Impact of Leakage on Data Harmonization in Machine Learning Pipelines in Class Imbalance Across Sites.

Nicolás Nieto[1,2,3,*], Simon B. Eickhoff[1,2], Christian Jung[3,4], Martin Reuter[5,6], Kersten Diers[5], for the Alzheimer's Disease Neuroimaging Initiative[+], Malte Kelm[3,4], Artur Lichtenberg[7], Federico Raimondo[1,2], and Kaustubh R. Patil[1,2]

[1]Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany
[2]Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[3]Department of Cardiology, Pulmonology and Vascular Medicine, University Hospital and Medical Faculty, Heinrich-Heine University, Duesseldorf, Germany
[4]Cardiovascular Research Institute Düsseldorf (CARID), Medical Faculty, Heinrich-Heine University, Duesseldorf, Germany
[5]Artificial Intelligence in Medical Imaging, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[6]Department of Radiology, Harvard Medical School, Boston, MA, USA
[7]Department of Cardiac Surgery, University Hospital and Medical Faculty, Heinrich-Heine University, Duesseldorf, Germany
[+]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: (`http://adni.loni.usc.edu/wp-content/uploads/how\protect_to\protect_ apply/ADNI\protect_Acknowledgement_List.pdf`)
[*]Correspondence: n.nieto@fz-juelich.de

Summary

Machine learning (ML) models benefit from large datasets. Collecting data in biomedical domains is costly and challenging, hence, combining datasets has become a common practice. However, datasets obtained under different conditions could present undesired site-specific variability. Data harmonization methods aim to remove site-specific variance while retaining biologically relevant information. This study evaluates the effectiveness of popularly used ComBat-based methods for harmonizing data in scenarios where the class balance is not equal across sites. We find that these methods struggle with data leakage issues. To overcome this problem, we propose a novel approach "PrettYharmonize", designed to harmonize data by pretending the target labels. We validate our approach using controlled datasets designed to benchmark the utility of harmonization. Finally, using real-world MRI and clinical data, we compare leakage-prone methods with "PrettYharmonize" and show that it achieves comparable performance while avoiding data leakage, particularly in site-target-dependence scenarios.

Keywords

Data Harmonization, ComBat, Data leakage, Machine learning, Medical Imaging, Magnetic Resonance Imaging, medical AI, clinical, ICU.

Introduction

Many research fields have greatly benefited from machine learning (ML) approaches. ML models can extract important values from large amounts of data. Having vast data benefits the model's classification performance and helps capture the underlying patterns, promoting better generalization to new unseen data. This makes combining multiple datasets an appealing approach, especially in domains where obtaining data in a uniform setting is challenging[1]. Moreover, small health or research centers that can not afford to collect a large number in-house data, using data acquired in different sites is the only possibility for train ML models. However, different datasets obtained under different conditions often present variability due to differences in the acquisition procedure that are unrelated to relevant biological information[2]. This undesired variability, also known as Effects of Site (EoS), can induce biased results if present or not correctly removed[3]. These differences may come from systematic differences, which can be corrected, or random variations, which can not be modeled or corrected by harmonization. This problem is of common occurrence in many biomedical domains. For example, clinical data is affected by the acquisition site, as different hospitals have different laboratory machines, procedures, and criteria. Another example is the medical imaging field, as images are affected by acquisition protocol, scanner drifts, and time of the day, just to name a few factors[3,4]. Within this field, Magnetic Resonance Imaging (MRI) images are particularly susceptible to this site-related variance, like the magnetic field strength, room temperature fluctuation or changes in the electromagnetic noise, which makes even images obtained from scanners with the same manufacturer and the same parameters exhibit different characteristics[5,6]. Many works showed that removing this undesired systematic variability, which is only related to the acquisition site and has no biological information, can benefit further analysis made with the data[7–11]. To this end, several Methods Aiming to Remove the Effects of Site (MAREoS) have been proposed and developed[4,12]. These MAREoS methods are typically used as a pre-processing step, where the site effects are removed and the "site-effect free" data, also known as harmonized data, is used for statistical analysis or to train and evaluate ML models.

Among these MAREoS, the ones based on "ComBat" are extensively used in several domains. The ComBat method was originally proposed for correcting batch differences in genomic data[13] and was later adapted to other domains like MRI data[7,14]. ComBat uses Bayesian regression to find additive (location) and multiplicative (scale) corrections for each feature in each site. Within the ComBat-based methods, "neuroHarmonize" was proposed[15] to allow for the preservation of non-linear covariate effects and has been widely used since[4,12,16,17]. Although ComBat and its derivations have been widely applied in medical imaging data, several concerns have been raised, mainly because ComBat's hypothesis and assumptions only hold for genomic data, where it was originally proposed, and may not be fulfilled in other applications fields[18]. Additionally, concerns had been raised on the integration of ComBat into ML models, as the location and scale parameters of the model could not be learned in a subset of data (train data) and then applied to a new unseen subset of data (test data)[19].

Early implementations of ComBat[14,20] used the whole dataset to estimate the model's parameters and create a harmonized dataset that is then used from all the downstream analyses. This approach was used in several works[7,8,21–24]. While this approach is valid when performing statistical analyses, it is not consistent with machine learning applications where the training and test data must be separate[19,25]. Specifically, the parameters of the models, including preprocessing models, must be obtained on a training set and then applied to the test set. This separation is important to get realistic estimates of generalization performance (e.g. using cross-validation) and to ensure deployability of the model in the real world where the test data is not yet available[26,27].

ComBat-MAGA[28], neuroHarmonize[15], and "harmonizer"[19], which is based on neuroHarmo-

nize, allow the estimation of the model's parameters in a training set and apply them to the test samples, however, a critical assumption of ComBat is that all variance not shared across sites is unwanted site-related variance. Consequently, ComBat removes any variance that is not common to all sites, including the relevant biological variance. This poses a new problem, as this assumption is broken when a class imbalance occurs across sites and a target-site dependence exists, for example when the control patients are acquired in one site and target patients in a different one[29]. This could also be extended to other possible biological information like co-morbidities or disease severity, for example, if more severe patients are consistently treated or acquired only in one site. In these cases, even though ComBat will provide harmonized data, it will remove the variance related to the target (control versus patient) as the assumption is that only non-relevant factors change between sites.

ComBat allows to retain the biologically relevant variables, for example, a diagnosis, the age, or the sex of a patient, by providing these variables as covariates to be retained. Nonetheless, this information is needed both when training the model and when applying the model to the test data. Thus, if target labels need to be preserved, this inevitably leads to the model requiring the test labels preventing the model's use in real-world applications where test labels are not available or known[18]. This phenomenon where information of the test set is presented to the models is commonly known as *data leakage*. It is also well described that leaking the test target information would produce overconfident results, which could be misleading and can jeopardize the progress of an entire research field, as researchers who avoid data leakage would not be able to outperform the models that present data leakage[25–27].

