Human Brain Mapping

Harmonization of diffusion MRI datasets with adaptive dictionary learning

Samuel St-Jean, Max A. Viergever, Alexander Leemans
Image Sciences Institute, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, the Netherlands

Abstract

Diffusion magnetic resonance imaging can indirectly infer the microstructure of tissues and provide metrics subject to normal variability in a population. Potentially abnormal values may yield essential information to support analysis of controls and patients cohorts, but subtle confounds could be mistaken for purely biologically driven variations amongst subjects. In this work, we propose a new harmonization algorithm based on adaptive dictionary learning to mitigate the unwanted variability caused by different scanner hardware while preserving the natural biological variability of the data. Our harmonization algorithm does not require paired training datasets, nor spatial registration or matching spatial resolution. Overcomplete dictionaries are learned iteratively from all datasets at the same time with an adaptive regularization criterion. removing variability attributable to the scanners in the process. The obtained mapping is applied directly in the native space of each subject towards a scanner-space. The method is evaluated with a public database which consists of two different protocols acquired on three different scanners. Results show that the effect size of the four studied diffusion metrics is preserved while removing variability attributable to the scanner. Experiments with alterations using a free water compartment, which is not simulated in the training data, shows that the modifications applied to the diffusion weighted images are preserved in the diffusion metrics after harmonization, while still reducing global variability at the same time. The algorithm could help multicenter studies pooling their data by removing scanner specific confounds, and increase statistical power in the process.

Keywords: Diffusion MRI, Harmonization, Scanner variability, Dictionary learning, Cross-validation, Akaike information criterion, Scanner-space

Corresponding author: Samuel St-Jean, email: samuel@isi.uu.nl

Data availability statement

The datasets from the CDMRI 2017 challenge are available from the organizers upon request at https://www.cardiff.ac.uk/cardiff-university-brain-research-imaging-centre/research/projects/cross-scanner-and-cross-protocol-diffusion-MRI-data-harmonisation. Code for the harmonization algorithm itself is available at https://github.com/samuelstjean/harmonization.

Funding statement

Samuel St-Jean is mainly supported by the Fonds de recherche du Québec - Nature et technologies (FRQNT) (Dossier 192865) and partly supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) [funding reference number BP–546283–2020] and the FRQNT (Dossier 290978). This research is supported by VIDI Grant 639.072.411 from the Netherlands Organization for Scientific Research (NWO). The funding agencies were not involved in the design, data collection nor interpretation of this study.

Conflict of interest statement

The authors have no conflict of interest to disclose.

Ethics approval statement

The original data acquisition was approved by Cardiff University School of Psychology ethics committee. Written informed consent was obtained from all subjects.

Permission to reproduce material from other sources

Some figures in the manuscript have been adapted from open access manuscripts available under the CC-BY license as indicated in their respective legend.

Acknowledgments

We would like to thank Chantal Tax for providing us with the evaluation masks and testing datasets from the challenge. The data were acquired at the UK National Facility for In Vivo MR Imaging of Human Tissue Microstructure located in CUBRIC funded by the EPSRC (grant EP/M029778/1), and The Wolfson Foundation. Acquisition and processing of the data was supported by a Rubicon grant from the NWO (680-50-1527), a Wellcome Trust Investigator Award (096646/Z/11/Z), and a Wellcome Trust Strategic Award (104943/Z/14/Z). This database was initiated by the 2017 and 2018 MICCAI Computational Diffusion MRI committees (Chantal Tax, Francesco Grussu, Enrico Kaden, Lipeng Ning, Jelle Veraart, Elisenda Bonet-Carne, and Farshid Sepehrband) and CUBRIC, Cardiff University (Chantal Tax, Derek Jones, Umesh Rudrapatna, John Evans, Greg Parker, Slawomir Kusmia, Cyril Charron, and David Linden).

ORCID

Samuel St-Jean - https://orcid.org/0000-0002-8092-2974
Alexander Leemans - https://orcid.org/0000-0002-9306-6126

1 Introduction

Diffusion weighted magnetic resonance imaging (dMRI) is a noninvasive imaging technique that can indirectly infer the microstructure of tissues based on the displacement of water molecules. As dMRI only offers an indirect way to study, e.g., the brain microstructure, analysis of dMRI datasets includes multiple processing steps to ensure adequate correction of acquisition artifacts due to subject motion or eddy current induced distortions, amongst others (Tournier, Mori, & Leemans, 2011). Quantitative scalar measures of diffusion can be extracted from the acquired datasets, such as the apparent diffusion coefficient (ADC) or fractional anisotropy (FA) as computed from diffusion tensor imaging (DTI) (P. Basser, Mattiello, & LeBihan, 1994; P. J. Basser & Pierpaoli, 1996), with a plethora of other measures and diffusion models nowadays available (Assemlal, Tschumperlé, Brun, & Siddiqi, 2011; Tournier, 2019). These measures are subject to normal variability across subjects and potentially abnormal values or features extracted from dMRI datasets may yield essential information to support analysis of controls and patients cohorts (Johansen-Berg & Behrens, 2009; Jones, 2011).

As small changes in the measured signal are ubiquitous due to differences in scanner hardware (Sakaie et al., 2018), software versions of the scanner or processing tools (Gronenschild et al., 2012; Sakaie et al., 2018), field strength of the magnet (Huisman et al., 2006) or reconstruction methods in parallel MRI and accelerated imaging (Dietrich et al., 2008; St-Jean, De Luca, Viergever, & Leemans, 2018), non-negligible effects may translate into small differences in the subsequently computed diffusion metrics. Subtle confounds affecting dMRI can even be due to measuring at different time points in the cardiac cycle, leading to changes in the measured values of pseudo-diffusion over the cardiac cycle (De Luca et al., 2019; Federau et al., 2013). In the presence of disease, these small variations in the measured signal are entangled in the genuine biological variability, which is usually the main criterion of interest to discover or analyze subsequently. This can lead to confounding effects and systematic errors that could be mistaken for purely biologically driven variations amongst subjects. To mitigate these issues, large-scale studies try to harmonize their acquisition protocols across centers to further reduce these potential sources of variability (Duchesne et al., 2019) or may only use a single scanner without upgrading it for long term studies (Hofman et al., 2015; Hofman, Grobbee, De Jong, & Van den Ouweland, 1991). The stability brought by keeping the same scanning hardware is however at the cost of potentially missing on improved, more efficient sequences or faster scanning methods becoming common in MRI (Feinberg et al., 2010; Larkman et al., 2001; Lustig, Donoho, & Pauly, 2007). Even by carefully controlling all these sources of variability as much as possible, there still remain reproducibility issues between scanners of the same model or in scan-rescan studies of dMRI metrics (Kristo et al., 2013; Magnotta et al., 2012; Vollmar et al., 2010). Over the years, many algorithms have been developed to mitigate the variability attributed to non-biological effects in dMRI, e.g., in order to combine datasets from multiple studies and increase statistical power, see e.g., (Pinto et al., 2020; Tax et al., 2019; Zhu, Moyer, Nir, Thompson, & Jahanshad, 2019) for reviews. Common approaches consist in harmonizing the dMRI datasets through the coefficients of a spherical harmonics representation (Blumberg et al., 2019; Cetin Karayumak et al., 2019; Mirzaalian et al., 2016) or the computed scalar metrics (Fortin et al., 2017; Pohl et al., 2016) to reduce vari-

ability between scanners. Recently, a dMRI benchmark database containing ten training subjects and four test subjects datasets acquired on three scanners with two acquisition protocols was presented at the computational diffusion MRI (CDMRI) 2017 challenge (Tax et al., 2019). The publicly available CDMRI database was previously used to compare five harmonization algorithms, including a previous version of the algorithm we present here, which we use for evaluation.

