm[c,d], Artur Lichtenberg[g], Federico Raimondo[a,b], Kaustubh R. Patil[a,b]

[a] Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany
[b] Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[c] Department of Cardiology, Pulmonology and Vascular Medicine, University Hospital and Medical Faculty, Heinrich-Heine University, Düsseldorf, Germany
[d] Cardiovascular Research Institute Düsseldorf (CARID), Medical Faculty, Heinrich-Heine University, Düsseldorf, Germany
[e] Artificial Intelligence in Medical Imaging, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[f] Department of Radiology, Harvard Medical School, Boston, MA, USA
[g] Department of Cardiac Surgery, University Hospital and Medical Faculty, Heinrich-Heine University, Düsseldorf, Germany

HIGHLIGHTS

- In class imbalance across sites, ComBat-based models require test labels to avoid signal loss, leading to data leakage.
- If no test labels are provided in class imbalance across sites, ComBat-based harmonization removes the signal of interest.
- PrettYharmonize integrates harmonization in ML pipelines in a leakage-free way by eliminating the need for test targets.

ARTICLE INFO

Communicated by C.M. Vong

Keywords:
Data harmonization
ComBat
Data leakage
Machine learning
Medical imaging
Magnetic resonance imaging
Medical AI
Clinical
ICU

ABSTRACT

Due to the cost and complexity of data collection in biomedical domains, it is a common practice to combine data from multiple sites to obtain large datasets required for machine learning. However, undesired site-specific variability presents challenges. Data harmonization aims to address this issue by removing site-specific variance while preserving biologically relevant information. We show that the widely used ComBat-based harmonization improvements are driven by data leakage due to illicit use of target information when class labels are imbalanced across sites, a common scenario in biomedical domains. We propose a novel approach, *PrettYharmonize*, which leverages subtle differences in data harmonized using different pretended target values. Using controlled benchmark datasets and real-world magnetic resonance imaging and clinical ICU data, we demonstrate that our leakage-free *PrettYharmonize* method achieves performance comparable to leakage-prone methods. As such, it is a viable method to integrate ComBat-based methods into machine learning applications.

1. Introduction

Many research fields have greatly benefited from machine learning (ML) approaches. ML relies on large datasets to learn generalizable models, as these datasets help capture robust underlying patterns. This makes combining multiple datasets particularly attractive, especially in domains where obtaining sufficient data from a single location is challenging. However, combining datasets remains a significant challenge, as datasets obtained from different locations, even with similar acquisition parameters, often contain variability unrelated to relevant biological information [1–3].

* Corresponding author at: Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany.
Email address: n.nieto@fz-juelich.de (N. Nieto).

https://doi.org/10.1016/j.neucom.2026.133146
Received 30 April 2025; Received in revised form 22 December 2025; Accepted 24 February 2026
Available online 25 February 2026

This undesired site-related variability may stem from systematic differences, which can be corrected, or random variations, which cannot be modeled or corrected. The systematic variability, known as Effects of Site (EoS), is prevalent in many biomedical domains and can lead to biased results if not properly addressed [4]. For example, clinical data are influenced by the acquisition site, as different hospitals use varying laboratory equipment, procedures, and criteria. Similarly, medical imaging data are affected by factors such as acquisition protocols, scanner drifts, and even the time of day [4,5]. Magnetic Resonance Imaging (MRI) is particularly susceptible to EoS, as variability can arise from differences in magnetic field strength, room temperature fluctuations, or electromagnetic noise—even when using scanners from the same manufacturer with identical parameters [6,7]. To address this issue, numerous data harmonization methods have been proposed [5,8,9]. Data harmonization is typically employed as a preprocessing step to generate *site-effect-free* data while preserving biological information, resulting in harmonized data that can enhance subsequent statistical and ML analyses [10–14].

ComBat-based harmonization is extensively used in several domains. Originally proposed to correct batch differences in genomic data [15], ComBat was later adapted to other domains, including MRI data [10,16]. ComBat employs a Bayesian approach to estimate additive (location) and multiplicative (scale) corrections for each feature across sites. In addition to removing site effects, ComBat can preserve the variance of biologically relevant variables when provided as *covariates*. The ComBat-GAM extension was developed to account for nonlinear covariate effects, and its associated "neuroHarmonize" software has been widely adopted [17]. Although concerns have been raised that the assumptions of ComBat, originally designed for genomic data, may not always hold for other domains [18], numerous empirical studies have demonstrated its effectiveness [5,8,19,20].

Some studies using ComBat within ML pipelines have failed to separate training and test data, incorrectly harmonizing the entire dataset before downstream analysis [10,11,16,21–24,57]. While valid for statistical analysis, this approach is inconsistent with ML principles, where data leakage can lead to invalid models and misleading interpretations if training and test data are not properly separated [25–27]. Failing to maintain this separation during harmonization can result in overly optimistic generalization performance estimates, particularly in cross-validation settings [28]. To address this, techniques have been developed to integrate ComBat into ML pipelines as a preprocessing step, where its parameters are learned on the training data and then applied to the training and unseen test data [17,28,30].

Further challenges arise due to the violation of a critical ComBat assumption–all variance not shared across sites is considered unwanted site-related variance. This becomes particularly problematic when biologically relevant variance is associated with the sites, such as in a diagnostic classification task where one site predominantly includes patients and another healthy controls [29], leading to class imbalance across sites. ComBat may eliminate variance related to the target in such cases of site-target dependence, resulting in a harmonized dataset that yields fewer or null findings in subsequent analyses. To mitigate this issue, the target variable is often included as a covariate in ComBat, ensuring its variance is preserved during harmonization. While preserving relevant variance is not inherently problematic, specifying the target as a covariate introduces a form of data leakage, as the target values of the test set are required to apply the harmonization model but are unavailable in real-world scenarios [25]. However, this approach is still commonly used in the literature.

In this work, we aimed to empirically demonstrate this limitation of ComBat-based harmonization in site-target dependence scenarios. To this end, we conducted controlled experiments for age regression and sex classification using MRI data from healthy individuals. Additionally, we performed two clinically relevant tasks: dementia and mild cognitive impairment (MCI) detection using MRI data, and hospital discharge prediction for septic patients using arterial blood gases, obtained from intensive care unit (ICU) data. To systematically evaluate the impact of site-target association, all experiments were conducted under both site-target dependence and independence conditions. Specifically, we investigated the performance of different harmonization schemes, both with and without explicitly preserving target variance.

Finally, to address the data leakage issue in ML pipelines while combining data from multiple sites, we propose a novel method called "PRETended Target Y harmonize" (*PrettYharmonize*). Using a stacking ensemble architecture [31], *PrettYharmonize* learns subtle differences in data harmonized using ComBat-based methods while varying the target values. It avoids data leakage by employing pretended target values for the test data. We validated our method using benchmark datasets [4] and demonstrated that *PrettYharmonize* performs competitively compared to other harmonization schemes in both site-target dependence and independence scenarios. We provide a comprehensive comparison of no-harmonization, leakage, and no-leakage harmonization schemes on real-world data. The corresponding Python package of the proposed method is publicly available at https://github.com/juaml/PrettYharmonize and also implemented as part of https://github.com/N-Nieto/UniHarmony. The code to replicate our results is also publicly available at https://github.com/juaml/harmonize_project.

2. Methods

2.1. PrettYharmonize

PrettYharmonize is based on the rationale that the output of covariate-adjusted harmonization methods, such as ComBat, is functionally dependent on the target values provided during the adjustment process. Specifically, a sample harmonized using its true label preserves target-relevant information essential for downstream tasks. Conversely, harmonizing a sample under an incorrect label assumption fundamentally alters the feature-target relationship. Consequently, the harmonized representation of a sample varies systematically based on the target value utilized as a covariate. *PrettYharmonize* exploits this phenomenon through a process we term *target pretending* (which is based on "imputing" target values), wherein samples are harmonized multiple times under exhaustive target assumptions. Crucially, this approach facilitates harmonization without requiring labels for test data, thereby preventing data leakage by design.

We start with a training dataset $\mathbf{D}_{train}$ which is composed of feature matrix $\mathbf{X}_{train}[N_{train} \times Features]$ and target values $\mathbf{y}_{train}[N_{train} \times 1]$. We define K as the number of unique target states. For a classification task, K equals the number of classes (e.g., $K = 2$ for binary classification), while for a regression task, the target range is sampled uniformly in K values.

The training algorithm of *PrettYharmonize* employs a stacking ensemble involving a primary predictive model and a secondary stacking model (Fig. 1). The process initiates with an internal cross-validation (Inner CV) procedure (Step 1), where $\mathbf{D}_{train}$ is partitioned into an inner training set $(\mathbf{X}_{in}, y_{in})$ and a validation set $(\mathbf{X}_{val}, y_{val})$.

In the second step, a Harmonization Model ($\mathcal{HM}$) is fitted on $\mathbf{X}_{in}$ preserving the target values $\mathbf{y}_{in}$ as a covariate (Step 2). Subsequently, $\mathcal{HM}$ is applied to harmonize the inner training data utilizing the actual target values, yielding $\tilde{\mathbf{X}}_{in}$ (Step 3). Following harmonization, a Predictive Model ($\mathcal{PM}$), denoted as $f : \mathbf{X} \rightarrow \hat{\mathbf{y}}$, is fitted on the harmonized inner training data $\tilde{\mathbf{X}}_{in}$ to predict the true targets $\mathbf{y}_{in}$ (Step 4).

