HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2023 July 01.

Published in final edited form as:

Neuroimage. 2023 July 01; 274: 120125. doi:10.1016/j.neuroimage.2023.120125.

Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization

Fengling Hu[a,*], **Andrew A. Chen**[a], **Hannah Horng**[a], **Vishnu Bashyam**[b], **Christos Davatzikos**[b], **Aaron Alexander-Bloch**[c,d,e], **Mingyao Li**[f], **Haochang Shou**[a,b], **Theodore D. Satterthwaite**[c,d,g], **Meichen Yu**[h,#], **Russell T. Shinohara**[a,b,#]

[a]Penn Statistics in Imaging and Visualization Endeavor (PennSIVE), Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Dr, Philadelphia, PA 19104, United States

[b]Center for Biomedical Image Computing and Analytics (CBICA), Perelman School of Medicine, United States

[c]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, United States

[d]Penn-CHOP Lifespan Brain Institute, United States

[e]Department of Child and Adolescent Psychiatry and Behavioral Science, Children's Hospital of Philadelphia, United States

[f]Statistical Center for Single-Cell and Spatial Genomics, Perelman School of Medicine, University of Pennsylvania, United States

[g]The Penn Lifespan Informatics and Neuroimaging Center, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, United States

[h]Indiana Alzheimer's Disease Research Center, Indiana University School of Medicine, United States

[*]Corresponding author. fengling.hu@pennmedicine.upenn.edu (F. Hu).
[#]Asterisks indicate equal contribution to this work.

Data and code availability statement

No data or code was used in the preparation of this manuscript.

Declaration of Competing Interest

RTS receives consulting income from Octave Bioscience and compensation for reviewership duties from the American Medical Association. The authors report no conflicts of interest.

Credit authorship contribution statement

Fengling Hu: Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Andrew A. Chen:** Investigation, Writing – original draft, Writing – review & editing, Visualization. **Hannah Horng:** Investigation, Writing – original draft, Writing – review & editing, Visualization. **Vishnu Bashyam:** Investigation, Writing – review & editing. **Christos Davatzikos:** Writing – review & editing, Supervision, Funding acquisition. **Aaron Alexander-Bloch:** Investigation, Writing – review & editing, Supervision. **Mingyao Li:** Investigation, Writing – review & editing, Supervision. **Haochang Shou:** Investigation, Writing – review & editing, Supervision. **Theodore D. Satterthwaite:** Investigation, Writing – review & editing, Supervision. **Meichen Yu:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Supervision. **Russell T. Shinohara:** Conceptualization, Investigation, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Abstract

Magnetic resonance imaging and computed tomography from multiple batches (e.g. sites, scanners, datasets, etc.) are increasingly used alongside complex downstream analyses to obtain new insights into the human brain. However, significant confounding due to batch-related technical variation, called batch effects, is present in this data; direct application of downstream analyses to the data may lead to biased results. Image harmonization methods seek to remove these batch effects and enable increased generalizability and reproducibility of downstream results. In this review, we describe and categorize current approaches in statistical and deep learning harmonization methods. We also describe current evaluation metrics used to assess harmonization methods and provide a standardized framework to evaluate newly-proposed methods for effective harmonization and preservation of biological information. Finally, we provide recommendations to end-users to advocate for more effective use of current methods and to methodologists to direct future efforts and accelerate development of the field.

1. Introduction

Brain imaging acquired via magnetic resonance imaging (MRI) or computed tomography (CT) from multiple batches, such as different sites or scanners, has shown promise in providing increased sample sizes for imaging-based neuroscience studies, prediction efforts, and more (Bethlehem et al., 2022; Casey et al., 2018; Choudhury et al., 2014; Di Martino et al., 2014; Horn et al., 2004; Marek et al., 2022; Mueller et al., 2005; Poldrack and Gorgolewski, 2014; van Erp et al., 2014; Van Essen et al., 2013). These multi-batch neuroimaging data are known to suffer from non-biological, technical variability between subjects from different batches, which we refer to as batch effects. Batch effects can be due to differences in acquisition protocol, magnetic field strength, scanner manufacturer, scanner drift, hardware imperfections, and more (Badhwar et al., 2020; Byrge et al., 2022; Cai et al., 2021; Han et al., 2006; Jovicich et al., 2006; Shinohara et al., 2017; Takao et al., 2014, 2011). These batch effects may explain, in part, challenges with reproducibility of neuroscience studies, generalizability of prediction algorithms, and incorporation of radiomics-derived imaging biomarkers in clinical practice (Crombé et al., 2021; Fournier et al., 2021; Mårtensson et al., 2020; Schwarz, 2021; Thieleking et al., 2021). Notably, batch effects have been shown to be significantly easier to detect than biological effects, both by statistical testing and machine learning algorithms (Bell et al., 2022; Fortin et al., 2018, 2017; Nielson et al., 2018). Additionally, due to the complex nature of batch effects, traditional statistical techniques for adjusting for confounders, such as inclusion of batch in a linear model as a mean effect, may be inadequate to sufficiently account for batch effects.

There is also growing interest in using neuroimaging to evaluate new treatments across a range of neurologic, psychiatric, and other clinical trials (Cash et al., 2014; Dercle et al., 2022; Polman et al., 2006; Saunders et al., 2016; Tariot et al., 2011; Tondelli et al., 2020; van Dyck et al., 2023). While clinical trial treatments are usually randomized within batches such that conclusions from unharmonized images are asymptotically unbiased, prespecified approaches to account for known confounders, including batch, allow for increased power and improved estimation of treatment effects (Hernández et al., 2006, 2004; Kent et al., 2009; Neuhaus, 1998; Optimising the Analysis of Stroke Trials (OAST) Collaboration

et al., 2009). This is especially important when randomized treatment assignments are not completely balanced within each batch. Ultimately, in clinical trials where imaging biomarkers are measured across multiple centers, addressing batch effects allows for the detection of smaller treatments effects while requiring fewer required subjects, minimizing participant burden, and reducing costs.

In observational settings where batch effects are present, such as when multiple small neuroimaging datasets are aggregated into one larger sample, addressing batch effects is even more important to obtain valid conclusions (Grech-Sollars et al., 2015; Keshavan et al., 2016; Stonnington et al., 2008; Takao et al., 2014). In these settings, failure to account for the known confounding of batch effects may lead to decreased power, less replicable findings, and potentially-biased findings. Effective removal of batch effects has been shown to enable detection of otherwise-undetected biological effects as well as increase the replicability of biological effects of interest in simulations of discovery-validation study designs (Bashyam et al., 2022; Bell et al., 2022; Carré et al., 2022; Fortin et al., 2017; Zhang et al., 2022; Zuo et al., 2021). Additionally, when batch-wise differences in participant populations are present, failure to address batch effects may result in biased conclusions (Suttorp et al., 2015).

Various solutions have been proposed and implemented to address this problem at different points in data collection and analysis pipelines. For example, in study design, batch effects can be minimized by collecting data from only one scanner, one manufacturer, one field strength, one acquisition protocol, or some combination of these criteria (Clarke et al., 2020; De Stefano et al., 2022; Ihalainen et al., 2004; Malyarenko et al., 2013; Meeter et al., 2017; Satterthwaite et al., 2014; van de Bank et al., 2015; Vogelbacher et al., 2021). However, when data collection is limited to only one batch, it is challenging to collect large sample sizes, and design-based solutions cannot address batch effects in data that has already been collected (Harms et al., 2018). Additionally, even when acquisition properties or scanner manufacturer are tightly controlled, batch effects can still arise due to residual differences, such as hardware imperfections, site or operator characteristics, software or hardware upgrades in long-running studies, or otherwise non-controllable scanner properties (Jovicich et al., 2016; Shinohara et al., 2017).

At other stages of the data analysis pipeline, such as during the image pre-processing step, standardization of images using methods for gradient distortion correction, bias field correction, and intensity normalization can also reduce batch effects (Brown et al., 2020; Fortin et al., 2016; Guan et al., 2022; Hellier, 2003; Jovicich et al., 2006; Nyúl and Udupa, 1999; Shinohara et al., 2014; Tustison et al., 2010; Wang et al., 1998; Wrobel et al., 2020). These normalization methods act on intersubject variability without explicitly modeling batch effects, and as a result, can only reduce batch effects that coincide with inter-subject variability.

Additionally, some approaches account for batch effects using batch-aware downstream statistical or machine learning analyses. For example, data aggregation can be carried out in post-analysis through the use of meta-analysis or mega-analysis techniques, where estimates of interest are first calculated within batches and then analyzed jointly (Jahanshad

et al., 2013). In certain settings, the simple approach of training models on large datasets across many batches can be considered, as these models are theoretically able to learn generalizable parameters that are invariant to batch, especially if the models are able to explicitly incorporate batch status. This approach has been used in normative modeling settings (Bayer et al., 2022a; Bethlehem et al., 2022; Kia et al., 2020; Kim et al., 2022). However, in many prediction or classification settings, complex machine learning algorithms are used that are not able to learn batch-invariant decision boundaries; in these settings, if outcome distributions differ across batches, models may incorrectly learn to use batch effects to make predictions. Here, transfer learning approaches have been used (Aderghal et al., 2020; Chen et al., 2020; Dar et al., 2020; He et al., 2021; Yang et al., 2019). In transfer learning, instead of reducing batch effects in the data itself, these methods seek to train deep learning models in a reference batch and then recalibrate these models for prediction in new batches.

Finally, batch effects can be explicitly modeled for and addressed in image pre-processing, such that raw data is mapped from multiple batches into one common batch and the resulting harmonized dataset can then be analyzed as if it originated from a common batch. We refer to this process as image harmonization, which is the focus of this review.

This review is broadly organized into four sections. In the first and second sections, we describe statistical harmonization methods and deep learning harmonization methods, respectively. These two sections are additionally subdivided based on whether methods are designed for retrospective or prospective study designs. We define prospective study designs as those where some subjects, commonly called “traveling subjects,” are purposefully scanned across multiple batches within a short time interval; these paired data across batches can then be used to facilitate harmonization of these batches at the time of analysis. In retrospective study designs, no such paired data are available. In the third section, we discuss the evaluation of harmonization methods, including the various domains under which harmonization should be evaluated as well as specific tests to perform that evaluation. Finally, in the fourth section, we provide recommendations to both end-users and methodologists. For end-users, we suggest harmonization methods for each data type and study design based on ease of use, theoretical behavior, and empirical validation. For methodologists, we provide guidance for further work in harmonization, a standardized framework of evaluation, and improved comparability of novel harmonization methods.

2. Literature search

We performed a literature search across the PubMed database using the following search term: (“magnetic resonance” OR “MRI”) AND (“harmonization” OR “harmonizing” OR “harmonize” OR “harmonisation” OR “harmonising” OR “harmonise” OR “scanner effect” OR “site effect” OR “batch effect” OR “batch correct” OR “domain effect” OR “domain transfer” OR “technical variability” OR “style transfer”).

This search returned 583 candidate publications, as of January 17th, 2023, which were screened by title and abstract. Publications were included if they proposed or validated a statistical or deep learning approach to image harmonization. Other literature the authors

were aware of, but were not found in this search, were also included as well as relevant citations from included publications.

Notably, we identified five relevant review articles on the topic (Bayer et al., 2022b; Bento et al., 2022; Da-Ano et al., 2020b; Pinto et al., 2020; Stamoulou et al., 2022). Da-Ano et al., (2020b); Bayer et al., (2022b), and Stamoulou et al., (2022) described statistical methods; Bento et al., (2022) described deep learning methods; and Pinto et al., (2020) described harmonization methods specifically for diffusion MRI. In this review, we seek to add to this literature by unifying statistical and deep learning methods for diffusion and non-diffusion MRI. Additionally, we describe common evaluation techniques for validating harmonization methods and provide a framework for proposing and evaluating new methods to direct future efforts in the field.

2.1. Statistical methods

Several overarching statistical models have been used for image harmonization, including linear models, basis representations, latent factor models, and others (Figure 1). In this review, we provide an overview of methods for harmonization of imaging features across known batch labels. These statistical methods can largely be divided into retrospective and prospective harmonization methods. Retrospective harmonization is performed after data collection and aims to mitigate biases due to scanner with the available data. Prospective harmonization needs to be integrated into the study design and often involves collecting repeated measures for downstream analyses.

2.2. Retrospective harmonization

2.2.1. ComBat—

Fortin et al., (2017) proposed that ComBat, a method first designed for batch effect correction in genomics, could be used to harmonize MRI images and derived features (Johnson et al., 2007). ComBat and its various extensions, discussed below, have been widely used in neuroimaging and are organized in Figure 2.

ComBat employs an empirical Bayes linear model framework, which we briefly review. Let y_{ijv}, $i = 1, 2, \ldots, M$, $j = 1, 2, \ldots, n_i$, $v = 1, 2, \ldots, V$ denote the V-dimensional vectors of observed data where i indexes site, j indexes subjects within sites, n_i is the number of subjects acquired on site i, and V is the number of features. The observed data can be measured across voxels, regions of interest, or any other parcellation of the brain. Our goal is to harmonize these features across the M sites. ComBat assumes that the data follow

$$y_{ijv} = \alpha_v + \mathrm{x}_{ij}^T \beta_v + \gamma_{iv} + \delta_{iv} e_{ijv}$$

where α_v is the intercept, x_{ij} is the vector of covariates, β_v is the vector of regression coefficients, γ_{iv} is the mean site effect, and δ_{iv} is the variance site effect. ComBat assumes that the errors e_{ijv} independently follow $e_{ijv} \sim N(0, \sigma_v^2)$. First, least-squares estimates $\hat{\alpha}_v$ and $\hat{\beta}_v$ are obtained for each feature. ComBat then assumes that the site effects follow the same distribution across features. That is, ComBat assumes the mean site effects γ_{iv} follow independent normal distributions and the variance site effects δ_{iv} follow independent inverse

gamma distributions. The empirical Bayes step estimates the hyperparameters via method of moments using data across all features. The empirical Bayes point estimates γ_{iv}^{*} and δ_{iv}^{*} are then obtained as the means of the posterior distributions. The ComBat-harmonized data are then obtained as

$$y_{ijv}^{ComBat} = \frac{y_{ijv} - \hat{\alpha}_v - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_v - \gamma_{iv}^{*}}{\delta_{iv}^{*}} + \hat{\alpha}_v + \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_v \quad (1)$$

ComBat was first applied to voxel-level fractional anisotropy (FA) values from two diffusion MRI datasets where, within each dataset, all subjects were imaged on the same scanner (Fortin et al., 2017). Subsequent studies validated ComBat on other neuroimaging features including cortical thickness and functional connectivity (Fortin et al., 2018; Yu et al., 2018). Since its publication and validation, ComBat has been widely validated and used in the field of MRI imaging (Acquitter et al., 2022; Barth et al., 2022; Bourbonne et al., 2021; Campello et al., 2022; Castaldo et al., 2022; P. Chen et al., 2022; A. Crombé et al., 2020; Dai et al., 2022; Haddad et al., 2022; Ingalhalikar et al., 2021; Leithner et al., 2022; Liu et al., 2022; Luna et al., 2021; Meyers et al., 2022; Onicas et al., 2022; Orlhac et al., 2021; Pagani et al., 2023; Radua et al., 2020; Saint Martin et al., 2021; Verma et al., 2019; Wengler et al., 2021; Whitney et al., 2021; H.M. 2020; Xia et al., 2022, 2019; Zavaliangos-Petropulu et al., 2019).

ComBat was also shown to be effective in magnetic resonance spectroscopy, and its applications to radiomics have been recently reviewed (Bell et al., 2022; Da-Ano et al., 2020b). To study its robustness, analyses have evaluated how ComBat behaves at various sample sizes (Parekh et al., 2022) and validated ComBat correction against correction based on traveling phantoms (Treit et al., 2022). ComBat has been recommended to use for harmonizing large-scale open-source neuroimaging datasets, such as the UK Biobank (Bijsterbosch et al., 2020; Bordin et al., 2021), ABIDE (Horien et al., 2021), ENIGMA (Hatton et al., 2020; Radua et al., 2020), ADNI (Ma et al., 2019), and ABCD (Hagler et al., 2019; Marek et al., 2019) datasets. Limitations of ComBat have been previously described in the field of genomics (T. Li et al., 2021; Nygaard et al., 2016; Zindler et al., 2020). These limitations are described in-depth in the “Recommendations for End-Users” section of the Discussion.

2.2.2. ComBat extensions—Extensions of the standard ComBat model have sought to relax certain model-based assumptions. Many of these methods and their methodological details are covered in a recent review (Bayer et al., 2022b). One popular extension is ComBat-GAM, which allows for preservation of non-linear covariate effects through use of the generalized additive model (GAM) (Pomponio et al., 2020). Such estimation of non-linear covariate effects has been shown to be necessary in certain data settings, such as in diffusion MRI (Cetin-Karayumak et al., 2020b). Another model-based extension incorporates Gaussian mixture models (GMM) into GMM-ComBat to account for multimodal feature distributions (Horng et al., 2022b).

Other extensions of ComBat retain the original model but modify its construction and estimation. A recent study used a fully Bayesian approach with Monte Carlo sampling

in the ComBat model for estimating posterior distributions and found that fully-Bayesian ComBat could provide more accurate harmonization results and unconstrained posterior distributions compared to the standard Empirical-Bayes ComBat model (Reynolds et al., 2022). B-ComBat and BM-ComBat estimate site parameters via bootstrapping and allow for robust harmonization to the pooled feature distribution or a reference batch, respectively (Da-Ano et al., 2020a). TL-ComBat provides an algorithm for applying ComBat parameters learned on training data to new subjects from a known batch (Da-Ano et al., 2021). Another study found that applying intensity normalization via RAVEL followed by ComBat provides greater removal of batch effects (Eshaghzadeh Torbati et al., 2021).

ComBat has been adapted to various study designs. In longitudinal studies where subjects may be imaged one or more times, Longitudinal ComBat accounts for intra-subject correlation by incorporating random effects into the model (Beer et al., 2020). The ComBat framework has also been independently extended by two groups to work in a distributed data setting via Decentralized ComBat/Distributed ComBat (D-ComBat), where data is collected across multiple sites but data-privacy concerns only allow summary statistics from each site to be shared (Bostami et al., 2022b; A. A. Chen et al., 2022b). Many of the above ComBat extensions have been externally validated and used in applied studies (Bostami et al., 2022a; Richter et al., 2022; Saponaro et al., 2022; Singh et al., 2022; Sun et al., 2022; Tafuri et al., 2022).

Finally, methodologists have extended the ComBat model to settings where batch status could be defined by multiple batch covariates, or an unseen batch must be harmonized to a set of known batches. ComBatPC proposed that secondary batch variables to remove could be modeled as additional mean effects in the ComBat model, while the primary batch variable remained in the model as both a mean and variance effect (Wachinger et al., 2021). Additionally, borrowing from the field of genome-wide association studies (GWAS), they showed that including first principal component as one of the secondary batch variables could capture unobserved subpopulations and therefore improve harmonization performance. Applicable to similar settings, OPNested ComBat, an extension of Nested ComBat, learns an optimal order for correcting multiple batch variables and then performs iterative correction for each batch variable individually via the ComBat or GMM-ComBat model (Horng et al., 2022b; Horng et al., 2022a). AutoComBat sidesteps the issue of multiple batches by clustering subjects into automatically-identified batches, implicitly learning which combinations of metadata, such as image acquisition tags or image summary statistics, best define batch status before applying the standard ComBat model (Carré et al., 2022). For settings where an unseen batch must be harmonized to a set of known batches, NeuroHarmony has also been proposed to learn to predict appropriate ComBat parameters for correcting the unseen batch using scanner-associated image quality metrics (Garcia-Dias et al., 2020).

2.2.3. Basis representation—Several harmonization approaches represent the original data using basis vectors or functions estimated from the data then remove batch effects from the representation. Compared to methods that treat features individually, basis representations can capture more complex batch effects and enable harmonization while preserving joint structure among features. The basis chosen varies depending on the imaging

modality but includes principal components, independent components, and spherical harmonics.

Correcting Covariance Batch Effects (CovBat) performs multivariate harmonization by projecting residuals from ComBat onto their principal component axes and applying batch-specific shifts in the principal component space. (A. A. Chen et al., 2022a). This study was the first to show that batch effects are present not only in individual features, but also in the covariance structure between features. CovBat first employs standard ComBat to globally shift and scale each feature, but additionally harmonizes in the principal component space to shift batch-specific covariance matrices towards the global covariance matrix. CovBat was shown to outperform existing harmonization methods in both multivariate statistical evaluations and prediction-based machine learning metrics in cortical structure measurements from the ADNI (A. A. Chen et al., 2022a). In functional connectivity harmonization, CovBat was shown to more effectively harmonize community structure, when compared to ComBat, in sites from the iSTAGING consortium as well as based on information theoretic metrics in the ABIDE, IMPAC, and ADHD-2020 studies (A. A. Chen et al., 2022c; Roffet et al., 2022). CovBat has also been shown to remove batch effects in the cortical and volumetric measures in the ENIGMA study and diffusion tensor imaging features from the ADNI study (Larivière et al., 2022; Sinha et al., 2021; Thomopoulos et al., 2021).

Independent component analysis (ICA) has been a widely used data-driven approach to identify and remove structured noise components, such as head motion-related, physiological, and scanner-induced noise, from fMRI signals (McKeown et al., 2003; Mckeown et al., 1998). Specifically, one study (Feis et al., 2015) used the Functional Magnetic Resonance Imaging of the Brain Centre's (FMRIB's) ICA-based X-noiseifier (FIX, Griffanti et al., 2014; Salimi-Khorshidi et al., 2014) implemented in FMRIB's Software Library (FSL) to reduce scanner-related effects in resting-state networks (RSNs). This study found that ICA-based FIX was useful to remove separate noise components in individual subjects' ICA, but it cannot deal with hardware differences in sensitivity to RSNs (in relation to configurations) or RSN spatial variability (in relation to head coils). Additionally, ICA-based FIX cannot remove scanner-related differences in the magnitude of the BOLD effect. A recently developed linked ICA method was shown to outperform standard general linear model and ICA in removing batch effects from multimodal MRI data collected on the same scanner, but with hardware and software upgrades and different acquisition parameters. Linked ICA used data fusion of multiple MRI modalities to identify and remove scanner-related noise components in multimodal spatial maps. It has yet to be shown whether linked ICA is efficient for removing batch effects from data collected from different scanners.

For diffusion tensor imaging (DTI), voxel-wise signal intensity can be represented in a spherical harmonics (SH) basis, which is an orthonormal basis for functions defined on a unit sphere. Projection of the original intensities into the SH basis yield rotation invariant spherical harmonic (RISH) features. Harmonization from a target batch to reference batch has been proposed by representing complex batch effects as mean shifts in RISH features, often referred to as RISH harmonization (Mirzaalian et al., 2015). Extensions of the RISH

harmonization method have been proposed (Cetin Karayumak et al., 2019; Mirzaalian et al., 2018; Mirzaalian et al., 2016) and covered in a recent review (Pinto et al., 2020). Recent studies have compared statistical and deep learning SH-based harmonization methods, finding that the methods effectively mitigate batch effects but vary in performance on different metrics (Ning et al., 2020; Tax et al., 2019). A recent study found that RISH harmonization outperformed ComBat for preservation of biological effects in large-scale multi-center studies (de Brito Robalo et al., 2022, 2021). RISH harmonization has also been validated in traveling subjects studies (De Luca et al., 2022; Ning et al., 2020) and several major studies (Cetin Karayumak et al., 2019; Cetin-Karayumak et al., 2020a).

2.2.4. Latent factor modeling—Another approach to retrospective harmonization uses latent factors to model biological or batch effects in order to separate wanted and unwanted variation. A latent factor model was first used in Removal of Artificial Voxel Effect by Linear regression (RAVEL) for neuroimaging normalization to model technical variability as latent factors estimated using a set of control voxels not associated with biological variables of interest (Fortin et al., 2016). RAVEL assumes that the $V \times n$ matrix of features Y follows

$$Y = \beta X^T + \theta Z^T + E \tag{2}$$

where X is the $n \times p$ matrix of known covariates, β is the $V \times p$ matrix of regression coefficients, Z is the $n \times b$ matrix of unwanted latent factors, and θ is the $V \times b$ coefficient matrix associated with Z. For a subset of voxels Y_c where there is no association between the voxels and X, an estimate of Z can be obtained by performing factor analysis on Y_c. Then, estimates for θ are obtained by fitting separate linear regressions for each voxel under the model in (2), and the RAVEL-corrected features are obtained as $Y^{RAVEL} = Y - \hat{Z}\hat{\theta}^T$.