In this work, we propose a new algorithm based on adaptive dictionary learning to mitigate the unwanted variability caused by different scanner hardware while preserving the natural biological variability present in the data. The algorithm is applied directly on the dMRI datasets themselves without using an alternative representation and can be used on datasets acquired at different spatial resolutions or with a different set of diffusion sensitizing gradients (i.e., b-vectors). Expanding upon the methodology presented in St-Jean, Coupé, and Descoteaux (2016), St-Jean, Viergever, and Leemans (2017), overcomplete dictionaries are learned automatically from the data with an automatic tuning of the regularization parameter to balance the fidelity of the reconstruction with sparsity of the coefficients at every iteration. These dictionaries are either constructed using the data from a given source scanner and used to reconstruct the data from a different target scanner (first set of experiments) or learned using datasets coming from multiple scanners at once—creating a "scanner space" in the process (second set of experiments). One of the improvements of the algorithm is the ability to harmonize datasets acquired with multiple scanners, without explicitly needing to define a source and target scanner as is usually done. This new formulation also does not need to match the gradient directions (i.e., the b-vectors) of the other datasets. In the first set of experiments, these dictionaries are used to reconstruct the data with a dictionary from a different target scanner, removing variability present in the source scanner in the process. Mapping across different spatial resolutions can be obtained by adequate subsampling of the dictionary. In the second set of experiments, the test datasets are altered with simulations mimicking edema while the training datasets are left untouched. We show that the harmonization algorithm preserves the natural variability of the data, even if these alterations are not part of the training datasets. This is done by mapping all the datasets towards a global scanner-space, which can be done for multiple scanners at once *without* paired datasets or spatial registration of subjects to do so. Removing the prerequisite of paired datasets for training makes the algorithm easy to apply for hard to acquire datasets (e.g., patients with Alzheimer's, Parkinson's or Huntington's disease) or when pooling datasets from unrelated studies that are acquired in separate centers. This makes our proposed method readily applicable for pre-existing and ongoing studies that would like to remove variability caused by non-biological or systematic effects in their data analyzes.

2 Theory

2.1 The dictionary learning algorithm

Dictionary learning (Elad & Aharon, 2006; Mairal, Bach, Ponce, & Sapiro, 2010) aims to find a set of basis elements to efficiently approximate a given set of input vectors. We follow here in general our previous formulation from (Tax et al., 2019) which optimizes both the representation $\boldsymbol{D}$ (called the dictionary or the set of atoms) and the coefficients

$\boldsymbol{\alpha}$ of that representation (called the sparse codes) as opposed to using a fixed basis (e.g., Fourier, wavelets, spherical harmonics). A dictionary can be chosen to be overcomplete (i.e., more column than rows) as the algorithm is designed to only select a few atoms to approximate the input vector with a penalization on the ℓ_1-norm of $\boldsymbol{\alpha}$ to promote a sparse solution. Applications in computer vision with the goal to reduce visual artifacts include demosaicking (Mairal, Bach, Ponce, Sapiro, & Zisserman, 2009), inpainting (Mairal et al., 2010) and upsampling (Yang, Wang, Lin, Cohen, & Huang, 2012; Yang, Wright, Huang, & Ma, 2010) amongst others.

In practice, local windows are used to extract spatial and angular neighborhoods of diffusion weighted images (DWIs) inside a brain mask to create the set of vectors required for dictionary learning as in St-Jean et al. (2016). This is done by first extracting a small 3D region from a single DWI, which we now refer to as a patch. To include angular information, a set of patches is taken at the same spatial location across DWIs in an angular neighborhood (as defined by the angle between their associated b-vector on the sphere). This considers that patches from different DWIs at the same spatial location, but which are in fact not too far on the sphere, exhibit self-similarity that can be exploited by dictionary learning. Once this process is done, every set of patches is concatenated to a single vector $\boldsymbol{X}$. All of these vectors $\boldsymbol{X}_n$ are then put in a 2D matrix $\Omega = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \ldots\}$, where n denotes one of the individual set of patches.

Once the set of patches Ω has been extracted, $\boldsymbol{D}$ can be initialized by randomly selecting N vectors from Ω (Mairal et al., 2010). With this initial overcomplete dictionary, a sparse vector $\boldsymbol{\alpha}_n$ can be computed for each $\boldsymbol{X}_n$ such that $\boldsymbol{D}$ is a good approximation to reconstruct $\boldsymbol{X}_n$, that is $\boldsymbol{X}_n \approx \boldsymbol{D}\boldsymbol{\alpha}_n$. This initial approximation can be refined iteratively by sampling randomly N new vectors $\boldsymbol{X}_n \in \Omega$ and updating $\boldsymbol{D}$ to better approximate those vectors. At the next iteration, a new set $\boldsymbol{X}_n \in \Omega$ is randomly drawn and $\boldsymbol{D}$ is updated to better approximate this new set of vectors. This iterative process can be written as

$$\underset{\boldsymbol{D},\boldsymbol{\alpha}}{\arg\min} \ \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{2}\|\boldsymbol{X}_n - \boldsymbol{D}\boldsymbol{\alpha}_n\|_2^2 + \lambda_i\|\boldsymbol{\alpha}_n\|_1\right) \ \text{s.t.} \ \|\boldsymbol{D}_{.p}\|_2^2 = 1 \tag{1}$$

with $\boldsymbol{\alpha}_n \in \mathbb{R}^{p\times 1}$ an array of sparse coefficients and $\boldsymbol{D}$ the dictionary where each column is constrained to unit ℓ_2-norm to prevent degenerated solutions. λ_i is a regularization parameter used at iteration i (which is further detailed in Section 2.2) to balance the ℓ_2-norm promoting data similarity and the ℓ_1-norm promoting sparsity of the coefficients $\boldsymbol{\alpha}_n$. Iterative updates using Equation (1) alternate between refining $\boldsymbol{D}$ (and holding $\boldsymbol{\alpha}$ fixed) and computing $\boldsymbol{\alpha}$ (with $\boldsymbol{D}$ held fixed) for the current set of $\boldsymbol{X}_n$. As updating $\boldsymbol{\alpha}$ needs an optimization scheme, this can be done independently for each $\boldsymbol{\alpha}_n$ using coordinate descent (Friedman, Hastie, & Tibshirani, 2010). For updating $\boldsymbol{D}$, we use the parameter-free closed form update from Mairal et al. (2010, Algorithm 2), which only requires storing intermediary matrices of the previous iteration using $\boldsymbol{\alpha}$ and $\boldsymbol{X}_n$ to update $\boldsymbol{D}$. Building dictionaries for the task at hand has been used previously in the context of diffusion MRI for denoising (Gramfort, Poupon, & Descoteaux, 2014; St-Jean et al., 2016) and compressed sensing (Gramfort et al., 2014; Merlet, Caruyer, Ghosh, & Deriche, 2013; Schwab, Vidal, & Charon, 2018) amongst other tasks. Note that it is also possible to design dictionaries based on products of fixed basis or adding additional constraints such as positivity or

spatial consistency to Equation (1), see e.g., (Schwab et al., 2018; Vemuri et al., 2019) and references therein for examples pertaining to diffusion MRI.