In this work, we aim to empirically demonstrate a shortcoming of ComBat-based harmonization in site-target dependence scenarios, i.e., that the model can properly harmonize the data only when test labels are used and data leakage happens. To do so, we performed controlled experiments for age regression and sex classification using real MRI data for healthy control individuals. Also using MRI data, a dementia and mild cognitive impairment (MCI) classification experiment was performed. Additionally, an outcome prediction of septic patients was performed using clinical data. All experiments were conducted both in site-target dependence and independence scenarios. Several harmonization schemes were used and compared, allowing and not allowing leakage, to harmonize the data.

Finally, to overcome the aforementioned problem, we propose a new harmonization method, called PRETended Target Y Harmonize (*PrettYHarmonize*), which allows the users to integrate ComBat in an ML pipeline, harmonizing the data and generating a prediction without using the test labels and thus avoiding data leakage. We validated our method using benchmark datasets[3]. Additionally, the proposed method was compared with the other harmonization schemes on the site-target dependence and independence scenarios to comprehensively compare no harmonization, leakage, and no-leakage methods. The corresponding Python package is publicly available via GitHub `https://github.com/juaml/PrettYharmonize`.

Results

0.1 PrettYharmonize validation

The proposed PrettYharmonize method is based on a neuroHarmonize model[15] but uses the combination of two ML models, a Predictive and a Stack model to harmonize the data without using test labels, preventing data leakage by design. To avoid using the test data labels, the proposed methods rest on the use of *pretended* target values, which are used to harmonize the test data and generate a prediction with the Predictive model. Our main assumption is that when harmonizing the test data with the correct label, e.g. when the pretended label matches

the real test label, the neuroHarmonize model will preserve the relevant information. On the contrary, when the pretended label doesn't correspond with the real label, the all information will be removed, both effects of site and biological information, as it was harmonized under the wrong test label assumption. Then, the Predictive model will generates a more accurate and Using these predictions as input features, the Stack model generates a final unique prediction for each sample. As test labels are pretended and not used, predictions can be generated without requiring test target values. A detailed description of the method workflow is presented in the Method-PrettYharmonize section.

PrettYharmonize was validated using the datasets specially designed to validate MAREoS[3]. This dataset consists of eight internal datasets simulating eighteen MRI features (cortical thickness, cortical surface area, or subcortical volumes). Four datasets contain a "True" signal and four only contain an Effect of Site ("EoS") signal related to a binary target, which the MAREoS should remove to avoid fraudulent classification performance. For each kind of signal, two "Simple" and "Interaction" datasets are proposed, for linear and no-linear relationships between the features and the target. For each internal dataset, 1000 samples coming from eight different sites are simulated. An extended data description of this dataset is available in the "Data-MAREoS" section. On the datasets that presented True signal and no EoS, a Baseline model (Random Forest), trained on the unharmonized data, obtained a balanced Accuracy (bACC) of around 80%, as expected (colum. This model also obtained a close to 80% bACC on the datasets that only contained the EoS signal, but this time fraudulently used the EoS signal to perform the classification (Table 1).

The model successfully removed the EoS in all datasets which only present an EoS signal. Furthermore, in those datasets where only the True signal was presented, the method did not degrade the real signal while aiming to remove EoS, which in the True datasets are not presented (Table 1). Finally, we repeated this analysis to check robustness using three different Predictive models; Gaussian Process Classifier (GP), Support Vector Machine with Radial basis kernel (SVM), and least absolute shrinkage and selection operator (LASSO). These yielded similar results (Tables Supp 1, 2, and 3).

Table 1: PrettYharmonize and Baseline (RF model without harmonization) performance on the MAREoS dataset (bACC [%]: mean of 10 folds).

Dataset Name	Baseline	PrettYharmonize	Expected	Difference
True Simple 1 (no site effect)	72.86	72.07	As Baseline	0.79
True Simple 2 (no site effect)	82.72	82.86	As Baseline	0.06
True Interaction 1 (no site effect)	79.43	79.46	As Baseline	0.03
True Interaction 2 (no site effect)	72.23	70.72	As Baseline	1.51
EoS Simple 1 (no real effect)	76.11	54.18	Chance (50)	5.18
EoS Simple 2 (no real effect)	75.35	52.35	Chance (50)	2.35
EoS Interaction 1 (no real effect)	77.48	56.20	Chance (50)	6.2
EoS Interaction 2 (no real effect)	82.79	58.81	Chance (50)	8.81

0.2 Forced site-target dependence and independence.

To investigate the impact of the harmonization when site and target are dependent or independent, different scenarios were generated by thoroughly sampling data from different datasets. To force site-target dependence, for each site, we retained the majority of samples from one class and a small number of samples from the other classes. For example, let's assume a binary classification problem where only two sites are presented. In this case, from site A, mainly

samples from class one were retained, and only a few samples from class two. For site B, an opposite sampling strategy was used, retaining mainly samples from class two and just a few from class one. The small number of samples from the minority class were retained to avoid singular matrices, which the algorithms cannot support. In this case, we hypothesize that the traditional harmonization will remove important information, as the biological variance is related (or dependent) to the sites, unless test labels are leaked to the harmonization model. In contrast, site-target independence scenarios were generated by retaining the same amount of samples for each class sampled from each site. We hypothesize that in this case, the harmonization scheme will not remove important information, as the biological variance is common to all sites.

A total of seven datasets were used in our experiments. Five datasets [AOMID-ID1000, eNKI, CamCAN, SALD, and 1000Brains] containing MRI data from healthy control participants were sampled for age regression and sex classification problems. Using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset[30], where the data was collected in different sites, a classification of healthy controls versus mild cognitive impairment (MCI) / dementia was performed. Finally, using a publicly available multi-site clinical dataset [eICU][31,32] was used to perform a a classification of hospital discharge status of septic patients, as this is an important and extensively studied problem in the literature[33–35]. An extensive and detailed explanation of all the datasets and the sampling method used to force site-target dependence and independence is presented in Section *Data description*.