A critical step occurs in the processing of the validation set. Since target labels are unavailable during the inference phase, we treat the validation targets as missing and impute them. Specifically, $\mathbf{X}_{val}$ is harmonized multiple times by *pretending* the targets correspond to each possible target value $k \in K$ (Step 5). For the binary classification task illustrated in Fig. 1, this yields two distinct harmonized versions of the validation set: $\tilde{\mathbf{X}}_{val}^{(0)}$ (assuming class 0) and $\tilde{\mathbf{X}}_{val}^{(1)}$ (assuming class 1).

The trained $\mathcal{PM}$ is then used to generate predictions for the validation samples across all harmonized versions. For every $k \in K$, we compute $\hat{\mathbf{y}}^{(k)} = \mathcal{PM}(\tilde{\mathbf{X}}_{val}^{(k)})$ (Step 6). This results in K vectors of

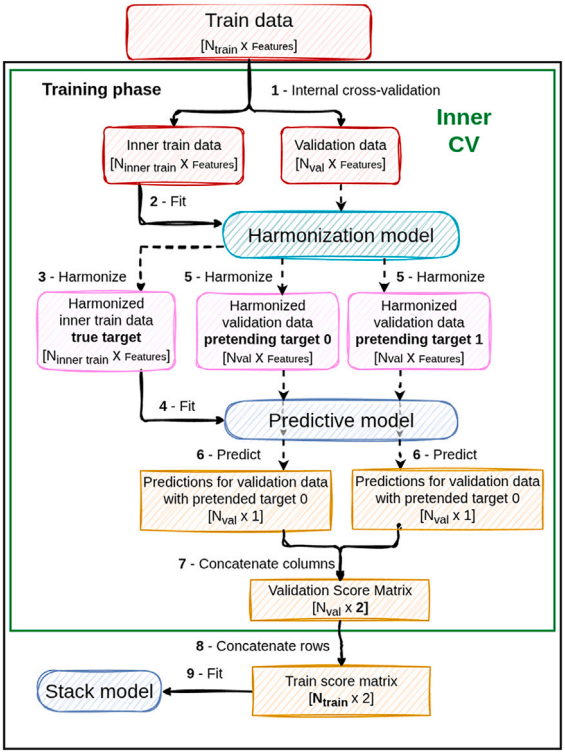


Fig. 1. The workflow illustrates the procedure for a binary classification problem ($K = 2$). Data dimensions are provided in brackets. Red and pink rounded boxes represent unharmonized ($\mathbf{X}$) and harmonized ($\tilde{\mathbf{X}}$) data, respectively. Cyan and blue ellipses represent the Harmonization Model ($\mathcal{HM}$) and Predictive Model ($\mathcal{PM}$), respectively. Yellow rectangles represent model predictions and score matrices. Solid arrows represent data flow used to fit the models, while dashed arrows represent data transformed by the models. Step 1: The training data ($\mathbf{D}_{train}$) is split via Internal Cross-Validation (Inner CV) into inner training ($\mathbf{X}_{in}$) and validation sets ($\mathbf{X}_{val}$). Steps 2-4: An $\mathcal{HM}$ is fitted on $\mathbf{X}_{in}$ and applied to harmonize it ($\tilde{\mathbf{X}}_{in}$). A $\mathcal{PM}$ is then trained on this harmonized data to predict the true targets. Steps 5-7: The validation data $\mathbf{X}_{val}$ is harmonized multiple times by pretending different target values (e.g., target 0 and target 1), creating multiple harmonized versions ($\tilde{\mathbf{X}}_{val}^{(k)}$). The trained $\mathcal{PM}$ generates predictions for each version, which are concatenated column-wise to create the Validation Score Matrix. Steps 8-9: Once the Inner CV is complete, the validation score matrices are concatenated row-wise to form the Train Score Matrix ($\mathbf{S}_{train}$). Finally, a Stack Model ($\mathcal{F}_{stack}$) is fitted on $\mathbf{S}_{train}$ to predict the true target $\mathbf{y}_{train}$.

predictions, each of dimension $N_{val} \times 1$. By concatenating these prediction vectors, a *Validation Score Matrix* of dimensions $N_{val} \times K$ is constructed (Step 7). This matrix encapsulates the score (or likelihood) of a sample belonging to a class, given that the data was harmonized assuming it belongs to that specific class.

This procedure is repeated until all folds of the Inner CV are processed. The resulting *Validation Score Matrices* are concatenated row-wise to form the *Train Score Matrix*, denoted as $\mathbf{S}_{train}$ $[N_{train} \times K]$ (Step 8). A meta-learner, the Stack Model ($\mathcal{F}_{stack}$), denoted as $g : \mathbf{S} \rightarrow \hat{\mathbf{y}}$, is then fitted on $\mathbf{S}_{train}$ to predict the actual labels $\mathbf{y}_{train}$ (Step 9). By learning the relationship between the predictions derived from different harmonization assumptions and the true labels, the *Stack Model* effectively serves as an ensemble combining the conditional predictions of the *Predictive Model*. Finally, for deployment on test data, $\mathcal{HM}$ and $\mathcal{PM}$ are re-trained on the full $\mathbf{D}_{train}$.

In summary, *PrettYharmonize* is a two-level stacking ensemble model where the first level predictions on target-pretended harmonized data obtained using the *Predictive* model are combined by the *Stack* model (Fig. 1).

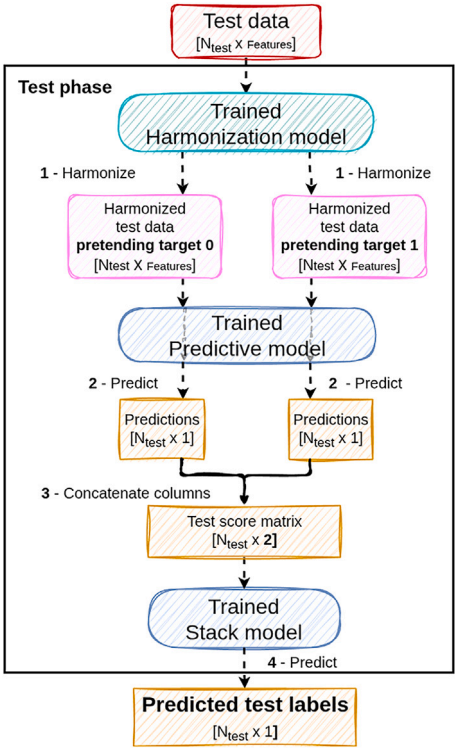


Fig. 2. PrettYharmonize test workflow. The workflow illustrates the inference procedure for a binary classification problem ($K = 2$). Data dimensions are provided in brackets. Red and pink rounded boxes represent unharmonized ($\mathbf{X}$) and harmonized ($\tilde{\mathbf{X}}$) data, respectively. Cyan and blue ellipses represent the trained Harmonization Model ($\mathcal{HM}$) and Predictive Model ($\mathcal{PM}$), while the solid blue shape represents the Stack Model ($\mathcal{F}_{stack}$). Yellow rectangles represent predictions and matrices. Step 1: The unlabelled test data $\mathbf{X}_{test}$ is harmonized by the trained $\mathcal{HM}$ multiple times, pretending all possible target labels (e.g., assuming class 0 and class 1). This yields K harmonized representations ($\tilde{\mathbf{X}}_{test}^{(0)}, \tilde{\mathbf{X}}_{test}^{(1)}$). Steps 2-3: The trained $\mathcal{PM}$ generates predictions for each of these harmonized datasets. These predictions are concatenated column-wise to form the Test Score Matrix ($\mathbf{S}_{test}$). Step 4: The trained $\mathcal{F}_{stack}$ takes $\mathbf{S}_{test}$ as input and generates the final Predicted test labels ($\hat{\mathbf{y}}_{test}$).

The inference workflow for new test data is illustrated in Fig. 2. The raw test data $\mathbf{X}_{test}$ $[N_{test} \times \text{Features}]$ is harmonized using the trained $\mathcal{HM}$ by imputing all possible target values as covariates. This generates K harmonized representations (e.g., $\tilde{\mathbf{X}}_{test}^{(0)}$ and $\tilde{\mathbf{X}}_{test}^{(1)}$ for binary problems) (Step 1). The trained $\mathcal{PM}$ generates predictions for each harmonized representation (Step 2). These predictions are concatenated to form the test score matrix $\mathbf{S}_{test}$ $[N_{test} \times K]$) (Step 3). Finally, the trained $\mathcal{F}_{stack}$ processes this matrix ($\hat{\mathbf{y}}_{test} = g(\mathbf{S}_{test})$) to output the predicted test labels $\hat{\mathbf{y}}_{test}$ $[N_{test} \times 1]$ (Step 4).

Of note, *PrettYharmonize* currently only uses the harmonized data internally and does not create a harmonized dataset as a final outcome, though the data harmonized using the predicted class can be extracted for inspection. While the proposed method allows for the selection of different harmonization backbones, in the experiments presented in this paper, neuroHarmonize [17] was used as the Harmonization Model.

2.2. Harmonization schemes

To benchmark the proposed *PrettYharmonize* model, which internally performs leakage-free data harmonization, five other harmonization schemes were evaluated. By comparing these schemes, we aimed to evaluate the trade-offs between preserving biological signals, avoiding data leakage, and maintaining model performance.