The model in (2) was adapted as a Bayesian harmonization method by representing wanted variation through the latent factors, including known batch indicators in the linear model, and yielding harmonized low-dimensional features as the estimated latent factors (Avalos-Pacheco et al., 2022). Their model extends (2) by including a known $n \times (M - 1)$ batch indicator matrix B via

$$Y = \beta X^T + \gamma B^T + \theta Z^T + E \tag{3}$$

where M is the number of batches and γ is the $V \times (M - 1)$ coefficient matrix associated with B. In contrast to RAVEL, this model also allows the variance of E to vary by batch. They develop a non-local spike-and-slab prior to induce sparsity on the factor loadings θ. The authors then develop an expectation maximization algorithm for estimation of the posterior distribution Z, and the harmonized reconstruction are obtained from the mean of the posterior. In an application to gene expression data, they demonstrate that their method performs dimension reduction while adjusting for distinct covariance patterns across batches and benefits downstream survival analyses.

The UNIFAC harmonization method proposes a generalization of the latent factor model, allowing for flexible removal of multivariate batch effects (Zhang et al., 2022). Their main

assumption is that the batch effects are low-rank and represented as matrix-valued shifts. Similar to ComBat and CovBat, UNIFAC harmonization first fits a linear model with known covariates and batch indicators, standardizes the data to have homogenous variance, and obtains standardized data $Y^* = [Y_1^*; Y_2^*; , \ldots; Y_M^*]$ where Y_j^* denotes data from batch j, $j = 1, 2, \ldots, M$. The method then assumes that Y^* follows

$$Y^* = R^* + [I_1^*; I_1^*; \ldots; I_M^*] + [\delta_1 E_1; \delta_2 E_2; \ldots; \delta_M E_M]$$

where R^* is $p \times n$ low-ranked latent structure, I_j^* are low-rank latent patterns associated with batch, E_j are full-rank noise matrices with unit variance, and δ_j capture batch-specific scale shifts. UNIFAC harmonization estimates these latent patterns by optimizing a loss function with a nuclear norm penalty, which yields low-rank structures.

The UNIFAC-harmonized data are defined as

$$Y^{UNIFAC} = \hat{\delta}_j \hat{R}_j^* + \hat{\delta}\left(Y_j^* - \hat{R}_j^* - \hat{I}_j^*\right)$$

where $\hat{\delta}$ is the estimated population variance from the standardization step. Unlike ComBat and CovBat, the UNIFAC harmonization method can capture multivariate batch effects that differ across subjects within the same batch. Compared to CovBat, UNIFAC harmonization can model batch effects that are not constrained to principal component directions. The authors compare UNIFAC harmonization to existing methods in a schizophrenia study conducted across three sites. They show that UNIFAC harmonization outperforms ComBat, CovBat, and several multivariate harmonization approaches on reducing differences in covariance, obscuring prediction of site, and statistical power in detection age-bydisease interactions.

2.3. Prospective harmonization

2.3.1. Traveling subjects linear models—Typical multi-center neuroimaging studies collect separate subjects from each study center, which leads to challenges in separating biological and technical variability. A recent study design addresses this issue by recruiting a subset of participants to travel to every scanner used in the study, often referred to as traveling subjects (Noble et al., 2017). Subsequent studies demonstrated that linear models effectively estimated and removed scanner-related biases from the traveling subjects subset (Yamashita et al., 2019). Increasingly, this study design has been employed in several large-scale multi-site studies (Hawco et al., 2022; Tanaka et al., 2021).

In these traveling subjects studies, N subjects, are acquired multiple times across M scanners. Let y_{ijv}, $i = 1, 2, \ldots, M$, $j = 1, 2, \ldots, N$, $v = 1, 2, \ldots, V$ denote the observed data where i indexes site, j indexes subject, and v indexes feature. Furthermore, let z_j denote a Q-dimensional vector of participant factors, which can include indicators for each participant, diagnosis labels, sample, or any other relevant label. The traveling-subject harmonization model, TS-GLM, assumes that batch effects can be modeled as mean shifts within subjects across batches (Yamashita et al., 2019). Notably, unlike many of the

retrospective harmonization methods described above, TS-GLM does not model batch effect as a scale component in the variance of the residuals. The model is expressed as

$$y_{ijv} = z_j^T \theta_v + \gamma_{iv} + e_{ijv}$$

where θ_v is the vector of regression coefficients, γ_{iv} is the mean site effect, and e_{ijv} are errors assumed to independently follow $e_{ijv} \sim N(0, \sigma_v^2)$. Depending on the choice of indicators in z_j, this model can have many more parameters than observations. Identifiability of the parameters in this model requires constraints on the estimators $\hat{\theta}_v$ and $\hat{\gamma}_{iv}$. In the simple case where z_j is a N-dimensional vector of participant indicators, the constraints are $\sum_{q=1}^{Q} \hat{\theta}_{vq} = 0$ and $\sum_{i=1}^{M} \hat{\gamma}_{iv} = 0$ for each v. Once estimates are obtained, the mean site parameters γ_{iv} can be applied to any subject acquired on scanner i, even those not included in the traveling subjects dataset. This model has been applied and validated across multiple studies (Koike et al., 2021; Yamashita et al., 2021; A. 2020).

ComBat has been extended to the traveling subjects study design, accounting for batch effects in the scale of measurements and leveraging information across features in parameter estimation (Maikusa et al., 2021). This traveling subjects ComBat (TS-ComBat) model is formulated as

$$y_{ijv} = z_j^T \theta_v + \gamma_{iv} + \delta_{iv} e_{ijv}$$

where δ_{iv} is the variance scanner effect. As in ComBat, the model assumes the mean batch effects γ_{iv} follow independent normal distributions and the variance batch effects δ_{iv} follow independent inverse gamma distributions. Estimation also requires identifiability constraints on θ_v and $\hat{\gamma}_{iv}$. The batch effects are obtained as empirical Bayes point estimates γ_{iv}^* and δ_{iv}^* are then obtained as the means of the posterior distributions. Comparison of TS-ComBat and the model in Yamashita et al., (2019) showed that both models performed well in multiple harmonization tasks, but TS-ComBat is superior in smaller sample sizes.

Limitations of TS-GLM and TS-ComBat restrict applicability to common scenarios. Both models require that sufficient subjects are scanned on all scanners in order to ensure that batch effects are not confounded with biological effects. Furthermore, these models do not account for time of scan, so any batch effects may also be driven by changes in imaging measurements over time. Since participants may be lost to follow-up and are acquired at multiple distant time points, these limitations are often relevant and impact the results of harmonization.

2.3.2. Longitudinal ComBat—An alternative for harmonization in traveling subjects studies is Longitudinal ComBat, which flexibly models repeated measures across time (Beer et al., 2020). Compared to other models, Longitudinal ComBat efficiently captures subject effects as random intercepts and incorporates time of scan into the harmonization. While this method was originally designed for longitudinal studies, it has recently been applied in a traveling subjects study to effectively mitigate batch effects (Richter et al., 2022).

Let $y_{ijv}(t)$, $i = 1, 2, \ldots, M$, $j = 1, 2, \ldots, N$, $v = 1, 2, \ldots, V$, denote the observed data where i indexes site, j indexes subject, v indexes feature, and t is a continuous or categorical time variable. The Longitudinal ComBat model is expressed as

$$y_{ijv}(t) = \alpha_v + \mathrm{x}_j(t)^T \beta_v + \eta_{jv} + \gamma_{iv} + \delta_{iv} e_{ijv}(t)$$

where α_v is the mean of feature v at baseline, γ_{iv} is the mean scanner effect, δ_{iv} is the variance scanner effect, $\mathrm{x}_j(t)$ is a potentially time-varying vector of covariates, β_v is a vector of regression coefficients, and η_{jv} is a subject-specific random intercept. The errors $e_{ijv}(t) \sim N(0, \sigma_v^2)$ are assumed to be independent from the random intercepts η_{jv}. ComBat assumptions are placed on the mean and variance scanner parameters, and estimation proceeds through standard mixed model estimation followed by a modified empirical Bayes step.

3. Deep learning methods

In recent years, a wide range of deep learning methods have been proposed as powerful and flexible tools to correct batch effects. These methods have especially shown promise for harmonization of unstructured data, such as images themselves, and for harmonization jointly across multivariate feature matrices. In the unpaired subject setting, popular approaches have used unpaired image-to-image translation frameworks as well as autoencoder networks designed to embed subjects into batch-invariant latent spaces. In paired subject data, methods have used specialized U-Net architectures adapted to imaging data as well as autoencoder methods to estimate direct mappings from one batch to another. Methods are categorized in Figure 3.

3.1. Retrospective harmonization

3.1.1. Cycle-consistency GANs (Image-level)—Zhu et al., (2017) proposed the cycle-consistent generative-adversarial network (CycleGAN) to address the problem of unpaired image-to-image translation. The goal of this network is to learn a mapping between two image batches, A and B, using two generator-discriminator pairs. One generator, G_A, seeks to learn a mapping $G_A(\cdot): A \rightarrow B$ such that its corresponding discriminator, D_B, cannot distinguish the distribution of images from $G(A)$ from that of images from B. Similarly, generator G_B and discriminator D_A learn the inverse mapping $G_B(\cdot) = B \rightarrow A$. Finally, a cycle-consistency loss is introduced as an additional constraint to push the network to preserve image-level features, $\mathscr{L}_{cycle}(G_A, G_B) = \mathbb{E}_A\{\|G_B(G_A(A)) - A\|_1\} + \mathbb{E}_B\{\|G_A(G_A(G_B(B)) - B\|_1\}$. This cycle-consistency loss enforces that an image translated from batch A to batch B and then back to batch A should resemble the untranslated image. Thus, classical CycleGAN attempts to minimize the following objective function: $\mathscr{L}_{total}(G_A, G_B, D_B, D_A) = \mathscr{L}_{GAN}(G_A, D_B, A, B) + \mathscr{L}_{GAN}(G_B, D_A, B, A) + \alpha\mathscr{L}_{cycle}(G_A, G_B)$, where α is a hyperparameter controlling relative importance of the loss components.

In image harmonization, this architecture has been leveraged for unpaired image-to-image translation in many contexts with minor additions to the original CycleGAN loss function and architecture (Dar et al., 2019; Hognon et al., 2019; Kieselmann et al., 2021; Liu et al., 2020; Sinha et al., 2021; Tixier et al., 2021; Zhao et al., 2019; Zhong et al., 2020). Zhao et al., (2019) proposed surface-to-surface GAN (S2SGAN), a variation of CycleGAN using spherical U-Net layers instead of standard convolutional layers, in order to perform harmonization on subject-wise cortical thicknesses projected to a spherical surface. Additionally, they added a cycle-consistency correlation loss component to the original CycleGAN loss such that corresponding vertices between input and cycled images are highly correlated. Dar et al., (2019) demonstrated that a CycleGAN network could generate T1-weighted images from T2-weighted images, and vice versa. Hognon et al., (2019) and Tixier et al., (2021) developed a two-stage framework, where the original CycleGAN network is first used with early stopping criteria to generate “pseudo-paired” data and then a pix2pix network is used on this “pseudo-paired” data to learn the final source-to-reference batch mapping. This two-stage framework differs markedly from other CycleGAN-based approaches; the authors claimed that it allows for better preservation of content information in their data setting where all reference batch subjects were controls while a significant subset of source batch subjects had anatomical pathologies. To validate the beneficial effects of CycleGAN on performance of downstream tasks, Liu et al., (2020) demonstrated that use of the standard CycleGAN model across a multi-batch dataset drastically increased the performance of a fully-convolutional segmentation neural network trained on reference batch images; however, they noted that post-harmonization performance remained substantially lower compared to performance on reference batch images.

Other adaptations of CycleGAN have imposed additional assumptions on the nature of batch effects – namely, that there should be no distortions in anatomy across batches. Previous studies have described distortions in anatomical features across batches, such as cortical thicknesses (Fortin et al., 2018), so the validity of this assumption depends on whether these previously described anatomical differences are actually due to true distortions or instead due to errors in automated segmentation because of batch-wise intensity differences. For example, Kieselmann et al., (2021) added a cycle-consistency geometric loss, where binary geometric masks (1 inside the brain and 0 otherwise) generated from input and cycled images are encouraged to be similar. Meanwhile, Chang et al., (2022) proposed semi-supervised harmonization (SSH), a variation of CycleGAN that uses a two-stage framework to perform harmonization in a manner similar to intensity normalization. In the first stage, the standard CycleGAN model is used to generate an initial harmonized image for each raw image. In the second stage, these initial harmonized images are used along with raw data to perform intensity normalization – that is, histogram matching is used to match each raw intensity to its corresponding initial harmonized intensity. Finally, to generate the output harmonized image, the raw intensities within the raw image are swapped out for their corresponding initial harmonized intensities. Thus, SSH can maintain the high resolution and anatomical fidelity of the raw image, but with brightness and contrast characteristics of the desired reference batch. The authors showed that SSH was able to improve the performance, when compared to ComBat and standard CycleGAN, of a cervical cancer classifier that was trained on subjects from the reference batch and tested on subjects from

the source batch that were harmonized to the reference batch. The authors did not compare SSH performance against standard intensity normalization techniques (Nyúl and Udupa, 1999; Shinohara et al., 2014).

3.1.2. Attention-Mechanism GANs (Image-level)—A further extension of the CycleGAN network called attention-guided GAN (AG-GAN) incorporated attention guidance in both generators and discriminators, where the network is able to learn which parts of an image are most different between batches and focus its attention on accurately translating these parts (Tang et al., 2019). It has been applied to the image harmonization setting with minimal alterations (Sinha et al., 2021). This model leverages the same cycle-consistency idea as CycleGAN, but additionally seeks to decompose generated images into an attention-weighted linear combination of the input image and a restyled image, such that voxels that do not differ between batches can be left mostly unchanged. The attention-guided discriminators then focus on the regions of the generated image that are most artificial. The AGGAN loss function consists of the original CycleGAN loss with additional attention-guided adversarial components, a pixel-wise loss to minimize unnecessary pixel-wise changes, and an attention mask loss to prevent attention masks from globally saturating to 1. Thus, in AG-GAN, the regions of generated images that are similar between batches A and B are largely reconstructed from the input image, allowing generator-discriminator pairs to focus on style transfer in the regions that differ. Other CycleGAN-based models that include attention mechanisms have also been introduced by Selim et al., (2022) and Gutierrez et al., (2023).

3.1.3. Style-conditional GANs (Image-level)—While CycleGAN-based methods perform style transfer conditional only on an input image, adaptations to the CycleGAN framework allow for GAN-based style transfer that is conditional on both an input image as well as a desired output style (Bashyam et al., 2022; Choi et al., 2020; Fetty et al., 2020; Karras et al., 2019; Liu et al., 2021; Tian et al., 2022; Yao et al., 2022). These methods implicitly learn continuous style features such that subtle batch features, like different acquisitions parameters within the same manufacturer, can potentially be corrected. Additionally, since these models include no explicit constraints to disentangle batch from non-batch style features, such as age and sex, nonbatch styles may also be incorporated into style representations. Notably, style-conditional GANs share key characteristics with other broad classes of methods described in this review; these methods incorporate cycle-consistency loss components, similarly to CycleGAN, and also attempt to learn a latent representation of data where content and style information are disentangled, similarly to autoencoder-based models discussed further below.

Qin et al., (2022) draw strongly from the original CycleGAN framework and perform harmonization between two batches using two paired style-conditional GANs, which they call style transfer conditional GAN (ST-cGAN). In each pair, an encoder takes two images as input – one image is encoded into a content representation while the other is encoded into a style representation. Then, these two components are fused via adaptive instance normalization (AdaIN, Huang and Belongie, 2017) by the generator to create an output with the content of the first image and style of the second. The loss function involves

the cycle-consistency and paired discrimination loss components along with an additional constraint of identity loss, which enforces that "harmonization" of an image directly to its own true batch should reproduce itself.

Meanwhile, other style-conditional GANs deviate more from the CycleGAN. One such model, StyleGAN, was proposed by Karras et al., (2019) and later applied to imaging data by Fetty et al., (2020) and Liu et al., (2021). StyleGAN consists of one style-mapping network, one generator, one image discriminator, and one style discriminator. First, StyleGAN uses the style-mapping network to create a style representation from a random-noise latent space. Then, the generator encodes an image, combines it with this style representation using adaptive instance normalization, and attempts to generate a new image in that style, such that the image discriminator cannot tell the image is generated and the style discriminator can recover the input style representation. Since this generative process is under-constrained, a cycle-consistency loss component is added as well as a style diversification loss component. Thus, the network learns to sample diverse styles, generate realistic images in those styles that retain content, and implicitly learn the original style of each image.

A similar concept is employed by StarGANv2 and has been used in the multi-batch image harmonization setting (Bashyam et al., 2022; Choi et al., 2020). This model incorporates a style encoder that directly learns style representations from training images, in contrast to the StyleGAN mapping network which generates style representations from noise and then associates these randomly-generated style representations with relevant images. Once style representations as well as realistic image generation are learned by StarGAN, style transfer can be achieved by combining content representations with desired style representations. Again, both cycle-consistency and style diversification loss components are used. Harmonization using this model has been shown to improve out-of-sample performance of an age-prediction network trained in the reference batch. A model based on similar style-disentangling mechanisms has been shown to improve the performance of a 3D segmentation network trained on the reference batch when applied to source batch images (Yao et al., 2022). Notably, like autoencoder-based models, StyleGAN, StarGANv2, and the model by Yao et al. rely on one common generator that is able to take any content representation and combine it with any style representation.

3.1.4. Autoencoder models (Feature-level)—In 2015, Sohn et al., (2015) introduced the conditional variational autoencoder (CVAE) in order to generate new data conditional on additional covariates. This model can be best understood through its predecessor, the variational autoencoder (VAE), which in turn, builds on the standard autoencoder, a simple neural network architecture that seeks to learn a non-linear, low-dimensional representation of input data that contains sufficient information for reconstruction (Kingma and Welling, 2014). The VAE architecture and loss function, discussed below, allow for additional constraints compared to the standard autoencoder and seek to improve organization of the latent space as well as reduce potential for overfitting. In this model, the encoder seeks to embed the input data into a lower-dimensional latent distribution, $q(z \mid a)$, which approximates some pre-specified "prior" distribution, $p(z)$. In practice, $p(z)$ is usually chosen to be the standard multivariate normal distribution. The probabilistic

decoder, $p(z \mid a)$ then takes a random sample from this distribution, $Z \sim q(z \mid a)$ and attempts to reconstruct the data using this sample. The VAE seeks to minimize the loss function $\mathscr{L}_{total} = \mathbb{E}_A(\|a - p(a \mid z\|_2) + \mathrm{KLD}(q(z \mid a), p(z))$, where $\mathrm{KLD}(,\mid \cdot)$ is the Kullback-Leibler divergence between the latent distribution and prior distribution. The reconstruction loss component encourages latent-space distributions to efficiently retain information, while the Kullback-Leibler divergence component creates a trade-off that encourages representations to coexist around the origin as well as inject noise. Together, these constraints organize the latent space such that nearby points produce similar reconstructions.

CVAE builds on the VAE architecture by concatenating additional covariates, c, onto the inputs for both the encoder and the decoder in order to condition the latent space on these covariates. In this model, since the decoder has necessary information from additional covariates readily available for reconstruction, the encoder no longer benefits from encoding covariate-dependent information in the latent space.

At the feature-level, a number of methodologies have harnessed CVAE ideas to learn a latent-space representation that is independent of the imaging batch and the corresponding batch-conditioned encoder-decoder pair (An et al., 2022; Moyer et al., 2020). Then, these methods perform harmonization by first encoding samples into the batch-invariant latent space using each samples' actual batch, and then decoding those latent-space representations using the desired output batch.

Moyer et al., (2020) leveraged a deep learning model using the CVAE structure to perform unsupervised image-based harmonization on diffusion MRI images. First, this model maps diffusion-weighted imaging (DWI) signal for each voxel to a vector of spherical harmonics representations. Then, for each voxel, spherical harmonics vectors from itself and its six immediate neighbors are concatenated along with the batch covariate and fed into the CVAE to learn the batch-invariant latent representation. The loss function consists of the standard VAE loss; a reconstruction error for the projection of spherical harmonics vectors back into DWI space; an adversarial loss for detecting batch on the reconstruction as estimated by a discriminator; and a penalty on the mutual information between the latent space and batch, enforced via the sum of pairwise Kullback-Leibler divergences between latent-space representations.

An extension of this model, called goal-specific conditional variational autoencoder (gcVAE), has been proposed to perform harmonization on image-derived features that is explicitly aware of desired downstream applications – in this case, the prediction of Alzheimer disease diagnosis and Mini-Mental State Examination (MMSE) scores (An et al., 2022). gcVAE seeks to trains two neural networks independently – first, a CVAE model is pre-trained to learn a conditionally-independent latent-space representation and the corresponding conditional decoders. Additionally, a generic feed-forward prediction network is trained on reference batch data to predict Alzheimer disease diagnoses and MMSE scores from unharmonized features, and its weights are frozen. Finally, data from both batches are harmonized through the pre-trained CVAE and then fed through the frozen prediction network; the loss function for this step seeks to minimize the error in prediction network outputs. This loss is used along with a small learning rate and limited training epochs to

fine-tune the CVAE model to retain information relevant to diagnosis and MMSE prediction in the harmonized reconstruction.

3.1.5. Autoencoder models (Image-level)—In image-level harmonization, methods have used ideas from the CVAE as well as from the standard autoencoder to disentangle content information from batch and other style features (Cackowski et al., 2021; Cao et al., 2022; Fatania et al., 2022; Zuo et al., 2021). These methods seek to decompose images into low-dimensional style-invariant content representations in the encoding step, and then in the generation step, inject these content representations with style information.

Zuo et al., (2021) introduced a harmonization method named Contrast Anatomy Learning and Analysis for MR Intensity Translation and Integration (CALAMITI) that uses similar tools to CVAE as well as style-conditional GANs. This model was based on previous work by the same group (Dewey et al., 2020). However, CALAMITI additionally leverages the fact that neuroimaging subjects are often imaged under multiple contrasts, such as T1-weighted and T2-weighted acquisitions. These intra-subject contrast pairs can be thought to share identical anatomical content with differing styles. Meanwhile, intra-batch images – those taken under the same contrast and scanner, but on different subjects – can be thought to share identical style but differing anatomical content. CALAMITI uses these two sets of pseudo-paired data to train a content encoder, style encoder, generator, and batch discriminator. Content representations within intra-subject pairs are constrained to be interchangeable and independent of batch as assessed by the batch discriminator. Style representations necessary to reconstruct a given image are obtained entirely from a random intra-batch image with no shared content. Harmonization is then performed by providing a trained decoder with image-specific content representations along with style representations from the desired reference batch. Finally, to account for the 3D structure of the brain despite using 2D slices, this procedure is performed in axial, coronal, and sagittal directions and the three “directional” brain volumes are unified into a final image through a 3D fusion network, an idea borrowed from DeepHarmony, described below (Dewey et al., 2019).

CALAMITI has been validated by Shao et al., (2022), who showed that training a 3D thalamus-segmentation network on images harmonized to the reference batch resulted in better out-of-sample performance on true images from the reference batch when compared to the same segmentation network trained on unharmonized images. Meanwhile, in-sample performance of the network did not decrease after harmonization, suggesting minimal degradation of anatomy. Additionally, the direct predecessor to CALAMITI, proposed by Dewey et al., has been shown to allow for improved harmonization, when compared to CycleGAN, of diffusion MRI across multiple batches as well as simultaneously allow for estimation of multi-shell diffusion MRI from single-shell data (Dewey et al., 2020; Hansen et al., 2022).

Inspired by the use of imaging data structure in CALAMITI, ImUnity sought to apply these ideas to the harmonization of not only batches available in the training dataset, but also unseen batches (Cackowski et al., 2021). At each training iteration, ImUnity takes two random slices, S_1 and S_2, from the same image as input, such that the slices can be thought to have different content but share the same style. Next, both S_1 and S_2 are modified to

S_1^γ and S_2^γ, respectively, using the gamma transformation, an image processing function that changes the relative intensity of gray colors. Slice S_1 is then embedded into a latent content representation, slice S_2^γ is embedded into a style representation, and these content and style representations are used to reconstruct slice S_1^γ, which should have the same content as S_1 and same style as S_2^γ. Additionally, this model applies both a batch discriminator and optional biological information classifier to the latent content representation which serve to promote the removal of batch information and maintenance of biological information, respectively. Through this process, content information can be disentangled from style in a self-supervised manner without additional imaging contrasts, and image harmonization can be carried out by inputting source batch slices to the content encoder and reference batch slices to the style encoder. If unseen batches are similar enough to training batches such that the content encoder can appropriately embed slices from unseen batches, the model can be easily extended to these settings.

StyleMapper also takes advantage of the ability to apply various image transformation functions to raw images in order to generate images that are known to have the same content but different styles (Cao et al., 2022). In this approach, each raw image is transformed to seven different styles using the following transformation functions: original, negative, logarithmic, gamma transformation, piecewise linear, Sobel X filter, and Sobel Y filter. Then, for each iteration, two raw images and two randomly-sampled corresponding transformed images (both using the same transformation function) are fed to a model consisting of one content encoder, one style encoder, and one generator, where the generator seeks reconstruct an image with desired style using the content and style representations. Notably, no discriminator is used in the StyleMapper model. To constrain this process, a number of loss function components are used: reconstruction of both raw images; reconstruction of both transformed images; similarity of style representations between raw images; similarity of style representations between transformed images; similarity of content representations between raw images and their corresponding transformed image; and cross-reconstruction, where swapping content and style representations between across input images should result in an output image that is similar to the corresponding “ground-truth” image. Thus, StyleMapper is able to create pseudo-paired data with the same content but different styles, learn to disentangle content and style within this dataset, and perform harmonization, given that differences across batches are somewhat similar to the transformations used in training.