Five harmonization schemes were evaluated. The proposed PrettYharmonize model performs leakage-free harmonization while pretending labels (see section Experimental procedures-PrettYharmonize). Further, two schemes that induce data leakage but have been used in the literature were implemented. A "Whole Data Harmonization" (WDH) scheme, where the pooled data from all sites was used to train a neuroHarmonize model (based on ComBat) and to obtain a harmonized dataset before splitting the data into train and test folds. Additionally, a "Test Target Leakage" (TTL) scheme was used, where a neuroHarmonize model learned its parameters only on the training data while retaining target variance and thus requires the test labels to perform harmonization. Additionally, a scheme called "No Target" was proposed, where the harmonization model learns its parameters on the train data but without explicitly retaining the target variance, and thus the test labels were not used for test set harmonization. Finally, a baseline model without harmonization was implemented, where the pooled data were used *unharmonized*.

0.2.1 Age prediction

For the site-target dependence scenario, from four MRI datasets that contain healthy participants [AOMIC, eNKI, CamCAN, and 1000Brains] 200 images were extracted from each site in disjoint age ranges, forcing a site-target dependence. The same proportion of male and female participants was retained in each age range. The unharmonized method obtained a Mean Average Error (MAE) of 6.20 (Table 2), which is in an expected range according to the literature[36]. The predictions using the WDH and TTL schemes showed an improvement in the performance of about 2 years, compared with the unharmonized scheme (Table 2). The harmonization scheme that does not use the test labels (No Target) showed the highest error. As expected, the harmonization process removed the age-related signal in the features and therefore the ML model was unable to learn the feature-target relationship and failed to generate accurate predictions. In this case, the model just predicts the mean population age for all individuals, predicting the individuals to be older in the AOMIC and eNKI datasets and younger in the CamCAN and 1000Brains datasets (Figure 1a). PrettYharmonize made better predictions, on average, compared with the unharmonized and No Target methods, improving the MAE, R2, and age bias, and without inducing leakage (Table 2). PrettYharmonize's performance was similar to the two methods that

allowed leakage. The individuals' predictions can be found in the Supplementary Information Figures Supp 1-5.

For the site-target independence scenario, three datasets [eNKI, CamCAN, and SALD] containing healthy controls were used. The same number of images from each dataset was retained in the 18-80 age range, maintaining an equal proportion of males/females. The unharmonized model obtained an MAE of 6.3144, similar to the performance obtained in the site-target dependence scenario (Table 2). Neither PrettYharmonize nor any of the leakage-prone harmonization schemes (WDH and TTL) showed a performance improvement for any of the sites, compared to the unharmonized model (Figure 1b). Moreover, the average performance was similar for all the harmonization schemes including the No Target scheme (Table 2). This result suggests that the EoS signal was also discarded by the ML model, as it was not related to the target. Notably, consistent with our hypothesis, the No Target scheme did not remove important biological information, as this biologically relevant variance was presented across all sites.
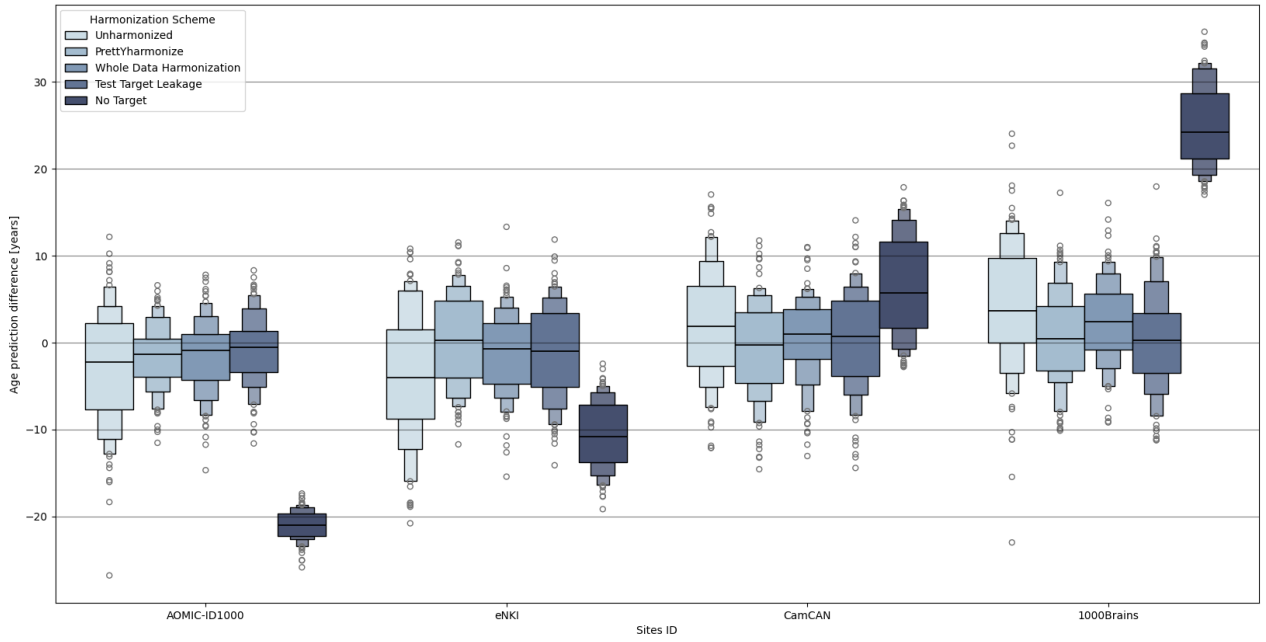
Table 2: Comparison of performance metrics across different harmonization schemes.