The first scheme was *Whole Data Harmonization* (*WDH*), which involves training a neuroHarmonize model on the pooled data from all sites to generate a harmonized dataset before splitting the data into training and test folds. This approach inherently leads to data leakage, as the test data is included in the harmonization training process [26,28]. Another scheme, *Test Target Leakage* (*TTL*), trains a neuroHarmonize model on the training data while explicitly retaining target variance by providing the target labels as covariates. Although this scheme does not use the test target to train the harmonization model, it requires test labels to transform (harmonize) the test data, resulting in another type of data leakage. A third scheme, *No Target*, trains a neuroHarmonize model on the training data without explicitly retaining target variance. As a result, test labels are not used for transforming (harmonizing) the test set, avoiding leakage, but potentially removing biologically relevant information. An *Unharmonized* approach was implemented as a baseline, where the original data were pooled and used without any harmonization.

We would like to point out that, while *Whole Data Harmonization* and *Test Target Leakage* are implemented for the sake of comparison, only the "*No Target*" harmonization scheme does not show any leakage; thus, leading to a valid comparison with the proposed *PrettYharmonize* method.

2.3. Forcing site-target (in)dependence

To systematically evaluate the impact of harmonization frameworks and the impact of confounding factors, we curated experimental scenarios that enforce specific dependencies between the acquisition site and the target variable. We defined two primary conditions: *Site-Target Dependence* and *Site-Target Independence*.

In the **Site-Target Dependence** condition, we generated scenarios where the site variable acts as a confounder. This was achieved by skewing the class balance within each site. For instance, in a binary classification task, Site A would predominantly contain samples from Class 0, while Site B would predominantly contain samples from Class 1. We hypothesize that in this scenario, standard ComBat-based methods will inadvertently suppress relevant biological variance, as the target signal is intrinsically entangled with site-specific effects. Consequently, explicit preservation of the target as a covariate becomes critical. Furthermore, we expect machine learning (ML) models trained on unharmonized data in this regime to fraudulently inflate performance metrics by learning spurious correlations between the site identifier and the target, rather than learning true biological features.

Conversely, in the **Site-Target Independence** condition, we balanced class proportions or target distributions within all the sites, enforcing independence between the site and the target. In this scenario, we hypothesize that harmonization will effectively remove site-related noise without compromising target-related information, even absent explicit target protection, as the biological variance is distributed across the sites. Here, ML models trained on unharmonized data are expected to show no performance advantage derived from site effects, as the site variable provides no predictive information regarding the target.

2.4. Data and experimental setup

To evaluate the harmonization capabilities of the harmonization scheme using the pretended target of *PrettYharmonize*, and to assess the performance of the other five harmonization schemes proposed, a total of eight datasets were utilized in the experiments. First, we benchmarked the proposed method on the datasets proposed in [4] (Called MAREoS dataset in the following), which are specifically designed to evaluate harmonization methods. Using five MRI datasets (AOMIC-ID1000, eNKI, CamCAN, SALD, and 1000Brains), containing data from healthy control participants, an age regression and sex classification tasks were performed. Additionally, also using MRI data, the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [32], which includes data collected from multiple sites, was used to detect separate healthy controls from cognitive impairment (MCI) or dementia patients.

Finally, the eICU [33–35] dataset was used to classify hospital discharge (Expired or Alive) in septic patients, which is a known and important problem in the literature [36–40]. In this experiment, no brain features were used, as the model relies solely on arterial blood gases (ABG) features.

For all the experiments, a 5-times repeated 5-fold stratified cross-validation was used. The only exception was the experiments with the MAREoS dataset, where the folds (10) are already provided by the authors, to improve reproducibility.

For the classification problems, the Area Under the Receiver Operating Characteristic Curve (AUC), balanced accuracy (bACC), and F1 score were calculated on the test sets. For the age regression problem, the Mean Absolute Error (MAE), coefficient of determination (R^2), and age bias (Pearson's correlation between the true age and the difference between the predicted and true age) were calculated on the test sets.

In the following, the description of the particular dataset and the ML model used in the corresponding experiments is provided, along with the sampling strategies to obtain the site-target (in)dependence.

2.4.1. MAREoS dataset

To evaluate the harmonization capabilities of the harmonization scheme using the pretended target of *PrettYharmonize*, a dataset specifically designed to benchmark data harmonization methods was used [4]. These datasets simulate eighteen MRI features, including cortical thickness, cortical surface area, and subcortical volumes, across eight internal datasets. Among these, four datasets contain a "True" signal, while the remaining four contain only an "EoS" signal related to a binary target. The EoS signal is designed to test whether data harmonization methods can remove it, thereby preventing fraudulent classification performance. For each type of signal (True and EoS), two variations are provided: "Simple" and "Interaction" datasets, representing linear and non-linear relationships between the features and the target, respectively. This makes a total of 8 datasets (True and EoS, both with 2 Simple and 2 Interaction types of signals). Each of these eight datasets comprises 1000 samples simulated from eight different sites. The datasets are provided as 10 train and test fold pairs. For the dataset containing only the EoS, the methods should be able to remove this effect, and the classification performance should be at the chance level, i.e., balanced accuracy (bACC) of 50%. On the other hand, in the dataset with only the True signal, the harmonization models should not degrade the signal, and the bACC is expected to be a high value (bACC $\approx$ 80%).

In these experiments, a Random Forest model (RF) [41], with default sklearn parameters, was used as the Predictive Model, and a Logistic Regression (LG) [42] was used as the Stack Model. The same RF model was used to train a model with the original data to obtain a classification baseline for each dataset (Baseline model).

2.4.2. MRI data

To empirically compare different harmonization schemes with and without site-target dependence, age regression, and sex classification were performed using MRI data. These targets were used as they are highly reliable and can be easily obtained. For all T1-weighted MR images, Voxel-Based Morphometry was performed using CAT12.8 [43] to obtain modulated gray matter (GM) volume, which was then linearly resampled to 8x8x8 mm3 voxels, resulting in 3747 voxels that were used as features. Five datasets were used: Amsterdam Open MRI Collection (AOMIC-ID1000) [44], The Enhanced Nathan Kline Institute (eNKI) [45], Cambridge Centre for Ageing Neuroscience (CamCAN) [46], 1000Brains [47], and the Southwest University Adult Lifespan Dataset (SALD) [48]. These datasets were selected as the data within each dataset was acquired only at one site, thus avoiding additional confounding. The demographic information of these datasets is presented in Table 1.

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

Dataset Name	N Images	Mean Age	Std Age	Min Age	Max Age	% Female
AOMIC-ID1000	928	22.85	1.71	19	26	52%
eNKI	818	46.90	17.73	19	85	65%
CamCAN	651	54.27	18.59	18	88	50%
1000Brains	1144	61.84	12.39	21	85	55%
SALD	494	45.18	17.44	19	80	62%

2.4.2.1 Age regression. For the age regression problems, Relevance Vector Regression [49] with a polynomial kernel of degree 1 (RVR) was used as a Predictive and Stack model for *PrettYharmonize*. This specific model was selected, as it showed the best performance in the age regression task [50]. The RVR model was also used in the other harmonization schemes.

Forced site-target dependence. To induce a strong correlation between age and acquisition site, four datasets were subsampled to cover distinct, non-overlapping age intervals: AOMIC-ID1000 ($N = 118$, age range: 19–26), eNKI ($N = 118$, age range: 27-40), CamCAN ($N = 118$, age range: 41–60), and 1000Brains ($N = 118$, age range: 61–79). Subsampling was constrained to ensure sex-balance within each site; thus, no sex-related signal can be associated with the site or age (Supplementary Table 1).

Forced site-target independence. To break the correlation between site and target, we selected three datasets capable of covering a broad, overlapping age spectrum: CamCAN ($N = 288$, range: 19–77), eNKI ($N = 300$, range: 19–77), and SALD ($N = 200$, range: 19–77). AOMIC and 1000Brains were excluded due to their restricted age coverage (primarily young and elderly cohorts, respectively). Each dataset was balanced for both sex and age. This balance was achieved by stratifying the data into 10 equally distributed age bins (spanning the minimum to maximum age of each dataset) and retaining an equal number of male and female subjects per bin (Supplementary Table 2).

2.4.2.2 Sex classification. For the sex classification, an LG, with default sklearn hyperparameters, was used as a Predictive and a Stack model. The same LG model was used for the rest of the harmonization schemes.

Forced site-target dependence. For the binary sex classification task, the eNKI ($N = 295$, age range: 19–84) and CamCAN ($N = 295$, age range: 18–88) datasets were selected due to their broad and comparable age distributions. Dependence was forced by manipulating the sex percentages: the female proportion was set to 5% in eNKI and 95% in CamCAN (Supplementary Table 3), creating a near-complete separation of class by site. The total sample size was balanced across sites, and the age distributions remained fully overlapping.

Forced site-target independence. The dataset constructed for the age regression independence scenario was reused for sex classification. This dataset is inherently suitable as it was explicitly balanced for both age and sex across all sites (Supplementary Table 2).

2.5. Dementia and mild cognitive impairment classification

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

An LG model with default sklearn hyperparameters was used as a Predictive and a Stack model for PrettYharmonize. The same LG model was used for the rest of the harmonization schemes.

Forced site-target dependence. Using the ADNI dataset, we simulated site-dependency by inducing class imbalances. From one acquisition site, we randomly selected 100 patients (Dementia/MCI) and 10 controls; conversely, from a second site, we selected 100 controls and 10 patients (Supplementary Table 4). All MRI scans were processed using FreeSurfer [51], extracting cortical thickness measures from 74 cerebral and subcortical structures as input features.