Finally, HarMOnAE removes batch effects using style transfer within a standard convolutional autoencoder (Fatania et al., 2022). In this model, style representations are explicitly defined as the batch covariate and directly injected into the decoder via adaptive instance normalization. To enforce the learning of batch-invariant content representations, an adversarial loss is imposed on the content representation space.

3.1.6. Batch-unlearning classifiers (Other)—Related to standard harmonization methods, some deep learning methods have been developed to simultaneously perform harmonization and downstream classification tasks, such that classification should be robust to batch effects (Dinsdale et al., 2021; Hong et al., 2022). Notably, unlike other

harmonization methods described in this review, these batch-unlearning classifiers do not attempt to produce a harmonized output dataset that can then be used for any generic downstream analysis.

Dinsdale et al., (2021) proposed a domain-adaptation classifier that could be used to improve the generalizability of age predictions across multiple batches where age distributions differed. The three-module network consists of a convolutional feature extractor, a batch discriminator, and a main task classifier, where the goal of the feature extractor is to learn a latent space representation of raw images that is useful for the main task classifier and can simultaneously fool the batch discriminator. Thus, the feature extractor learns to extract batch-invariant features, and the main task classifier learns generalizable decision boundaries. Importantly, the batch-unlearning classifier is trained using a subsample of the data where the outcome of interest is balanced across batches in order to avoid confounding. The authors showed this strategy is especially useful in settings where one batch makes up a large majority of the dataset and the distribution of the outcome of interest differs greatly in this batch compared to others. The method also improved performance of age prediction in an unseen batch. Similarly, Hong et al., (2022) showed a non-convolutional version of this network, which they call scanner-generalization neural network (SGNN), could be used to improve prediction of general psychopathology factors (Caspi and Moffitt, 2018) using functional connectivity matrices within the ABCD study.

3.2. Prospective harmonization

3.2.1. Direct mapping—In specially-curated multi-batch studies where traveling subjects are available, the “ground truth” batch-specific scans for these subjects are known under the assumption that all differences between these scans are entirely due to technical artifacts. This allows for a class of much more powerful and accurate methods that leverage this unique pairing of data to learn a mapping from one batch to another. Then, this mapping can be applied to unpaired images to remove batch effects, under the assumption that data from traveling subjects are a representative sample of those from unpaired subjects. However, despite the benefits of prospective harmonization methods, datasets where the required traveling subjects are available are expensive to obtain and can be limited in terms of subjects. Additionally, the assumption that traveling subjects are representative of all subjects should be verified; traveling subjects could, for example, be healthier or wealthier than non-traveling subjects.

Dewey et al., (2019) proposed DeepHarmony, a convolutional U-Net-based architecture could be applied to 2D patches across multiple contrasts from twelve subjects each scanned under each of two batches in order to directly harmonize the images themselves. In this architecture, the network attempts to jointly use multiple contrasts (T1-weighted, T2-weighted, FLAIR, and proton density) from each subject collected under one protocol. These multiple contrasts are used simultaneously to reconstruct the corresponding contrasts for that subject collected under another protocol. This “many-to-many” reconstruction approach can be thought of as allowing for the use of complementary information across contrasts. Additionally, DeepHarmony slightly modifies the vanilla U-Net architecture such that, in the final convolutional layer, the input contrasts are concatenated to the final

feature map. Thus, instead of having to recreate reference contrasts entirely from scratch, the network can instead focus on learning an appropriate transform of the input data to reconstruct the intended output. Finally, as with CALAMITI, DeepHarmony sought to learn three independent image-to-image mappings for slices in each of the axial, sagittal, and coronal directions. These "directional" images are then aggregated using voxel-wise medians to produce a final harmonized image.

For diffusion imaging, Tong et al., (2020) showed that deep learning can be applied to pre-processed DWI images across traveling subjects in order to estimate derived diffusional kurtosis imaging (DKI) measures that are harmonized across batches. This study leveraged a 3D hierarchical-structured convolutional neural network (H–CNN) designed to take 3 × 3 × 3 voxel patches as input and jointly produce eight scalar DKI measures as output (axial diffusivity, radial diffusivity, mean diffusivity, fractional anisotropy, axial kurtosis, radial kurtosis, mean kurtosis, kurtosis fractional anisotropy) (Li et al., 2019). To perform harmonization, Tong et al. used DWI images from traveling subjects in the reference batch to calculate DKI measures for each image using an iteratively-reweighted linear least squares method. Then, these DKI measures were non-linearly registered to corresponding paired DWI images in source batches to create a training dataset, where the input is a DWI image from a source batch while the output is the set of DKI measures extracted from the paired image in the reference batch. Next, H–CNN is trained on this dataset in order to learn a mapping from source batch DWI images to reference batch DKI measures. Finally, this trained H–CNN was applied to other DWI images from the source batches in order to estimate DKI measures harmonized to the reference batch.

3.2.2. Content-style disentanglement—Another approach for directly harmonizing images, Multi-scanner Image harmonization via Structure Preserving Embedding Learning (MISPEL), was introduced by Torbati et al., (2022). Unlike DeepHarmony, MISPEL hopes to perform harmonization across m batches, where m can be more than two, through the use of a set of m batch-specific convolutional autoencoders that are trained via a two-step algorithm. Importantly, the encoders are allowed to be deep networks while the decoders merely perform a linear combination of the latent-space representations. In step one, MISPEL seeks to train each batch-specific encoder to embed slices from its batch into a common latent space and then train the corresponding decoder to use those latent-space representations to reconstruct slices in the style of its batch. To do so, MISPEL trains each batch-specific autoencoder separately in a self-supervised fashion using a reconstruction loss and additionally enforces a common latent space between all autoencoders through a representation similarity loss, which penalizes high variance across all latent-space representations. In step two, all encoders are frozen and only the decoders are updated such that all decoders produce similar harmonized output slices and the outputs are also similar to the input slice. Thus, intuitively, MISPEL can be thought of as disentangling images into content and style representations, where the latent-space representations contain content information and differences in how those representations are linearly combined by the decoder describe style differences.

Tian et al., (2022) address the setting of paired data in a multiple-batch setting via their model, DeRed. This model can be thought of as an adaptation of CycleGAN and especially

ST-cGAN, discussed in the style-conditional GAN section. Similarly to ST-cGAN, DeRed uses paired GANs to perform harmonization – however, to adapt the paired-GAN framework to the multiple-batch setting, DeRed trains a separate style encoder and generator for each batch-to-batch harmonization task, such that each set of networks harmonizes images either to or from the reference batch. Then, DeRed is able to harmonize any batch to the reference batch by combining a source-batch content representation with a reference-batch style representation. Additionally, harmonization to any source batch can be achieved through a two-step process, where all other source batches are first harmonized to the reference batch and then these generated reference-batch images are harmonized to the desired source batch. Data from paired subjects is taken advantage of in the loss function, which consists of four components: 1) batch consistency, where style representations should be similar within each batch; 2) content consistency, where content representations should be similar within paired subjects even from different batches; 3) reconstruction, where content and style representations from the same image should result in reconstruction of that image; and 4) cross-reconstruction, where content and style representations from different images of the same subject should result in reconstruction of the image that corresponds to the style representation.

4. Evaluation metrics

Increasing interest in the development and application of harmonization methods requires standardized and effective metrics that quantify performance. Harmonization evaluation metrics can largely be grouped into two categories, harmonization performance metrics and predictive performance metrics (Figure 4). Harmonization performance metrics aim to detect or quantify batch effects and can be separated into metrics measured at the feature level and at the image level. These metrics can often be interpreted as summary statistics, requiring accompanying visualizations to complement their findings. Predictive performance metrics measure the effects of harmonization on performance in downstream analyses. Importantly, effective harmonization methods should reduce detectable batch effects in the data while preserving performance in downstream analyses.

4.1. Harmonization performance

4.1.1. Feature-level metrics—Evaluation approaches for methods that perform feature-level harmonization can be broadly grouped into four general paradigms: statistical testing for differences in distribution across batches, predictive modeling of batch, assessing feature dispersion and similarity, and qualitative visualization.

Features can be interpreted as each having their own distribution that can be split along batch variables such that in the absence of batch effects, these sub-distributions should be identical. Harmonization methods can thus be evaluated based on their ability to remove differences in feature distribution across batch groups. This can be evaluated using statistical testing, where the test used depends on the assumed form of the distributional differences. Location effects can be assessed using tests for differences in mean (e.g. students and paired t-tests, ANOVA, linear regression to control for covariates, Wilcoxon rank-sum and signed rank tests, and Kruskal-Wallis test) while scale effects can be detected using tests

for differences in variance (e.g. Bartlett's sphericity test) (Fortin et al., 2018; Y. Li et al., 2021; Wengler et al., 2021; Yu et al., 2018). To test for more general differences in distribution beyond disparity in mean and variance, the Kolmogorov-Smirnov or Anderson-Darling tests can be used (Da-Ano et al., 2020a; Fatania et al., 2022; H.M. Whitney et al., 2020). These tests are all completed at the feature-level such that if harmonization is effective, significant differences in distribution due to batch will be detected before but not after harmonization. This result would indicate that the harmonization tool has removed differences in distribution associated with batch variables. In settings where a p-value would be inappropriate, effect size measures (e.g. Cohen's d, Hedge's g) can be used (Radua et al., 2020; Reardon et al., 2021). In the specific setting of functional connectivity matrices, which can be studied from the network theory perspective, Roffet et al., (2022) demonstrated the utility of the Kruskal-Wallis test on batch-wise differences between Normalized Network Shannon Entropy and Normalized Network Fisher Information metrics.

If biological covariates are imbalanced across batches, it may be expected that this imbalance may lead to differences in marginal batch-wise feature means that should not be corrected by harmonization. In these settings, it is instead important to evaluate harmonized outputs for differences in biological-covariate-conditional batch-wise feature means. One common approach is to use linear regression or linear mixed effects regression, where batch and biological covariates (e.g. age, sex) are used to jointly model the feature. The estimated regression coefficients for batch and biological covariates can be tested for significant effects on each feature, where a significant regression coefficient for the batch covariate corresponds to statistically-detectable batch effects (Badhwar et al., 2020; Bell et al., 2022; Wengler et al., 2021; Zavaliangos-Petropulu et al., 2019). Notably, this approach will provide a valid assessment of batch effects even if the biological covariates are not imbalanced across batches. Looking beyond batch, this evaluation procedure allows for simultaneous assessment of preservation of biological covariates; comparing regression coefficients for biological covariates before and after harmonization can provide insight into whether biological information is preserved.

Another approach uses features as predictors in a machine learning classifier – random forests, support vector machines (SVM), AdaBoost, and others – in order to predict batch as an outcome. If harmonization is effective, there will be reduced signal from batch in the data and therefore reduced classifier performance (An et al., 2022; A. A. Chen et al., 2022a; Saponaro et al., 2022). While this approach is more general than using a linear model, this comes at the cost of interpretability. When using a statistical test for differences in distribution or on linear model regression coefficients, there is a clear null hypothesis about the nature of batch effects – that is, whether they are differences in mean, variance, or distribution. This is contrasts with the machine learning classifier approach, where detection of batch effects is easy, but understanding the nature of these detected batch effects is challenging. While there are methods for measuring feature importance for machine learning classifiers, further visualization is necessary to fully characterize batch effects. Additionally, it is challenging to account for confounders when using this machine learning approach; for example, if there is significant imbalance in a biological covariate such that batch can be easily predicted by this biological covariate, preservation of biological information in

the harmonized data would also result in predictability of batch, even if batch effects were perfectly removed.

A more direct metric for identifying variation associated with batch in feature-level data is the coefficient of variation (CoV). The CoV is the ratio of the mean to the standard deviation and can be used to measure between-batch variability by calculating the CoV within each batch for each feature (Cai et al., 2021; Garcia-Dias et al., 2020; Treit et al., 2022). The resulting set of CoV values is then described using summary statistics, and if harmonization is effective, the differences in CoV distributions between batch groups will be reduced post-harmonization.

In traveling subject studies or other datasets where matched-subject data is available, another direct metric for measuring feature similarity across batches is correlation coefficients, including the intra-class correlation coefficient (ICC), Spearman's correlation, and Pearson's correlation. If batch effects are not present in the data, then a feature extracted from scans associated with the same subject under different acquisition protocols should be more similar across protocols (Crombé et al., 2021; A. 2020; Kurokawa et al., 2021). Effective harmonization tools should increase the correlation coefficient for a given feature across batch groups provided the scans are from the same subject. Additionally, the discriminability statistic may also be a reasonable metric for this data setting, though this statistic has not yet been used in the context of harmonization (Bridgeford et al., 2021).

Finally, visualizations are an essential tool for characterizing batch effects more comprehensively than summary metrics. Visualization methods pertinent to harmonization can be broadly grouped into decomposition-based approaches and displays of feature distributions. Decomposition-based approaches condense high-dimensional data into a two to three-dimensional space suitable for visualization and include methods such as principal components analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP). In low-dimensional space, batch effects can be seen as increased distances between points of differing batch groups. Harmonization should reduce these distances and bring points of different batch closer together (Acquitter et al., 2022; A. A. Chen et al., 2022c; Guan et al., 2021).

However, decomposition-based methods condense information from all features into a single figure, necessitating visualizations of univariate or bivariate feature distributions to further characterize distributional differences affiliated with batch (e.g. feature density plots, box-plots, scatterplots etc.). Effective harmonization should reduce visual differences in distribution across batch groups (Bethlehem et al., 2022; Clarke et al., 2020; Da-Ano et al., 2021; Saint Martin et al., 2021). These visualizations can also be used to identify cases in which distributional assumptions of model-based methods are violated (e.g. non-Gaussian for ComBat) and further troubleshoot harmonization methods by providing comprehensive information regarding the effects of harmonization on feature distributions (Horng et al., 2022b).

4.1.2. Image-level metrics—Applications of deep learning to harmonize image-level data have emerged as promising approaches for correcting unstructured data. Consequently,

their evaluation requires metrics that quantify the effects of harmonization at the image level. Because the goal of image-level harmonization can be viewed as mapping an image from one batch to another, the resulting evaluation is often based around measuring the distance between images of different batches.

When paired data are available, this distance can be directly quantified as the voxel-level difference between the harmonized image and the true image from the reference batch using metrics such as Mean Absolute Error (MAE) or Mean Squared Error (MSE). Also included in this category is peak signal to noise ratio (PSNR), a measure of image quality that takes the ratio of the maximum image value and the root MSE. For example, Dewey et al., (2019) use the MAE as a component of their loss function as well as a final measure of image similarity to compare paired images from the same subject scanned with different MRI acquisition protocols. While this approach likely provides the most accurate quantification of image differences associated with batch, it is not as commonly used because datasets of sufficient sample size to train deep learning algorithms that also contain paired samples from each batch are rare. A possible solution to this problem is to use unpaired data for training and use a more limited paired dataset for testing and evaluation (Denck et al., 2021).

The scenario of unpaired data is more common, but this setting requires more indirect measures of image similarity because no “ground truth” is available. The two most common metrics used in this context are the structural similarity index measure (SSIM) and Fréchet Inception Distance (FID) (Heusel et al., 2018; Wang et al., 2004). SSIM, as the name implies, measures the degree to which structures are preserved post-transformation. While historically used in paired data, SSIM can be applied in unpaired data under the assumption that key structures are largely the same between subjects. FID is a common evaluation metric for GANs that measures the distance between the ground truth and generated image distributions as opposed to the images themselves. Both FID and SSIM have been employed in the evaluation of adversarial networks used for image-level harmonization (Liu et al., 2021; Sinha et al., 2021). Notably, while SSIM measures presence of similar anatomy and FID measures “realism” of generated images – both important metrics for assessing the quality of generated images – neither explicitly evaluates whether generated images match the distribution of the reference batch or how well the images are harmonized. Additionally, FID is based on features learned on natural scenes from the ImageNet database; such features may not be applicable to medical images, so FID may not be a reliable measure of realism in this setting (Deng et al., 2009).

Finally, qualitative visualizations may include side-by-side image slices representing unharmonized slices, harmonized slices, and reference slices. Importantly, “directionality” of visualized slices (i.e. axial, coronal, sagittal) is important, since many image-level methods correct images at the individual slice level. Thus, visualization using slices in the same direction as the harmonization as well as slices in different directions may be revealing.

While these metrics are commonly used in the evaluation of image-level harmonization, recent work by Ravano et al., (2022) suggests that image-level metrics are poor indicators of cross-batch consistency and robustness in downstream analyses. While predictive

performance should not be the sole evaluation metric for harmonization methods, as will be discussed below, these findings indicate image-level metrics should be interpreted with caution and that increases in image similarity do not guarantee improved robustness. Therefore, additional evaluation may be carried out by extracting select features, such as voxel intensities or measures of structural characteristics, and assessing feature-level harmonization performance using the techniques described in the above section. Evaluation of the distributions of extracted features may also be useful in assessing for mode collapse, where GAN-based methods and CVAE-based methods only generate a small subset of the original variability in harmonized images.

4.2. Downstream analysis performance

For many applications, the primary goal of harmonization is not necessarily to remove batch effects from the data, but instead to improve robustness or overall performance in some downstream analysis, such as inference or prediction. Inference tasks tend to be associated with feature-level data and can be viewed as seeking to precisely estimate the magnitude and direction of biological effects of interest. These tasks involve regression of feature-level data on biological covariates, and successful harmonization is often assessed as removal of batch effects while statistical power for detecting such biological effects is preserved but not artificially biased or inflated. Many studies have suggested harmonization can improve inference when biological covariates are explicitly controlled for in the model; however, it remains a challenge to validate such claims as ground-truth biological effects are unavailable in real data, and simulation of realistic batch-confounded data is unsolved (An et al., 2022; A. A. Chen et al., 2022a; Fortin et al., 2018; Yu et al., 2018). Additionally, it is important to keep in mind that, in cases where batch status and biological effects are highly correlated, unbiased removal of true batch effect may correctly reduce observed biological effects.

In the harmonization literature, post-harmonization prediction evaluation can be broadly grouped into three major categories: segmentation, classification, and regression. Segmentation involves the separation of regions of interest (ROIs) from the surrounding anatomy, a task often affected by the differences in intensity associated with differences in image acquisition. Segmentation is an essential task for many downstream analyses, as the resulting regions can be used in the extraction of quantitative features for predictive modeling. Many studies have already demonstrated that image-level harmonization can improve downstream segmentation performance (Dewey et al., 2019; Dinsdale et al., 2021; He et al., 2021; B. Li et al., 2021; Shao et al., 2022). The performance of segmentation algorithms can be quantified using metrics such as the Dice coefficient, Mean Surface Distance (MSD), Hausdorff distance, and others. Classification and regression use a matrix of quantitative features to predict discrete and continuous outcomes, respectively. In these contexts, batch effects may introduce additional noise that can obscure signal, result in models that learn batch-confounded parameters, as well as induce overfitting that reduces the ability of models to generalize to unseen data from other batches. To this end, many studies have applied harmonization techniques to demonstrate improved predictive performance and model robustness in the prediction of a variety of outcomes, including malignancy, age, survival, neurodegenerative disease, and more (Fortin et al., 2018; Tixier et al., 2021; H.M. Whitney et al., 2020; Zavaliangos-Petropulu et al., 2019). Classification

performance is typically evaluated using metrics such as accuracy, sensitivity, specificity, and area under receiver operating curve (AUROC) (Ingalhalikar et al., 2021; Sinha et al., 2021; Whitney et al., 2021). Evaluation for regression methods involves measuring the distance between the observed and predicted outcome vectors using metrics such as mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) (Bashyam et al., 2022; Chen et al., 2020).

4.2.1. Accounting for confounders—Notably, evaluation of harmonization performance and downstream analysis performance in the presence of confounding by biological covariates of interest remains an active challenge. Depending on the strength and nature of such confounding, naive application of the above evaluation metrics may incorrectly show harmonization is performing poorly even if it is working perfectly, or incorrectly show harmonization is performing well even if it is working poorly. The same is true for downstream analyses.

For example, imbalance of biological covariates across batches may result in seemingly poor harmonization performance even in the setting of theoretically-perfect batch effect removal. In imbalanced datasets, biological information will and should remain correlated with batch status after harmonization. Therefore, accurate preservation of biological information will result in marginal differences in imaging data across batches that will be detectable by statistical and machine learning methods that do not condition on these covariates. Notably, even evaluation approaches that do condition on biological covariates, such as linear regression, may provide inaccurate conclusions if the model is mis-specified with respect to the relationship between biological covariates, batch, and the imaging data.

In the opposite direction, imbalance of biological covariates may also induce incorrect removal of biological information that the harmonization method views as batch effects. For example, if age is imbalanced across batches but not appropriately accounted for by the harmonization methods, age-related differences between batches that should be preserved will instead be attributed to batch effects and removed. Additionally, in this setting, naive approaches for evaluating harmonization performance will incorrectly show the harmonization method is performing well, since marginal batch-wise differences may be removed when they should be preserved.

While downstream analysis performance is a key priority in the wider imaging community, it is critical to distinguish this performance from the specific goal of harmonization: the removal of batch effects from data. Evaluating within-sample performance does not provide explicit information regarding harmonization performance, nor vice versa, particularly in settings where biological and batch variables are associated (Dinsdale et al., 2021; Horng et al., 2022a).

For example, consider a hypothetical study in which most patients with a cancer diagnosis are imaged at a tertiary referral hospital, while most patients without a cancer diagnosis are imaged at a primary care hospital. Because of this imbalance, the batch variable of hospital type becomes highly associated with the outcome of cancer diagnosis. In this setting, a theoretically-perfect harmonization method will eliminate this association,

therefore resulting in reduced within-sample performance. In a different example, if there is minimal confounding between batch status and an outcome of interest, removal of batch-related noise may increase the relative signal of the outcome of interest, and within-sample performance may improve.

While harmonization is not guaranteed to improve overall predictive performance, the removal of batch effects can result in increased predictive model robustness and generalizability. This can be evaluated by measuring predictive performance on out-of-sample testing data in the harmonized output space. For example, such external validation has been applied as test-retest analyses (Mirzaalian et al., 2016; van de Bank et al., 2015), out-of-sample cross-validation procedures (Dinsdale et al., 2021), or true out-of-sample test datasets (Chang et al., 2022; Liu et al., 2020; Shao et al., 2022). Improved performance on external, out-ofsample data would indicate that a predictive model trained on harmonized data is more robust to differences in image acquisition and is overfitting less on batch-related noise.

5. Discussion

5.1. Recommendations for end-users

Image harmonization methods have been proposed for a wide variety of data structures and study designs. Optimal selection of the state-of-the-art harmonization method for each study is thus highly dependent on these characteristics as well as on the ease-of-use of available methods. In this section, we provide our recommendations to users seeking to apply existing harmonization methods to their own datasets in order to best reduce bias and improve generalizability of results.

Generally, for both feature-level and image-level data, we recommend that image harmonization should be used as a final correction step. That is, raw imaging data should first be pre-processed using available non-harmonization methods designed to minimize technical artifacts, including bias field correction (Tustison et al., 2010), intensity normalization (Shinohara et al., 2014), and if applicable, other steps like brain extraction (Smith, 2002), registration to a common template (Avants et al., 2008). In the setting of functional MRI, additional preprocessing steps should also be used, if necessary, such as motion correction (Ciric et al., 2017; Jenkinson et al., 2002) or spatial smoothing (Mikl et al., 2008). Notably, small differences in both functional and structural pre-processing pipelines can induce marked variation in downstream analyses (Cetin-Karayumak et al., 2020b). Consensus as to how to perform such pre-processing is critical in multi-batch studies if pre-processing is conducted independently within sites (Li et al., 2022). Finally, once all standard pre-processing steps have been implemented in order to reduce technical noise, remaining batch effects can be addressed via harmonization.

For feature-level data from studies without traveling subjects, ComBat and its various extensions should still be considered state-of-the-art despite recent advances in deep learning methods. Specifically, CovBat is a strong choice when batch effects are suspected in the covariance structure of the linear model residuals (A. A. Chen et al., 2022a), ComBat-GAM should be used when non-linear covariate or batch effects may be at play (Pomponio

et al., 2020), and FC–CovBat is recommended for the specific application to functional connectivity values (A. A. Chen et al., 2022c). In datasets where at least one batch has a small sample size, the standard ComBat model likely outcompetes more complex methods – in these settings, estimation of higher-order biological and batch effects may be imprecise and reduce harmonization performance (Fortin et al., 2017; Nygaard et al., 2016; Zindler et al., 2020). In these settings, the principal component decomposition step of CovBat and the GAM estimation step of ComBat-GAM may be highly variable and therefore unreliable. For study designs with longitudinal data and therefore non-independent observations, Longitudinal ComBat should be used (Beer et al., 2020). In the presence of privacy-preserving constraints, D-ComBat yields equivalent results as standard ComBat without the need to have the full dataset at a single location (Bostami et al., 2022b; A. A. Chen et al., 2022b).