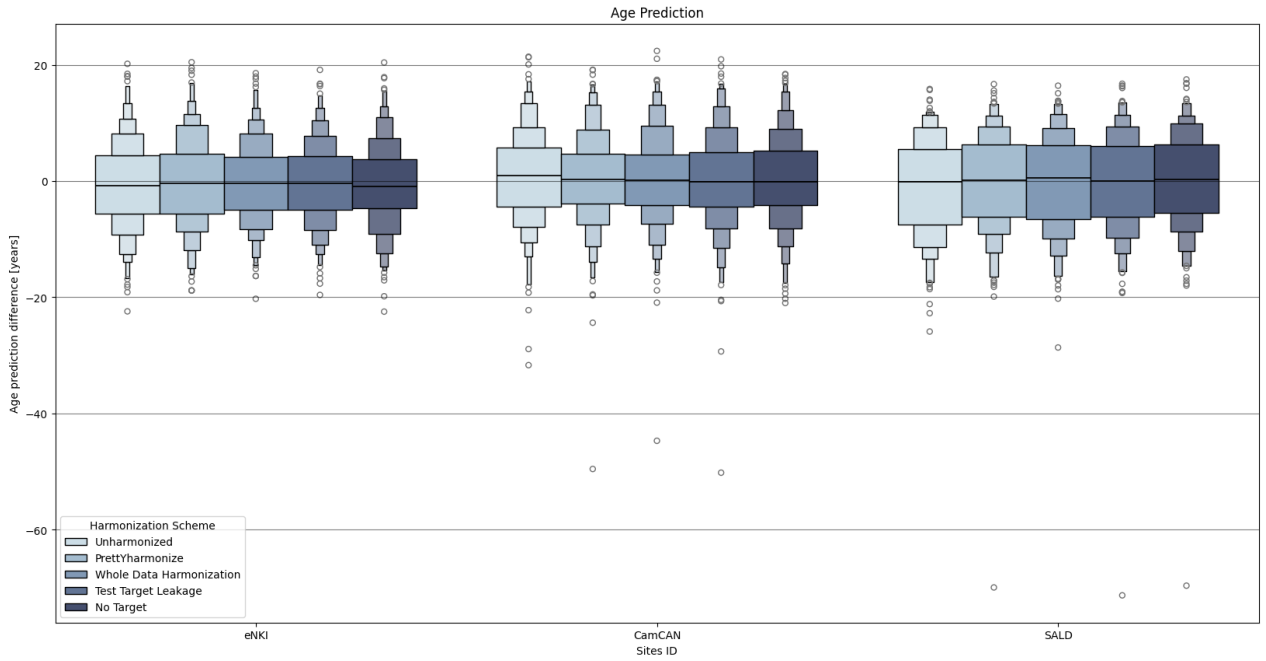| Metric	Unharmonized	PrettYharmonize	WDH	TTL	No Target
	Site-target dependence scenario				
MAE	6.20	4.12	3.82	4.28	15.93
R^2	0.81	0.919	0.925	0.912	-0.007
Age Bias	-0.43	-0.26	-0.32	-0.23	-0.998
	Site-target independence scenario				
MAE	6.314	6.306	6.034	6.153	6.036
R^2	0.785	0.769	0.803	0.775	0.790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

0.2.2 Sex classification

Two datasets [eNKI and CamCAN] containing healthy controls were used in this experiment were used for the forced site-target dependence. From the first one, 95% of females were retained, whereas for the second dataset only 5%, forcing a site-target dependence. In this case, the age range of the individuals was completely overlapped. The unharmonized scheme obtained a high performance (AUC = 0.97), which is similar to the performance reported in the literature[37]. The harmonization schemes that allow leakage (WDH and TTL) and PrettYharmonize did not show improvement compared to the unharmonized scheme (Table 3). This can be mainly because the features carry a strong signal related to the participant's sex, yielding high performance, even using unharmonized data. Consistent with the age regression experiment, the harmonization scheme without test labels (No Target), removed sex-related information significantly reducing the model's classification performance (Table 3).

Using the same generated dataset as in the age regression problem for site-target independence, a sex classification experiment was performed. The unharmonized scheme obtained a slightly lower (AUC=0.918) classification performance compared to the sex classification experiment with site-target dependence (Table 3). The harmonization schemes did not show classification improvement compared with the unharmonized model (Table 3). The No Target scheme does not remove target-related variance while harmonizing the features, explaining the similar performance of this model compared to the other schemes.

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

(b) Age regression on site-target dependence scenario

Figure 1: Age regression **a**) Site desegregated performance in site-target dependence scenarios. **b**) Site desegregated performance in site-target independence scenarios.

Table 3: Comparison of sex classification performance metrics across different harmonization schemes.

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

0.2.3 Dementia and mild cognitive impairment classification

For the site-target dependence scenarios, 100 dementia-MCI patients and 10 controls from one site and 100 controls and 10 dementia-MCI on a second site were selected from the ADNI dataset. The unharmonized method obtained an AUC of 0.81, consistent with other findings in the literature[38]. PrettYharmonize and the leakage-prone methods showed a slightly higher classification performance than the unharmonized method. As observed before, No Target removed important biological information, jeopardizing the ML model's performance (Table 4).

For the site-target independence scenario, the same number of control and patients were selected from both sites. In this scenario, all methods obtained a similar classification performance in all metrics (Table 4). Noteworthy, an important performance drop in all schemes is presented, compared with the site-target dependant experiment.

Table 4: Classification performance metrics for dementia-MCI prediction task in site-target dependent scenarios.

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

0.2.4 Discharge status prediction of septic patients

From the eICU dataset, 20 sites with more than 50 patients were selected. From these selected sites, the "Alive" patients were removed from 10 sites and the Expired patients were removed from the other 10 sites, forcing a site-target relationship. The unharmonized method obtained an AUC of 0.76, slightly lower than the one obtained in[39]. However, this difference is expected, as fewer patients were used in our experiments, compared with the reference study. PrettYharmonize obtained an important AUC performance improvement compared with all the benchmarked schemes. No Target scheme removed almost all the relevant information, obtaining a practically by-chance performance (Table 5).

For the site-target independence scenario, the same 20 sites selected were used, but the same number of Alive and Expired patients were retained in each site. All methods obtained the same classification performance in all metrics. The unharmonized method obtained a slightly lower classification performance (0.72 AUC) compared with the site-target dependence scenario (0.76 AUC) (Table 5). Furthermore, PrettYharmonize and the leakage-prone schemes (WDH and TTL) showed a great drop in classification performance, compared with the site-target dependence scenario. Finally, the No Target method did not remove important information while harmonizing the features, obtaining a similar performance as the rest of the benchmarked methods.

Table 5: Classification performance metrics for discharge status prediction task in site-target dependent and independent scenarios.

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

Discussion

Combining data from differing acquisitions is an appealing, and sometimes only, option for building ML models as they typically benefit from greater sample sizes. However, it is important to correctly integrate data harmonization methods, like the widely used ComBat, in ML pipelines[19]. In this study, we evaluated several ML pipelines incorporating ComBat-based data harmonization using a wide variety of biomedical data from different domains for both classification and regression tasks. Regarding the use of ComBat, feature distribution before and after harmonization is often analyzed to assess the effectiveness of the harmonization process. While this analysis can confirm that the features have been adjusted to exhibit more similar distributions, this analysis alone does not provide sufficient insight into how these adjustments impact the performance of ML models.

For all the evaluated scenarios, the ML pipelines using unharmonized data showed a performance close to the reported in the literature[36–39]. Importantly, for the site-target dependence scenarios, all the unharmonized models showed a better performance compared with the site-target independent scenarios. This is expected as the ML models can to pick EoS signal, which is related to the target in the site-target dependence scenarios, and use it to fraudulently increase the classification performance.