Forced site-target independence. Using the same feature set, we constructed a balanced scenario. From the largest available site in the ADNI dataset, 126 patients and 126 controls were randomly sampled. From the second-largest site, 56 patients and 56 controls were sampled (Supplementary Table 5). It is notable that while the class probability is balanced within each site, the total sample size (N) differs between sites ($N = 252$ vs. $N = 112$).

2.6. Discharge status prediction of septic patients

The eICU [33–35] dataset was used for the experiments, which contains 200,859 ICU stays from 139,367 patients in 208 different ICUs across the United States. We use a well-known problem of classified hospital discharge (Expired or Alive), in septic patients [36–40]. The approach described in [55] was followed for selecting the features and extracting the patient cohort. The features used were arterial blood gases: paO2, paCO2, pH, base excess, Hgb, glucose, bicarbonate, and lactate. No brain-derived features were used in this experiment. After the patients' selection, a final dataset of 496 Expired and 3021 Alive patients was retained. From this filtered dataset, we removed those sites with fewer than 50 patients, retaining 20 final sites from where the site-target (in)dependence scenarios were created.

LG was used as a Predictive and a Stack model for *PrettYharmonize*, and as a classification model for the other harmonization schemes.

Forced site-target dependence. This experiment utilized data from 20 clinical sites to predict discharge status ("Alive" vs. "Expired"). To create extreme dependence, we manipulated the data such that 10 sites contained exclusively "Alive" patients (with the exception of a single "Expired" outlier per site to permit variance estimation). Conversely, the remaining 10 sites contained exclusively "Expired" patients (with a single "Alive" outlier per site). This resulted in a total cohort of 249 "Expired" and 666 "Alive" patients. Unlike previous experiments, both the total sample size and the class proportions varied significantly across sites (Supplementary Table 6).

Forced site-target independence. In this scenario, we enforced a balanced class distribution (50% "Alive", 50% "Expired") within each of the previously selected 20 sites. The final dataset comprised 324 "Expired" and 324 "Alive" patients (Supplementary Table 7). While the class ratio was identical across sites, the total volume of patients varied per site, ranging from 8 to 128 subjects per site, reflecting realistic variations in site throughput while maintaining independence between site and outcome.

3. Results

3.1. PrettYharmonize validation on MAREoS dataset

On datasets containing the True signal (no EoS), a Baseline model (Random Forest) trained on unharmonized data achieved a balanced accuracy (bACC) of approximately 80%, as expected. However, the same model also achieved a bACC close to 80% on datasets containing only EoS signal (no True signal), fraudulently leveraging the EoS signal for classification (Table 2).

Table 2
PrettYharmonize and Baseline (RF model without harmonization) performance on the MAREoS dataset (bACC [%]: mean of 10 folds). True: Only "True" signal is presented, and no Effect of Site. EoS: Only the Effect of Site and no True signal is presented. Simple: Simulated linear relationship between features and target. Interaction: non-linear relationship between features and target. 1 and 2 represent different simulated signals.

Dataset Name	Baseline	PrettYharmonize	Expected	Difference
True Simple 1	72.86	72.07	As Baseline	0.79
True Simple 2	82.72	82.86	As Baseline	0.06
True Interaction 1	79.43	79.46	As Baseline	0.03
True Interaction 2	72.23	70.72	As Baseline	1.51
EoS Simple 1	76.11	54.18	Chance (50)	4.18
EoS Simple 2	75.35	52.35	Chance (50)	2.35
EoS Interaction 1	77.48	56.20	Chance (50)	6.2
EoS Interaction 2	82.79	58.81	Chance (50)	8.81

Table 3
Comparison of performance metrics across different harmonization schemes.

Harmonization Scheme	Site-target dependence			Site-target independence		
	MAE	R^2	Age Bias	MAE	R^2	Age Bias
Unharmonized	6.20	0.81	−0.43	6.314	0.785	−0.341
PrettYharmonize	4.12	0.919	−0.26	6.306	0.769	−0.423
WDH	3.82	0.925	−0.32	6.034	0.803	−0.366
TTL	4.28	0.912	−0.23	6.153	0.775	−0.319
No Target	15.93	−0.007	−0.998	6.036	0.790	−0.361

PrettYharmonize successfully removed the EoS signal in all datasets containing only EoS. Furthermore, in datasets where only the True signal was present, the method preserved the real signal while aiming to remove EoS (Table 2). To ensure robustness, we repeated this analysis using three additional Predictive models: Gaussian Process Classifier (GP), Support Vector Machine with a Radial basis kernel (SVM), and Least Absolute Shrinkage and Selection Operator (LASSO). These models yielded similar results (Supplementary Tables 8, 9, and 10).

3.2. Age prediction

3.2.1. Forced site-target dependence

The *Unharmonized* method achieved a Mean Absolute Error (MAE) of 6.20 (Table 3), which falls within the expected range according to the literature [50]. Predictions using the *WDH* and *TTL* schemes showed an improvement in performance of approximately 2 years compared to the *Unharmonized* scheme (Table 3).

As expected, *No Target* removed the age-related signal in the features, preventing the ML model from learning the feature-target relationship and resulting in inaccurate predictions. Specifically, the model predicted the mean population age for all individuals, leading to overestimations in the AOMIC and eNKI datasets and underestimations in the CamCAN and 1000Brains datasets (Fig. 3a, and Supplementary Fig. 5).

PrettYharmonize, on the other hand, achieved better predictions, on average, compared to both the *Unharmonized* and *No Target* methods, improving the MAE, R^2, and age bias without inducing data leakage (Table 3). Notably, *PrettYharmonize*'s performance was comparable to the two leakage-prone methods (*WDH* and *TTL*), but WDH showed a slightly better average performance. Detailed individual predictions and site-specific MAE are presented in Supplementary Figs. 1–5 (one for each harmonization scheme).

3.2.2. Forced site-target independence

The model that used unharmonized data achieved an MAE of 6.314, similar to the performance in the site-target dependence scenario (Table 3).

In this scenario, the average performance was similar across all harmonization schemes, including the *No Target* scheme (Table 3). This result suggests that the improvement in the signal-to-noise ratio made by removing the EoS signal was not sufficient to boost ML performance. Consistent with our hypothesis, the *No Target* scheme did not remove biologically relevant information, as this variance was shared across all sites, and the ML model showed a comparable performance to the other benchmarked schemes.

3.3. Sex classification

3.3.1. Forced site-target dependence

The *Unharmonized* scheme achieved a high performance (AUC = 0.97), consistent with results reported in the literature [52].

The harmonization schemes that allow data leakage (*WDH* and *TTL*) and *PrettYharmonize* did not show any improvement over the *Unharmonized* scheme (Table 4). This is likely due to the presence of a strong sex-related signal in the features, which enables high performance (AUC = 0.97) even without harmonization. Consistent with the findings from the age regression experiment, the *No Target* scheme removed sex-related information, resulting in a significant drop in classification performance (Table 4).

3.3.2. Forced site-target independence

For the site-target independence, the *Unharmonized* scheme achieved a lower performance (AUC = 0.918) compared to the site-target dependence scenario (Table 4). This is expected, as the ML model can not exploit the EoS signal to fraudulently improve its performance. None of the harmonization schemes demonstrated improved classification performance relative to the *Unharmonized* model (Table 4). Again, consistent with the age regression experiment, the *No Target* scheme did not eliminate target-related variance during harmonization, leading to performance similar to the other schemes.

3.4. Dementia and mild cognitive impairment classification

3.4.1. Forced site-target dependence

The *Unharmonized* method achieved an AUC of 0.81, consistent with findings reported in the literature [54]. *PrettYharmonize*, *WDH*, and *TTL* showed a slight improvement in classification performance compared to the *Unharmonized* method (Table 5). As observed in previous site-target dependence experiments, the *No Target* scheme removed biologically relevant information, significantly impairing the ML model's performance (Table 5).

3.4.2. Forced site-target independence

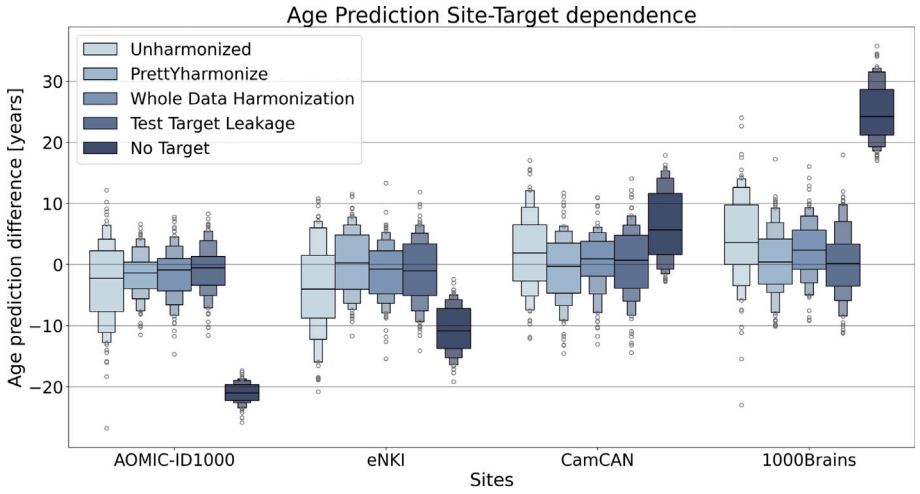
All benchmarked schemes achieved similar classification performance across all metrics (Table 5). Notably, a consistent performance drop was observed across all schemes compared to the site-target dependence scenario.