While it is unlikely that batch effects are perfectly modeled in these ComBat-style methods, these methods have been extensively validated in many datasets and data types including cortical thicknesses, fractional anisotropy values, functional connectivity values, and radiomic features. Even in the setting of data types that have not been previously validated, ComBat-style methods can be applied reliably; they perform principled model-based correction with minimal risk of overfitting and tend to err on the side of under-correction rather than over-correction. For multisite studies with small sample sizes, the simplicity of these models and the empirical Bayes estimation procedure allow for stable correction in settings where more sophisticated correction would be infeasible. Importantly, these methods also provide easy-to-use open-source code in R, Python, or both. However, because of the simplicity of these models, substantial multivariate batch effects will remain following correction, and model misspecification poses the potential for bias and increased false positives. While CVAE-based methods have been proposed for feature-level correction, such as Moyer et al., (2020) and gcVAE (An et al., 2022), these methods still require users to have considerable deep learning experience for hyperparameter tuning and evaluation, and the behavior has not yet been extensively validated by follow-up studies in different datasets or data types.

For feature-level data in the prospective setting where matched pairs are available, TS-GLM and Longitudinal ComBat have strong theoretical foundations in the linear model and random effects model framework, respectively (Beer et al., 2020; Yamashita et al., 2019). While TS-GLM has been used more often in this setting, Longitudinal ComBat is theoretically advantageous as this model can jointly use both paired and unpaired data in the estimation of batch effects.

For image-level harmonization, while ComBat-style methods can be applied on the voxel level, where subjects are registered to each other and represented by vectorized voxel intensities, ComBat is almost certainly inadequate. In this setting, deep learning methods are a much more reasonable choice. However, while image-level harmonization is almost certainly the ultimate goal for the field of harmonization, given the current state of the field, we recommend that, if possible, end-users should avoid image-level harmonization and instead seek to extract relevant features from the images and apply feature-level methods. This is because image-level methods have only been evaluated under ideal settings, require

extensive deep learning expertise and computational capacity, and may introduce bias in datasets where biological covariates confounders are present. These limitations are discussed in more depth below.

If image-level harmonization is necessary and unavoidable, we recommend the following methods. In studies where individuals are imaged under at least two modalities on the same scanner but no traveling subjects are used, CALAMITI has an elegant theoretical basis, has been validated in a few follow-up studies, and most importantly, provides readily-available code (Zuo et al., 2021). In the prospective setting, MISPEL should be considered, as it provides open-source code and has been internally validated to improve harmonization both in terms of images and image-extracted features when compared to a matched-pairs-aware version of CALAMITI; however, no follow-up studies have yet externally validated this model (Torbati et al., 2022). While many CycleGAN-based methods have been proposed and assessed, we do not recommend these methods. This is because the CycleGAN architecture is known to be under-constrained which could lead to potential anatomical distortions; GAN models can be challenging to train; and to our knowledge, no open-access code is available for proposed adaptations of the architecture or loss functions.

Despite the potential that CALAMITI and other deep learning methods have shown in correcting image-level data, we believe these methods are not yet ready for end-users to apply to their own imaging data. Firstly, from the resource perspective, this is partly due to the immense computational resources required for training and the extensive technical expertise necessary to troubleshoot code and perform hyperparameter tuning. Additionally, deep learning methods require that end-users thoroughly validate harmonization results – the flexibility of these networks can result in unexpected behavior that may break down in certain unknown settings. Secondly, from the technical perspective, since training these deep learning models require large sample sizes and three-dimensional convolutional models are computationally prohibitive, deep learning methods treat each axial slice as an independent sample, even when slices are from the same subject or from nearby planes; this process does not explicitly model the correlation between these slices and hopes the model can implicitly pick up on these relationships. Thirdly, while these methods have been shown to work well in their respective published manuscripts, limited follow-up studies have been published to validate these results in other datasets, so it is uncertain if the results are easily generalizable. Finally, for most studies, harmonization was also only validated in the image domain with the implicit assumption that, if the image is harmonized, then extracted features from these harmonized images will also be subsequently harmonized; explicit evaluation of whether this assumption holds will be important to strengthen the case for using these methods.

Across data types and study design settings, once a reasonable harmonization method is applied, the resulting harmonized dataset can be evaluated for harmonization performance and predictive performance. Evaluation for harmonization performance is especially important for more complex methods that are sensitive to changes in user-defined hyperparameters, as these methods may underperform if the hyperparameters not tuned appropriately. Note that such methods include CovBat and ComBat-GAM, since they

require the specification of the number of principal components to correct and the standard GAM hyperparameters, respectively.

End-users can also evaluate harmonization methods based on predictive performance, especially on out-of-sample data, such as that generated using cross-validation, train-test splits, or test-retest data. Effective harmonization should improve the generalizability of prediction models, so predictive performance on out-of-sample data may increase. However, end-users should be aware that predictive performance may decrease in training sample data, especially if batch status was correlated with the outcome of interest. Additionally, large increases in predictive performance might be observed if the harmonization method accidentally introduces biases or artifacts – end-users should be especially aware of this possibility if using less-constrained methods such as GAN-based models.

5.2. Limitations of harmonization

Importantly, end-users should be aware of two limitations of harmonization – namely, that removal of batch effects induces correlation between subjects and that removal of batch effects and preservation of biological effects depends on the ability to precisely estimate these effects (Bayer et al., 2022a; T. Li et al., 2021; Nygaard et al., 2016; Zindler et al., 2020). The studies below specifically describe these limitations within the context of the ComBat model, since this model is easily used and has been widely studied in the field of genomics for over a decade; however, these limitations are broadly true of any harmonization method.

Firstly, harmonization is used as a pre-processing step, where batch effects are estimated using the whole dataset under some model, and subsequently removed. The harmonized output is then used for any downstream inference or prediction analyses. This separation of harmonization from downstream analyses is advantageous – under this paradigm, harmonization methods can be as complex as necessary to adequately remove batch effects, and any downstream analysis model can be used afterwards. This contrasts with joint methods for inference that account for batch effects. For example, multiple linear regression where batch status is included as a covariate is a simple joint method; however, in this model, batch effects can only be accounted for as differences in expected mean, and the only downstream analysis possible is inference on the linear effect of biological covariates of interest.

However, separation of harmonization from downstream analyses also induces artificial correlation between originally-independent subjects (T. Li et al., 2021). This is because batch effects are estimated using all subjects in the dataset, and then this estimated batch effect is removed from each subject's data. As a result, each harmonized data point is some function of all the other data in the dataset and therefore correlated with each other. This limitation could lead to exaggerated or reduced findings in downstream analyses that do not account for this induced correlation. Li et al. provide a potential solution to this problem in the context of ComBat through their approach, ComBat+Cor. This model applies standard Combat for harmonization, but accounts for the induced correlation in downstream linear models. Notably, this approach would not be useful for downstream analyses that cannot account for sample correlation (i.e. machine learning models, qualitative visualizations,

etc.), and ComBat+Cor has only been validated in the genomics context. Additionally, Li et al. noted that ComBat+Cor was too conservative in settings with large variance batch effects, which may be common in neuroimaging data; in these settings, they recommended standard ComBat instead.

Secondly, harmonization methods may inaccurately remove batch effects in settings where it is challenging to accurately estimate batch effects (Nygaard et al., 2016; Zindler et al., 2020). For example, in datasets where biological covariates are heavily imbalanced across batches, there will be insufficient overlap of these biological covariates to independently estimate batch and biological effects. Instead, batch and biology can be thought to be a form of "multicollinear" which will result in unstable estimation for both batch and biological effects (Nygaard et al., 2016). Similar estimation issues occur in datasets with a large number of batches and a small number of subjects within each batch, as well as in settings where batch effects are extremely small or non-existent such that they are easily overfit (Zindler et al., 2020). In all these settings, harmonization will be carried out using only the point estimate for batch effects; the large estimation errors for batch effects will be ignored. If the magnitude of the original batch effects is greater than that of the estimation errors, harmonization may partially ameliorate the batch effects problem, but if the reverse is true, harmonization may make things worse. Additionally, when considered together, the combination of harmonization-induced correlation and inaccurately-estimated batch effects may result in increased false positives.

Ultimately, while it is important for end-users to be aware of these issues with harmonization as a whole, we still consider harmonization to be the state-of-the-art approach for addressing batch effects, since no better solution exists for removing complex batch effects while allowing the flexibility of using any downstream methods. However, end-users should exercise care to avoid blindly applying harmonization methods in settings where batch effects cannot be precisely estimated to reduce the risk of false positives. In these settings, end-users should reach for alternative approaches, such as joint methods for inference that account for batch effects, or consider consultation with neuroimaging statisticians. Harmonization-induced correlation is more challenging to avoid or take into account, but we believe that the increased generalizability of post-harmonization analyses outweighs the risk of exaggerated or diminished findings due to correlation-induced bias.

5.3. Recommendations for methodologists

As methodologists continue to propose novel ideas to improve both feature and image-level harmonization, we provide recommendations for a more standardized framework for describing evaluating, comparing, and releasing novel methods that we believe will help accelerate the advancement of the field.

5.3.1. Transparency in assumptions and limitations—Firstly, new methods should be explicit about the specific scenarios under which the method is intended to work, since use, evaluation, and comparison to similar methods all depend on the scenario. To do so, methods should define assumptions made about the data-generating process as well as describe assumptions about the availability of various information in their dataset. The need

for such transparency becomes clearer when harmonization is viewed as causal inference problem. Under the causal inference framework, different batches are different “treatments,” unharmonized data are “observed outcomes” under these treatments, and harmonization methods attempt to estimate “counterfactual outcomes” at the individual level – what the data would have looked like in a hypothetical scenario where all subjects were scanned in the same batch (Höfler, 2005; Rosenbaum and Rubin, 1983; Rothman et al., 2008). Notably, such estimation requires strong assumptions that may be relevant when end-users decide which harmonization method may be most reasonable for their dataset.

As an example of a common implicit assumption, prospective methods are defined by the assumption of paired data across batches; however, they also assume variation within pairs is entirely due to batch effects and that the batch effects estimated using this paired data is representative of batch effects in the rest of the sample. While such assumptions may be reasonable in some datasets, they may be unreasonable in others. The first assumption is violated if paired scans across batches are taken with a larger interval of time in between, since differences between scans may be due to changes in age or disease progression in addition to batch effects. The second assumption is violated if traveling subjects tend to be more able or willing to travel than non-traveling subjects, perhaps due to relatively younger age or better health. In this setting, if covariates that affect tendency to be a traveling subject also affect brain structure or function, estimation of batch effects in these traveling subjects may be non-representative.

In retrospective studies, these assumptions on paired data are not necessary. However, these methods instead make assumptions on the nature of batch effects and how confounders are controlled for. For example, ComBat relies heavily on an assumption of correct model specification; that is, batch effects can be fully captured by univariate shifts in mean and rescaling of error terms and that biological effects are confounders that can be controlled for linearly. Meanwhile, deep learning methods make minimal model specification assumptions, but data-based assumptions are encoded in model parameters based on biases present in the training data. For example, when deep learning methods do not account for biological covariates when performing harmonization; implicitly, they assume that batch status is independent of biological covariates. This may not be reasonable if, for example, sicker subjects tend to be scanned at tertiary care hospitals while healthier patients tend to be scanned in primary care hospitals. Thus, transparency in assumptions about confounders is necessary in understanding when methods can be applied.

Transparency of methods known to require more computational power, higher technical expertise, or larger sample sizes is also recommended. While harmonization methodologists may prioritize implementing interesting ideas to advance the field and improve our ability to remove batch effects, end-users may place less emphasis on using such “optimal” methods and instead look to apply methods that are more accessible yet still perform acceptably. Thus, methodologists should include a discussion of computational resources required, approximate run times, and approximate empirical lower bounds for sample size required so that subsequent readers can have a better sense of when/if the method is usable in their settings.

5.3.2. Standardized evaluation framework—Secondly, methods should be explicitly evaluated both in terms of removal of batch effects as well as preservation of biological effects. In feature-level data, evaluation of batch effects should consist of statistical testing for difference in means for individual features, prediction of batch using machine learning classifiers, and qualitative visualization of feature distributions using dimension reduction techniques as well as univariate and bivariate plotting. For statistical testing, we recommend use of the linear model, where batch and confounding covariates are the independent variables and feature data is the dependent variable, in order to estimate the mean batch effects when confounders are controlled for. For batch prediction, we recommend random forests or support vector machines as powerful multivariate methods that are easy to apply and robust to hyperparameter tuning. For qualitative visualization, we recommend UMAP or PCA for multivariate visualization, univariate/bivariate density plots across batches for a small number of randomly selected features, and scatterplots of unharmonized data against harmonized data for a small number of randomly selected features.

Evaluation of preservation of biological effects should be tested by choosing a few biological effects that may be of interest to end-users and using them as the covariate or outcome of interest in the above analyses. Note that in batch effects evaluation, less evidence of batch effects is desired, while in biological effects evaluation, more evidence of biological effects is better. For both batch effects and biological effects evaluation, additional evaluation can be added as appropriate, including other metrics highlighted in Figure 4. For example, if the primary goal of the harmonization method is to use a reference batch-trained prediction algorithm on source-batch data, improvement in test time performance of this prediction algorithm should be included as part of the evaluation. For all metrics, baseline comparison of outputs should be made to those from unharmonized data in addition to one or more previously validated methods designed for the same data setting.

To evaluate removal of batch effects in image-level data, we encourage the use of both image-level and feature-level metrics to fully characterize harmonization performance. At the image-level, evaluation should be conducted through both quantitative image metrics, such as SSIM and FID, as well as qualitative visualization of several comparable image slices. In prospective study designs, comparable image slices refer to paired data, and in retrospective designs, they refer to slices from individuals with similar pertinent covariates. For qualitative visualization, we encourage the inclusion of axial, coronal, and sagittal slices for each of unharmonized, harmonized, and reference images. We recognize that many harmonization methods on 3D neuroimaging data are limited to correction of axial slices, small 3D patches, or even individual voxels due to constraints in computational power, model complexity, and sample size, so coronal and sagittal slices may look distorted. However, we believe it is important to establish a baseline as to the extent and characteristics of such distortions.

For feature-level evaluation of image-level harmonization methods, we recommend that methodologists extract a small number of relevant image-derived features from both unharmonized and harmonized datasets. Then, the full set of metrics described above for evaluating feature-level harmonization can be applied to assess for effective harmonization and look out for signs of mode collapse. We argue that while image-level harmonization

should imply harmonization of downstream extracted features, this may not necessarily be the case in existing methods due to how challenging it is to estimate and remove batch effects in images. More thorough characterization of how image-level methods affect these subsequent features is necessary for methodologists to better understand areas for improvement and for end-users to assess the robustness of these methods.

As we encourage authors of image-level methods to include potentially distorted visualizations or sub-optimal evaluation results on image-derived features, we simultaneously encourage editors and reviewers to ask for such assessments in order to characterize the behavior of current state-of-the-art methods more comprehensively. Additionally, we hope these editors and reviewers recognize the immense challenge of image-level harmonization, and in doing so, publish manuscripts with interesting ideas or making encouraging progress despite distortions or bias that may be evident.

5.3.3. Code availability—Thirdly, we encourage methodologists of both image-level and feature-level methods to provide easy-to-use, open-source code so that novel harmonization methods can be compared to previously described methods, applied to real-world problems by neuroscientists, and understood at the code level. The lack of such available code is particularly evident in deep learning image-level methods, where most methods provide no code or refer readers to the original codebase the novel method was based on. Methods that do provide code tend to do so by uploading entire project directories with minimal curation, leaving subsequent users to parse through, edit, and re-implement the code themselves. Ideally, both deep learning and statistical methodologists should strive to write comprehensive tutorials, provide well-organized code, and create a small number of high-level wrapper functions such that subsequent users can run the method on their own data with only a few lines of user-written code. Software engineering principles would also be useful, including implementation of continuous integration tests, containerization of code, and reduction of dependencies.

Such standards are already widespread in similar fields, such as batch effect correction methods for single-cell RNA sequencing (scRNA-seq) analyses. In scRNA-seq batch effect correction, most statistical and deep learning methods have been proposed with the inclusion of easy-to-use code. As a result, comprehensive reviews have been conducted to assess method performance across different large datasets, allowing for empirical quantitative and qualitative comparison (Tran et al., 2020). A similar ability to comparatively assess a broad range of harmonization methods and establish a current gold-standard would be hugely impactful for the field. In application, improved accessibility to proposed harmonization methods will allow these methods to now only present interesting ideas for growth, but also provide useful and applicable methods for end-users.

Finally, code-level understanding is especially important in deep learning models. While descriptions of network architecture and theoretical loss functions illustrate the main ideas behind a model, there are numerous ways these design choices, and others, can be implemented. For example, there are many details that may be unimportant for theoretical understanding and therefore excluded from the manuscript text, but still have large empirical

impacts, including: choice of optimizer and optimizer parameters; hyperparameter-tuning algorithm and hyperparameter search ranges; minimization of mode collapse risk; and more.

5.3.4. Future work—In retrospective feature-level data, methodologists should seek to further develop statistical techniques for harmonization. While widely-used statistical approaches have largely relied on univariate modeling or strong assumptions about the nature of batch and biological effects, recently proposed multivariate harmonization methods such as CovBat and UNIFAC have been shown to greatly improve harmonization. However, these approaches continue to make strong assumptions and require more validation. For example, CovBat assumes multivariate batch effects is present only in the covariance matrix of the residuals while UNIFAC assumes multivariate batch effects can be estimated as low-rank latent patterns. Thus, further work in validating such methods as well as developing novel statistical methods to remove complex multivariate, non-linear batch effects in a theoretically-rigorous manner may be warranted.

Complementary work on applying deep learning methods to feature-level data is a promising next step, with the hope that an appropriately-designed neural network may be able to model and remove complex batch effects in a data-driven manner. In this vein, methods such as CVAE and gcVAE have been proposed. However, CVAE has been shown to have the unintended consequence of removing biological effects of interest along with batch effects. To address this consequence, gcVAE explicitly rewards the model for retaining biologic effects, which may introduce bias into the harmonized dataset; this consequence has not been empirically demonstrated. Additionally, like many image-level deep learning methods and unlike statistical methods, CVAE and gcVAE assume the complexity of their neural networks allow for near-perfect model fit, such that output can be directly treated as harmonized data without explicit reintroduction or modeling of error terms. Further work in deep learning harmonization of feature-level data should evaluate the validity of this assumption and its impact on downstream analyses.

Ultimately, efforts should be made to develop strong methodology that can easily and robustly perform harmonization on image-level data across a range of sample sizes, acquisition sequences, and study designs. To do so, methodologists should consider leveraging both statistical and deep learning ideas; statistical methods may allow for improved robustness and strong performance in smaller samples or when confounding is present, while deep learning models may better capture the complexity of image-level data, which pose serious challenges to traditional statistics. For all image-level harmonization methods, care must be taken to characterize harmonization performance both qualitatively and quantitatively, not only at the image level, but also for subsequent features extracted from these harmonized images; evaluation on extracted features is both sensitive and specific for poor harmonization performance, and performance on extracted features may additionally be of interest to end-users. Again, when reviewing image-level harmonization papers that include unfavorable results, we encourage editors and reviewers to note the difficulty of performing harmonization at the image level.

Finally, more work is necessary in evaluation. Firstly, further development of sensitive, covariate-aware multivariate evaluation metrics is important. While univariate feature-wise

regression approaches can detect batch effects conditional on confounding biological covariates; similar capabilities of conditioning should be developed or borrowed from other fields for multivariate machine learning approaches and validated in the context of harmonization. Additional qualitative and quantitative image-level metrics suited for retrospective datasets are also necessary to provide better assessment of image-level harmonization. To support this effort and demonstrate the validity of these newly proposed metrics as well as pre-existing ones, progress must be made in developing simulation studies with realistic batch effects and biologic effects or large traveling subject cohorts, such that "gold-standard" harmonization can be known. The availability of these datasets will also allow methodologists to confirm the behavior of newly developed methods.

Comprehensive comparative analyses of currently proposed harmonization methods under a wide range of data settings would also be hugely beneficial. In the current literature, novel methods tend to compare their harmonization outputs to a small set of similar methods using a limited number of evaluation metrics. This leads to challenges in comparing novel methods with one other and a less complete understanding of how each harmonization method succeeds or why it struggles. Thorough quantitative and qualitative comparison will allow for end-users to more confidently choose optimal methods and for methodologists to better focus their efforts on addressing underlying problems.

Conclusion

In neuroimaging, multi-batch data is increasingly necessary to obtain sufficient sample sizes and produce generalizable results. Furthermore, in these settings, end-users are more interested in applying powerful and flexible models to perform both inference and prediction. To enable these efforts, removal of batch effects via image harmonization is an important, but complex, pre-processing step.

In this review, we comprehensively discuss the growing set of statistical and deep learning image harmonization methods, categorizing these methods broadly to highlight common themes. We then summarize approaches for evaluating the effectiveness of harmonization in feature-level and image-level methods. Finally, we provide recommendations to neuroscientists and harmonization methodologists. For neuroscientists, we give suggestions on when to perform harmonization and which harmonization method to choose in each data and study design setting. We also discuss important limitations of harmonization and the settings where these limitations may be most relevant. For methodologists, we highlight critical methodological obstacles, advocate for a standardized evaluation framework, push for increased transparency in assumptions and code-availability, and provide guidance on possible future directions for the field. Overall, we hope these recommendations will allow for more effective and widespread application of current harmonization methods as well as accelerated progress towards thorough and precise removal of batch effects in increasingly complex neuroimaging data.

Acknowledgments

This study was supported by grants from the National Institute of Mental Health (R01MH123550, R01MH112847, R01MH120482, R37MH125829, R01MH113550), the National Institute on Aging (U19AG074879,

R01AG019771, P30AG072976, U01AG072177, U01AG068057, RF1AG054409), the National Institute of Neurological Disorders and Stroke (R01NS085211, U24-NS130411), the National Institute of Biomedical Imaging and Bioengineering (R01EB022573), and the Alzheimer's Association (AARF-22-722571). FH was supported by NIH Medical Scientist Training Program T32 GM07170. Funding sources were not involved in study design, data analysis, manuscript preparation, or submission decisions.

Data availability

No data was used for the research described in the article.

References

Acquitter C, Piram L, Sabatini U, Gilhodes J, Moyal Cohen-Jonathan E, Ken S, Lemasson B, 2022. Radiomics-based detection of radionecrosis using harmonized multiparametric MRI. Cancers 14, 286. doi:10.3390/cancers14020286. [PubMed: 35053450]

Aderghal K, Afdel K, Benois-Pineau J, Catheline G, 2020. Improving Alzheimer's stage categorization with Convolutional Neural Network using transfer learning and different magnetic resonance imaging modalities. Heliyon 6, e05652. doi:10.1016/j.heliyon.2020.e05652. [PubMed: 33336093]

An L, Chen J, Chen P, Zhang C, He T, Chen C, Zhou JH, Yeo BTT, 2022. Goal-specific brain MRI harmonization. Neuroimage 263, 119570. doi:10.1016/j.neuroimage.2022.119570. [PubMed: 35987490]

Avalos-Pacheco A, Rossell D, Savage RS, 2022. Heterogeneous large datasets integration using bayesian factor regression. Bayesian Anal. 17, 33–66. doi:10.1214/20-BA1240.

Avants BB, Epstein CL, Grossman M, Gee JC, 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal 12, 26–41. doi:10.1016/j.media.2007.06.004. [PubMed: 17659998]

Badhwar A, Collin-Verreault Y, Orban P, Urchs S, Chouinard I, Vogel J, Potvin O, Duchesne S, Bellec P, 2020. Multivariate consistency of resting-state fMRI connectivity maps acquired on a single individual over 2.5 years, 13 sites and 3 vendors. Neuroimage 205, 116210. doi:10.1016/j.neuroimage.2019.116210. [PubMed: 31593793]

Barth C, Kelly S, Nerland S, Jahanshad N, Alloza C, Ambrogi S, Andreassen OA, Andreou D, Arango C, Baeza I, Banaj N, Bearden CE, Berk M, Bohman H, Castro-Fornieles J, Chye Y, Crespo-Facorro B, de la Serna E, Díaz-Caneja CM, Gurholt TP, Hegarty CE, James A, Janssen J, Johannessen C, Jönsson EG, Karlsgodt KH, Kochunov P, Lois NG, Lundberg M, Myhre AM, Pascual-Diaz S, Piras F, Smelror RE, Spalletta G, Stokkan TS, Sugranyes G, Suo C, Thomopoulos SI, Tordesillas-Gutiérrez D, Vecchio D, Wedervang-Resell K, Wortinger LA, Thompson PM, Agartz I, 2022. In vivo white matter microstructure in adolescents with early-onset psychosis: a multi-site mega-analysis. Mol. Psychiatry doi:10.1038/s41380-022-01901-3.