We observed that ComBat-based harmonization struggles to provide benefits when the site and target variables are independent, even when allowing leakage and the target variance was preserved in ComBat modelling. None of the harmonization-based ML pipelines showed performance improvement over the baseline of pooling the data together with site-target independence. This can be explained because the harmonization may be removing a signal that it is also discarded by the ML models, as it is not related to the target. In this case, the benefit of removing the Effect of Site (EoS), did not improve the signal-to-noise ratio of the signal enough

to benefit the performance of the ML models. Despite the wide variety of tasks and data from different domains we tested, ComBat-based harmonization does not seem to provide an advantage. Nevertheless, it is possible that our data and task selection, albeit comprehensive, does not include cases where harmonization can be indeed beneficial.

On the other hand, on the site-target dependence senarios, a performance improvement was observed when when applying ComBat and allowing the target variance to be preserved. Problematically, in conventional ways to integrate ComBat in ML pipelines (Whole Data Harmonization and Test Target Leakage), this leads to data leakage as the test labels are used. Whereas not explicitly preserving the target (No Target), the performance was consistently worse. These observations can be explained as ComBat works on the assumption that the site-specific variance and variance of interest are independent. However, this assumption is violated when the site and target are dependent. Consequently, when the site and target are dependent, using ComBat without explicitly preserving the target variance can remove variance related to the target.

The proposed a new method called PrettYharmonize avoids leakage by design, as it relays in the harmonization using pretended target labels and a Stack model, which combines the predictions made with the different harmonized data. In this way, the method can circumvent the need for target values of the test samples. PrettYharmonize was validated on the MAREoS datasets which were specifically devised for this purpose. The solid results obtained in this dataset indicate that the proposed method effectively harmonizes data without compromising model integrity. The method provides a leakage-free pipeline which in our evaluations showed performs at par with leakage-prone pipelines. These findings suggest that PrettYharmonize holds promise for real-world deployment, particularly in contexts where data leakage is a critical concern. Therefore, we recommend PrettYharmonize for future use cases. Overall, our results suggest that future studies should carefully evaluate their ML pipelines and follow reproducible and open science practices for the benefit of the community.

Alternative approaches such as calculating ComBat parameters using phantoms[18] can harmonize data independent of biological variability. By doing so, the location and scale parameters specific to each MRI setup and parameter setting can be accurately estimated and applied to real data. This approach mitigates the risk of inadvertently removing meaningful biological variation during harmonization. However, such approaches are domain-specific and incur challenges such as measuring additional data.

Finally, although our study focused on widely used ComBat-based methods, it is important to acknowledge the existence of other harmonization techniques, such as deep learning-based methods like style-matching generative models or variational autoencoders[4]. These approaches may offer promising alternatives, particularly in complex scenarios where traditional methods may fall short though they tend to be data and compute-intensive.

Limitations of the study

This study has some limitations that should be considered when interpreting the results. First, our analysis focuses primarily on ComBat-based harmonization methods due to their widespread use; however, we did not extensively explore other emerging techniques such as deep learning, optimal transport, or style-matching generative models, which offer different strengths and weaknesses.

Second, the impact of harmonization on feature selection and model interpretability was not deeply explored, warranting further investigation into how these methods influence model behavior in different contexts.

Third, even though in a controlled fashion we simulated somewhat extreme site-target dependence and independence scenarios, it is safe to assume that any real-case scenario will fall in

between. Our aim in this study was to empirically demonstrate the problems that may occur while harmonizing the data without factoring in appropriate considerations. Further research should be conducted to better measure the relationship between the site-target dependence degree and the harmonization impact.

Finally, regarding PrettYharmonize, the proposed method is more computationally expensive than the traditional harmonization schemes because in the "pretending" process the data need to be harmonized several times.

Experimental procedures

Resource availability

Lead contact

Please send any inquiry to the corresponding author, Nicolás Nieto (n.nieto@fz-juelich.de)

Materials availability

No materials were used in this work.

Data and code availability

All used MRI datasets are publicly available possibly upon registration. For the eICU dataset, data 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 library is publicly available at: `https://github.com/juaml/PrettYharmonize`. The scripts to replicate the experiments and replicate the processing of the datasets are available at: `https://github.com/juaml/harmonize_project`

Data description

MAREoS dataset

To ensure the validity of PrettYharmonize, we benchmarked it in a classification problem using the datasets specifically designed to evaluate harmonization models[3]. This MAREoS dataset consists of eight datasets simulating 18 MRI features (cortical thickness, cortical surface area, or subcortical volumes). Four datasets contain a "True" signal and four only contain an Effect of Site ("EoS") signal related to a binary target. In that sense, an ML model that learns the "EoS" signal can fraudulently achieve a good classification performance. The signal, both the True or EoS, are called "Simple" and "Interactions", depending on a linear or non-linear effect, respectively. Within each dataset, approximately 1000 samples, coming from 8 sites, were simulated. 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%).

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[40] 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)[41], The Enhanced Nathan Kline Institute (eNKI)[42], Cambridge Centre for Ageing Neuroscience (CamCAN)[43], 1000Brains[44], and the Southwest University Adult Lifespan Dataset (SALD)[45]. These datasets were selected as the data within each dataset was acquired only in one site thus avoiding additional confounding. The demographic information of these datasets is presented in Table 6.

Table 6: 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%

Age regression

0.2.4.1 Forced site-target dependence.

Four datasets, AOMIC-ID1000, eNKI, CamCAN, and 1000Brains, were randomly subsampled in different age ranges forcing a site-target dependence. The subsample was performed to ensure the same amount of subjects by each sex in each dataset (Table 7).

Table 7: Dataset characteristics for site-target dependent age regression experiment.

Dataset Name	N Images	Mean Age	Std Age	Min Age	Max Age	% Female
AOMIC-ID1000	118	22.73	1.61	19	26	50%
eNKI	118	33.00	3.95	27	40	50%
CamCAN	118	50.09	6.06	41	60	50%
1000Brains	118	68.74	5.02	61	79	50%

0.2.4.2 Forced site-target independence.

Three datasets, CamCAN, eNIKI, and SALD, were used. The datasets were selected as they contain individuals covering a wide range of ages above 18. The AOMIC and 1000Brains datasets were excluded as those datasets mainly included young and old participants, respectively. Each dataset was balanced in terms of sex and age (Table 8). This was achieved by retaining the same number of subjects for each sex in 10 equally distributed age ranges, from the minimum to the maximum age in each dataset.

Table 8: Dataset characteristics for site-target independent age regression and sex classification experiments.