3.5. Discharge status prediction of septic patients

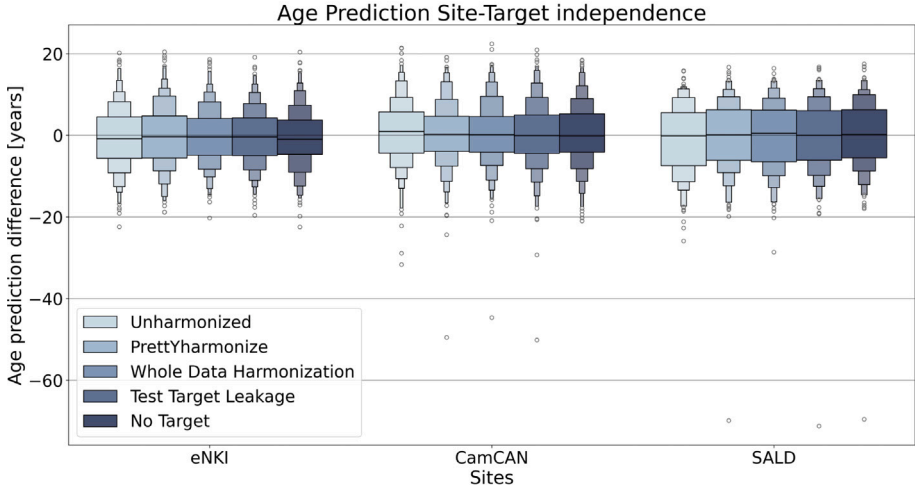
3.5.1. Forced site-target dependence

The *Unharmonized* scheme achieved an AUC of 0.76, slightly lower than values previously reported in the literature [55]. This difference is expected, as fewer patients were used in our experiments compared to the referenced study.

PrettYharmonize demonstrated an improvement in AUC performance compared to all benchmarked schemes, achieving an AUC of 0.86. In contrast, and consistent with the previous site-target dependence scenarios, the *No Target* scheme removed nearly all relevant information, resulting in performance close to chance (bACC = 51.66) (Table 6).



(a) Age regression on site-target dependence scenario.



(b) Age regression on site-target independence scenario

Fig. 3. Impact of Site-Target Dependence on Age Prediction Bias. The figure displays Letter-value plots, also known as boxen plots, representing the distribution of age prediction errors across different harmonization schemes. This representation was chosen as it is designed for skewed distributions [53]. **(a) Forced Site-Target Dependence:** In this scenario, datasets (AOMIC-ID1000, eNKI, CamCAN, 1000Brains) were subsampled to possess disjoint, site-specific age ranges. The "No Target" harmonization scheme (darkest blue), which corrects for site effects without preserving the target, severely jeopardizes the models' performance by removing age-related biological signal confounded with the site. In contrast, *PrettYharmonize* (light blue) achieved error distributions comparable to the data leakage schemes (WDH and TTL), and improved the performance compared to the "Unharmonized" baseline. **(b) Forced Site-Target Independence:** In this scenario, datasets (eNKI, CamCAN, SALD) were balanced to ensure fully overlapping age distributions. Consequently, site and target variables are not correlated, and all harmonization schemes, including "No Target", exhibit comparable performance. This confirms that target preservation strategies are specifically critical when site and target distributions are correlated.

3.5.2. Forced site-target independence

For the site-target independence scenario, an equal number of "Alive" and "Expired" patients were retained across all 20 sites. All methods achieved similar classification performance across all metrics. The *Unharmonized* method obtained a slightly lower AUC (0.72) compared to the site-target dependence scenario (0.76) (Table 6).

PrettYharmonize and the leakage-prone schemes (*WDH* and *TTL*) showed a drop in classification performance compared to the site-target dependence scenario. The *No Target* method, however, did not remove biologically relevant information during harmonization, achieving performance similar to the other benchmarked methods (AUC = 0.72).

Table 4
Comparison of sex classification performance metrics across different harmonization schemes.

Harmonization Scheme	Site-target dependence			Site-target independence		
	AUC	bACC [%]	F1	AUC	bACC [%]	F1
Unharmonized	0.969	92.64	0.923	0.918	84.94	0.851
PrettYharmonize	0.968	92.18	0.918	0.921	85.06	0.851
WDH	0.975	92.10	0.917	0.913	84.64	0.847
TTL	0.967	92.07	0.917	0.918	85.16	0.852
No Target	0.703	63.08	0.608	0.919	84.85	0.849

4. Discussion

In many clinical domains, data collected from a single location is insufficient for training generalizable ML models, as large and diverse datasets are needed to identify meaningful feature-target relationships. Combining data from multiple acquisition sites is thus a common and attractive strategy for developing ML models. However, this introduces unwanted variability due to acquisition idiosyncrasies. Data harmonization methods, such as ComBat, are often used to remove site-related

Table 5
Classification performance metrics for dementia-MCI prediction task in site-target dependent and independent scenarios.

Harmonization Scheme	Site-target dependence			Site-target independence		
	AUC	bACC [%]	F1	AUC	bACC [%]	F1
Unharmonized	0.8131	73.7273	0.7371	0.7092	65.68	0.6698
PrettYharmonize	0.8429	77.2727	0.7715	0.7089	65.31	0.6659
WDH	0.8385	76.6364	0.7644	0.7118	66.01	0.6755
TTL	0.8381	76.3636	0.7622	0.7103	65.85	0.6742
No Target	0.6384	60.1818	0.6054	0.7096	66.23	0.6794

Table 6
Classification performance metrics for discharge status prediction task in site-target dependent and independent scenarios.

Harmonization Scheme	Site-target dependence			Site-target independence		
	AUC	bACC [%]	F1	AUC	bACC [%]	F1
Unharmonized	0.7655	64.37	0.4571	0.7227	66.88	0.6250
PrettYharmonize	0.8588	66.25	0.4910	0.7101	66.14	0.6295
WDH	0.7995	63.39	0.4408	0.7029	65.25	0.6133
TTL	0.7897	63.91	0.4517	0.6907	64.75	0.6091
No Target	0.5723	51.66	0.0921	0.7198	66.42	0.6211

variability with promising results in ML applications [56]; however, integrating such methods into ML pipelines brings additional challenges. For instance, care must be taken to avoid data leakage by properly separating training and test sets [25,26,28], avoiding the creation of a "site-effect free" dataset that is then split into train and test sets.

Other sources of leakage and the effectiveness and practicality of data harmonization in ML workflows remain under-investigated, particularly in scenarios where the target variable and acquisition site are not independent, a situation in which harmonization methods typically struggle. In this study, we proposed a novel method, *PrettYharmonize*, which was first validated using the datasets designed for that purpose, as proposed in [4], and then benchmarked against five harmonization schemes using 7 medical datasets from various domains [32,33,44–48], specifically comparing their performance in cases of both site-target dependence and independence.

The proposed *PrettYharmonize* method is designed to avoid data leakage. It integrates a stacking ensemble architecture [31] with a ComBat-based model for the harmonization procedure. This design ensures that the harmonization process remains confined to the training phase and circumvents the need for actual target values from the test samples, ensuring a leakage-free ML pipeline. Importantly, the method does not detect data leakage but rather allows for a leakage-free integration of ComBat-based models and ML pipelines. The method first harmonizes the data using pretended labels and then uses the harmonized output to learn a Predictive model. The main idea behind *PrettYharmonize* is learning subtle differences caused by the use of correct or incorrect labels. Several key aspects of *PrettYharmonize* deserve emphasis. First, it is built around the ComBat-based model method, meaning that we do not propose a novel way of harmonizing the data but rather introduce an approach that integrates a stack ensemble architecture with harmonization models into ML pipelines while avoiding data leakage by design. This distinction is crucial, as the innovation lies not in altering the harmonization process but in structuring its application to ensure compatibility with ML workflows across both site-target dependence scenarios. Furthermore, this modular design of *PrettYharmonize* enables the use of other harmonization methods. Second, the output of *PrettYharmonize* is not a harmonized dataset but rather a final target prediction. Unlike traditional harmonization schemes, which produce a harmonized dataset for downstream use, our method uses harmonized data internally as an intermediate step to generate a prediction. By focusing on prediction, the method prioritizes practical utility in real-world applications where the primary goal is accurate target prediction rather than data transformation. Finally, *PrettYharmonize* will inherit the limitations of the chosen harmonization model. For example, its ability to capture non-linearities, or to handle imbalance in the number of samples by site, among others.

PrettYharmonize was rigorously validated on the MAREoS datasets, which were specifically designed to evaluate harmonization methods [4]. The strong performance observed on these benchmark datasets indicates that the internal harmonization scheme, based on pretended targets, successfully harmonizes data for the posterior use of the Predictive model. In addition, on real-world datasets *PrettYharmonize* achieved competitive performance comparable to leakage-prone pipelines, demonstrating its practical utility. These results suggest that the proposed method is a promising alternative for real-world applications, particularly in settings where data leakage poses a substantial risk. Consequently, we advocate for its adoption in future use cases.