Bashyam VM, Doshi J, Erus G, Srinivasan D, Abdulkadir A, Singh A, Habes M, Fan Y, Masters CL, Maruff P, Zhuo C, Völzke H, Johnson SC, Fripp J, Koutsouleris N, Satterthwaite TD, Wolf DH, Gur RE, Gur RC, Morris JC, Albert MS, Grabe HJ, Resnick SM, Bryan NR, Wittfeld K, Bülow R, Wolk DA, Shou H, Nasrallah IM, Davatzikos CiSTAGING and PHENOM consortia, 2022. Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors. J. Magnet. Reson. Imaging: JMRI 55, 908–916. doi:10.1002/jmri.27908. [PubMed: 34564904]

Bayer JMM, Dinga R, Kia SM, Kottaram AR, Wolfers T, Lv J, Zalesky A, Schmaal L, Marquand A, 2022a. Accommodating site variation in neuroimaging data using normative and hierarchical Bayesian models. Neuroimage 264, 119699. doi:10.1016/j.neuroimage.2022.119699. [PubMed: 36272672]

Bayer JMM, Thompson PM, Ching CRK, Liu M, Chen A, Panzenhagen AC, Jahanshad N, Marquand A, Schmaal L, Sämann PG, 2022b. Site effects how-to and when: an overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. Front. Neurol 13.

Beer JC, Tustison NJ, Cook PA, Davatzikos C, Sheline YI, Shinohara RT, Linn KA, 2020. Longitudinal ComBat: a method for harmonizing longitudinal multi-scanner imaging data. Neuroimage 220, 117129. doi:10.1016/j.neuroimage.2020.117129. [PubMed: 32640273]

Bell TK, Godfrey KJ, Ware AL, Yeates KO, Harris AD, 2022. Harmonization of multi-site MRS data with ComBat. Neuroimage 257, 119330. doi:10.1016/j.neuroimage.2022.119330. [PubMed: 35618196]

Bento M, Fantini I, Park J, Rittner L, Frayne R, 2022. Deep learning in large and multi-site structural brain MR imaging datasets. Front. Neuroinformat 15.

Bethlehem R.a.I., Seidlitz J, White SR, Vogel JW, Anderson KM, Adamson C, Adler S, Alexopoulos GS, Anagnostou E, Areces-Gonzalez A, Astle DE, Auyeung B, Ayub M, Bae J, Ball G, Baron-Cohen S, Beare R, Bedford SA, Benegal V, Beyer F, Blangero J, Blesa Cábez M, Boardman JP, Borzage M, Bosch-Bayard JF, Bourke N, Calhoun VD, Chakravarty MM, Chen C, Chertavian C, Chetelat G, Chong YS, Cole JH, Corvin A, Costantino M, Courchesne E, Crivello F, Cropley VL, Crosbie J, Crossley N, Delarue M, Delorme R, Desrivieres S, Devenyi GA, Di Biase MA, Dolan R, Donald KA, Donohoe G, Dunlop K, Edwards AD, Elison JT, Ellis CT, Elman JA, Eyler L, Fair DA, Feczko E, Fletcher PC, Fonagy P, Franz CE, Galan-Garcia L, Gholipour A, Giedd J, Gilmore JH, Glahn DC, Goodyer IM, Grant PE, Groenewold NA, Gunning FM, Gur RE, Gur RC, Hammill CF, Hansson O, Hedden T, Heinz A, Henson RN, Heuer K, Hoare J, Holla B, Holmes AJ, Holt R, Huang H, Im K, Ipser J, Jack CR, Jackowski AP, Jia T, Johnson KA, Jones PB, Jones DT, Kahn RS, Karlsson H, Karlsson L, Kawashima R, Kelley EA, Kern S, Kim KW, Kitzbichler MG, Kremen WS, Lalonde F, Landeau B, Lee S, Lerch J, Lewis JD, Li J, Liao W, Liston C, Lombardo MV, Lv J, Lynch C, Mallard TT, Marcelis M, Markello RD, Mathias SR, Mazoyer B, McGuire P, Meaney MJ, Mechelli A, Medic N, Misic B, Morgan SE, Mothersill D, Nigg J, Ong MQW, Ortinau C, Ossenkoppele R, Ouyang M, Palaniyappan L, Paly L, Pan PM, Pantelis C, Park MM, Paus T, Pausova Z, Paz-Linares D, Pichet Binette A, Pierce K, Qian X, Qiu J, Qiu A, Raznahan A, Rittman T, Rodrigue A, Rollins CK, Romero-Garcia R, Ronan L, Rosenberg MD, Rowitch DH, Salum GA, Satterthwaite TD, Schaare HL, Schachar RJ, Schultz AP, Schumann G, Schöll M, Sharp D, Shinohara RT, Skoog I, Smyser CD, Sperling RA, Stein DJ, Stolicyn A, Suckling J, Sullivan G, Taki Y, Thyreau B, Toro R, Traut N, Tsvetanov KA, Turk-Browne NB, Tuulari JJ, Tzourio C, Vachon-Presseau É, Valdes-Sosa MJ, Valdes-Sosa PA, Valk SL, van Amelsvoort T, Vandekar SN, Vasung L, Victoria LW, Villeneuve S, Villringer A, Vértes PE, Wagstyl K, Wang YS, Warfield SK, Warrier V, Westman E, Westwater ML, Whalley HC, Witte AV, Yang N, Yeo B, Yun H, Zalesky A, Zar HJ, Zettergren A, Zhou JH, Ziauddeen H, Zugman A, Zuo XN, Bullmore ET, Alexander-Bloch AF, 2022. Brain charts for the human lifespan. Nature 604, 525–533. doi:10.1038/s41586-022-04554-y. [PubMed: 35388223]

Bijsterbosch J, Harrison SJ, Jbabdi S, Woolrich M, Beckmann C, Smith S, Duff EP, 2020. Challenges and future directions for representations of functional brain organization. Nat. Neurosci 23, 1484–1495. doi:10.1038/s41593-020-00726-z. [PubMed: 33106677]

Bordin V, Bertani I, Mattioli I, Sundaresan V, McCarthy P, Suri S, Zsoldos E, Filippini N, Mahmood A, Melazzini L, Laganà MM, Zamboni G, Singh-Manoux A, Kivimäki M, Ebmeier KP, Baselli G, Jenkinson M, Mackay CE, Duff EP, Griffanti L, 2021. Integrating large-scale neuroimaging research datasets: harmonisation of white matter hyperintensity measurements across Whitehall and UK Biobank datasets. Neuroimage 237, 118189. doi:10.1016/j.neuroimage.2021.118189. [PubMed: 34022383]

Bostami B, Espinoza FA, van der Horn Harm J., van der Naalt J, Calhoun VD, Vergara VM, 2022a. Multi-site mild traumatic brain injury classification with machine learning and harmonization. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference 2022, pp. 537–540. doi:10.1109/EMBC48229.2022.9871869.

Bostami B, Hillary FG, van der Horn Harm Jan, van der Naalt J, Calhoun VD, Vergara VM, 2022b. A decentralized ComBat algorithm and applications to functional network connectivity. Front. Neurol 13, 826734. doi:10.3389/fneur.2022.826734. [PubMed: 35370895]

Bourbonne V, Jaouen V, Nguyen TA, Tissot V, Doucet L, Hatt M, Visvikis D, Pradier O, Valéri A, Fournier G, Schick U, 2021. Development of a radiomic-based model predicting lymph node involvement in prostate cancer patients. Cancers 13, 5672. doi:10.3390/cancers13225672. [PubMed: 34830828]

Bridgeford EW, Wang S, Wang Z, Xu T, Craddock C, Dey J, Kiar G, Gray-Roncal W, Colantuoni C, Douville C, Noble S, Priebe CE, Caffo B, Milham M, Zuo X-N, Vogelstein JTReproducibility, C. for R., 2021. Eliminating accidental deviations to minimize generalization error and maximize

replicability: applications in connectomics and genomics. PLoS Comput. Biol. 17, e1009279. doi:10.1371/journal.pcbi.1009279. [PubMed: 34529652]

Brown RA, Fetco D, Fratila R, Fadda G, Jiang S, Alkhawajah NM, Yeh EA, Banwell B, Bar-Or A, Arnold DLCanadian Pediatric Demyelinating Disease Network, 2020. Deep learning segmentation of orbital fat to calibrate conventional MRI for longitudinal studies. Neuroimage 208, 116442. doi:10.1016/j.neuroimage.2019.116442. [PubMed: 31821865]

Byrge L, Kliemann D, He Y, Cheng H, Tyszka JM, Adolphs R, Kennedy DP, 2022. Video-evoked fMRI BOLD responses are highly consistent across different data acquisition sites. Hum. Brain Mapp 43, 2972–2991. doi:10.1002/hbm.25830. [PubMed: 35289976]

Cackowski S, Barbier EL, Dojat M, Christen T, 2021. ImUnity: a generalizable VAEGAN solution for multicenter MR image harmonization

Cai LY, Yang Q, Kanakaraj P, Nath V, Newton AT, Edmonson HA, Luci J, Conrad BN, Price GR, Hansen CB, Kerley CI, Ramadass K, Yeh F-C, Kang H, Garyfallidis E, Descoteaux M, Rheault F, Schilling KG, Landman BA, 2021. MASiVar: multisite, multiscanner, and multisubject acquisitions for studying variability in diffusion weighted MRI. Magn. Reson. Med 86, 3304–3320. doi:10.1002/mrm.28926. [PubMed: 34270123]

Campello VM, Martín-Isla C, Izquierdo C, Guala A, Palomares JFR, Viladés D, Descalzo ML, Karakas M, Çavuş E, Raisi-Estabragh Z, Petersen SE, Escalera S, Segué S, Lekadir K, 2022. Minimising multi-centre radiomics variability through image normalisation: a pilot study. Sci. Rep 12, 12532. doi:10.1038/s41598-022-16375-0. [PubMed: 35869125]

Cao S, Konz N, Duncan J, Mazurowski MA, 2022. Deep Learning for Breast MRI Style Transfer with Limited Training Data. J. Digit. Imaging doi:10.1007/s10278-022-00755-z.

Carré A, Battistella E, Niyoteka S, Sun R, Deutsch E, Robert C, 2022. AutoComBat: a generic method for harmonizing MRI-based radiomic features. Sci. Rep 12, 12762. doi:10.1038/s41598-022-16609-1. [PubMed: 35882891]

Casey BJ, Cannonier T, Conley MI, Cohen AO, Barch DM, Heitzeg MM, Soules ME, Teslovich T, Dellarco DV, Garavan H, Orr CA, Wager TD, Banich MT, Speer NK, Sutherland MT, Riedel MC, Dick AS, Bjork JM, Thomas KM, Chaarani B, Mejia MH, Hagler DJ, Daniela Cornejo M, Sicat CS, Harms MP, Dosenbach NUF, Rosenberg M, Earl E, Bartsch H, Watts R, Polimeni JR, Kuperman JM, Fair DA, Dale AM, 2018. The Adolescent Brain Cognitive Development (ABCD) study: imaging acquisition across 21 sites. Dev. Cognit. Neurosci 32, 43–54. doi:10.1016/j.dcn.2018.03.001. [PubMed: 29567376]

Cash DM, Rohrer JD, Ryan NS, Ourselin S, Fox NC, 2014. Imaging endpoints for clinical trials in Alzheimer's disease. Alzheimer's Res. Ther 6, 87. doi:10.1186/s13195-014-0087-9. [PubMed: 25621018]

Caspi A, Moffitt TE, 2018. All for one and one for all: mental disorders in one dimension. Am. J. Psychiatry 175, 831–844. doi:10.1176/appi.ajp.2018.17121383. [PubMed: 29621902]

Castaldo R, Brancato V, Cavaliere C, Trama F, Illiano E, Costantini E, Ragozzino A, Salvatore M, Nicolai E, Franzese M, 2022. A framework of analysis to facilitate the harmonization of multicenter radiomic features in prostate cancer. J. Clin. Med 12, 140. doi:10.3390/jcm12010140. [PubMed: 36614941]

Cetin Karayumak S, Bouix S, Ning L, James A, Crow T, Shenton M, Kubicki M, Rathi Y, 2019. Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. Neuroimage 184, 180–200. doi:10.1016/j.neuroimage.2018.08.073. [PubMed: 30205206]

Cetin-Karayumak S, Di Biase MA, Chunga N, Reid B, Somes N, Lyall AE, Kelly S, Solgun B, Pasternak O, Vangel M, Pearlson G, Tamminga C, Sweeney JA, Clementz B, Schretlen D, Viher PV, Stegmayer K, Walther S, Lee J, Crow T, James A, Voineskos A, Buchanan RW, Szeszko PR, Malhotra AK, Hegde R, McCarley R, Keshavan M, Shenton M, Rathi Y, Kubicki M, 2020a. White matter abnormalities across the lifespan of schizophrenia: a harmonized multi-site diffusion MRI study. Mol. Psychiatry 25, 3208–3219. doi:10.1038/s41380-019-0509-y. [PubMed: 31511636]

Cetin-Karayumak S, Stegmayer K, Walther S, Szeszko PR, Crow T, James A, Keshavan M, Kubicki M, Rathi Y, 2020b. Exploring the limits of ComBat method for multi-site diffusion MRI harmonization. 10.1101/2020.11.20.390120

Chang X, Cai X, Dan Y, Song Y, Lu Q, Yang G, Nie S, 2022. Self-supervised learning for multi-center magnetic resonance imaging harmonization without traveling phantoms. Phys. Med. Biol 67. doi:10.1088/1361-6560/ac7b66.

Chen AA, Beer JC, Tustison NJ, Cook PA, Shinohara RT, Shou HAlzheimer's Disease Neuroimaging Initiative, 2022a. Mitigating site effects in covariance for machine learning in neuroimaging data. Hum. Brain Mapp 43, 1179–1195. doi:10.1002/hbm.25688. [PubMed: 34904312]

Chen AA, Luo C, Chen Y, Shinohara RT, Shou H, 2022b. Privacypreserving harmonization via distributed ComBat. Neuroimage 248, 118822. doi:10.1016/j.neuroimage.2021.118822. [PubMed: 34958950]

Chen AA, Srinivasan D, Pomponio R, Fan Y, Nasrallah IM, Resnick SM, Beason-Held LL, Davatzikos C, Satterthwaite TD, Bassett DS, Shinohara RT, Shou H, 2022c. Harmonizing functional connectivity reduces scanner effects in community detection. Neuroimage 119198. doi:10.1016/j.neuroimage.2022.119198.

Chen C-L, Hsu Y-C, Yang L-Y, Tung Y-H, Luo W-B, Liu C-M, Hwang T-J, Hwu H-G, Isaac Tseng W-Y, 2020. Generalization of diffusion magnetic resonance imagingbased brain age prediction model through transfer learning. Neuroimage 217, 116831. doi:10.1016/j.neuroimage.2020.116831. [PubMed: 32438048]

Chen P, Yao H, Tijms BM, Wang P, Wang D, Song C, Yang H, Zhang Z, Zhao K, Qu Y, Kang X, Du K, Fan L, Han T, Yu C, Zhang X, Jiang T, Zhou Y, Lu J, Han Y, Liu B, Zhou B, Liu YAlzheimer's Disease Neuroimaging Initiative, 2022. Four distinct subtypes of Alzheimer's disease based on resting-state connectivity biomarkers. Biol. Psychiatry doi:10.1016/j.biopsych.2022.06.019, S0006-3223(22)01368-3.

Choi Y, Uh Y, Yoo J, Ha J-W, 2020. StarGAN v2: diverse image synthesis for multiple domains. 10.48550/arXiv.1912.01865

Choudhury S, Fishman JR, McGowan ML, Juengst ET, 2014. Big data, open science and the brain: lessons learned from genomics. Front. Hum. Neurosci 8.

Ciric R, Wolf DH, Power JD, Roalf DR, Baum GL, Ruparel K, Shinohara RT, Elliott MA, Eickhoff SB, Davatzikos C, Gur RC, Gur RE, Bassett DS, Satterthwaite TD, 2017. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. Neuroimage 154, 174–187. doi:10.1016/j.neuroimage.2017.03.020. [PubMed: 28302591]

Clarke WT, Mougin O, Driver ID, Rua C, Morgan AT, Asghar M, Clare S, Francis S, Wise RG, Rodgers CT, Carpenter A, Muir K, Bowtell R, 2020. Multisite harmonization of 7 tesla MRI neuroimaging protocols. Neuroimage 206, 116335. doi:10.1016/j.neuroimage.2019.116335. [PubMed: 31712167]

Crombé A, Buy X, Han F, Toupin S, Kind M, 2021. Assessment of repeatability, reproducibility, and performances of T2 mapping-based radiomics features: a comparative study. J. Magn. Reson. Imaging 54, 537–548. doi:10.1002/jmri.27558. [PubMed: 33594768]

Crombé A, Kind M, Fadli D, Le Loarer F, Italiano A, Buy X, Saut O, 2020. Intensity harmonization techniques influence radiomics features and radiomics-based predictions in sarcoma patients. Sci. Rep 10, 15496. doi:10.1038/s41598-020-72535-0. [PubMed: 32968131]

Da-Ano R, Lucia F, Masson I, Abgral R, Alfieri J, Rousseau C, Mervoyer A, Reinhold C, Pradier O, Schick U, Visvikis D, Hatt M, 2021. A transfer learning approach to facilitate ComBat-based harmonization of multicentre radiomic features in new datasets. PLoS One 16, e0253653. doi:10.1371/journal.pone.0253653. [PubMed: 34197503]

Da-Ano R, Masson I, Lucia F, Doré M, Robin P, Alfieri J, Rousseau C, Mervoyer A, Reinhold C, Castelli J, De Crevoisier R, Rameé JF, Pradier O, Schick U, Visvikis D, Hatt M, 2020a. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. Sci. Rep 10, 10248. doi:10.1038/s41598-020-66110-w. [PubMed: 32581221]

Da-Ano R, Visvikis D, Hatt M, 2020b. Harmonization strategies for multicenter radiomics investigations. Phys. Med. Biol 65, 24TR02. doi:10.1088/1361-6560/aba798.

Dai P, Xiong T, Zhou X, Ou Y, Li Y, Kui X, Chen Z, Zou B, Li W, Huang ZThe Rest-Meta-Mdd Consortium, null, 2022. The alterations of brain functional connectivity networks in major depressive disorder detected by machine learning through multisite rs-fMRI data. Behav. Brain Res 435, 114058. doi:10.1016/j.bbr.2022.114058. [PubMed: 35995263]

Dar SUH, Özbey M, Çatlı AB, Çukur T, 2020. A transfer-learning approach for accelerated MRI using deep neural networks. Magn. Reson. Med 84, 663–685. doi:10.1002/mrm.28148. [PubMed: 31898840]

Dar SUH, Yurt M, Karacan L, Erdem A, Erdem E, Çukur T, 2019. Image synthesis in multi-contrast MRI with conditional generative adversarial networks. IEEE Trans. Med. Imaging 38, 2375–2388. doi:10.1109/TMI.2019.2901750. [PubMed: 30835216]

de Brito Robalo BM, Biessels GJ, Chen C, Dewenter A, Duering M, Hilal S, Koek HL, Kopczak A, Yin Ka Lam B, Leemans A, Mok V, Onkenhout LP, van den Brink H, de Luca A, 2021. Diffusion MRI harmonization enables joint-analysis of multicentre data of patients with cerebral small vessel disease. NeuroImage: Clin. 32, 102886. doi:10.1016/j.nicl.2021.102886. [PubMed: 34911192]

de Brito Robalo BM, de Luca A, Chen C, Dewenter A, Duering M, Hilal S, Koek HL, Kopczak A, Lam BYK, Leemans A, Mok V, Onkenhout LP, van den Brink H, Biessels GJ, 2022. Improved sensitivity and precision in multicentre diffusion MRI network analysis using thresholding and harmonization. NeuroImage. Clinical 36, 103217. doi:10.1016/j.nicl.2022.103217. [PubMed: 36240537]

De Luca A, Karayumak SC, Leemans A, Rathi Y, Swinnen S, Gooijers J, Clauwaert A, Bahr R, Sandmo SB, Sochen N, Kaufmann D, Muehlmann M, Biessels G-J, Koerte I, Pasternak O, 2022. Cross-site harmonization of multi-shell diffusion MRI measures based on rotational invariant spherical harmonics (RISH). Neuroimage 259, 119439. doi:10.1016/j.neuroimage.2022.119439. [PubMed: 35788044]

De Stefano N, Battaglini M, Pareto D, Cortese R, Zhang J, Oesingmann N, Prados F, Rocca MA, Valsasina P, Vrenken H, Gandini Wheeler-Kingshott CAM, Filippi M, Barkhof F, Rovira ÀMAGNIMS Study Group, 2022. MAGNIMS recommendations for harmonization of MRI data in MS multicenter studies. NeuroImage. Clin 34, 102972. doi:10.1016/j.nicl.2022.102972. [PubMed: 35245791]

Denck J, Guehring J, Maier A, Rothgang E, 2021. MR-contrast-aware image-to-image translations with generative adversarial networks. Int. J. Comput. Assisted Radiol. Surg 16, 2069–2078. doi:10.1007/s11548-021-02433-x.

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L, 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Dercle L, Zhao B, Gönen M, Moskowitz CS, Firas A, Beylergil V, Connors DE, Yang H, Lu L, Fojo T, Carvajal R, Karovic S, Maitland ML, Goldmacher GV, Oxnard GR, Postow MA, Schwartz LH, 2022. Early readout on overall survival of patients with melanoma treated with immunotherapy using a novel imaging analysis. JAMA Oncol. 8, 385–392. doi:10.1001/jamaoncol.2021.6818. [PubMed: 35050320]

Dewey BE, Zhao C, Reinhold JC, Carass A, Fitzgerald KC, Sotirchos ES, Saidha S, Oh J, Pham DL, Calabresi PA, van Zijl PCM, Prince JL, 2019. DeepHarmony: a deep learning approach to contrast harmonization across scanner changes. Magn. Reson. Imaging 64, 160–170. doi:10.1016/j.mri.2019.05.041. [PubMed: 31301354]

Dewey BE, Zuo L, Carass A, He Y, Liu Y, Mowry EM, Newsome S, Oh J, Calabresi PA, Prince JL, 2020. A disentangled latent space for cross-site MRI harmonization. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L (Eds.), Medical Image Computing and Computer Assisted Intervention MICCAI 2020. Springer International Publishing, Cham, pp. 720–729. doi:10.1007/978-3-030-59728-3_70.