2.2 Automatic regularization selection

Equation (1) relies on a regularization term λ_i which can be different for each set of vectors $\boldsymbol{X}_n$ at iteration i. It is, however, common to fix λ_i for all $\boldsymbol{X}_n$ depending on some heuristics such as the size of $\boldsymbol{X}_n$ (Mairal et al., 2010), the local noise variance (St-Jean et al., 2016) or through a grid search (Gramfort et al., 2014). In the present work, we instead rely on an automatic tuning criterion since datasets acquired on multiple scanners are subject to different local noise properties and of various signal-to-noise ratio (SNR) spatially. In addition, the datasets do not need to be at the same spatial resolution; defining a single scalar value for the regularization parameter as done in previous works is therefore not straightforward anymore. In this work, a search through a sequence of candidates $\{\lambda_0, \ldots, \lambda_s, \ldots, \lambda_{\text{last}}\}$, which is automatically determined for each individual $\boldsymbol{X}_n$, is instead employed. The optimal value of λ is chosen by minimizing the Akaike information criterion (AIC) (Akaike, 1974; Zou, Hastie, & Tibshirani, 2007) as in (Tax et al., 2019) or additionally by using either 3-fold cross-validation (CV) and minimizing the mean squared error. For the AIC, the number of non-zero coefficients in $\boldsymbol{\alpha}_n$ provides an unbiased estimate of degrees of freedom for the model (Tibshirani & Taylor, 2012; Zou et al., 2007). We use the AIC for normally distributed errors in least-squares problems from Burnham and Anderson (2004)

$$AIC_{\lambda_i} = \arg\min_{\lambda_s} \; m \log\left(\frac{\|\boldsymbol{X}_n - \boldsymbol{D}\boldsymbol{\alpha}_{\lambda_s}\|_2^2}{m}\right) + 2\text{df}(\boldsymbol{\alpha}_{\lambda_s}) \tag{2}$$

with m the number of elements of $\boldsymbol{X}_n$. In practice, this sequence of λ_s is chosen automatically on a log scale starting from λ_0 (providing the null solution $\boldsymbol{\alpha}_{\lambda_0} = 0$) up to $\lambda_{\text{last}} = \epsilon > 0$ (providing the regular least squares solution) (Friedman et al., 2010). The solution $\boldsymbol{\alpha}_n$ at λ_s is then used as a starting estimate for the next value of λ_{s+1}. The process can be terminated early if the cost function Equation (1) does not change much (e.g., the difference between the solution at λ_s and λ_{s+1} is below 10^{-5}) for decreasing values of λ_s, preventing computation of similar solutions.

3 Methods

In this section, we detail how a dictionary can be learned to create an implicit mapping between scanners. This is done by first constructing a target dictionary with datasets acquired on at least one or multiple scanners. After this target dictionary is constructed, a set of coefficients using the data from a given source scanner is computed, keeping the precomputed target dictionary fixed during the process. The resulting reconstructed datasets have implicit features specifically captured by the initial target scanner, without reconstructing the features only found in the source scanner used to acquire the data initially.

3.1 Building an optimal representation across scanners

For harmonization based on dictionary learning, all 3D patches of small spatial and angular local neighborhoods inside a brain mask were extracted from the available training

datasets for a given scanner as done in (St-Jean et al., 2016; Tax et al., 2019). Since different patch sizes are used depending on the reconstruction task, Sections 3.2 and 3.5 detail each case that we study in this manuscript. Only patches present inside a brain mask were used for computation and reconstruction. These patches were reorganized as column arrays $\Omega = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \ldots\}$ with each $\boldsymbol{X}_n \in \mathbb{R}^{m\times 1}$ represented as vectors of size m. Each volume was mean subtracted and each patch $\boldsymbol{X}_n$ was scaled to have unit variance (Friedman et al., 2010). Subsequently, features were automatically created from the target scanner datasets using dictionary learning as detailed in Section 2.1. A dictionary $\boldsymbol{D} \in \mathbb{R}^{m\times p}$ was initialized with p vectors $\boldsymbol{X}_{m\times 1} \in \Omega$ randomly chosen, where $\boldsymbol{D}$ is set to have twice as many columns as rows (i.e., $p = 2m$) as previously done in St-Jean et al. (2016), St-Jean et al. (2017). Updates using Equation (1) were carried for 500 iterations using a batchsize of $N = 32$. The coefficients $\boldsymbol{\alpha}_n$ were unscaled afterwards.

Once a dictionary $\boldsymbol{D}$ has been computed, the new, harmonized representation (possibly from a *different* scanner) can be obtained by computing $\boldsymbol{\alpha}_n$ for every $\boldsymbol{X}_n \in \Omega$. As $\boldsymbol{D}$ was created to reconstruct data from a chosen target scanner, it contains generic features tailored to this specific target scanner that are not necessarily present in the set of patches Ω extracted from a different scanner. As such, reconstruction using $\boldsymbol{D}_{\text{target}}$ created from Ω_{target} can be used to map Ω_{source} towards Ω_{target}, that is $\boldsymbol{X}_{n_{\text{harmonized}}} = \boldsymbol{D}_{\text{target}}\boldsymbol{\alpha}_n$ by using $\boldsymbol{X}_{n_{\text{source}}}$ and holding $\boldsymbol{D}_{\text{target}}$ fixed while solving Equation (1) for $\boldsymbol{\alpha}_n$. These specially designed features from Ω_{target} are not necessarily present in Ω_{source}, therefore eliminating the source scanner specific effects, as they are not contained in $\boldsymbol{D}_{\text{target}}$.

Downsampling $\boldsymbol{D}_{\text{target}}$ into $\boldsymbol{D}_{\text{small}}$ can also be used to reconstruct data at a different resolution than initially acquired by creating an implicit mapping between two different spatial resolutions. This is done by finding the coefficients $\boldsymbol{\alpha}$ by holding $\boldsymbol{D}_{\text{small}}$ fixed when solving Equation (1), but using $\boldsymbol{D}_{\text{target}}$ for the final reconstruction such that $\boldsymbol{X}_{n_{\text{harmonized}}} = \boldsymbol{D}_{\text{target}}\boldsymbol{\alpha}_n$. This reconstruction with the full sized dictionary provides an upsampled version of $\boldsymbol{X}_n$, the implicit mapping being guaranteed by sharing the same coefficients $\boldsymbol{\alpha}_n$ for both reconstructions. A similar idea has been exploited previously for the 3D reconstruction of T1w images by Rueda, Malpica, and Romero (2013) and in diffusion MRI by St-Jean et al. (2017) in the context of single image upsampling. The general reconstruction process for the harmonization of datasets between scanners is illustrated in Figure 1. Our implementation of the harmonization algorithm is detailed in Appendix A and also available in both source form and as a Docker container at https://github.com/samuelstjean/harmonization (St-Jean, Viergever, & Leemans, 2019).