The ML pipelines pooling data without any harmonization showed performance consistent with values reported in the literature for all evaluated tasks [50,52,54,55], indicating correct application of ML models. As expected, all models using pooled unharmonized data performed better in site-target dependence scenarios, as the ML models could exploit the EoS signal, which in this case is related to the target, illicitly increasing their classification performance.

In site-target dependence scenarios, where classes are imbalanced across sites, we observed performance improvements when applying ComBat-based models (particularly neuroHarmonize) with the target specified as a covariate, which preserves its variance (Tables 3, 5 and 6). However, using the target as a covariate inevitably leads to data leakage as the target values of the test set are used in both Whole Data Harmonization (*WDH*) and Test Target Leakage (*TTL*) schemes. In the *WDH* scheme, all available data is used to train the harmonization model, and the whole dataset is transformed before splitting it into train and test sets, leading to a form of leakage known as "preprocessing on training and test sets" [26]. Although in the TTL scheme the test data is not used in the harmonization training process, the ComBat model still requires the target values of the test set to correctly transform the test data, a form of "target leakage" [25]. As expected, using ComBat without the target as a covariate (*No Target*) avoids data leakage but removes biologically relevant information, as evidenced by the marked and consistent drop in the performance across all evaluated scenarios (Tables 3–6). Aligned with the observations in [17], our experiments demonstrate that the target range (e.g., the age range) must overlap so the harmonization method does not eliminate relevant information or leak the test labels. Our results empirically demonstrate the impact of the site-target dependence, which violates ComBat's assumption that relevant variance, i.e., target variance, is shared across sites.

Conversely, in site-target independence scenarios, despite testing a wide range of tasks and datasets from different domains, none of the harmonization schemes improved performance over the baseline approach of simply pooling the data. This also includes our proposed *PrettYharmonize* method, which, while not being a harmonization method itself, uses a novel way to leverage harmonization without causing data leakage. This lack of performance improvement may be due to the fact that the harmonization process did not sufficiently enhance the signal-to-noise ratio to allow the models to generate better predictions. Nevertheless, we acknowledge that it is possible that our selection of data and tasks, while comprehensive, may not encompass all scenarios where harmonization could prove beneficial. Some potential scenarios where *PrettYharmonize* could prove useful are when a higher number of sites are used, or when small samples are available per site, where the ML models are not able to extract a robust signal that generalizes across sites.

ComBat-based methods use the covariates to estimate and preserve their variance. In machine learning pipelines, the target-related variance of the data is the signal of interest. Original ComBat was only able to estimate linear relationships, but ComBat-GAM improved upon this

limitation to allow the model to capture non-linear relationships. Thus, harmonization model selection is heavily dependent on the target type.

An alternative approach to training harmonization models is to use phantoms or traveling subjects [14,18]. This can allow for accurate estimation of location and scale parameters specific to each acquisition setup and parameter setting, which can then be applied to real data. This approach mitigates the risk of inadvertently removing meaningful biological variation during harmonization. However, it is domain-specific and requires additional data collection, which is costly, as well as accounting for other challenges such as temporal shifts in equipment.

This is the first study, to the best of our knowledge, to apply harmonization to features derived from arterial blood gases (ABGs). Although these features violate ComBat's assumption of features presenting similar value ranges, *PrettYharmonize* yielded the largest performance gain in the site-target dependence scenario. Further research is needed to separate the benefits of the stacked ensemble architecture from those of data harmonization. It is important to distinguish this form of harmonization, which aims to remove EoS from features, from the more common use of the term in the clinical domain, where "harmonization" typically refers to the standardization of feature names and units across sites [58].

Taken together, our findings underscore the need for meticulous evaluation when using data harmonization together with machine learning and encourage the adoption of reproducible, open science practices to advance the field and benefit the wider community.

5. Conclusion

This study highlights three key findings regarding the use of harmonization methods in ML pipelines. First, while ComBat-based methods do not intrinsically cause data leakage, they might require that target variance be preserved as a covariate during training when a site-target relationship exists. As demonstrated, this requirement precludes their use in real-world ML applications, as it requires target values of the test samples that are not available. Our results empirically demonstrate that failing to preserve target variance leads to the removal of relevant signals, significantly undermining the effectiveness of harmonization. Second, harmonization did not lead to meaningful performance improvement when the site and targets are independent. This suggests that the benefits of harmonization may be limited in such cases, as the removal of site-related variability might not sufficiently enhance the signal-to-noise ratio for ML models. Finally, the proposed *PrettYharmonize* method advances the field by adapting ComBat-based methods in ML pipelines, effectively integrating harmonization while eliminating the need for test targets during training or transformation. This approach not only prevents data leakage by design but also achieves encouraging results, ensuring that the harmonization process remains both robust and practical for real-world applications.

6. Limitations of the study

While this study provides valuable insights into harmonization methods, several limitations must be acknowledged. First, although our analysis focused on widely used ComBat-based techniques, other harmonization approaches, such as deep learning-based methods (e.g., style-matching generative models or variational autoencoders [5,9]), were not explored. These methods may offer promising alternatives, particularly in complex scenarios where traditional approaches fall short, although they often require significant computational resources and large datasets.

Second, the impact of harmonization on feature selection and model interpretability has not been thoroughly investigated. Future research should explore how harmonization methods affect model behavior, especially in different domains and contexts, to better understand their broader implications for ML pipelines.

Third, while we simulated extreme site-target dependence and independence scenarios in a controlled manner, real-world cases are likely to fall somewhere along this spectrum. Our goal was to highlight potential problems that may arise from applying harmonization without proper consideration. Further studies are needed to investigate the effectiveness of harmonization methods at varying degrees of site-target dependence.

Finally, the proposed method introduces additional computational complexity compared to training a harmonization model, as the "pretending" process requires multiple harmonization transformations. However, the extra burden of this process is limited to *applying* a trained harmonization model (neuroHarmonize in our experiments) several times (one harmonization transformation for each presented class), which is less time-consuming than training the model. Importantly, the computational cost of the pretending process does not scale with the number of sites to be harmonized, mitigating some of the practical limitations.

CRediT authorship contribution statement

Nicolás Nieto: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Simon B. Eickhoff:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Christian Jung:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Martin Reuter:** Writing – review & editing, Resources, Funding acquisition. **Kersten Diers:** Writing – review & editing, Resources. **Malte Kelm:** Writing – review & editing, Resources, Funding acquisition. **Artur Lichtenberg:** Writing – review & editing, Resources, Funding acquisition. **Federico Raimondo:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Kaustubh R. Patil:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was 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. It was also supported by Helmholtz Portfolio Theme Supercomputing and Modeling for the Human Brain. This work was partly supported by the Helmholtz-AI project DeGen (ZT-I-PF-5-078).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies;

Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix A. Supplementary data

Supplementary data for this article can be found online at doi: 10.1016/j.neucom.2026.133146.

Data and code availability

All used MRI datasets are publicly available, possibly upon registration. The eICU dataset is publicly available at https://physionet.org/content/eicu-crd/2.0/ after registration. Registration includes the completion of a training course in research with human individuals at https://about.citiprogram.org/ and signing of a data use agreement mandating responsible handling of the data and adhering to the principle of collaborative research.

The *PrettYharmonize* library is publicly available at: https://github.com/juaml/PrettYharmonize and also implemented as part of https://github.com/N-Nieto/UniHarmony. The scripts to replicate the experiments and to process the datasets are available at: https://github.com/juaml/harmonize_project.