Di Martino A, Yan C-G, Li Q, Denio E, Castellanos FX, Alaerts K, Anderson JS, Assaf M, Bookheimer SY, Dapretto M, Deen B, Delmonte S, Dinstein I, Ertl-Wagner B, Fair DA, Gallagher L, Kennedy DP, Keown CL, Keysers C, Lainhart JE, Lord C, Luna B, Menon V, Minshew NJ, Monk CS, Mueller S, Müller R-A, Nebel MB, Nigg JT, O'Hearn K, Pelphrey KA, Peltier SJ, Rudie JD, Sunaert S, Thioux M, Tyszka JM, Uddin LQ, Verhoeven JS, Wenderoth N, Wiggins JL, Mostofsky SH, Milham MP, 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry 19, 659–667. doi:10.1038/mp.2013.78. [PubMed: 23774715]

Dinsdale NK, Jenkinson M, Namburete AIL, 2021. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. Neuroimage 228, 117689. doi:10.1016/j.neuroimage.2020.117689. [PubMed: 33385551]

Eshaghzadeh Torbati M, Minhas DS, Ahmad G, O'Connor EE, Muschelli J, Laymon CM, Yang Z, Cohen AD, Aizenstein HJ, Klunk WE, Christian BT, Hwang SJ, Crainiceanu CM, Tudorascu DL, 2021. A multi-scanner neuroimaging data harmonization using RAVEL and ComBat. Neuroimage 245, 118703. doi:10.1016/j.neuroimage.2021.118703. [PubMed: 34736996]

Fatania K, Clark A, Frood R, Scarsbrook A, Al-Qaisieh B, Currie S, Nix M, 2022. Harmonisation of scanner-dependent contrast variations in magnetic resonance imaging for radiation oncology, using style-blind auto-encoders. Phys. Imaging Radiat. Oncol 22, 115–122. doi:10.1016/j.phro.2022.05.005. [PubMed: 35619643]

Feis RA, Smith SM, Filippini N, Douaud G, Dopper EGP, Heise V, Trachtenberg AJ, van Swieten JC, van Buchem MA, Rombouts SARB, Mackay CE, 2015. ICA-based artifact removal diminishes scan site differences in multi-center resting-state fMRI. Front. Neurosci 9, 395. doi:10.3389/fnins.2015.00395. [PubMed: 26578859]

Fetty L, Bylund M, Kuess P, Heilemann G, Nyholm T, Georg D, Löfstedt T, 2020. Latent space manipulation for high-resolution medical image synthesis via the StyleGAN. Zeitschr. Med. Phys 30, 305–314. doi:10.1016/j.zemedi.2020.05.001. [PubMed: 32564924]

Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, Adams P, Cooper C, Fava M, McGrath PJ, McInnis M, Phillips ML, Trivedi MH, Weissman MM, Shinohara RT, 2018. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167, 104–120. doi:10.1016/j.neuroimage.2017.11.024. [PubMed: 29155184]

Fortin J-P, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, Roalf DR, Satterthwaite TD, Gur RC, Gur RE, Schultz RT, Verma R, Shinohara RT, 2017. Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161, 149–170. doi:10.1016/j.neuroimage.2017.08.047. [PubMed: 28826946]

Fortin J-P, Sweeney EM, Muschelli J, Crainiceanu CM, Shinohara RT, 2016. Removing inter-subject technical variability in magnetic resonance imaging studies. Neuroimage 132, 198–212. doi:10.1016/j.neuroimage.2016.02.036. [PubMed: 26923370]

Fournier L, Costaridou L, Bidaut L, Michoux N, Lecouvet FE, de Geus-Oei L-F, Boellaard R, Oprea-Lager DE, Obuchowski NA, Caroli A, Kunz WG, Oei EH, O'Connor JPB, Mayerhoefer ME, Franca M, Alberich-Bayarri A, Deroose CM, Loewe C, Manniesing R, Caramella C, Lopci E, Lassau N, Persson A, Achten R, Rosendahl K, Clement O, Kotter E, Golay X, Smits M, Dewey M, Sullivan DC, van der Lugt A, deSouza NMEuropean Society Of Radiology, 2021. Incorporating radiomics into clinical trials: expert consensus endorsed by the European Society of Radiology on considerations for data-driven compared to biologically driven quantitative biomarkers. Eur. Radiol 31, 6001–6012. doi:10.1007/s00330-020-07598-8. [PubMed: 33492473]

Garcia-Dias R, Scarpazza C, Baecker L, Vieira S, Pinaya WHL, Corvin A, Redolfi A, Nelson B, Crespo-Facorro B, McDonald C, Tordesillas-Gutiérrez D, Cannon D, Mothersill D, Hernaus D, Morris D, Setien-Suero E, Donohoe G, Frisoni G, Tronchin G, Sato J, Marcelis M, Kempton M, van Haren NEM, Gruber O, McGorry P, Amminger P, McGuire P, Gong Q, Kahn RS, Ayesa-Arriola R, van Amelsvoort T, Ortiz-García de la Foz V, Calhoun V, Cahn W, Mechelli A, 2020. Neuroharmony: a new tool for harmonizing volumetric MRI data from unseen scanners. Neuroimage 220, 117127. doi:10.1016/j.neuroimage.2020.117127. [PubMed: 32634595]

Grech-Sollars M, Hales PW, Miyazaki K, Raschke F, Rodriguez D, Wilson M, Gill SK, Banks T, Saunders DE, Clayden JD, Gwilliam MN, Barrick TR, Morgan PS, Davies NP, Rossiter J, Auer DP, Grundy R, Leach MO, Howe FA, Peet AC, Clark CA, 2015. Multi-centre reproducibility of diffusion MRI parameters for clinical sequences in the brain. NMR Biomed. 28, 468–485. doi:10.1002/nbm.3269. [PubMed: 25802212]

Griffanti L, Salimi-Khorshidi G, Beckmann CF, Auerbach EJ, Douaud G, Sexton CE, Zsoldos E, Ebmeier KP, Filippini N, Mackay CE, Moeller S, Xu J, Yacoub E, Baselli G, Ugurbil K, Miller KL, Smith SM, 2014. ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. Neuroimage 95, 232–247. doi:10.1016/j.neuroimage.2014.03.034. [PubMed: 24657355]

Guan H, Liu S, Lin W, Yap P-T, Liu M, 2022. Fast Image-Level MRI Harmonization via Spectrum Analysis. In: Machine learning in medical imaging. MLMI (Workshop), 13583, pp. 201–209. doi:10.1007/978-3-031-21014-3_21.

Guan H, Liu Y, Yang E, Yap P-T, Shen D, Liu M, 2021. Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. Med. Image Anal 71, 102076. doi:10.1016/j.media.2021.102076. [PubMed: 33930828]

Gutierrez A, Tuladhar A, Wilms M, Rajashekar D, Hill MD, Demchuk A, Goyal M, Fiehler J, Forkert ND, 2023. Lesion-preserving unpaired image-to-image translation between MRI and CT from ischemic stroke patients. Int. J. Comput. Assisted Radiol. Surg doi:10.1007/s11548-022-02828-4.

Haddad E, Pizzagalli F, Zhu AH, Bhatt RR, Islam T, Ba Gari I, Dixon D, Thomopoulos SI, Thompson PM, Jahanshad N, 2022. Multisite test-retest reliability and compatibility of brain metrics derived from FreeSurfer versions 7.1, 6.0, and 5.3. Hum. Brain Mapp doi:10.1002/hbm.26147.

Hagler DJ, Hatton S, Cornejo MD, Makowski C, Fair DA, Dick AS, Sutherland MT, Casey BJ, Barch DM, Harms MP, Watts R, Bjork JM, Garavan HP, Hilmer L, Pung CJ, Sicat CS, Kuperman J, Bartsch H, Xue F, Heitzeg MM, Laird AR, Trinh TT, Gonzalez R, Tapert SF, Riedel MC, Squeglia LM, Hyde LW, Rosenberg MD, Earl EA, Howlett KD, Baker FC, Soules M, Diaz J, de Leon OR, Thompson WK, Neale MC, Herting M, Sowell ER, Alvarez RP, Hawes SW, Sanchez M, Bodurka J, Breslin FJ, Morris AS, Paulus MP, Simmons WK, Polimeni JR, van der Kouwe A, Nencka AS, Gray KM, Pierpaoli C, Matochik JA, Noronha A, Aklin WM, Conway K, Glantz M, Hoffman E, Little R, Lopez M, Pariyadath V, Weiss SR, Wolff-Hughes DL, DelCarmen-Wiggins R, Feldstein Ewing SW, Miranda-Dominguez O, Nagel BJ, Perrone AJ, Sturgeon DT, Goldstone A, Pfefferbaum A, Pohl KM, Prouty D, Uban K, Bookheimer SY, Dapretto M, Galvan A, Bagot K, Giedd J, Infante MA, Jacobus J, Patrick K, Shilling PD, Desikan R, Li Y, Sugrue L, Banich MT, Friedman N, Hewitt JK, Hopfer C, Sakai J, Tanabe J, Cottler LB, Nixon SJ, Chang L, Cloak C, Ernst T, Reeves G, Kennedy DN, Heeringa S, Peltier S, Schulenberg J, Sripada C, Zucker RA, Iacono WG, Luciana M, Calabro FJ, Clark DB, Lewis DA, Luna B, Schirda C, Brima T, Foxe JJ, Freedman EG, Mruzek DW, Mason MJ, Huber R, McGlade E, Prescot A, Renshaw PF, Yurgelun-Todd DA, Allgaier NA, Dumas JA, Ivanova M, Potter A, Florsheim P, Larson C, Lisdahl K, Charness ME, Fuemmeler B, Hettema JM, Maes HH, Steinberg J, Anokhin AP, Glaser P, Heath AC, Madden PA, Baskin-Sommers A, Constable RT, Grant SJ, Dowling GJ, Brown SA, Jernigan TL, Dale AM, 2019. Image processing and analysis methods for the adolescent brain cognitive development study. Neuroimage 202, 116091. doi:10.1016/j.neuroimage.2019.116091. [PubMed: 31415884]

Han X, Jovicich J, Salat D, van der Kouwe A, Quinn B, Czanner S, Busa E, Pacheco J, Albert M, Killiany R, Maguire P, Rosas D, Makris N, Dale A, Dickerson B, Fischl B, 2006. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. Neuroimage 32, 180–194. doi:10.1016/j.neuroimage.2006.02.051. [PubMed: 16651008]

Hansen CB, Schilling KG, Rheault F, Resnick S, Shafer AT, Beason-Held LL, Landman BA, 2022. Contrastive semi-supervised harmonization of single-shell to multi-shell diffusion MRI. Magn. Reson. Imaging 93, 73–86. doi:10.1016/j.mri.2022.06.004. [PubMed: 35716922]

Harms MP, Somerville LH, Ances BM, Andersson J, Barch DM, Bastiani M, Bookheimer SY, Brown TB, Buckner RL, Burgess GC, Coalson TS, Chappell MA, Dapretto M, Douaud G, Fischl B, Glasser MF, Greve DN, Hodge C, Jamison KW, Jbabdi S, Kandala S, Li X, Mair RW, Mangia S, Marcus D, Mascali D, Moeller S, Nichols TE, Robinson EC, Salat DH, Smith SM, Sotiropoulos SN, Terpstra M, Thomas KM, Tisdall MD, Ugurbil K, van der Kouwe A, Woods RP, Zöllei L, Van Essen DC, Yacoub E, 2018. Extending the human connectome project across ages: imaging protocols for the lifespan development and aging projects. Neuroimage 183, 972–984. doi:10.1016/j.neuroimage.2018.09.060. [PubMed: 30261308]

Hatton SN, Huynh KH, Bonilha L, Abela E, Alhusaini S, Altmann A, Alvim MKM, Balachandra AR, Bartolini E, Bender B, Bernasconi N, Bernasconi A, Bernhardt B, Bargallo N, Caldairou B, Caligiuri ME, Carr SJA, Cavalleri GL, Cendes F, Concha L, Davoodi-Bojd E, Desmond PM, Devinsky O, Doherty CP, Domin M, Duncan JS, Focke NK, Foley SF, Gambardella A, Gleichgerrcht E, Guerrini R, Hamandi K, Ishikawa A, Keller SS, Kochunov PV, Kotikalapudi R, Kreilkamp BAK, Kwan P, Labate A, Langner S, Lenge M, Liu M, Lui E, Martin P, Mascalchi M, Moreira JCV, Morita-Sherman ME, O'Brien TJ, Pardoe HR, Pariente JC, Ribeiro LF, Richardson

MP, Rocha CS, Rodríguez-Cruces R, Rosenow F, Severino M, Sinclair B, Soltanian-Zadeh H, Striano P, Taylor PN, Thomas RH, Tortora D, Velakoulis D, Vezzani A, Vivash L, von Podewils F, Vos SB, Weber B, Winston GP, Yasuda CL, Zhu AH, Thompson PM, Whelan CD, Jahanshad N, Sisodiya SM, McDonald CR, 2020. White matter abnormalities across different epilepsy syndromes in adults: an ENIGMA-Epilepsy study. Brain: J. Neurol 143, 2454–2473. doi:10.1093/brain/awaa200.

Hawco C, Dickie EW, Herman G, Turner JA, Argyelan M, Malhotra AK, Buchanan RW, Voineskos AN, 2022. A longitudinal multi-scanner multimodal human neuroimaging dataset. Sci. Data 9, 332. doi:10.1038/s41597-022-01386-3. [PubMed: 35701471]

He Y, Carass A, Zuo L, Dewey BE, Prince JL, 2021. Autoencoder based self-supervised test-time adaptation for medical image analysis. Med. Image Anal 72, 102136. doi:10.1016/j.media.2021.102136. [PubMed: 34246070]

Hellier P, 2003. Consistent intensity correction of MR images. In: Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429), pp. I–1109. doi:10.1109/ICIP.2003.1247161.

Hernández AV, Eijkemans MJC, Steyerberg EW, 2006. Randomized controlled trials with time-to-event outcomes: how much does prespecified covariate adjustment increase power? Ann. Epidemiol 16, 41–48. doi:10.1016/j.annepidem.2005.09.007. [PubMed: 16275011]

Hernández AV, Steyerberg EW, Habbema JDF, 2004. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. J. Clin. Epidemiol 57, 454–460. doi:10.1016/j.jclinepi.2003.09.014. [PubMed: 15196615]

Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S, 2018. GANs trained by a two time-scale update rule converge to a local nash equilibrium. 10.48550/arXiv.1706.08500

Höfler M, 2005. Causal inference based on counterfactuals. BMC Med. Res. Method 5, 28. doi:10.1186/1471-2288-5-28.

Hognon C, Tixier F, Gallinato O, Colin T, Visvikis D, Jaouen V, 2019. Standardization of multicentric image datasets with generative adversarial networks. IEEE Nuclear Science Symposium and Medical Imaging Conference 2019.

Hong J, Hwang J, Lee J-H, 2022. General psychopathology factor (p-factor) prediction using resting-state functional connectivity and a scanner-generalization neural network. J. Psychiatr. Res 158, 114–125. doi:10.1016/j.jpsychires.2022.12.037. [PubMed: 36580867]

Horien C, Noble S, Greene AS, Lee K, Barron DS, Gao S, O'Connor D, Salehi M, Dadashkarimi J, Shen X, Lake EMR, Constable RT, Scheinost D, 2021. A hitchhiker's guide to working with large, open-source neuroimaging datasets. Nat. Hum. Behav 5, 185–193. doi:10.1038/s41562-020-01005-4. [PubMed: 33288916]

Horn JDV, Grafton ST, Rockmore D, Gazzaniga MS, 2004. Sharing neuroimaging studies of human cognition. Nat. Neurosci 7, 473–481. doi:10.1038/nn1231. [PubMed: 15114361]

Horng H, Singh A, Yousefi B, Cohen EA, Haghighi B, Katz S, Noël PB, Kontos D, Shinohara RT, 2022a. Improved generalized ComBat methods for harmonization of radiomic features. Sci. Rep 12, 19009. doi:10.1038/s41598-022-23328-0. [PubMed: 36348002]

Horng H, Singh A, Yousefi B, Cohen EA, Haghighi B, Katz S, Noël PB, Shinohara RT, Kontos D, 2022b. Generalized ComBat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects. Sci. Rep 12, 4493. doi:10.1038/s41598-022-08412-9. [PubMed: 35296726]

Huang X, Belongie S, 2017. Arbitrary style transfer in real-time with adaptive instance normalization. 10.48550/arXiv.1703.06868

Ihalainen T, Sipilä O, Savolainen S, 2004. MRI quality control: six imagers studied using eleven unified image quality parameters. Eur. Radiol 14, 1859–1865. doi:10.1007/s00330-004-2278-4. [PubMed: 14997335]

Ingalhalikar M, Shinde S, Karmarkar A, Rajan A, Rangaprakash D, Deshpande G, 2021. Functional connectivity-based prediction of autism on site harmonized ABIDE dataset. IEEE Trans. Biomed. Eng 68, 3628–3637. doi:10.1109/TBME.2021.3080259. [PubMed: 33989150]

Jahanshad N, Kochunov PV, Sprooten E, Mandl RC, Nichols TE, Almasy L, Blangero J, Brouwer RM, Curran JE, de Zubicaray GI, Duggirala R, Fox PT, Hong LE, Landman BA, Martin NG, McMahon KL, Medland SE, Mitchell BD, Olvera RL, Peterson CP, Starr JM, Sussmann JE, Toga AW, Wardlaw JM, Wright MJ, Hulshoff Pol HE, Bastin ME, McIntosh AM, Deary IJ, Thompson PM, Glahn DC, 2013. Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: a pilot project of the ENIGMA working group. Neuroimage 81, 455–469. doi:10.1016/j.neuroimage.2013.04.061. [PubMed: 23629049]

Jenkinson M, Bannister P, Brady M, Smith S, 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17, 825–841. doi:10.1016/s1053-8119(02)91132-8. [PubMed: 12377157]

Johnson WE, Li C, Rabinovic A, 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127. doi:10.1093/biostatistics/kxj037. [PubMed: 16632515]

Jovicich J, Czanner S, Greve D, Haley E, van der Kouwe A, Gollub R, Kennedy D, Schmitt F, Brown G, Macfall J, Fischl B, Dale A, 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage 30, 436–443. doi:10.1016/j.neuroimage.2005.09.046. [PubMed: 16300968]

Jovicich J, Minati L, Marizzoni M, Marchitelli R, Sala-Llonch R, Bartrés-Faz D, Arnold J, Benninghoff J, Fiedler U, Roccatagliata L, Picco A, Nobili F, Blin O, Bombois S, Lopes R, Bordet R, Sein J, Ranjeva J-P, Didic M, Gros-Dagnac H, Payoux P, Zoccatelli G, Alessandrini F, Beltramello A, Bargalló N, Ferretti A, Caulo M, Aiello M, Cavaliere C, Soricelli A, Parnetti L, Tarducci R, Floridi P, Tsolaki M, Constantinidis M, Drevelegas A, Rossini PM, Marra C, Schönknecht P, Hensch T, Hoffmann K-T, Kuijer JP, Visser PJ, Barkhof F, Frisoni GB, Consortium, PharmaCog, 2016. Longitudinal reproducibility of defaultmode network connectivity in healthy elderly participants: a multicentric resting-state fMRI study. Neuroimage 124, 442–454. doi:10.1016/j.neuroimage.2015.07.010. [PubMed: 26163799]

Karras T, Laine S, Aila T, 2019. A style-based generator architecture for generative adversarial networks. 10.48550/arXiv.1812.04948

Kent DM, Trikalinos TA, Hill MD, 2009. Are unadjusted analyses of clinical trials inappropriately biased toward the null? Stroke 40, 672–673. doi:10.1161/STROKEAHA.108.532051. [PubMed: 19164784]

Keshavan A, Paul F, Beyer MK, Zhu AH, Papinutto N, Shinohara RT, Stern W, Amann M, Bakshi R, Bischof A, Carriero A, Comabella M, Crane JC, D'Alfonso S, Demaerel P, Dubois B, Filippi M, Fleischer V, Fontaine B, Gaetano L, Goris A, Graetz C, Gröger A, Groppa S, Hafler DA, Harbo HF, Hemmer B, Jordan K, Kappos L, Kirkish G, Llufriu S, Magon S, Martinelli-Boneschi F, McCauley JL, Montalban X, Mühlau M, Pelletier D, Pattany PM, Pericak-Vance M, Cournu-Rebeix I, Rocca MA, Rovira A, Schlaeger R, Saiz A, Sprenger T, Stecco A, Uitdehaag BMJ, Villoslada P, Wattjes MP, Weiner H, Wuerfel J, Zimmer C, Zipp F, Hauser SL, Oksenberg JR, Henry RG, 2016. Power estimation for non-standardized multisite studies. Neuroimage 134, 281–294. doi:10.1016/j.neuroimage.2016.03.051. [PubMed: 27039700]

Kia SM, Huijsdens H, Dinga R, Wolfers T, Mennes M, Andreassen OA, Westlye LT, Beckmann CF, Marquand AF, 2020. Hierarchical Bayesian regression for multi-site normative modeling of neuroimaging data. 10.48550/arXiv.2005.12055

Kieselmann JP, Fuller CD, Gurney-Champion OJ, Oelfke U, 2021. Cross-modality deep learning: contouring of MRI data from annotated CT data only. Med. Phys 48, 1673–1684. doi:10.1002/mp.14619. [PubMed: 33251619]

Kim S, Kim S-W, Noh Y, Lee PH, Na DL, Seo SW, Seong J-K, 2022. Harmonization of multicenter cortical thickness data by linear mixed effect model. Front. Aging Neurosci 14.

Kingma DP, Welling M, 2014. Auto-encoding variational Bayes

Koike S, Tanaka SC, Okada T, Aso T, Yamashita A, Yamashita O, Asano M, Maikusa N, Morita K, Okada N, Fukunaga M, Uematsu A, Togo H, Miyazaki A, Murata K, Urushibata Y, Autio J, Ose T, Yoshimoto J, Araki T, Glasser MF, Van Essen DC, Maruyama M, Sadato N, Kawato M, Kasai K, Okamoto Y, Hanakawa T, Hayashi T, 2021. Brain/MINDS beyond human brain MRI project: a protocol for multi-level harmonization across brain disorders throughout the lifespan. NeuroImage: Clin. 30, 102600. doi:10.1016/j.nicl.2021.102600. [PubMed: 33741307]

Kurokawa R, Kamiya K, Koike S, Nakaya M, Uematsu A, Tanaka SC, Kamagata K, Okada N, Morita K, Kasai K, Abe O, 2021. Cross-scanner reproducibility and harmonization of a diffusion MRI structural brain network: a traveling subject study of multi-b acquisition. Neuroimage 245, 118675. doi:10.1016/j.neuroimage.2021.118675. [PubMed: 34710585]

Larivière S, Royer J, Rodríguez-Cruces R, Paquola C, Caligiuri ME, Gambardella A, Concha L, Keller SS, Cendes F, Yasuda CL, Bonilha L, Gleichgerrcht E, Focke NK, Domin M, von Podewills F, Langner S, Rummel C, Wiest R, Martin P, Kotikalapudi R, O'Brien TJ, Sinclair B, Vivash L, Desmond PM, Lui E, Vaudano AE, Meletti S, Tondelli M, Alhusaini S, Doherty CP, Cavalleri GL, Delanty N, Kälviäinen R, Jackson GD, Kowalczyk M, Mascalchi M, Semmelroch M, Thomas RH, Soltanian-Zadeh H, Davoodi-Bojd E, Zhang J, Winston GP, Griffin A, Singh A, Tiwari VK, Kreilkamp BAK, Lenge M, Guerrini R, Hamandi K, Foley S, Rüber T, Weber B, Depondt C, Absil J, Carr SJA, Abela E, Richardson MP, Devinsky O, Severino M, Striano P, Tortora D, Kaestner E, Hatton SN, Vos SB, Caciagli L, Duncan JS, Whelan CD, Thompson PM, Sisodiya SM, Bernasconi A, Labate A, McDonald CR, Bernasconi N, Bernhardt BC, 2022. Structural network alterations in focal and generalized epilepsy assessed in a worldwide ENIGMA study follow axes of epilepsy risk gene expression. Nat. Commun 13, 4320. doi:10.1038/s41467-022-31730-5. [PubMed: 35896547]

Leithner D, Schoder H, Haug AR, Vargas HA, Gibbs P, Häggström I, Rausch I, Weber M, Becker AS, Schwartz J, Mayerhoefer ME, 2022. Impact of ComBat harmonization on PET radiomics-based tissue classification: a dualcenter PET/MR and PET/CT study. J. Nucl. Med.: Off. Publ., Soc. Nucl. Med doi:10.2967/jnumed.121.263102, jnumed.121.263102.

Li B, You X, Wang J, Peng Q, Yin S, Qi R, Ren Q, Hong Z, 2021. IASNET: joint intraclassly adaptive GAN and segmentation network for unsupervised cross-domain in neonatal brain MRI segmentation. Med. Phys 48, 6962–6975. doi:10.1002/mp.15212. [PubMed: 34494276]

Li T, Zhang Y, Patil P, Johnson WE, 2021. Overcoming the impacts of two-step batch effect correction on gene expression estimation and inference. Biostat. kxab039 doi:10.1093/biostatistics/kxab039.

Li X, Ai L, Giavasis S, Jin H, Feczko E, Xu T, Clucas J, Franco A, Heinsfeld AS, Adebimpe A, Vogelstein JT, Yan C-G, Esteban O, Poldrack RA, Craddock C, Fair D, Satterthwaite T, Kiar G, Milham MP, 2022. Moving beyond processing and analysis-related variation in neuroscience. 10.1101/2021.12.01.470790

Li Y, Ammari S, Balleyguier C, Lassau N, Chouzenoux E, 2021. Impact of preprocessing and harmonization methods on the removal of scanner effects in brain MRI radiomic features. Cancers 13, 3000. doi:10.3390/cancers13123000. [PubMed: 34203896]

Li Z, Gong T, Lin Z, He H, Tong Q, Li C, Sun Y, Yu F, Zhong J, 2019. Fast and robust diffusion Kurtosis parametric mapping using a three-dimensional convolutional neural network. IEEE Access 7, 71398–71411. doi:10.1109/ACCESS.2019.2919241.

Liu F, Xu J, Guo L, Qin W, Liang M, Schumann G, Yu C, 2022. Environmental neuroscience linking exposome to brain structure and function underlying cognition and behavior. Mol. Psychiatry doi:10.1038/s41380-022-01669-6.

Liu M, Maiti P, Thomopoulos S, Zhu A, Chai Y, Kim H, Jahanshad N, 2021. Style transfer using generative adversarial networks for multi-site MRI harmonization. In: Medical image computing and computer-assisted intervention: MICCAI … International Conference on Medical Image Computing and Computer-Assisted Intervention 12903, pp. 313–322. doi:10.1007/978-3-030-87199-4_30.

Liu Y, Nacewicz BM, Zhao G, Adluru N, Kirk GR, Ferrazzano PA, Styner MA, Alexander AL, 2020. A 3D fully convolutional neural network with top-down attention-guided refinement for accurate and robust automatic segmentation of amygdala and its subnuclei. Front. Neurosci 14.