3.2 Reconstruction tasks of the challenge

For the reconstruction in task 1 (matched resolution scanner-to-scanner mapping), the dictionary $\boldsymbol{D}_{\text{target}}$ was created using patches of size $3 \times 3 \times 3$ with 5 angular neighbors and one randomly chosen $b = 0$ s/mm^2 image in each block. We chose these parameters as they have previously been shown to offer a good trade-off between accuracy and computation time in a previously published denoising task (St-Jean et al., 2016). The angular patch size (i.e., how many DWIs are included across gradients) is chosen to include all volumes at the same angular distance on the sphere as in St-Jean et al. (2016). Optimization for constructing $\boldsymbol{D}_{\text{target}}$ with Equation (1) was performed using 3-fold CV and reconstruction of the final harmonized datasets was done with either CV or minimizing the AIC with Equa-

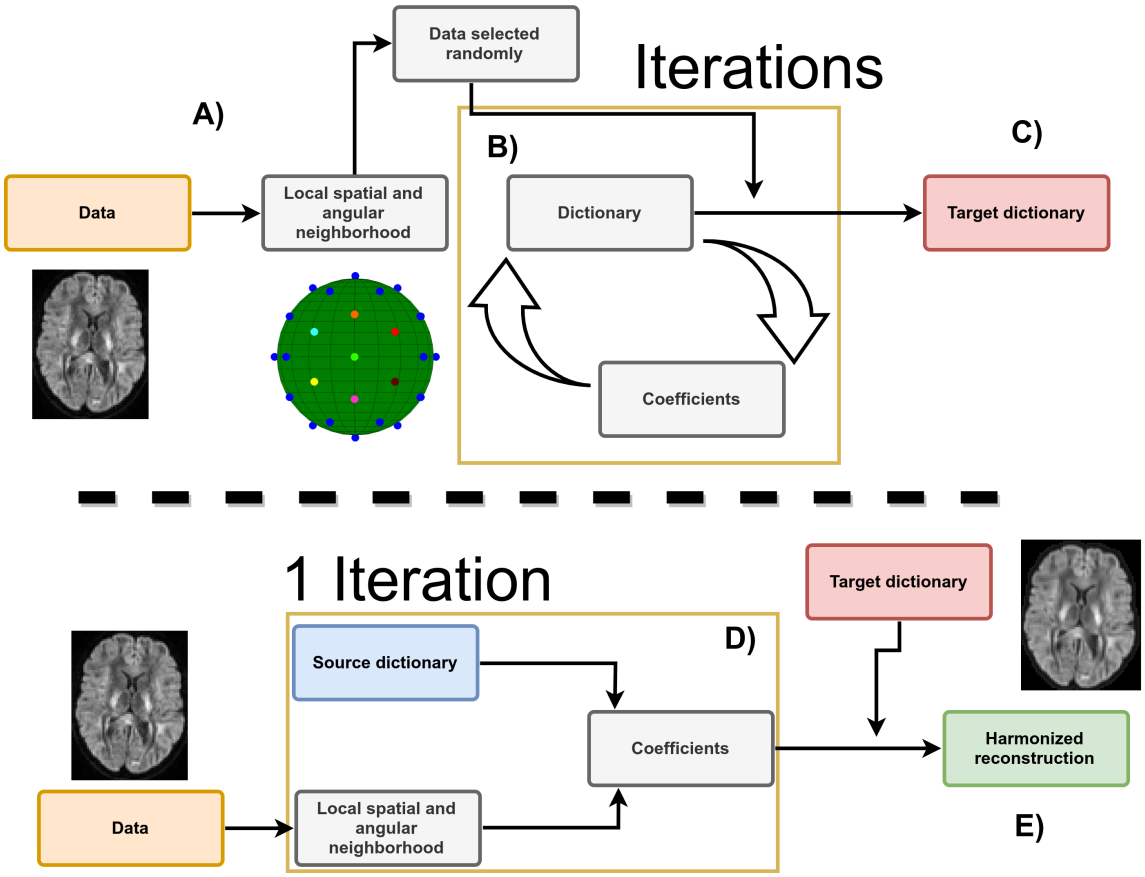

Figure 1. Schematic representation of the harmonization between scanners with adaptive dictionary learning. **A)** Local patches are decomposed into vectors $\boldsymbol{X}_n$ and a random subset is used to initialize the dictionary $\boldsymbol{D}$. For harmonization to a scanner-agnostic space, $\boldsymbol{D}$ is initialized with data drawn from all scanners. **B)** A new set of patches is drawn at every iteration and the dictionary is refined iteratively by alternating updates for the coefficients $\boldsymbol{\alpha}$ and the dictionary $\boldsymbol{D}$ using Equation (1). **C)** After a set number of iterations, this target dictionary $\boldsymbol{D}$ can now be used to reconstruct data from a potentially different dataset. **D)** A set of coefficients is computed for each patch $\boldsymbol{X}_n$ of the input dataset with a source dictionary. For harmonization tasks, the source dictionary is the target dictionary obtained from a different scanner in step **C)** and of the same size. For upsampling tasks, the source dictionary is a downsampled version of the target dictionary. When $\boldsymbol{D}$ is constructed from datasets acquired on multiple scanners, the reconstruction step removes variability intrinsic to a given dataset which is not present in the remaining scanners as $\boldsymbol{D}$ would not have captured this variability. **E)** The harmonized reconstruction for each patch $\boldsymbol{X}_n$ is obtained by multiplying the target dictionary $\boldsymbol{D}$ and the coefficients $\boldsymbol{\alpha}_n$.

tion (2) in two separate experiments. The datasets from the GE scanner were reconstructed using the dictionary built from the Prisma or Connectom scanner datasets for their respective harmonization task. For the reconstruction in task 2 (spatial and angular resolution enhancement), patches of different spatial sizes were extracted from the images at higher resolution (patches of size $5 \times 5 \times 5$ for the Prisma scanner and $6 \times 6 \times 6$ for the Connectom scanner) and used for the dictionary learning algorithm as described in Section 2.1. In this task, the target datasets patch size was chosen so that the ratio between the original patch size of $3 \times 3 \times 3$ and the target patch size matches the ratio in spatial resolution between the harmonized datasets as previously done in St-Jean et al. (2017). Under the hypothesis that a larger patch is a good representation for its lower resolution counterpart when downsampled, each column of the optimized dictionary $\boldsymbol{D}_{\text{target}}$ was resized to a spatial dimension of $3 \times 3 \times 3$ and the coefficients $\boldsymbol{\alpha}$ computed for this lower resolution dictionary $\boldsymbol{D}_{\text{small}}$. The patches were finally reconstructed by multiplying the original dictionary $\boldsymbol{D}_{\text{target}}$ with the coefficients $\boldsymbol{\alpha}$. This creates a set of upsampled patches from the GE scanner that are both harmonized and at the same spatial resolution as either the Prisma or the Connectom datasets. All reconstruction tasks were computed overnight on our computing server using 100 cores running at 2.1 GHz. On a standard desktop with a 4 cores 3.5 GHz processor, rebuilding one dataset took approximately two hours and 30 minutes with the AIC criterion.