References

[1] J. Chen, J. Liu, V. Calhoun, A. Arias-Vasquez, M. Zwiers, C. Gupta, B. Franke, J. Turner, Exploration of scanning effects in multi-site structural MRI studies, J. Neurosci. Methods. 230 (2014) 37–50.
[2] J. Bayer, P. Thompson, C. Ching, M. Liu, A. Chen, A. Panzenhagen, N. Jahanshad, A. Marquand, L. Schmaal, P. Sämann, Site effects how-to and when: an overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses, Front. Neurol. 13 (2022) 923988.
[3] R. Botvinik-Nezer, T. Wager, Reproducibility in neuroimaging analysis: challenges and solutions, Biol. Psychiatry Cogn. Neurosci. Neuroimaging. 8 (2023) 780–788.
[4] A. Solanes, C. Gosling, L. Fortea, M. Ortuño, E. Lopez-Soley, S. Llufriu, S. Madero, E. Martinez-Heras, E. Pomarol-Clotet, E. Solana, et al., Removing the effects of the site in brain imaging machine-learning–measurement and extendable benchmark, NeuroImage 265 (2023) 119800.
[5] F. Hu, A. Chen, H. Horng, V. Bashyam, C. Davatzikos, A. Alexander-Bloch, M. Li, H. Shou, T. Satterthwaite, M. Yu, et al., Image harmonization: a review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization, NeuroImage 274 (2023) 120125.
[6] H. Li, S. Smith, S. Gruber, S. Lukas, M. Silveri, K. Hill, W. Killgore, L. Nickerson, Denoising scanner effects from multimodal MRI data using linked independent component analysis, Neuroimage. 208 (2020) 116388.
[7] C. Wachinger, A. Rieckmann, S. Pölsterl, A. Initiative, et al., Detect and correct bias in multi-site neuroimaging datasets, Med. Image Anal. 67 (2021) 101879.
[8] R. Da-Ano, D. Visvikis, M. Hatt, Harmonization strategies for multicenter radiomics investigations, Phys. Med. Biol. 65 (2020) 24TR02.
[9] S. Abbasi, H. Lan, J. Choupan, N. Sheikh-Bahaei, G. Pandey, B. Varghese, Deep learning for the harmonization of structural MRI scans: a survey, BioMed. Eng. OnLine. 23 (2024) 90.
[10] J. Fortin, N. Cullen, Y. Sheline, W. Taylor, I. Aselcioglu, P. Cook, P. Adams, C. Cooper, M. Fava, P. McGrath, et al., Harmonization of cortical thickness measurements across scanners and sites, Neuroimage 167 (2018) 104–120.
[11] C. Acquitter, L. Piram, U. Sabatini, J. Gilhodes, E. Moyal Cohen-Jonathan, S. Ken, B. Lemasson, Radiomics-based detection of radionecrosis using harmonized multiparametric MRI, Cancers 14 (2022) 286.
[12] Y. Li, S. Ammari, C. Balleyguier, N. Lassau, E. Chouzenoux, Impact of preprocessing and harmonization methods on the removal of scanner effects in brain MRI radiomic features, Cancers 13 (2021) 3000.
[13] M. Ingalhalikar, S. Shinde, A. Karmarkar, A. Rajan, D. Rangaprakash, G. Deshpande, Functional connectivity-based prediction of autism on site harmonized ABIDE dataset, IEEE Trans. Biomed. Eng. 68 (2021) 3628–3637.
[14] N. Maikusa, Y. Zhu, A. Uematsu, A. Yamashita, K. Saotome, N. Okada, K. Kasai, K. Okanoya, O. Yamashita, S. Tanaka, et al., Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics, Hum. Brain Mapp. 42 (2021) 5278–5287.
[15] W. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics 8 (2007) 118–127.
[16] J. Fortin, D. Parker, B. Tunç, T. Watanabe, M. Elliott, K. Ruparel, D. Roalf, T. Satterthwaite, R. Gur, R. Gur, et al., Harmonization of multi-site diffusion tensor imaging data, Neuroimage. 161 (2017) 149–170.
[17] R. Pomponio, G. Erus, M. Habes, J. Doshi, D. Srinivasan, E. Mamourian, V. Bashyam, I. Nasrallah, T. Satterthwaite, Y. Fan, et al., Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan, NeuroImage 208 (2019) 116450.
[18] A. Ibrahim, S. Primakov, M. Beuque, H. Woodruff, I. Halilaj, G. Wu, T. Refaee, R. Granzier, Y. Widaatalla, R. Hustinx, et al., Radiomics for precision medicine: current challenges, future prospects, and the proposal of a new framework, Methods. 188 (2021) 20–29.
[19] M. Yu, K. Linn, P. Cook, M. Phillips, M. McInnis, M. Fava, M. Trivedi, M. Weissman, R. Shinohara, Y. Sheline, Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data, Hum. Brain Mapp. 39 (2018) 4213–4227.
[20] J. Dudley, T. Maloney, J. Simon, G. Atluri, S. Karalunas, M. Altaye, J. Epstein, L. Tamm, Abcd_harmonizer: an open-source tool for mapping and controlling for scanner induced variance in the adolescent brain cognitive development study, Neuroinform. 21 (2023) 323–337.
[21] V. Bourbonne, V. Jaouen, T. Nguyen, V. Tissot, L. Doucet, M. Hatt, D. Visvikis, O. Pradier, A. Valéri, G. Fournier, et al., Development of a radiomic-based model predicting lymph node involvement in prostate cancer patients, Cancers 13 (2021) 5672.
[22] V. Campello, C. Martín-Isla, C. Izquierdo, A. Guala, J. Palomares, D. Viladés, M. Descalzo, M. Karakas, E. Cavus, Z. Raisi-Estabragh, et al., Minimising multi-centre radiomics variability through image normalisation: a pilot study, Sci. Rep. 12 (2022) 12532.
[23] P. Chen, H. Yao, B. Tijms, P. Wang, D. Wang, C. Song, H. Yang, Z. Zhang, K. Zhao, Y. Qu, et al., Four distinct subtypes of alzheimer's disease based on resting-state connectivity biomarkers, Biol. Psychiatry. 93 (2023) 759–769.
[24] B. Bostami, F. Espinoza, H. Horn, J. Van Der Naalt, V. Calhoun, V. Vergara, Multi-site mild traumatic brain injury classification with machine learning and harmonization, in: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2022, pp. 537–540.
[25] L. Sasse, E. Nicolaisen-Sobesky, J. Dukart, S.B. Eickhoff, M. Götz, S. Hamdan, V. Komeyer, A. Kulkarni, J.M. Lahnakoski, B.C. Love, et al., Overview of leakage scenarios in supervised machine learning, J. Big Data 12 (1) (2025) 135.
[26] S. Kapoor, A. Narayanan, Leakage and the reproducibility crisis in machine-learning-based science, Patterns. 4 (2023).
[27] M. Lones, How to avoid machine learning pitfalls: a guide for academic researchers, ArXiv Preprint ArXiv:2108.02497, 2021.
[28] C. Marzi, M. Giannelli, A. Barucci, C. Tessa, M. Mascalchi, S. Diciotti, Efficacy of MRI data harmonization in the age of machine learning: a multicenter study across 36 datasets, Sci. Data. 11 (2024) 115.
[29] V. Nygaard, E. Rødland, E. Hovig, Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses, Biostatistics. 17 (2016) 29–39.
[30] J. Radua, E. Vieta, R. Shinohara, P. Kochunov, Y. Quidé, M. Green, C. Weickert, T. Weickert, J. Bruggemann, T. Kircher, et al., Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA, Neuroimage. 218 (2020) 116956.
[31] D. Wolpert, Stacked generalization, Neural Netw. 5 (1992) 241–259.
[32] C. Jack Jr, M. Bernstein, N. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. Britson, L. Whitwell, J. Ward, et al., The alzheimer's disease neuroimaging initiative (ADNI): MRI methods, J. Magn. Reson. Imaging 27 (2008) 685–691.
[33] T. Pollard, A. Johnson, J. Raffa, L. Celi, O. Badawi, R. Mark, eICU collaborative research database (version 2.0), PhysioNet. 10 (2019) C2WM1R.
[34] T. Pollard, A. Johnson, J. Raffa, L. Celi, R. Mark, O. Badawi, The eICU collaborative research database, a freely available multi-center database for critical care research, Sci. Data. 5 (2018) 1–13.
[35] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C. Peng, H. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circ. 101 (2000) e215–e220.
[36] N. Hou, M. Li, L. He, B. Xie, L. Wang, R. Zhang, Y. Yu, X. Sun, Z. Pan, K. Wang, Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost, J. Transl. Med. 18 (2020) 1–14.
[37] H. Deng, M. Sun, Y. Wang, J. Zeng, T. Yuan, T. Li, D. Li, W. Chen, P. Zhou, Q. Wang, et al., Evaluating machine learning models for sepsis prediction: a systematic review of methodologies, Iscience. 25 (2022).
[38] Z. Yang, X. Cui, Z. Song, Predicting sepsis onset in ICU using machine learning models: a systematic review and meta-analysis, BMC Infect. Dis. 23 (2023) 635.
[39] M. Wu, X. Du, R. Gu, J. Wei, Artificial intelligence for clinical decision support in sepsis, Front. Med. 8 (2021) 665464.
[40] Y. Zhang, W. Xu, P. Yang, A. Zhang, Machine learning for the prediction of sepsis-related death: a systematic review and meta-analysis, BMC Med. Inform. Decis. Mak. 23 (2023) 283.
[41] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[42] J. Tolles, W. Meurer, Logistic regression: relating patient characteristics to outcomes, JAMA. 316 (2016) 533–534.
[43] C. Gaser, R. Dahnke, P. Thompson, F. Kurth, E. Luders, A. Initiative, CAT–a computational anatomy toolbox for the analysis of structural MRI data, Biorxiv pp. 2022–06 (2022).
[44] L. Snoek, M. Miesen, T. Beemsterboer, A. Van Der Leij, A. Eigenhuis, H. Steven Scholte, The Amsterdam open MRI collection, a set of multimodal MRI datasets for individual difference analyses, Sci. Data. 8 (2021) 85.
[45] K. Nooner, S. Colcombe, R. Tobe, M. Mennes, M. Benedict, A. Moreno, L. Panek, S. Brown, S. Zavitz, Q. Li, et al., The NKI-rockland sample: a model for accelerating the pace of discovery science in psychiatry, Front. Neurosci. 6 (2012) 152.
[46] M. Shafto, L. Tyler, M. Dixon, J. Taylor, J. Rowe, R. Cusack, A. Calder, W. Marslen-Wilson, J. Duncan, T. Dalgleish, et al., The Cambridge centre for ageing and Neuroscience (CAM-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing, BMC Neurol. 14 (2014) 1–25.
[47] S. Caspers, S. Moebus, S. Lux, N. Pundt, H. Schütz, T. Mühleisen, V. Gras, S. Eickhoff, S. Romanzetti, T. Stöcker, et al., Studying variability in human brain aging in a population-based German cohort—rationale and design of 1000brains, Front. Aging Neurosci. 6 (2014) 149.
[48] D. Wei, K. Zhuang, L. Ai, Q. Chen, W. Yang, W. Liu, K. Wang, J. Sun, J. Qiu, Structural and functional brain scans from the cross-sectional southwest university adult lifespan dataset, Sci. Data. 5 (2018) 1–10.
[49] M. Tipping, The relevance vector machine, Adv. Neural Inf. Process. Syst. 12 (1999).
[50] S. More, G. Antonopoulos, F. Hoffstaedter, J. Caspers, S. Eickhoff, K. Patil, A. Initiative, et al., Brain-age prediction: a systematic comparison of machine learning workflows, NeuroImage 270 (2023) 119947.
[51] B. Fischl, FreeSurfer, Neuroimage 62 (2012) 774–781.
[52] C. Flint, K. Förster, S. Koser, C. Konrad, P. Zwitserlood, K. Berger, M. Hermesdorf, T. Kircher, I. Nenadic, A. Krug, et al., Biological sex classification with structural MRI data shows increased misclassification in transgender women, Neuropsychopharmacology 45 (2020) 1758–1765.
[53] M. Hubert, E. Vandervieren, An adjusted boxplot for skewed distributions, Comput. Stat. Data Anal. 52 (2008) 5186–5201.
[54] T. Illakiya, R. Karthik, Automatic detection of alzheimer's disease using deep learning models and neuro-imaging: current trends and future perspectives, Neuroinformatics 21 (2023) 339–364.
[55] B. Wernly, B. Mamandipoor, P. Baldia, C. Jung, V. Osmani, Machine learning predicts mortality in septic patients using only routinely available ABG variables: a multi-centre evaluation, Int. J. Med. Inform. 145 (2021) 104312.
[56] F. Orlhac, J. Eertink, A. Cottereau, J. Zijlstra, C. Thieblemont, M. Meignan, R. Boellaard, I. Buvat, A guide to ComBat harmonization of imaging biomarkers in multicenter studies, J. Nucl. Med. 63 (2022) 172–179.
[57] R. Castaldo, V. Brancato, C. Cavaliere, F. Trama, E. Illiano, E. Costantini, A. Ragozzino, M. Salvatore, E. Nicolai, M. Franzese, A framework of analysis to facilitate the harmonization of multicenter radiomic features in prostate cancer, J. Clin. Med. 12 (2022) 140.
[58] M. Plebani, Harmonization in laboratory medicine: requests, samples, measurements and reports, Crit. Rev. Clin. Lab. Sci. 53 (2016) 184–196.