Dataset	N Images	Mean Age	Std Age	Min Age	Max Age	% Female
SALD	200	48.99	16.97	19	77	50%
eNKI	300	47.75	17.43	18	78	50%
CamCAN	288	48.60	18.00	18	78	50%

0.2.5 Sex classification

For this experiment, only the eNKI and CamCAN datasets were used, as those present a broad and similar age range. In this case, the percentages of females in each dataset were forced to be 95% in eNKI and 5% in CamCAN (Table 9). Additionally, the same number of images for each dataset was retained.

Table 9: Dataset characteristics for site-target dependent sex classification experiment.

Dataset Name	N Images	Mean Age	Std Age	Min Age	Max Age	% Female
eNKI	295	45.13	18.93	19	84	5%
CamCAN	295	53.77	18.51	18	88	95%

0.2.5.1 Forced site-target independence

The same dataset generated in the site-target independence scenario for age regression was used for sex classification.

0.2.6 Dementia and mild cognitive impairment classification

0.2.6.1 Forced site-target dependence

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 assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

In our experiments using the ADNI dataset, where the data were acquired in different sites, we selected 100 dementia-MCI patients and 10 controls from one site. We selected 100 control patients and 10 dementia-MCi patients from another site, again forcing the site-target relationship (Table 10). The images were processed with FreeSurfer[46]. The thickness from 74 cerebral and sub-cerebral structures were extracted as features.

Table 10: Datasets characteristics for site-target dependent Dementia-MCI classification experiment

Dataset name	N Images	Mean Age	Std Age	Min Age	Max Age	% Dementia-MCI
Site 1	110	75.182	6.667	59	92	9 %
Site 2	110	72.331	5.560	60	97	91 %

Table 11: Datasets characteristics for site-target independent dementia-MCI classification experiment

Dataset ID	N Images	Mean Age	Std Age	Min Age	Max Age	% Dementia-MCI
Site 1	252	73.72	6.582	59	93	50 %
Site 2	114	72.68	6.448	56	96	50 %

0.2.6.2 Forced site-target independence

Using the ADNI dataset the same extracted features were used. From the dataset, 126 dementia-MCI and control patients were randomly selected from the first site while 56 dementia-MCI and control patients were randomly selected from the second site (Table 11).

0.2.7 Discharge status prediction of septic patients

0.2.8 Forced site-target dependence

The eICU[31,32,47] dataset was used for the experiments, which contains 200859 ICU stays from 139367 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[48,49]. The approach described in[39] 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. After the patients' selection, a final dataset of 496 Expired and 3021 Alive patients was retained. From this filtered dataset, we remove those sites with less than 50 stays, retaining 20 final sites.

From 20 of these sites, all "Alive" patients, except for one, were removed for the 10 sites with more "Expired" patients. Contrary, all the "Expired" patients, except for one, were removed from the 10 sites with less number of "Expired" patients. Note that in this case, the classes (Alive and Expired) are represented in several sites and are not the same in each site (Table 12), compared with the previous classification experiments performed. A total of 249 Expired and 666 Alive patients were used in this experiment.

0.2.9 Forced site-target independence

From the eICU, the same features extraction and patient selection were made (496 Expired and 3021 Alive patients). The same 20 sites, with more than 50 images, were used.

From all sites, the same number of "Alive" and "Expired" patients were retained. A total of 324 Expired and 324 Alive patients were used in this experiment (Table 13).

Methods

PrettYHarmonize

The proposed method rests on the use of "pretended" target values on the harmonization process, thereby enabling predictions without requiring test target values (Figure 2). For our experiments, the available data is divided into train and test, simulating a real use case. The training fold is further divided into inner train and validation folds. A neuroHarmonize model is trained on the inner training data to learn to remove the site's effect. The inner training data is harmonized and a Predictive model is trained on the harmonized inner train data to predict the target. This model needs to be chosen as the best possible model to solve the problem at hand.

Table 12: Datasets characteristics for site-target dependent outcome prediction on septic patients experiment

Dataset ID	N Images	Alive Count	Expired Count	% Expired
Site 79	92	91	1	1.08 %
Site 148	69	68	1	1.14 %
Site 15	56	55	1	1.17 %
Site 157	14	1	13	92.85 %
Site 165	15	1	14	93.33 %
Site 167	21	1	20	95.54 %
Site 176	111	110	1	0.90 %
Site 188	32	1	31	96.87 %
Site 248	55	54	1	1.82 %
Site 252	27	1	26	96.30 %
Site 264	17	1	16	94.12 %
Site 300	69	68	1	1.45 %
Site 345	55	54	1	1.82 %
Site 365	58	57	1	1.72 %
Site 416	15	1	14	93.33 %
Site 420	65	1	64	98.46 %
Site 443	58	57	1	1.72 %
Site 449	18	1	17	94.44 %
Site 452	43	42	1	2.33 %
Site 458	24	1	23	95.83 %
Total	915	666	249	27.21 %

Table 13: Datasets characteristics for site-target independent outcome prediction on septic patients experiment

Dataset ID	N Images	Alive Count	Expired Count	% Expired
Site 79	20	10	10	50 %
Site 148	24	12	12	50 %
Site 154	18	9	9	50 %
Site 157	26	13	13	50 %
Site 165	28	14	14	50 %
Site 167	42	21	21	50 %
Site 176	14	7	7	50 %
Site 188	62	31	31	50 %
Site 248	26	13	13	50 %
Site 252	52	26	26	50 %
Site 264	32	16	16	50 %
Site 300	12	6	6	50 %
Site 345	8	4	4	50 %
Site 365	8	4	4	50 %
Site 416	28	14	14	50 %
Site 420	128	64	64	50 %
Site 443	22	11	11	50 %
Site 449	34	17	17	50 %
Site 452	18	9	9	50 %
Site 458	46	23	23	50 %
Total	648	324	324	50 %

Figure 2: PrettYharmonize training workflow. The workflow showcases the training workflow for a binary classification problem.

Using the trained neuroHarmonize model, the validation samples are harmonized while "pretending" their target value. For example, for a binary classification problem, all the validation labels are set as the first class, pretending that all validation samples belong to the first class. Using these "pretended" labels, the validation data is harmonized and a prediction is made using the trained Predictive model. Later, the validation labels are set to the second class and the validation data is harmonized again and a new prediction is generated. In general, for a classification task, the set of available classes is pretended, while for a regression task, the values are linearly sampled in the target's range. All the predictions, generated with the pretended harmonized data, are concatenated and a "Score matrix" is created. This matrix has a dimension of number of validation samples times the number of labels. To effectively utilize the training dataset, a K-fold cross-validation (CV) procedure was employed, generating out-of-sample predictions for the entire dataset. Using this Score matrix as input features, a "Stack" model is trained to predict the target and give a final prediction.