3.3 Evaluation framework of the challenge

The original challenge requested the participants to match the original gradient directions of the source to the target datasets and evaluated various scalar metrics on the diffusion weighted images. In our original submission, this matching was done with the truncated spherical harmonics (SH) basis of order 6 (Descoteaux, Angelino, Fitzgibbons, & Deriche, 2007) on the source dataset and sampling the basis at the gradient directions from the target scanner. In the present manuscript, we chose instead to evaluate the metrics directly in the original gradient directions as they are rotationally invariant, saving one interpolation step in the process as it could potentially introduce unwanted blurring of the data. The metrics used in the original evaluation were the apparent diffusion coefficient (ADC) and the fractional anisotropy (FA) from diffusion tensor imaging (DTI) and the rotationally invariant spherical harmonic (RISH) features of order 0 (RISH 0) and order 2 (RISH 2) of the SH basis, see Tax et al. (2019) for additional details. As our evaluation framework is slightly different, a direct numerical comparison with the results previously reported in the CDMRI challenge is not possible, even if we use the same metrics, as the exact way to compute the metrics was not made available to the participants. This unfortunately prevents us from replicating exactly the challenge or to exclude poorly performing regions as was done in the original evaluation, making a comparison between the previously reported results impossible. We compare our new approach using automatic regularization with both the AIC and CV criterion against our initial version of the harmonization algorithm (which included interpolation of the DWIs using the SH basis) and a baseline reference prediction created by trilinear interpolation from the source to the target scanner in the spirit of the original challenge.

3.4 Datasets and experiments

We used the datasets from the CDMRI 2017 harmonization challenge (Tax et al., 2019), consisting of ten training subjects and four test subjects acquired on three different scanners (GE, Siemens Prisma and Siemens Connectom) using different gradient strength (40 mT/m, 80 mT/m and 300 mT/m, respectively) with two acquisition protocols. The study was originally approved by Cardiff University School of Psychology ethics committee and written informed consent was obtained from all subjects. Experiments are only reported for the four test subjects, which are later on denoted as subjects 'H', 'L', 'M' and 'N'. The standard protocol (ST) consists of 30 DWIs acquired at 2.4 mm isotropic with a b-value of $b = 1200$ s/mm^2, 3 $b = 0$ s/mm^2 images for the GE datasets, 4 $b = 0$ s/mm^2 images for the Siemens datasets and TE = 98 ms. Note that the TR is cardiac gated for the GE datasets while the Siemens datasets both use TR = 7200 ms. The state-of-the-art (SA) protocol for the Siemens scanners contains 60 DWIs with a b-value of $b = 1200$ s/mm^2 and 5 $b = 0$ s/mm^2 images. The Prisma datasets were acquired with a spatial resolution of 1.5 mm isotropic and TE / TR = 80 ms / 4500 ms. The Connectom datasets were acquired with a spatial resolution of 1.2 mm isotropic and TE / TR = 68 ms / 5400 ms. Most of the acquisition parameters were shared for the SA protocol which are listed in Table 1 with full details of the acquisition available in Tax et al. (2019). Standard preprocessing applied by the challenge organizers on the datasets includes motion correction, EPI distortions corrections and image registration for each subject across scanners. The SA protocols were additionally corrected for gradient nonlinearity distortions. These datasets are available upon request from the organizers at https://www.cardiff.ac.uk/cardiff-university-brain-research-imaging-centre/research/projects/cross-scanner-and-cross-protocol-diffusion-MRI-data-harmonisation. Figure 2 shows an example of the acquired datasets for a single subject.

Scanner	GE 40 mT/m	Siemens Prisma 80 mT/m		Siemens Connectom 300 mT/m	
Protocol	Standard (ST)	Standard (ST)	State-of-the-art (SA)	Standard (ST)	State-of-the-art (SA)
Sequence	TRSE	PGSE	PGSE	PGSE	PGSE
# directions per b-value	30	30	60	30	60
TE [ms]	89	89	80	89	68
TR [ms]	Cardiac gated	7200	4500	7200	5400
Δ/δ [ms]		41.4/26.0	38.3/19.5	41.8/28.5	31.1/8.5
$\delta_1 = \delta_4/\delta_2 = \delta_3$ [ms]	11.23/17.84				
Acquired voxel size [mm^3]	2.4 x 2.4 x 2.4	2.4 x 2.4 x 2.4	1.5 x 1.5 x 1.5	2.4 x 2.4 x 2.4	1.2 x 1.2 x 1.2
Reconstructed voxel size [mm^3]	1.8 x 1.8 x 2.4	1.8 x 1.8 x 2.4	1.5 x 1.5 x 1.5	1.8 x 1.8 x 2.4	1.2 x 1.2 x 1.2
SMS factor	1	1	3	1	2
Parallel imaging	ASSET 2	GRAPPA 2	GRAPPA 2	GRAPPA 2	GRAPPA 2
Bandwidth [Hz/Px]	3906	2004	1476	2004	1544
Partial Fourier	5/6	—	6/8	6/8	6/8
Coil combine		Adaptive combine	Sum of Squares	Adaptive combine	Adaptive combine
Head coil	8 channel	32 channel	32 channel	32 channel	32 channel

Table 1

Acquisition parameters of the datasets for the three different scanners. TRSE: twice-refocused spin-echo, PGSE: pulsed-gradient spin-echo, TE: echo time, TR: repetition time, SMS: Simultaneous multi-slice, Hz/Px: Hertz/Pixel. The table is adapted from Tax et al. (2019), available under the CC-BY 4.0 license.

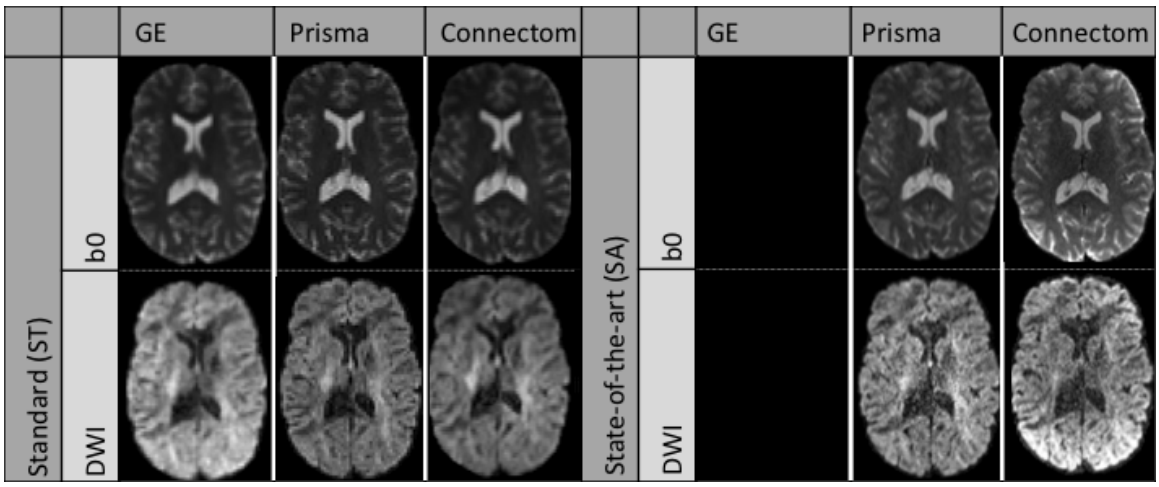

Figure 2. Example $b = 0$ s/mm^2 images (top row) and $b = 1200$ s/mm^2 images (bottom row) for a single subject acquired on the three scanners after preprocessing. The standard protocol (ST) is shown on the left and the state-of-the-art protocol (SA) is shown on the right. Note that the challenge asked participants to harmonize the GE ST protocol towards the two other scanners, but no SA protocol is available for the GE scanner. The figure is adapted from Tax et al. (2019), available under the CC-BY 4.0 license.