Author biography

Nicolás Nieto is a Biomedical engineer and currently a Postdoctoral Fellow in the Applied Machine Learning group at the Research Center Jülich. He holds a PhD in Computational Intelligence and Signal Processing from Universidad Nacional del Litoral. His research focuses on machine learning and data harmonization, with a particular interest in the ethics of AI and the real-world integration of machine learning models.

Simon B. Eickhoff studied medicine in Aachen, Sheffield, Sydney and London. He received his doctorate degree in neuroanatomy in 2006, following work on brain histology and structure-function correlations at the Heinrich-Heine University in Düsseldorf. He went on to work as a post-doctoral fellow in functional neuroanatomy at the Research Center Jülich, Germany before being appointed as assistant professor for Psychiatry at the RWTH Aachen in 2009. From 2011–2017 he served as a professor for cognitive neuroscience at the Heinrich-Heine University in Düsseldorf and group leader of the Brain Network Modeling group at the Institute of Neuroscience and Medicine in Jülich. From 2017 Simon Eickhoff is a full professor and chair of the Institute for Systems Neuroscience at the Heinrich-Heine University in Düsseldorf and the director of the Institute of Neuroscience and Medicine (INM-7, Brain and Behavior) at the Forschungszentrum Jülich. Working at the interface between neuroanatomy, data-science and brain medicine, he aims to obtain a more detailed characterization of the organization of the human brain and its inter-individual variability in order to better understand its changes in advanced age as well as neurological and psychiatric disorders.

Christian Jung is a professor for interventional cardiology and cardiovascular critical care medicine at the Heinrich-Heine-University, Düsseldorf, Germany. He and his group have a broad approach in basic, translational and clinical research with a specific focus on hemodynamics, endothelial function and microcirculation. Consequently, cardiogenic shock is a major research interest. In these different fields, novel digital technologies are applied with the ultimate goal to improve patient care. Prof. Jung is the co-PI of INDICATE (https://indicate-europe.eu) a new EU-funded project in intensive care medicine. INDICATE aims to advance patient-centered care and promote ethically responsible data use and the development and implementation of trustworthy AI models.

Martin Reuter received the Diplom degree in Mathematics from Leibniz University Hannover, Germany, and the Ph.D. degree in Computer Science and Applied Mathematics from the same institution. He is Director of AI in Medical Imaging at the German Center for Neurodegenerative Diseases (DZNE) in Bonn and Principal Investigator of the Deep Medical Imaging Lab, as well as Assistant Professor of Radiology and Neurology at Harvard Medical School and the Athinoula A. Martinos Center for Biomedical Imaging. His research focuses on artificial intelligence and deep learning for medical imaging, computational neuroimaging, and computational geometry, with applications to automated brain MRI analysis and neurodegenerative diseases. He has co-developed widely used methods and software, including ShapeDNA, FreeSurfer, and FastSurfer, and has authored highly cited work on longitudinal image analysis, shape analysis, and deep-learning–based neuroimaging.

Kersten Diers graduated from TU Dresden, Germany, with a diploma in Psychology and from University of Heidelberg, Germany, with a M.Sc. degree in Medical Biometry/Biostatistics. At the German Center of Neurodegenerative Diseases in Bonn, Germany, his work is at the intersection of applied methods development and empirical research, with a particular focus on shape analysis and statistical modeling of neuroimaging data.

Malte Kelm completed his medical studies and received his MD in 1986 in Cologne, where he subsequently did his specialisation in internal medicine and took on academic leading positions until his appointment as director of the university clinic Aachen, which he left for Düsseldorf in 2009, finally becoming the director of the clinic for cardiology, angiology and pneumology. There, he initially cofounded the Cardiovascular Research Center Düsseldorf (CARID) in 2013 and in 2022 the CARDDIAB Center for cardiovascular research in Diabetes. His scientific focus areas are the pathophysiology of coronary perfusion, interventional cardiology with focus on coronary artery disease, AMI and structural heart disease, regulation of coronary blood flow, ischemia-reperfusion injury, endothelial (dys)function, circulating NO pool and anemia. This research contributed to the body of evidence demonstrating that plasma and red blood cells store, transport and produce nitric oxide (NO) metabolites and intermediates, thus contributing to the regulation of blood flow and vascular tone in physiology and pathophysiology.

Artur Lichtenberg, after completing his medical degree, began his residency in cardiac surgery at the University of Tübingen in 1993, where he completed his doctoral thesis in 1995. After residencies in thoracic and cardiovascular surgery at the University Hospital Tübingen, the Heart Center Lahr/Baden, and the Medical School Hannover, he obtained board certification in Cardiac Surgery in 2000 and subsequently served as staff surgeon at the Medical University of Hannover, where he also received the venia legendi for Cardiac Surgery. In 2009 he became Vice Head of the Cardiac Surgery Department at the University Hospital Heidelberg. In 2009, he was appointed W3-Professor and Director of the Department of Thoracic and Cardiac Surgery at the University Hospital Jena, and since August 2009 he has been W3-Professor, Chairman, and Chief of the Department of Cardiac Surgery at the University Hospital Düsseldorf. His clinical expertise covers the full spectrum of adult cardiac surgery with a focus on surgical therapy of end-stage heart failure (VAD, heart transplantation), minimally invasive coronary surgery (OPCAB, MIDCAB), minimally invasive mitral and aortic valve reconstructive procedures, and aortic surgery. He is a member of several professional societies, including the German Association for Thoracic and Cardiovascular Surgery (DGTHG), the European Association for Cardio-Thoracic Surgery (EACTS), the German Cardiac Society (DGK), the European Society of Cardiology (ESC), and CTS-Net.

Federico Raimondo received the Licenciado degree in Computer Science (equivalent to B.Sc. and M.Sc.) from the University of Buenos Aires, Argentina, in 2011, and a joint Ph.D. in Computer Science and Cognitive Neuroscience from the University of Buenos Aires and Sorbonne University, France, in 2018. His doctoral work focused on machine-learning approaches for the diagnosis of disorders of consciousness using electrophysiological brain signals. He was a Postdoctoral Fellow at the University of Liège, Belgium (2018–2020), and a Postdoctoral Researcher at the Institute of Neuroscience and Medicine – Brain and Behaviour (INM-7), Forschungszentrum Jülich, Germany (2020–2023). Since 2023, he leads the Research and Infrastructural Software Engineering for/in Machine Learning (RISE-ML) group at Forschungszentrum Jülich. His research focuses on developing trustworthy and explainable AI methods for neuroscience, particularly in the study of consciousness and cognition.

Kaustubh R. Patil received the B.E. degree in Electronics Engineering from Shivaji University, India, in 2003, the M.Sc. degree in Artificial Intelligence and Intelligent Systems from the University of Porto, Portugal, in 2007, and the Ph.D. degree in Computer Science from the Max Planck Institute for Informatics, Germany, in 2013. From 2013 to 2016, he was a Postdoctoral Fellow at University College London (UCL) and the Massachusetts Institute of Technology (MIT). Since 2017, he has been with Forschungszentrum Jülich (FZJ), where since 2019 he leads the Applied Machine Learning Group. Since 2022 Kaustubh is a visiting professor at IIT Bombay. His research focuses on developing and applying machine learning methods to advance the understanding of biological systems. He is a member of IEEE and currently serves as an Associate Editor of IEEE Access.