At the test time, when the test label is not available, the same procedure is followed. For example, in a binary classification problem, a test sample will be harmonized using the neuroHarmonize model first pretending that the test label belongs to the first class. The Predict model will generate a prediction using the harmonized data and the process will be repeated pretending the test sample belongs to the second class. Both predictions generated by the Predictive model are concatenated and a test Score matrix is built. This matrix is used by the Stack model, which generates the final prediction.

0.3 Machine learning model

For the binary classification problem using the synthetic data (MAREoS datasets), a Random Forest model (RF)[50], with default sklearn parameters, was used as a Predictive Model, and a Logistic Regression (LG)[51] was used as a 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).

For the age regression problems using real MRI data, Relevance Vector Regression[52] with a polynomial kernel of degree 1 (RVR) was used as a Predictive and Stack model for PrettYharmonize. The RVR model was used for the rest of the harmonization schemes[36]. A 5-fold cross-validation scheme was used, and the Mean Absolute Error (MAE), coefficient of determination (R2), 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.

For the sex classification using real MRI data, RVR was used as a Predictive and Stack model. A 5 times repeated 5-fold stratified cross-validation scheme was used, and the Area under the receive operation curve (AUC), balanced accuracy (bACC) and were calculated on the test

Supplemental information

In the Supplementary Information Tables S1-S3 and Figures S1-S5 and their legends are presented.

Acknowledgments

Author contributions

NN: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing – original draft, and writing – review and editing

SE: conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, writing – review and editing

CJ: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, visualization, writing – review and editing

MR: funding acquisition, resources, review and editing

KD: resources, review and editing

MK: funding acquisition, resources, review and editing

AL: funding acquisition, resources, review and editing

FR: conceptualization, formal analysis, investigation, methodology, resources, software, supervision, validation, visualization, writing – original draft, and writing – review and editing

KP: conceptualization, funding acquisition, formal analysis, investigation, methodology, project administration, supervision, validation, visualization, writing – original draft, and writing – review and editing

Declaration of interests

The authors declare no competing interests.

References

1. Hosseini, S. A., Shiri, I., Ghaffarian, P., Hajianfar, G., Avval, A. H., Seyfi, M., Servaes, S., Rosa-Neto, P., Zaidi, H., and Ay, M. R. (2024). The effect of harmonization on the variability of pet radiomic features extracted using various segmentation methods. Annals of Nuclear Medicine (1–15).

2. Chen, J., Liu, J., Calhoun, V. D., Arias-Vasquez, A., Zwiers, M. P., Gupta, C. N., Franke, B., and Turner, J. A. (2014). Exploration of scanning effects in multi-site structural mri studies. Journal of neuroscience methods *230*, 37–50.

3. Solanes, A., Gosling, C. J., Fortea, L., Ortuño, M., Lopez-Soley, E., Llufriu, S., Madero, S., Martinez-Heras, E., Pomarol-Clotet, E., Solana, E. et al. (2023). Removing the effects of the site in brain imaging machine-learning–measurement and extendable benchmark. NeuroImage *265*, 119800.

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

5. Li, H., Smith, S. M., Gruber, S., Lukas, S. E., Silveri, M. M., Hill, K. P., Killgore, W. D., and Nickerson, L. D. (2020). Denoising scanner effects from multimodal mri data using linked independent component analysis. Neuroimage *208*, 116388.

6. Wachinger, C., Rieckmann, A., Pölsterl, S., Initiative, A. D. N. et al. (2021). Detect and correct bias in multi-site neuroimaging datasets. Medical Image Analysis *67*, 101879.

7. Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J. et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. Neuroimage *167*, 104–120.

8. Acquitter, C., Piram, L., Sabatini, U., Gilhodes, J., Moyal Cohen-Jonathan, E., Ken, S., and Lemasson, B. (2022). Radiomics-based detection of radionecrosis using harmonized multiparametric mri. Cancers *14*, 286.

9. Li, Y., Ammari, S., Balleyguier, C., Lassau, N., and Chouzenoux, E. (2021). Impact of pre-processing and harmonization methods on the removal of scanner effects in brain mri radiomic features. Cancers *13*, 3000.

10. Ingalhalikar, M., Shinde, S., Karmarkar, A., Rajan, A., Rangaprakash, D., and Deshpande, G. (2021). Functional connectivity-based prediction of autism on site harmonized abide dataset. IEEE Transactions on Biomedical Engineering *68*, 3628–3637.

11. Maikusa, N., Zhu, Y., Uematsu, A., Yamashita, A., Saotome, K., Okada, N., Kasai, K., Okanoya, K., Yamashita, O., Tanaka, S. C. et al. (2021). Comparison of traveling-subject and combat harmonization methods for assessing structural brain characteristics. Human brain mapping *42*, 5278–5287.

12. Da-Ano, R., Visvikis, D., and Hatt, M. (2020). Harmonization strategies for multicenter radiomics investigations. Physics in Medicine & Biology *65*, 24TR02.

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

14. Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E. et al. (2017). Harmonization of multi-site diffusion tensor imaging data. Neuroimage *161*, 149–170.

15. Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I., Satterthwaite, T., Fan, Y. et al. (2019). Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. neuroimage, 208, article 116450.

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

17. Dudley, J. A., Maloney, T. C., Simon, J. O., Atluri, G., Karalunas, S. L., Altaye, M., Epstein, J. N., and Tamm, L. (2023). Abcd_harmonizer: An open-source tool for mapping and controlling for scanner induced variance in the adolescent brain cognitive development study. Neuroinformatics *21*, 323–337.

18. Ibrahim, A., Primakov, S., Beuque, M., Woodruff, H., Halilaj, I., Wu, G., Refaee, T., Granzier, R., Widaatalla, Y., Hustinx, R. et al. (2021). Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. Methods *188*, 20–29.

19. Marzi, C., Giannelli, M., Barucci, A., Tessa, C., Mascalchi, M., and Diciotti, S. (2024). Efficacy of mri data harmonization in the age of machine learning: a multicenter study across 36 datasets. Scientific Data *11*, 115.

20. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.

21. Barth, C., Kelly, S., Nerland, S., Jahanshad, N., Alloza, C., Ambrogi, S., Andreassen, O. A., Andreou, D., Arango, C., Baeza, I. et al. (2023). In vivo white matter microstructure in adolescents with early-onset psychosis: a multi-site mega-analysis. Molecular Psychiatry *28*, 1159–1169.