3.5 Simulations beyond the challenge

To further make our proposed harmonization algorithm widely applicable, we designed additional experiments beyond the challenge to harmonize data towards a new scanner-space. As the CDMRI challenge focused on harmonization of datasets from a source scanner to a target scanner, the organizers essentially provided matching datasets of all subjects across all scanners. This data collection would be appropriate, for example, in a longitudinal study design with scanner hardware upgrades during the study and subsequent data analysis. However, such an experimental setup might not be available in practice when harmonizing datasets from multiple centers or studies where data collection is done only once per subject e.g., to reduce costs associated with scan time or reduce traveling of the participants.

The additional experiments create a new harmonization space by randomly sampling datasets from the three scanners at once to build the target dictionary instead of matching the GE datasets to a particular target scanner as in the previous experiments. To ensure that the scanner effects are properly removed, the test datasets were additionally altered in a small region with a simulated free water compartment as described in Section 3.6, creating additional test datasets contaminated with simulated edema. These newly created datasets were never "seen" by the harmonization algorithm, making it possible to quantify if the induced effects are properly reconstructed without discarding the natural biological variability of the datasets, as these alterations were not present in the training set in the first place. This experiment is similar to creating a common space on a larger set of healthy subjects and finally harmonizing data from the remaining healthy subjects and "patients" towards this common space. In our current setup, the harmonization algorithm is not aware

that the datasets are in fact from matched subjects and, by design, could also be used on unpaired training datasets.

3.6 Alterations of the original datasets

To create the altered version of the test datasets, a region of 3000 voxels ($15 \times 20 \times 10$ voxels) in the right hemisphere was selected at the same spatial location in image space. The size of the region is kept constant throughout experiments and subjects to facilitate statistical analysis and comparisons. Every voxel in the selected region was separately affected by a free water compartment to mimic infiltration of edema according to

$$S_{b_{\text{altered}}} = S_b + f S_0 \exp(-b D_{\text{csf}}) \tag{3}$$

with $S_{b_{\text{altered}}}$ the new signal in the voxel, S_b the original signal in the voxel at b-value b and S_0 the signal in the $b = 0$ s/mm^2 image, f is the fraction of the free water compartment, which is drawn randomly for every voxel from a uniform distribution $U(0.7, 0.9)$ and $D_{\text{csf}} = 3 \times 10^{-3}$ mm^2/s is the nominal value of diffusivity for free water (e.g., cerebrospinal fluid (CSF)) at 37 ℃ (Pasternak, Sochen, Gur, Intrator, & Assaf, 2009; Pierpaoli & Jones, 2004). As the individual subjects across scanners are only registered to their counterpart across scanners, the affected region will be approximately (up to errors due to registration) at the same spatial location in each subject. This location will however be slightly different between subjects, which introduces normal variability in terms of the number of white matter and gray matter voxels that would be affected by edema and their location in a cohort of patients.

3.7 Evaluation metrics

Error and accuracy of predicted metrics. We reproduced parts of the analyses conducted in the original CDMRI challenge from Tax et al. (2019), namely the per voxel error for each metric as computed by the mean normalized error (MNE) and the voxelwise error. Denoting the target data to be reproduced as *acquired* (Prisma or Connectom scanners) and the source data to be harmonized as *predicted* (GE scanner), the MNE is defined as MNE = |(predicted - acquired)| / acquired and the error is defined as error = predicted - acquired. The MNE is a relative metric, penalizing more when the error is large relative to the target value, while the error itself only measures the magnitude of the mistake, but can indicate global under or overestimation with the sign of the metric. A small error with a large MNE would likely indicate that most of the mistakes committed by an algorithm are in regions where the metric of interest is low. The original challenge reports values taken either globally in a brain mask, in FreeSurfer regions of interest (ROI) and excluding poorly performing regions or the median value computed in sliding windows. Since the masks of these ROIs were not released for the challenge, we instead report boxplots of the two metrics using the brain masks from the challenge as this reports the global median error in addition to the global mean error and additional quantiles of their distribution. To prevent outliers from affecting the boxplots (particularly located at the edges of the brain masks), we clip the MNE and error values at the lowest 0.1% and largest 99.9% for each dataset separately.

Kullback-Leibler divergence as a measure of similarity. As the voxelwise difference may not be fully indicative of the global trend of the harmonization procedure between datasets (e.g., due to registration errors), we also computed the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) between the distributions of each harmonized dataset from the GE scanner and its counterpart from the target scanner for each of the four metrics. The KL divergence is a measure of similarity between two probability distributions $P(x)$ and $Q(x)$ where lower values indicate a higher similarity and $\mathrm{KL}(P, Q) = 0$ when $P(x) = Q(x)$. In its discrete form, the Kullback-Leibler divergence is given by

$$\mathrm{KL}(P, Q) = \sum_k P_k \log\left(\frac{P_k}{Q_k}\right), \tag{4}$$

where P_k is the "candidate" probability distribution, Q_k the true probability distribution and k represents the number of discrete histogram bins. The measure is not symmetric, that is $\mathrm{KL}(P, Q) \neq \mathrm{KL}(Q, P)$ in general. We instead use the symmetric version of the KL divergence as originally defined by Kullback and Leibler (1951)

$$\mathrm{KL}_{sym} = \mathrm{KL}(P, Q) + \mathrm{KL}(Q, P). \tag{5}$$

In practice, a discrete distribution can be constructed from a set of samples by binning and counting the data. By normalizing each bin so that their sum is 1, we obtain a (discrete) probability mass function. For each metric, the discrete distribution was created with $k = 100$ equally spaced bins. We also remove all elements with probability 0 from either P_k or Q_k (if any) to prevent division by 0 in Equation (4). As the binning procedure does not share the same bins between scanners, the results can not be compared directly between the Connectom and Prisma scanners.

Statistical testing and effect size in the presence of alteration. To evaluate quantitatively if the harmonization algorithm did not remove signal attributable to genuine biological variability, we computed the percentage difference between the harmonized test datasets in the affected region of 3000 voxels before alteration and after alteration as given by Equation (6)

$$100 \times \left| \left(\frac{\text{harmonized} - \text{baseline}}{\text{baseline}}\right) - \left(\frac{\text{harmonized_altered} - \text{baseline_altered}}{\text{baseline_altered}}\right) \right|, \tag{6}$$

where baseline (resp. harmonized) denotes the datasets before (resp. after) harmonization and the suffix altered indicates the datasets altered with simulated edema. A value close to 0 therefore indicates that the harmonization procedure performed similarly in reducing variability attributable to differences in the scanner for harmonization of the regular datasets and in the presence of alteration. To investigate the magnitude of these differences, we conducted Student's t-test for paired samples for each subject separately (Student, 1908). This was done on both the normal datasets (testing between scanners) and the altered datasets (testing between scanners and additionally between the normal and altered datasets). The p-values from the tests were subsequently corrected for the false discovery rate (FDR) at a level of $\alpha = 0.05$ (Benjamini & Hochberg, 1995). In addition, we

also report the effect size of those paired t-tests as computed by Hedges' g (Hedges, 1981; Lakens, 2013), which we redefine as

$$g = \frac{|\mu_1 - \mu_2|}{(\sigma_1 + \sigma_2)/2} \times \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right), \tag{7}$$

where μ_i, σ_i and n_i are the mean, the standard deviation, and the size of sample i, respectively. A value of $g = 1$ indicates that the difference between the means is of one standard deviation, with larger values indicating larger effect sizes as reported by the difference in the group means. In the original definition of Hedges (1981), g is not enforced to be positive. We instead report the absolute value of g as we do not know *a priori* which mean is larger than the other, but are only interested in the magnitude of the effect rather than its sign. With this definition, values of g reported for the test between a given subject for two different scanners which are lower than the reference method indicate an improvement by removing scanner specific effects. On the other hand, similar values of g between the reference and the harmonized dataset for a given subject and its altered counterpart on the same scanner indicates preservation of the simulated effects as it is the *only* difference between these two datasets by construction.