Luna A, Bernanke J, Kim K, Aw N, Dworkin JD, Cha J, Posner J, 2021. Maturity of gray matter structures and white matter connectomes, and their relationship with psychiatric symptoms in youth. Hum. Brain Mapp 42, 4568–4579. doi:10.1002/hbm.25565. [PubMed: 34240783]

Ma D, Popuri K, Bhalla M, Sangha O, Lu D, Cao J, Jacova C, Wang L, Beg MFAlzheimer's Disease Neuroimaging Initiative, 2019. Quantitative assessment of field strength, total intracranial volume, sex, and age effects on the goodness of harmonization for volumetric analysis on the ADNI database. Hum. Brain Mapp 40, 1507–1527. doi:10.1002/hbm.24463. [PubMed: 30431208]

Maikusa N, Zhu Y, Uematsu A, Yamashita A, Saotome K, Okada N, Kasai K, Okanoya K, Yamashita O, Tanaka SC, Koike S, 2021. Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics. Hum. Brain Mapp 42, 5278–5287. doi:10.1002/hbm.25615. [PubMed: 34402132]

Malyarenko D, Galbán CJ, Londy FJ, Meyer CR, Johnson TD, Rehemtulla A, Ross BD, Chenevert TL, 2013. Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. J. Magnet. Reson. Imaging: JMRI 37, 1238–1246. doi:10.1002/jmri.23825. [PubMed: 23023785]

Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, Donohue MR, Foran W, Miller RL, Hendrickson TJ, Malone SM, Kandala S, Feczko E, Miranda-Dominguez O, Graham AM, Earl EA, Perrone AJ, Cordova M, Doyle O, Moore LA, Conan GM, Uriarte J, Snider K, Lynch BJ, Wilgenbusch JC, Pengo T, Tam A, Chen J, Newbold DJ, Zheng A, Seider NA, Van AN, Metoki A, Chauvin RJ, Laumann TO, Greene DJ, Petersen SE, Garavan H, Thompson WK, Nichols TE, Yeo BTT, Barch DM, Luna B, Fair DA, Dosenbach NUF, 2022. Reproducible brain-wide association studies require thousands of individuals. Nature 1–7. doi:10.1038/s41586-022-04492-9.

Marek S, Tervo-Clemmens B, Nielsen AN, Wheelock MD, Miller RL, Laumann TO, Earl E, Foran WW, Cordova M, Doyle O, Perrone A, Miranda-Dominguez O, Feczko E, Sturgeon D, Graham A, Hermosillo R, Snider K, Galassi A, Nagel BJ, Ewing SWF, Eggebrecht AT, Garavan H, Dale AM, Greene DJ, Barch DM, Fair DA, Luna B, Dosenbach NUF, 2019. Identifying reproducible individual differences in childhood functional brain networks: an ABCD study. Dev. Cognit. Neurosci 40, 100706. doi:10.1016/j.dcn.2019.100706. [PubMed: 31614255]

Mårtensson G, Ferreira D, Granberg T, Cavallin L, Oppedal K, Padovani A, Rektorova I, Bonanni L, Pardini M, Kramberger MG, Taylor J-P, Hort J, Snædal J, Kulisevsky J, Blanc F, Antonini A, Mecocci P, Vellas B, Tsolaki M, Kłoszewska I, Soininen H, Lovestone S, Simmons A, Aarsland D, Westman E, 2020. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. Med. Image Anal 66, 101714. doi:10.1016/j.media.2020.101714. [PubMed: 33007638]

McKeown MJ, Hansen LK, Sejnowsk TJ, 2003. Independent component analysis of functional MRI: what is signal and what is noise? Curr. Opin. Neurobiol 13, 620–629. doi:10.1016/j.conb.2003.09.012. [PubMed: 14630228]

Mckeown MJ, Makeig S, Brown GG, Jung T-P, Kindermann SS, Bell AJ, Sejnowski TJ, 1998. Analysis of fMRI data by blind separation into independent spatial components. Hum. Brain Mapp 6, 160–188 10.1002/(SICI)1097-0193(1998)6:3<160::AID-HBM5>3.0.CO;2-1. [PubMed: 9673671]

Meeter LH, Kaat LD, Rohrer JD, van Swieten JC, 2017. Imaging and fluid biomarkers in frontotemporal dementia. Nat. Rev. Neurol 13, 406–419. doi:10.1038/nrneurol.2017.75. [PubMed: 28621768]

Meyers B, Lee VK, Dennis L, Wallace J, Schmithorst V, Votava-Smith JK, Rajagopalan V, Herrup E, Baust T, Tran NN, Hunter J, Licht DJ, Gaynor JW, Andropoulos DB, Panigrahy A, Ceschin R, 2022. Harmonization of multicenter diffusion tensor tractography in neonates with congenital heart disease: optimizing post-processing and application of ComBat. Neuroimage. Rep 2, 100114. doi:10.1016/j.ynirp.2022.100114. [PubMed: 36258783]

Mikl M, Mareček R, Hluštík P, Pavlicová M, Drastich A, Chlebus P, Brázdil M, Krupa P, 2008. Effects of spatial smoothing on fMRI group inferences. Magn. Reson. Imaging 26, 490–503. doi:10.1016/j.mri.2007.08.006. [PubMed: 18060720]

Mirzaalian H, de Pierrefeu A, Savadjiev P, Pasternak O, Bouix S, Kubicki M, Westin C-F, Shenton ME, Rathi Y, 2015. Harmonizing diffusion MRI data across multiple sites and scanners. In: Medical image computing and computer-assisted intervention: MICCAI … International Conference on Medical Image Computing and Computer-Assisted Intervention, 9349, pp. 12–19. doi:10.1007/978-3-319-24553-9_2.

Mirzaalian H, Ning L, Savadjiev P, Pasternak O, Bouix S, Michailovich O, Grant G, Marx CE, Morey RA, Flashman LA, George MS, McAllister TW, Andaluz N, Shutter L, Coimbra R, Zafonte RD, Coleman MJ, Kubicki M, Westin CF, Stein MB, Shenton ME, Rathi Y, 2016. Inter-site

and inter-scanner diffusion MRI data harmonization. Neuroimage 135, 311–323. doi:10.1016/j.neuroimage.2016.04.041. [PubMed: 27138209]

Mirzaalian H, Ning L, Savadjiev P, Pasternak O, Bouix S, Michailovich O, Karmacharya S, Grant G, Marx CE, Morey RA, Flashman LA, George MS, McAllister TW, Andaluz N, Shutter L, Coimbra R, Zafonte RD, Coleman MJ, Kubicki M, Westin C-F, Stein MB, Shenton ME, Rathi Y, 2018. Multi-site harmonization of diffusion MRI data in a registration framework. Brain Imaging Behav. 12, 284–295. doi:10.1007/s11682-016-9670-y. [PubMed: 28176263]

Moyer D, Ver Steeg G, Tax CMW, Thompson PM, 2020. Scanner invariant representations for diffusion MRI harmonization. Magn. Reson. Med 84, 2174–2189. doi:10.1002/mrm.28243. [PubMed: 32250475]

Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, Trojanowski JQ, Toga AW, Beckett L, 2005. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's Dement. 1, 55–66. doi:10.1016/j.jalz.2005.06.003. [PubMed: 17476317]

Neuhaus JM, 1998. Estimation efficiency with omitted covariates in generalized linear models. J. Am. Statist. Assoc 93, 1124–1129. doi:10.1080/01621459.1998.10473773.

Nielson DM, Pereira F, Zheng CY, Migineishvili N, Lee JA, Thomas AG, Bandettini PA, 2018. Detecting and harmonizing scanner differences in the ABCD study - annual release 1.0. 10.1101/309260

Ning L, Bonet-Carne E, Grussu F, Sepehrband F, Kaden E, Veraart J, Blumberg SB, Khoo CS, Palombo M, Kokkinos I, Alexander DC, Coll-Font J, Scherrer B, Warfield SK, Karayumak SC, Rathi Y, Koppers S, Weninger L, Ebert J, Merhof D, Moyer D, Pietsch M, Christiaens D, Gomes Teixeira RA, Tournier J-D, Schilling KG, Huo Y, Nath V, Hansen C, Blaber J, Landman BA, Zhylka A, Pluim JPW, Parker G, Rudrapatna U, Evans J, Charron C, Jones DK, Tax CMW, 2020. Cross-scanner and cross-protocol multi-shell diffusion MRI data harmonization: algorithms and results. Neuroimage 221, 117128. doi:10.1016/j.neuroimage.2020.117128. [PubMed: 32673745]

Noble S, Scheinost D, Finn ES, Shen X, Papademetris X, McEwen SC, Bearden CE, Addington J, Goodyear B, Cadenhead KS, Mirzakhanian H, Cornblatt BA, Olvet DM, Mathalon DH, McGlashan TH, Perkins DO, Belger A, Seidman LJ, Thermenos H, Tsuang MT, van Erp TGM, Walker EF, Hamann S, Woods SW, Cannon TD, Constable RT, 2017. Multisite reliability of MR-based functional connectivity. Neuroimage 146, 959–970. doi:10.1016/j.neuroimage.2016.10.020. [PubMed: 27746386]

Nygaard V, Rødland EA, Hovig E, 2016. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics 17, 29–39. doi:10.1093/biostatistics/kxv027. [PubMed: 26272994]

Nyúl LG, Udupa JK, 1999. On standardizing the MR image intensity scale. Magn. Reson. Med 42, 1072–1081 10.1002/(sici)1522-2594(199912)42:6<1072::aid-mrm11>3.0.co;2-m. [PubMed: 10571928]

Onicas AI, Ware AL, Harris AD, Beauchamp MH, Beaulieu C, Craig W, Doan Q, Freedman SB, Goodyear BG, Zemek R, Yeates KO, Lebel C, 2022. Multisite harmonization of structural DTI networks in children: an A-CAP study. Front. Neurol 13, 850642. doi:10.3389/fneur.2022.850642. [PubMed: 35785336]

Gray LJ, Bath PMW, Collier TOptimising the Analysis of Stroke Trials (OAST) Collaboration, 2009. Should stroke trials adjust functional outcome for baseline prognostic factors? Stroke 40, 888–894. doi:10.1161/STROKEAHA.108.519207. [PubMed: 19164798]

Orlhac F, Lecler A, Savatovski J, Goya-Outi J, Nioche C, Charbonneau F, Ayache N, Frouin F, Duron L, Buvat I, 2021. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. Eur. Radiol 31, 2272–2280. doi:10.1007/s00330-020-07284-9. [PubMed: 32975661]

Pagani E, Storelli L, Pantano P, Petsas N, Tedeschi G, Gallo A, De Stefano N, Battaglini M, Rocca MA, Filippi MINNI Network, 2023. Multicenter data harmonization for regional brain atrophy and application in multiple sclerosis. J. Neurol 270, 446–459. doi:10.1007/s00415-022-11387-2. [PubMed: 36152049]

Parekh P, Vivek Bhalerao G, John JP, Venkatasubramanian GADBS consortium, 2022. Sample size requirement for achieving multisite harmonization using structural brain MRI features. Neuroimage 264, 119768. doi:10.1016/j.neuroimage.2022.119768. [PubMed: 36435343]

Pinto MS, Paolella R, Billiet T, Van Dyck P, Guns P-J, Jeurissen B, Ribbens A, den Dekker AJ, Sijbers J, 2020. Harmonization of brain diffusion MRI: concepts and methods. Front. Neurosci 14.

Poldrack RA, Gorgolewski KJ, 2014. Making big data open: data sharing in neuroimaging. Nat. Neurosci 17, 1510–1517. doi:10.1038/nn.3818. [PubMed: 25349916]

Polman CH, O'Connor PW, Havrdova E, Hutchinson M, Kappos L, Miller DH, Phillips JT, Lublin FD, Giovannoni G, Wajgt A, Toal M, Lynn F, Panzara MA, Sandrock AW, 2006. A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis. N. Engl. J. Med 354, 899–910. doi:10.1056/NEJ-Moa044397. [PubMed: 16510744]

Pomponio R, Erus G, Habes M, Doshi J, Srinivasan D, Mamourian E, Bashyam V, Nasrallah IM, Satterthwaite TD, Fan Y, Launer LJ, Masters CL, Maruff P, Zhuo C, Völzke H, Johnson SC, Fripp J, Koutsouleris N, Wolf DH, Gur R, Gur R, Morris J, Albert MS, Grabe HJ, Resnick SM, Bryan RN, Wolk DA, Shinohara RT, Shou H, Davatzikos C, 2020. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. Neuroimage 208, 116450. doi:10.1016/j.neuroimage.2019.116450. [PubMed: 31821869]

Qin Z, Liu Z, Zhu P, Ling W, 2022. Style transfer in conditional GANs for crossmodality synthesis of brain magnetic resonance images. Comput. Biol. Med 148, 105928. doi:10.1016/j.compbiomed.2022.105928. [PubMed: 35952543]

Radua J, Vieta E, Shinohara R, Kochunov P, Quidé Y, Green MJ, Weickert CS, Weickert T, Bruggemann J, Kircher T, Nenadić I, Cairns MJ, Seal M, Schall U, Henskens F, Fullerton JM, Mowry B, Pantelis C, Lenroot R, Cropley V, Loughland C, Scott R, Wolf D, Satterthwaite TD, Tan Y, Sim K, Piras F, Spalletta G, Banaj N, Pomarol-Clotet E, Solanes A, Albajes-Eizagirre A, Canales-Rodríguez EJ, Sarro S, Di Giorgio A, Bertolino A, Stäblein M, Oertel V, Knöchel C, Borgwardt S, du Plessis S, Yun J-Y, Kwon JS, Dannlowski U, Hahn T, Grotegerd D, Alloza C, Arango C, Janssen J, Díaz-Caneja C, Jiang W, Calhoun V, Ehrlich S, Yang K, Cascella NG, Takayanagi Y, Sawa A, Tomyshev A, Lebedeva I, Kaleda V, Kirschner M, Hoschl C, Tomecek D, Skoch A, van Amelsvoort T, Bakker G, James A, Preda A, Weideman A, Stein DJ, Howells F, Uhlmann A, Temmingh H, López-Jaramillo C, Díaz-Zuluaga A, Fortea L, Martinez-Heras E, Solana E, Llufriu S, Jahanshad N, Thompson P, Turner J, van Erp TENIGMA Consortium collaborators, 2020. Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. Neuroimage 218, 116956. doi:10.1016/j.neuroimage.2020.116956. [PubMed: 32470572]

Ravano V, Démonet J-F, Damian D, Meuli R, Piredda GF, Huelnhagen T, Maréchal B, Thiran J-P, Kober T, Richiardi J, 2022. Neuroimaging harmonization using cGANs: image similarity metrics poorly predict cross-protocol volumetric consistency. In: Abdulkadir A, Bathula DR, Dvornek NC, Habes M, Kia SM, Kumar V, Wolfers T (Eds.), Machine Learning in Clinical Neuroimaging, Lecture Notes in Computer Science. Springer Nature Switzerland, Cham, pp. 83–92. doi:10.1007/978-3-031-17899-3_9.

Reardon AM, Li K, Hu XP, 2021. Improving between-group effect size for multi-site functional connectivity data via site-wise de-meaning. Front. Computat. Neurosci 15, 762781. doi:10.3389/fncom.2021.762781.

Reynolds M, Chaudhary T, Torbati ME, Tudorascu DL, Batmanghelich K, Initiative, the A.D.N., 2022. ComBat harmonization: empirical bayes versus fully Bayes approaches. 10.1101/2022.07.13.499561

Richter S, Winzeck S, Correia MM, Kornaropoulos EN, Manktelow A, Outtrim J, Chatfield D, Posti JP, Tenovuo O, Williams GB, Menon DK, Newcombe VFJ, 2022. Validation of cross-sectional and longitudinal ComBat harmonization methods for magnetic resonance imaging data on a travelling subject cohort. Neuroimage. Rep 2. doi:10.1016/j.ynirp.2022.100136, None.

Roffet F, Delrieux C, Patow G, 2022. Assessing multi-site rs-fMRI-based connectomic harmonization using information theory. Brain Sci. 12, 1219. doi:10.3390/brainsci12091219. [PubMed: 36138956]

Rosenbaum PR, Rubin DB, 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55. doi:10.1093/biomet/70.1.41.

Rothman KJ, Greenland S, Lash TL, 2008. Modern Epidemiology. Lippincott Williams & Wilkins.

Saint Martin M-J, Orlhac F, Akl P, Khalid F, Nioche C, Buvat I, Malhaire C, Frouin F, 2021. A radiomics pipeline dedicated to Breast MRI: validation on a multiscanner phantom study. MAGMA 34, 355–366. doi:10.1007/s10334-020-00892-y. [PubMed: 33180226]

Salimi-Khorshidi G, Douaud G, Beckmann CF, Glasser MF, Griffanti L, Smith SM, 2014. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. Neuroimage 90, 449–468. doi:10.1016/j.neuroimage.2013.11.046. [PubMed: 24389422]

Saponaro S, Giuliano A, Bellotti R, Lombardi A, Tangaro S, Oliva P, Calderoni S, Retico A, 2022. Multi-site harmonization of MRI data uncovers machine-learning discrimination capability in barely separable populations: an example from the ABIDE dataset. NeuroImage: Clin. 35, 103082. doi:10.1016/j.nicl.2022.103082. [PubMed: 35700598]

Satterthwaite TD, Elliott MA, Ruparel K, Loughead J, Prabhakaran K, Calkins ME, Hopson R, Jackson C, Keefe J, Riley M, Mensh FD, Sleiman P, Verma R, Davatzikos C, Hakonarson H, Gur RC, Gur RE, 2014. Neuroimaging of the Philadelphia neurodevelopmental cohort. Neuroimage 86, 544–553. doi:10.1016/j.neuroimage.2013.07.064. [PubMed: 23921101]

Saunders KEA, Cipriani A, Rendell J, Attenburrow M-J, Nelissen N, Bilderbeck AC, Vasudevan SR, Churchill G, Goodwin GM, Nobre AC, Harmer CJ, Harrison PJ, Geddes JR, 2016. Oxford Lithium Trial (OxLith) of the early affective, cognitive, neural and biochemical effects of lithium carbonate in bipolar disorder: study protocol for a randomised controlled trial. Trials 17, 116. doi:10.1186/s13063-016-1230-7. [PubMed: 26936776]

Schwarz CG, 2021. Uses of human MR and PET imaging in research of neurodegenerative brain diseases. Neurotherapeutics 18, 661–672. doi:10.1007/s13311-021-01030-9. [PubMed: 33723751]

Selim M, Zhang J, Fei B, Zhang G-Q, Ge GY, Chen J, 2022. Cross-vendor CT image data harmonization using CVH-CT. In: AMIA Annual Symposium Proceedings 2021, pp. 1099–1108. [PubMed: 35308983]

Shao M, Zuo L, Carass A, Zhuo J, Gullapalli RP, Prince JL, 2022. Evaluating the impact of MR image harmonization on thalamus deep network segmentation. In: Proceedings of SPIE–the International Society for Optical Engineering 12032 doi:10.1117/12.2613159.

Shinohara RT, Oh J, Nair G, Calabresi PA, Davatzikos C, Doshi J, Henry RG, Kim G, Linn KA, Papinutto N, Pelletier D, Pham DL, Reich DS, Rooney W, Roy S, Stern W, Tummala S, Yousuf F, Zhu A, Sicotte NL, Bakshi R, Cooperative, the N., 2017. Volumetric analysis from a harmonized multisite brain MRI study of a single subject with multiple sclerosis. Am. J. Neuroradiol 38, 1501–1509. doi:10.3174/ajnr.A5254. [PubMed: 28642263]

Shinohara RT, Sweeney EM, Goldsmith J, Shiee N, Mateen FJ, Calabresi PA, Jarso S, Pham DL, Reich DS, Crainiceanu CM, 2014. Statistical normalization techniques for magnetic resonance imaging. NeuroImage: Clin. 6, 9–19. doi:10.1016/j.nicl.2014.08.008. [PubMed: 25379412]

Singh A, Horng H, Chitalia R, Roshkovan L, Katz SI, Noël P, Shinohara RT, Kontos D, 2022. Resampling and harmonization for mitigation of heterogeneity in image parameters of baseline scans. Sci. Rep 12, 21505. doi:10.1038/s41598-022-26083-4. [PubMed: 36513760]

Sinha S, Thomopoulos SI, Lam P, Muir A, Thompson PM, 2021. Alzheimer's disease classification accuracy is improved by MRI harmonization based on attention-guided generative adversarial networks. In: Proceedings of SPIE–the International Society for Optical Engineering 12088, 120880L doi:10.1117/12.2606155.

Smith SM, 2002. Fast robust automated brain extraction. Hum. Brain Mapp 17, 143–155. doi:10.1002/hbm.10062. [PubMed: 12391568]

Sohn K, Lee H, Yan X, 2015. Learning structured output representation using deep conditional generative models. Advances in Neural Information Processing Systems. Curran Associates, Inc.

Stamoulou E, Spanakis C, Manikis GC, Karanasiou G, Grigoriadis G, Foukakis T, Tsiknakis M, Fotiadis DI, Marias K, 2022. Harmonization strategies in multicenter MRI-based radiomics. J. Imaging 8, 303. doi:10.3390/jimaging8110303. [PubMed: 36354876]

Stonnington CM, Tan G, Klöppel S, Chu C, Draganski B, Jack CR, Chen K, Ashburner J, Frackowiak RSJ, 2008. Interpreting scan data acquired from multiple scanners: a study

with Alzheimer's disease. Neuroimage 39, 1180–1185. doi:10.1016/j.neuroimage.2007.09.066. [PubMed: 18032068]

Sun D, Rakesh G, Haswell CC, Logue M, Baird CL, O'Leary EN, Cotton AS, Xie H, Tamburrino M, Chen T, Dennis EL, Jahanshad N, Salminen LE, Thomopoulos SI, Rashid F, Ching CRK, Koch SBJ, Frijling JL, Nawijn L, van Zuiden M, Zhu X, Suarez-Jimenez B, Sierk A, Walter H, Manthey A, Stevens JS, Fani N, van Rooij SJH, Stein M, Bomyea J, Koerte IK, Choi K, van der Werff SJA, Vermeiren RRJM, Herzog J, Lebois LAM, Baker JT, Olson EA, Straube T, Korgaonkar MS, Andrew E, Zhu Y, Li G, Ipser J, Hudson AR, Peverill M, Sambrook K, Gordon E, Baugh L, Forster G, Simons RM, Simons JS, Magnotta V, Maron-Katz A, du Plessis S, Disner SG, Davenport N, Grupe DW, Nitschke JB, deRoon-Cassini TA, Fitzgerald JM, Krystal JH, Levy I, Olff M, Veltman DJ, Wang L, Neria Y, De Bellis MD, Jovanovic T, Daniels JK, Shenton M, van de Wee NJA, Schmahl C, Kaufman ML, Rosso IM, Sponheim SR, Hofmann DB, Bryant RA, Fercho KA, Stein DJ, Mueller SC, Hosseini B, Phan KL, McLaughlin KA, Davidson RJ, Larson CL, May G, Nelson SM, Abdallah CG, Gomaa H, Etkin A, Seedat S, Harpaz-Rotem I, Liberzon I, van Erp TGM, Quidé Y, Wang X, Thompson PM, Morey RA, 2022. A comparison of methods to harmonize cortical thickness measurements across scanners and sites. Neuroimage 261, 119509. doi:10.1016/j.neuroimage.2022.119509. [PubMed: 35917919]

Suttorp MM, Siegerink B, Jager KJ, Zoccali C, Dekker FW, 2015. Graphical presentation of confounding in directed acyclic graphs. Nephrol. Dialy. Transplant 30, 1418–1423. doi:10.1093/ndt/gfu325.