22. Bourbonne, V., Jaouen, V., Nguyen, T. A., Tissot, V., Doucet, L., Hatt, M., Visvikis, D., Pradier, O., Valéri, A., Fournier, G. et al. (2021). Development of a radiomic-based model predicting lymph node involvement in prostate cancer patients. Cancers *13*, 5672.

23. Campello, V. M., Martín-Isla, C., Izquierdo, C., Guala, A., Palomares, J. F. R., Viladés, D., Descalzo, M. L., Karakas, M., Çavuş, E., Raisi-Estabragh, Z. et al. (2022). Minimising multi-centre radiomics variability through image normalisation: a pilot study. Scientific reports *12*, 12532.

24. Chen, P., Yao, H., Tijms, B. M., Wang, P., Wang, D., Song, C., Yang, H., Zhang, Z., Zhao, K., Qu, Y. et al. (2023). Four distinct subtypes of alzheimer's disease based on resting-state connectivity biomarkers. Biological Psychiatry *93*, 759–769.

25. Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., Eickhoff, S. B., Götz, M., Hamdan, S., Komeyer, V., Kulkarni, A., Lahnakoski, J., Love, B. C. et al. (2023). On leakage in machine learning pipelines. arXiv preprint arXiv:2311.04179.

26. Lones, M. A. (2021). How to avoid machine learning pitfalls: a guide for academic researchers. arXiv preprint arXiv:2108.02497.

27. Kapoor, S., and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns *4*.

28. Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quidé, Y., Green, M. J., Weickert, C. S., Weickert, T., Bruggemann, J., Kircher, T. et al. (2020). Increased power by harmonizing structural mri site differences with the combat batch adjustment method in enigma. Neuroimage *218*, 116956.

29. Nygaard, V., Rødland, E. A., and Hovig, E. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics *17*, 29–39.

30. Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C. et al. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine *27*, 685–691.

31. Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. (2018). The eicu collaborative research database, a freely available multi-center database for critical care research. Scientific data *5*, 1–13.

32. Pollard, T., Johnson, A., Raffa, J., Celi, L. A., Badawi, O., and Mark, R. (2019). eicu collaborative research database (version 2.0). PhysioNet *10*, C2WM1R.

33. Yang, Z., Cui, X., and Song, Z. (2023). Predicting sepsis onset in icu using machine learning models: a systematic review and meta-analysis. BMC infectious diseases *23*, 635.

34. Wu, M., Du, X., Gu, R., and Wei, J. (2021). Artificial intelligence for clinical decision support in sepsis. Frontiers in Medicine *8*, 665464.

35. Zhang, Y., Xu, W., Yang, P., and Zhang, A. (2023). Machine learning for the prediction of sepsis-related death: a systematic review and meta-analysis. BMC Medical Informatics and Decision Making *23*, 283.

36. More, S., Antonopoulos, G., Hoffstaedter, F., Caspers, J., Eickhoff, S. B., Patil, K. R., Initiative, A. D. N. et al. (2023). Brain-age prediction: A systematic comparison of machine learning workflows. NeuroImage *270*, 119947.

37. Flint, C., Förster, K., Koser, S. A., Konrad, C., Zwitserlood, P., Berger, K., Hermesdorf, M., Kircher, T., Nenadic, I., Krug, A. et al. (2020). Biological sex classification with structural mri data shows increased misclassification in transgender women. Neuropsychopharmacology *45*, 1758–1765.

38. Illakiya, T., and Karthik, R. (2023). Automatic detection of alzheimer's disease using deep learning models and neuro-imaging: current trends and future perspectives. Neuroinformatics *21*, 339–364.

39. Wernly, B., Mamandipoor, B., Baldia, P., Jung, C., and Osmani, V. (2021). Machine learning predicts mortality in septic patients using only routinely available abg variables: a multi-centre evaluation. International journal of medical informatics *145*, 104312.

40. Gaser, C., Dahnke, R., Thompson, P. M., Kurth, F., Luders, E., and Initiative, A. D. N. (2022). Cat–a computational anatomy toolbox for the analysis of structural mri data. biorxiv (2022–06).

41. Snoek, L., van der Miesen, M. M., Beemsterboer, T., Van Der Leij, A., Eigenhuis, A., and Steven Scholte, H. (2021). The amsterdam open mri collection, a set of multimodal mri datasets for individual difference analyses. Scientific data *8*, 85.

42. Nooner, K. B., Colcombe, S. J., Tobe, R. H., Mennes, M., Benedict, M. M., Moreno, A. L., Panek, L. J., Brown, S., Zavitz, S. T., Li, Q. et al. (2012). The nki-rockland sample: a model for accelerating the pace of discovery science in psychiatry. Frontiers in neuroscience *6*, 152.

43. Shafto, M. A., Tyler, L. K., Dixon, M., Taylor, J. R., Rowe, J. B., Cusack, R., Calder, A. J., Marslen-Wilson, W. D., Duncan, J., Dalgleish, T. et al. (2014). The cambridge centre for ageing and neuroscience (cam-can) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. BMC neurology *14*, 1–25.

44. Caspers, S., Moebus, S., Lux, S., Pundt, N., Schütz, H., Mühleisen, T. W., Gras, V., Eickhoff, S. B., Romanzetti, S., Stöcker, T. et al. (2014). Studying variability in human brain aging in a population-based german cohort—rationale and design of 1000brains. Frontiers in aging neuroscience *6*, 149.

45. Wei, D., Zhuang, K., Ai, L., Chen, Q., Yang, W., Liu, W., Wang, K., Sun, J., and Qiu, J. (2018). Structural and functional brain scans from the cross-sectional southwest university adult lifespan dataset. Scientific data *5*, 1–10.

46. Fischl, B. (2012). Freesurfer. Neuroimage *62*, 774–781.

47. Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. circulation *101*, e215–e220.

48. Hou, N., Li, M., He, L., Xie, B., Wang, L., Zhang, R., Yu, Y., Sun, X., Pan, Z., and Wang, K. (2020). Predicting 30-days mortality for mimic-iii patients with sepsis-3: a machine learning approach using xgboost. Journal of translational medicine *18*, 1–14.

49. Deng, H.-F., Sun, M.-W., Wang, Y., Zeng, J., Yuan, T., Li, T., Li, D.-H., Chen, W., Zhou, P., Wang, Q. et al. (2022). Evaluating machine learning models for sepsis prediction: A systematic review of methodologies. Iscience *25*.

50. Breiman, L. (2001). Random forests. Machine learning *45*, 5–32.

51. Tolles, J., and Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. Jama *316*, 533–534.

52. Tipping, M. (1999). The relevance vector machine. Advances in neural information processing systems *12*.