4 Results

4.1 Results from the challenge

Mapping between scanners for matched acquisition protocols. Figure 3 shows the KL symmetric divergence as presented in Section 3.7 for the standard protocol. In general, the baseline has a higher KL value than the other methods on the Connectom scanner. The CV based method is generally tied or outperforms the AIC based method. For the Prisma scanner, results show that the AIC performs best with the CV based method following the baseline reference. In the case of the ADC metric, our initial algorithm outperforms the three other methods for some subjects.

Figure 4 shows the distribution (as boxplots) in the absolute mean normalized error and mean error of the four metrics for the standard protocol. The MNE is almost tied or slightly higher for the baseline method than the alternatives for both scanners. For the FA and RISH 2 metrics, the baseline error is tied or larger than the other methods. For the voxelwise error, all methods underestimate the ADC and overestimate the RISH 0 on average while the FA and RISH 2 metrics show a different pattern depending on the scanner. For the Connectom scanner, the CV based method generally has an average error around 0 for the FA while the AIC and our initial algorithm generally overestimate the metric. The baseline is on the other spectrum and generally underestimates the FA. On the Prisma scanner, the effect is reversed; there is a general overestimation of the FA while the error committed by the AIC based method is in general close to 0. The RISH 2 error follows the same pattern as the FA error on both scanners for the four compared methods.

Mapping between scanners across spatial resolutions. Figure 5 shows the KL symmetric divergence for the second task of the challenge, mapping the GE ST protocol datasets to the SA protocols of the Prisma or Connectom scanners. For the Connectom scanner, the AIC based algorithm and our initial algorithm, which is also AIC based, performs best in most cases. The CV based algorithm also outperforms the baseline method

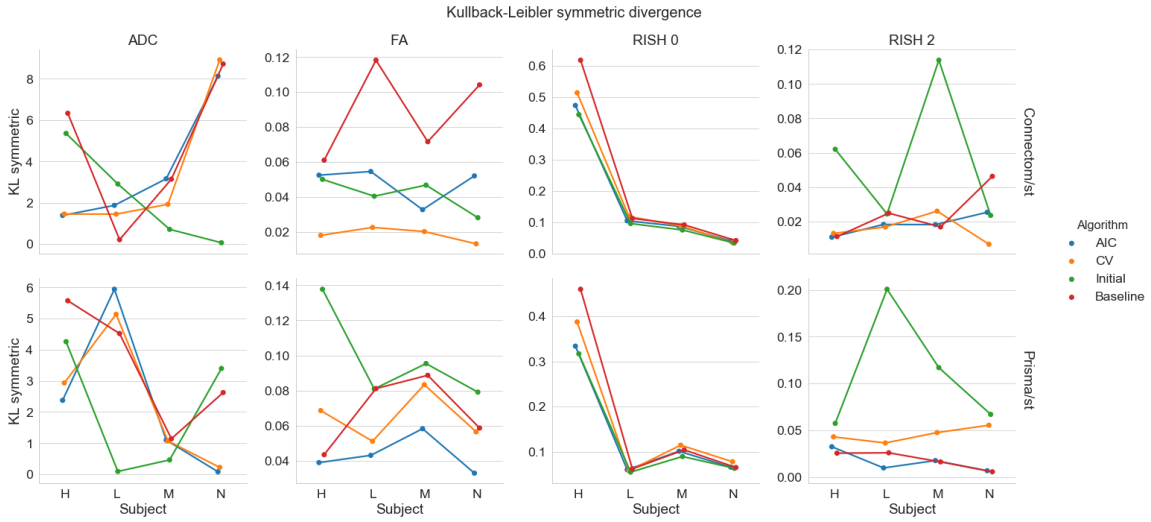

Figure 3. KL symmetric divergence (where lower is better) for the harmonization task at the same resolution between the GE ST datasets and the Connectom ST (top row) or the Prisma ST (bottom row) datasets on the four test subjects ('H', 'L', 'M' and 'N'). Each metric is organized by column (ADC, FA, RISH 0 and RISH 2) for the four compared algorithms (AIC in blue, CV in orange, our initial version of the harmonization algorithm in green and the baseline comparison in red).

for the ADC and RISH 0 metrics. For the Prisma scanner, the AIC outperforms most of the compared methods or is tied with the CV. Notably, the baseline ranks second for the FA and RISH 2 metrics, but is the worst performer for the ADC and the RISH 0 metrics.

Figure 6 shows results for the absolute mean normalized error and mean error for all algorithms on harmonizing the SA protocol. For the Connectom scanner, the baseline ranks last for most subjects on the isotropy metrics (ADC and RISH 0) while it only performs slightly better than the CV based algorithm for the anisotropy metrics (FA and RISH 2). On the Prisma scanner, results are similar for the ADC and RISH 0 metrics. For the FA metrics, the best performance is obtained with the AIC based method while the baseline is better for harmonizing the RISH 2 metric for three of the subjects.

Now looking at the mean error, results show that the ADC metric is underestimated for all methods and on both scanners with the three methods usually outperforming the baseline comparison. The FA, RISH 0 and RISH 2 metrics are instead overestimated. For the FA metric, the AIC and our initial algorithm commit less error on average than the baseline on the Connectom scanner. On the Prisma scanner, only the AIC has an average error lower than the baseline. All methods perform better or almost equal on average to the baseline comparison for the RISH 0 metric. The RISH 2 metric shows a scanner dependent pattern; on the Connectom scanner, the best performing method is our initial algorithm followed by the AIC based algorithm while on the Prisma scanner, the lowest error is achieved by the AIC based method.