Tafuri B, Lombardi A, Nigro S, Urso D, Monaco A, Pantaleo E, Diacono D, De Blasi R, Bellotti R, Tangaro S, Logroscino G, 2022. The impact of harmonization on radiomic features in Parkinson's disease and healthy controls: a multicenter study. Front. Neurosci 16, 1012287. doi:10.3389/fnins.2022.1012287. [PubMed: 36300169]

Takao H, Hayashi N, Ohtomo K, 2014. Effects of study design in multiscanner voxel-based morphometry studies. Neuroimage 84, 133–140. doi:10.1016/j.neuroimage.2013.08.046. [PubMed: 23994315]

Takao H, Hayashi N, Ohtomo K, 2011. Effect of scanner in longitudinal studies of brain volume changes. J. Magn. Reson. Imaging 34, 438–444. doi:10.1002/jmri.22636. [PubMed: 21692137]

Tanaka SC, Yamashita A, Yahata N, Itahashi T, Lisi G, Yamada T, Ichikawa N, Takamura M, Yoshihara Y, Kunimatsu A, Okada N, Hashimoto R, Okada G, Sakai Y, Morimoto J, Narumoto J, Shimada Y, Mano H, Yoshida W, Seymour B, Shimizu T, Hosomi K, Saitoh Y, Kasai K, Kato N, Takahashi H, Okamoto Y, Yamashita O, Kawato M, Imamizu H, 2021. A multi-site, multi-disorder resting-state magnetic resonance image database. Sci. Data 8, 227. doi:10.1038/s41597-021-01004-8. [PubMed: 34462444]

Tang H, Xu D, Sebe N, Yan Y, 2019. Attention-guided generative adversarial networks for unsupervised image-to-image translation. 10.48550/arXiv.1903.12296

Tariot PN, Schneider LS, Cummings J, Thomas RG, Raman R, Jakimovich LJ, Loy R, Bartocci B, Fleisher A, Ismail MS, Porsteinsson A, Weiner M, Jack CR, Thal L, Aisen PSAlzheimer's Disease Cooperative Study Group, 2011. Chronic divalproex sodium to attenuate agitation and clinical progression of Alzheimer disease. Arch. Gen. Psychiatry 68, 853–861. doi:10.1001/archgenpsychiatry.2011.72. [PubMed: 21810649]

Tax CMW, Grussu F, Kaden E, Ning L, Rudrapatna U, John Evans C, St-Jean S, Leemans A, Koppers S, Merhof D, Ghosh A, Tanno R, Alexander DC, Zappalà S, Charron C, Kusmia S, Linden DEJ, Jones DK, Veraart J, 2019. Cross-scanner and cross-protocol diffusion MRI data harmonisation: a benchmark database and evaluation of algorithms. Neuroimage 195, 285–299. doi:10.1016/j.neuroimage.2019.01.077. [PubMed: 30716459]

Thieleking R, Zhang R, Paerisch M, Wirkner K, Anwander A, Beyer F, Villringer A, Witte AV, 2021. Same brain, different look?-The impact of scanner, sequence and preprocessing on diffusion imaging outcome parameters. J. Clin. Med 10, 4987. doi:10.3390/jcm10214987. [PubMed: 34768507]

Thomopoulos SI, Nir TM, Villalon-Reina JE, Zavaliangos-Petropulu A, Maiti P, Zheng H, Nourollahimoghadam E, Jahanshad N, Thompson PM, 2021. Diffusion MRI metrics and their relation to dementia severity: effects of harmonization approaches. In: 17th

International Symposium on Medical Information Processing and Analysis. SPIE, pp. 166–179. doi:10.1117/12.2606337.

Tian D, Zeng Z, Sun X, Tong Q, Li H, He H, Gao J-H, He Y, Xia M, 2022. A deep learning-based multisite neuroimage harmonization framework established with a traveling-subject dataset. Neuroimage 257, 119297. doi:10.1016/j.neuroimage.2022.119297. [PubMed: 35568346]

Tixier F, Jaouen V, Hognon C, Gallinato O, Colin T, Visvikis D, 2021. Evaluation of conventional and deep learning based image harmonization methods in radiomics studies. Phys. Med. Biol 66, 245009. doi:10.1088/1361-6560/ac39e5.

Tondelli M, Vaudano AE, Sisodiya SM, Meletti S, 2020. Valproate use is associated with posterior cortical thinning and ventricular enlargement in epilepsy patients. Front. Neurol 11, 622. doi:10.3389/fneur.2020.00622. [PubMed: 32714274]

Tong Q, Gong T, He H, Wang Z, Yu W, Zhang J, Zhai L, Cui H, Meng X, Tax CWM, Zhong J, 2020. A deep learning-based method for improving reliability of multicenter diffusion kurtosis imaging with varied acquisition protocols. Magn. Reson. Imaging 73, 31–44. doi:10.1016/j.mri.2020.08.001. [PubMed: 32822818]

Torbati ME, Minhas DS, Laymon CM, Maillard P, Wilson JD, Chen C-L, Crainiceanu CM, DeCarli CS, Hwang SJ, Tudorascu DL, 2022. MISPEL: a deep learning approach for harmonizing multi-scanner matched neuroimaging data. 10.1101/2022.07.27.501786

Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, Chen J, 2020. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 21, 12. doi:10.1186/s13059-019-1850-9. [PubMed: 31948481]

Treit S, Stolz E, Rickard JN, McCreary CR, Bagshawe M, Frayne R, Lebel C, Emery D, Beaulieu C, 2022. Lifespan volume trajectories from non-harmonized T1-weighted MRI do not differ after site correction based on traveling human phantoms. Front. Neurol 13, 826564. doi:10.3389/fneur.2022.826564. [PubMed: 35614930]

Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC, 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29, 1310–1320. doi:10.1109/TMI.2010.2046908. [PubMed: 20378467]

van de Bank BL, Emir UE, Boer VO, van Asten JJA, Maas MC, Wijnen JP, Kan HE, Oz G, Klomp DWJ, Scheenen TWJ, 2015. Multi-center reproducibility of neurochemical profiles in the human brain at 7 T. NMR Biomed. 28, 306–316. doi:10.1002/nbm.3252. [PubMed: 25581510]

van Dyck CH, Swanson CJ, Aisen P, Bateman RJ, Chen C, Gee M, Kanekiyo M, Li D, Reyderman L, Cohen S, Froelich L, Katayama S, Sabbagh M, Vellas B, Watson D, Dhadda S, Irizarry M, Kramer LD, Iwatsubo T, 2023. Lecanemab in early Alzheimer's disease. N. Engl. J. Med 388, 9–21. doi:10.1056/NEJMoa2212948. [PubMed: 36449413]

van Erp TGM, Greve DN, Rasmussen J, Turner J, Calhoun VD, Young S, Mueller B, Brown GG, McCarthy G, Glover GH, Lim KO, Bustillo JR, Belger A, McEwen S, Voyvodic J, Mathalon DH, Keator D, Preda A, Nguyen D, Ford JM, Potkin SG, Fbirn, null, 2014. A multi-scanner study of subcortical brain volume abnormalities in schizophrenia. Psychiatry Res. 222, 10–16. doi:10.1016/j.pscychresns.2014.02.011. [PubMed: 24650452]

Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K, 2013. The WU-Minn human connectome project: an overview. Neuroimage 80, 62–79. doi:10.1016/j.neuroimage.2013.05.041. [PubMed: 23684880]

Verma R, Swanson RL, Parker D, Ould Ismail AA, Shinohara RT, Alappatt JA, Doshi J, Davatzikos C, Gallaway M, Duda D, Chen HI, Kim JJ, Gur RC, Wolf RL, Grady MS, Hampton S, Diaz-Arrastia R, Smith DH, 2019. Neuroimaging findings in US government personnel with possible exposure to directional phenomena in Havana, Cuba. JAMA 322, 336–347. doi:10.1001/jama.2019.9269. [PubMed: 31334794]

Vogelbacher C, Sommer J, Schuster V, Bopp MHA, Falkenberg I, Ritter PS, Bermpohl F, Hindi Attar C, Rauer L, Einenkel KE, Treutlein J, Gruber O, Juckel G, Flasbeck V, Mulert C, Hautzinger M, Pfennig A, Matura S, Reif A, Grotegerd D, Dannlowski U, Kircher T, Bauer M, Jansen A, 2021. The German research consortium for the study of bipolar disorder (BipoLife): a magnetic resonance imaging study protocol. Int. J. Bipolar Disord 9, 37. doi:10.1186/s40345-021-00240-6. [PubMed: 34786613]

Wachinger C, Rieckmann A, Pölsterl S, 2021. Detect and correct bias in multi-site neuroimaging datasets. Med. Image Anal 67, 101879. doi:10.1016/j.media.2020.101879. [PubMed: 33152602]

Wang L, Lai HM, Barker GJ, Miller DH, Tofts PS, 1998. Correction for variations in MRI scanner sensitivity in brain studies with histogram matching. Magn. Reson. Med 39, 322–327. doi:10.1002/mrm.1910390222. [PubMed: 9469718]

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP, 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process 13, 600–612. doi:10.1109/TIP.2003.819861. [PubMed: 15376593]

Wengler K, Cassidy C, van der Pluijm M, Weinstein JJ, Abi-Dargham A, van de Giessen E, Horga G, 2021. Cross-scanner harmonization of neuromelanin-sensitive MRI for multisite studies. J. Magn. Reson. Imaging 54, 1189–1199. doi:10.1002/jmri.27679. [PubMed: 33960063]

Whitney HM, Li H, Ji Y, Liu P, Giger ML, 2021. Multi-stage harmonization for robust AI across breast MR databases. Cancers 13, 4809. doi:10.3390/cancers13194809. [PubMed: 34638294]

Whitney HM, Li H, Ji Y, Liu P, Giger ML, 2020. Harmonization of radiomic features of breast lesions across international DCE-MRI datasets. J. Med. Imaging 7, 012707. doi:10.1117/1.JMI.7.1.012707.

Wrobel J, Martin ML, Bakshi R, Calabresi PA, Elliot M, Roalf D, Gur RC, Gur RE, Henry RG, Nair G, Oh J, Papinutto N, Pelletier D, Reich DS, Rooney WD, Satterthwaite TD, Stern W, Prabhakaran K, Sicotte NL, Shinohara RT, Goldsmith J, Cooperative, NAIMS, 2020. Intensity warping for multisite MRI harmonization. Neuroimage 223, 117242. doi:10.1016/j.neuroimage.2020.117242. [PubMed: 32798678]

Xia M, Liu J, Mechelli A, Sun X, Ma Q, Wang X, Wei D, Chen Y, Liu B, Huang CC, Zheng Y, Wu Y, Chen T, Cheng Y, Xu X, Gong Q, Si T, Qiu S, Lin CP, Cheng J, Tang Y, Wang F, Qiu J, Xie P, Li L, Working Group, DIDA-MDD, He Y, 2022. Connectome gradient dysfunction in major depression and its association with gene expression profiles and treatment outcomes. Mol. Psychiatry 27, 1384–1393. doi:10.1038/s41380-022-01519-5. [PubMed: 35338312]

Xia M, Si T, Sun X, Ma Q, Liu B, Wang L, Meng J, Chang M, Huang X, Chen Z, Tang Y, Xu K, Gong Q, Wang F, Qiu J, Xie P, Li L, He YDIDA-Major Depressive Disorder Working Group, 2019. Reproducibility of functional brain alterations in major depressive disorder: evidence from a multisite resting-state functional MRI study with 1,434 individuals. Neuroimage 189, 700–714. doi:10.1016/j.neuroimage.2019.01.074. [PubMed: 30716456]

Yamashita A, Sakai Y, Yamada T, Yahata N, Kunimatsu A, Okada N, Itahashi T, Hashimoto R, Mizuta H, Ichikawa N, Takamura M, Okada G, Yamagata H, Harada K, Matsuo K, Tanaka SC, Kawato M, Kasai K, Kato N, Takahashi H, Okamoto Y, Yamashita O, Imamizu H, 2021. Common brain networks between major depressive-disorder diagnosis and symptoms of depression that are validated for independent cohorts. Front. Psychiatry 12, 667881. doi:10.3389/fpsyt.2021.667881. [PubMed: 34177657]

Yamashita A, Sakai Y, Yamada T, Yahata N, Kunimatsu A, Okada N, Itahashi T, Hashimoto R, Mizuta H, Ichikawa N, Takamura M, Okada G, Yamagata H, Harada K, Matsuo K, Tanaka SC, Kawato M, Kasai K, Kato N, Takahashi H, Okamoto Y, Yamashita O, Imamizu H, 2020. Generalizable brain network markers of major depressive disorder across multiple imaging sites. PLoS Biol. 18, e3000966. doi:10.1371/journal.pbio.3000966. [PubMed: 33284797]

Yamashita A, Yahata N, Itahashi T, Lisi G, Yamada T, Ichikawa N, Takamura M, Yoshihara Y, Kunimatsu A, Okada N, Yamagata H, Matsuo K, Hashimoto R, Okada G, Sakai Y, Morimoto J, Narumoto J, Shimada Y, Kasai K, Kato N, Takahashi H, Okamoto Y, Tanaka SC, Kawato M, Yamashita O, Imamizu H, 2019. Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. PLoS Biol. 17, e3000042. doi:10.1371/journal.pbio.3000042. [PubMed: 30998673]

Yang J, Dvornek NC, Zhang F, Chapiro J, Lin M, Duncan JS, 2019. Unsupervised domain adaptation via disentangled representations: application to cross-modality liver segmentation. In: Medical Image Computing and Computer-Assisted Intervention: MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention 11765, pp. 255–263. doi:10.1007/978-3-030-32245-8_29.

Yao K, Su Z, Huang K, Yang X, Sun J, Hussain A, Coenen F, 2022. A novel 3D unsupervised domain adaptation framework for cross-modality medical image segmentation. IEEE J. Biomed. Health Informat. PP doi:10.1109/JBHI.2022.3162118.

Yu M, Linn KA, Cook PA, Phillips ML, McInnis M, Fava M, Trivedi MH, Weissman MM, Shinohara RT, Sheline YI, 2018. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Hum. Brain Mapp 39, 4213–4227. doi:10.1002/hbm.24241. [PubMed: 29962049]

Zavaliangos-Petropulu A, Nir TM, Thomopoulos SI, Reid RI, Bernstein MA, Borowski B, Jack CR Jr., Weiner MW, Jahanshad N, Thompson PM, 2019. Diffusion MRI indices and their relation to cognitive impairment in brain aging: the updated multi-protocol approach in ADNI3. Front. Neuroinformat 13.

Zhang R, Oliver LD, Voineskos AN, Park JY, 2022. A structured multivariate approach for removal of latent batch effects. 10.1101/2022.08.01.502396

Zhao F, Wu Z, Wang L, Lin W, Xia S, Shen D, Li G, 2019. Harmonization of infant cortical thickness using surface-to-surface cycle-consistent adversarial networks. In: Medical image computing and computer-assisted intervention: MICCAI … International Conference on Medical Image Computing and Computer-Assisted Intervention 11767, pp. 475–483. doi:10.1007/978-3-030-32251-9_52.

Zhong J, Wang Y, Li J, Xue X, Liu S, Wang M, Gao X, Wang Q, Yang J, Li X, 2020. Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: application to neonatal white matter development. Biomed. Eng. Online 19, 4. doi:10.1186/s12938-020-0748-9. [PubMed: 31941515]

Zhu J-Y, Park T, Isola P, Efros AA, 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251. doi:10.1109/ICCV.2017.244.

Zindler T, Frieling H, Neyazi A, Bleich S, Friedel E, 2020. Simulating ComBat: how batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies. BMC Bioinf. 21, 271. doi:10.1186/s12859-020-03559-6.

Zuo L, Dewey BE, Liu Y, He Y, Newsome SD, Mowry EM, Resnick SM, Prince JL, Carass A, 2021. Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. Neuroimage 243, 118569. doi:10.1016/j.neuroimage.2021.118569. [PubMed: 34506916]

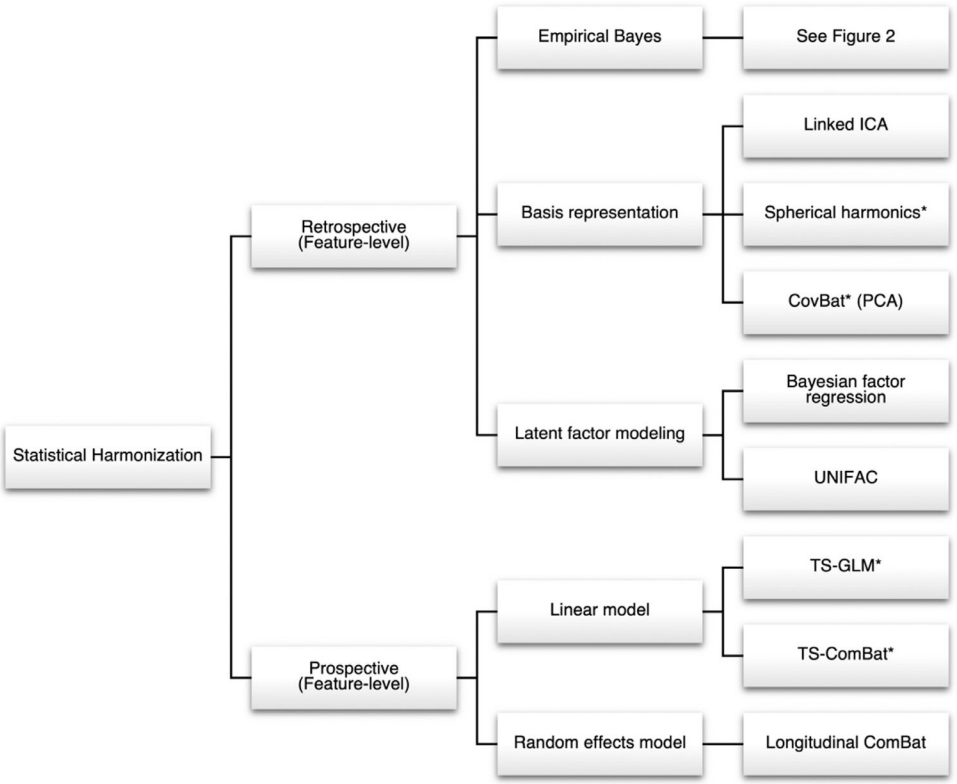

Fig. 1.

Flowchart of statistical models organized by study design and underlying model class. Asterisks indicate methods that have been evaluated in more than one study.

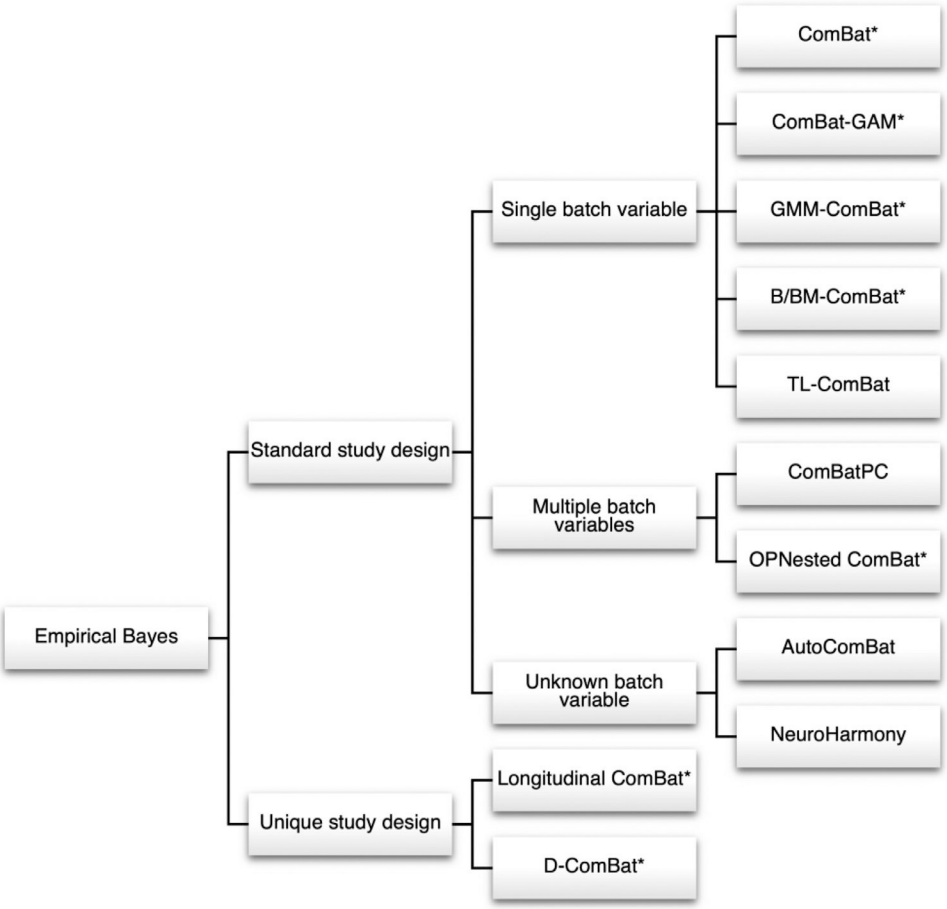

Fig. 2.

Flowchart of ComBat-based models organized by study design and underlying model class. All models presented in this figure perform feature-level harmonization in retrospective settings. Asterisks indicate methods that have been evaluated in more than one study.

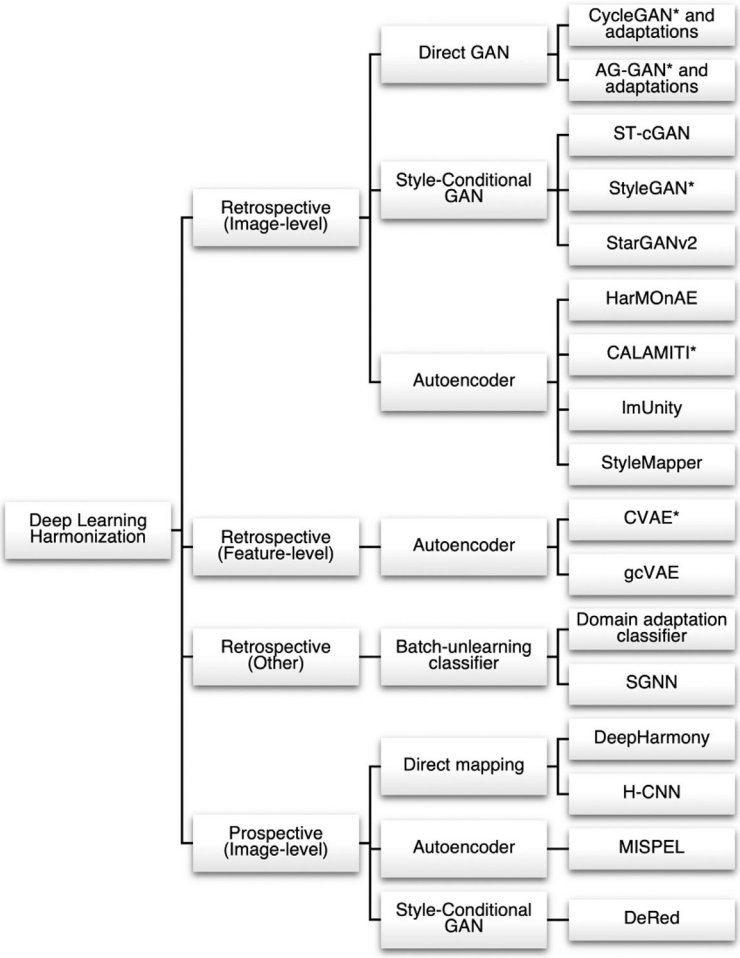

Fig. 3.

Flowchart of deep learning models organized by study design and underlying model class. Asterisks indicate methods that have been evaluated in more than one study.

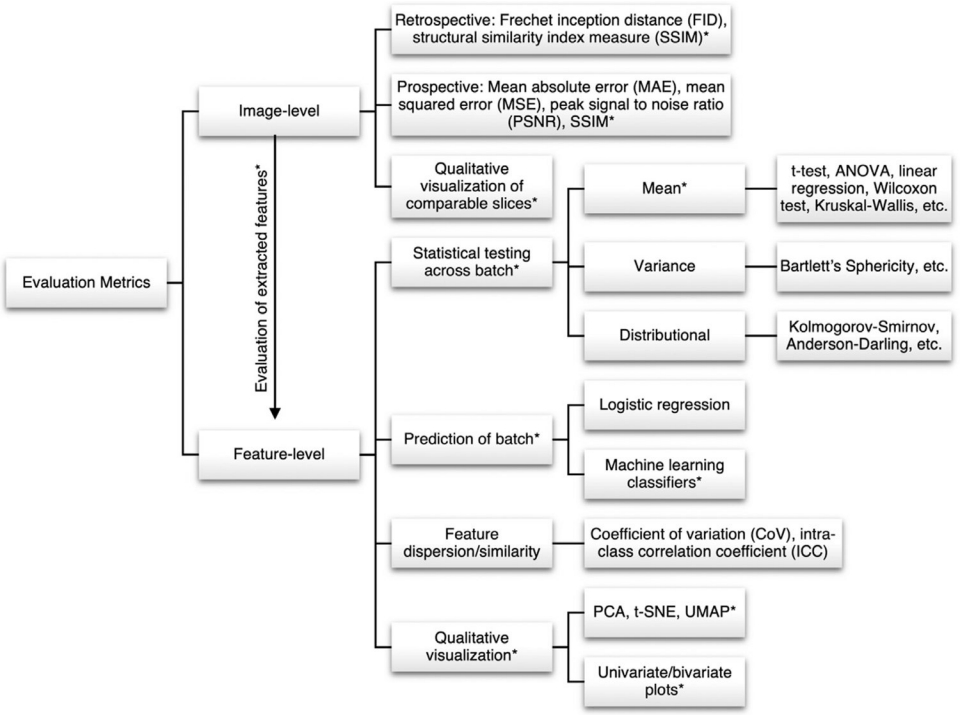

Fig. 4.

Flowchart of evaluation metrics for harmonization organized by data type and evaluation types. Asterisks indicate the set of standardized evaluation types that we believe should be included in the evaluation of novel harmonization methods, depending on data type and study design. Note that metrics included here are only for evaluating harmonization and do not include metrics for evaluating performance in downstream analyses.