In general, results show that the isotropy metrics (ADC and RISH 0) are subject to global scanner effects while the anisotropy metrics (FA and RISH 2) may be subject to

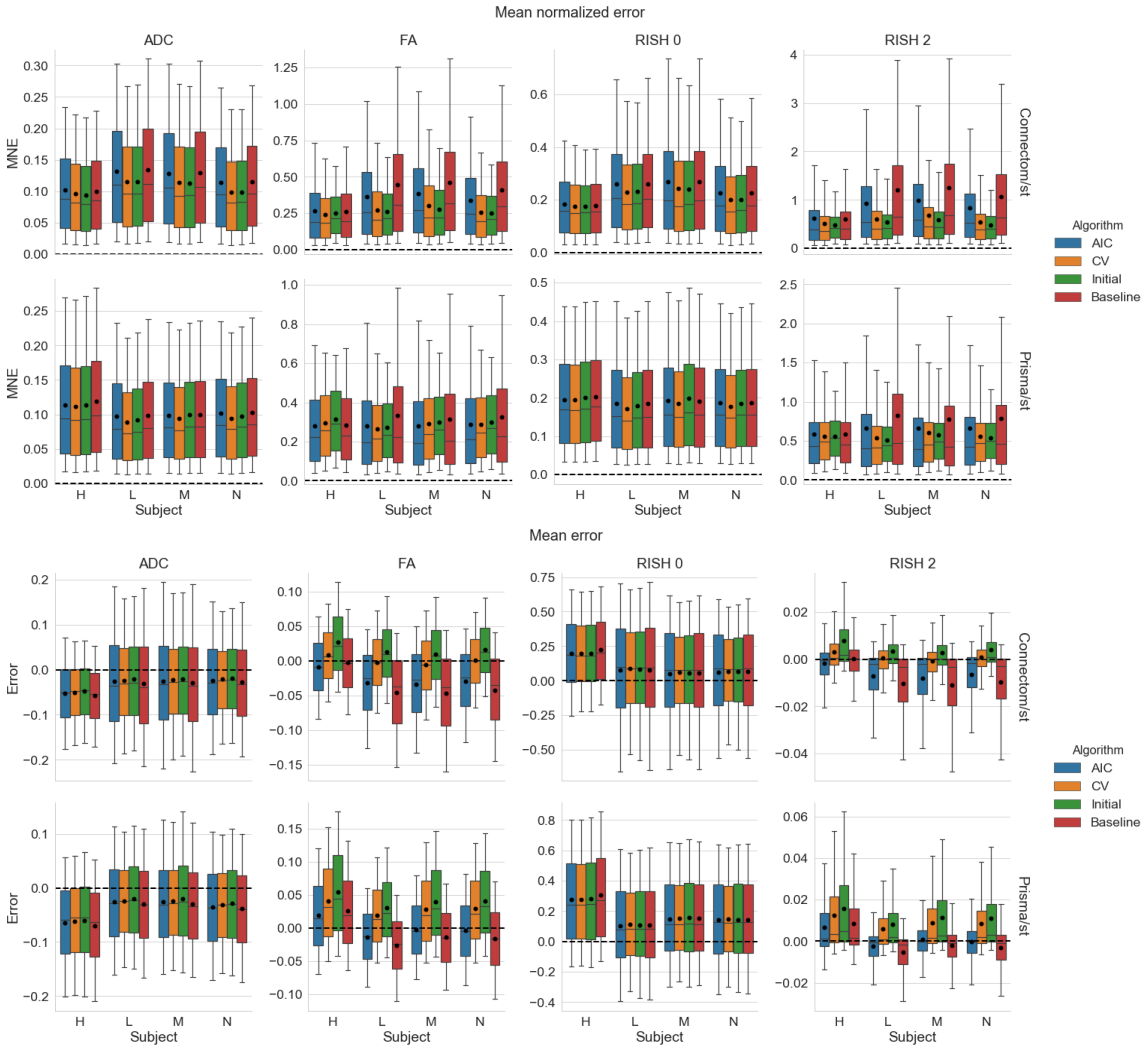

Figure 4. Boxplots of the voxelwise mean normalized error (top) and error (bottom) for each metric, following the same conventions detailed in Figure 3. The black dot shows the mean error and the dashed line indicates an error of 0, representing a perfect match between the harmonized GE dataset and the dataset for the target scanner.

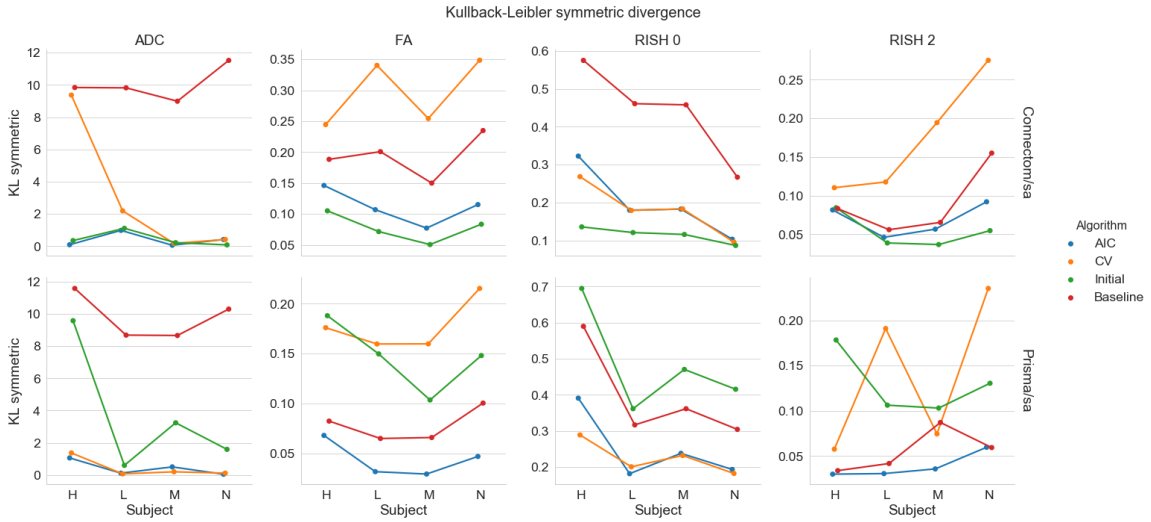

Figure 5. Symmetric KL divergence (where lower is better) for the harmonization task across resolution between the GE ST datasets and the Connectom SA (top row) or the Prisma SA (bottom row) datasets. The organization is the same as previously used in Figure 3.

orientation dependent effects. These effects are also likely different for each scanner since the gradient strength and timings are different, even if the b-values are matched. In these experiments, the target scanner is untouched and therefore still contains its own scanner effect when computing the voxelwise error of each harmonization algorithm.

4.2 Mapping original and altered datasets towards a common space

In these experiments, alterations were made to the test set as previously described in Section 3.6. As these altered datasets were never used for training, we can quantify the removal of scanner effects and the preservation of the alterations by comparing solely the altered regions with their original counterpart in each subject, free of processing effects. In these experiments, the baseline comparison is to not process the datasets at all since the datasets are altered versions of themselves, therefore not requiring any interpolation or resampling. As these experiments are outside of the challenge's scope, they are not covered by our initial algorithm and therefore the "initial" category is not presented in this section. Figure 7 shows the original and altered metrics for one subject on the raw data and after harmonization with the AIC and CV based algorithms and Figure 8 shows the relative percentage difference between the raw datasets and their harmonized counterpart. We define the relative percentage difference as difference $= 100 \times$ (harmonized - raw) / raw. The alterations are mostly visible on the $b = 0$ s/mm^2 image while the $b = 1200$ s/mm^2 image is only slightly affected due to the high diffusivity of the CSF compartment. However, the differences are visible on the diffusion derived maps, seen as an increase in ADC and a reduction for the FA, RISH 0 and RISH 2 metrics. Visually, harmonized datasets do not seem different from their original counterpart, but the difference maps show that small differences are present with the CV method generally showing larger differences than the AIC

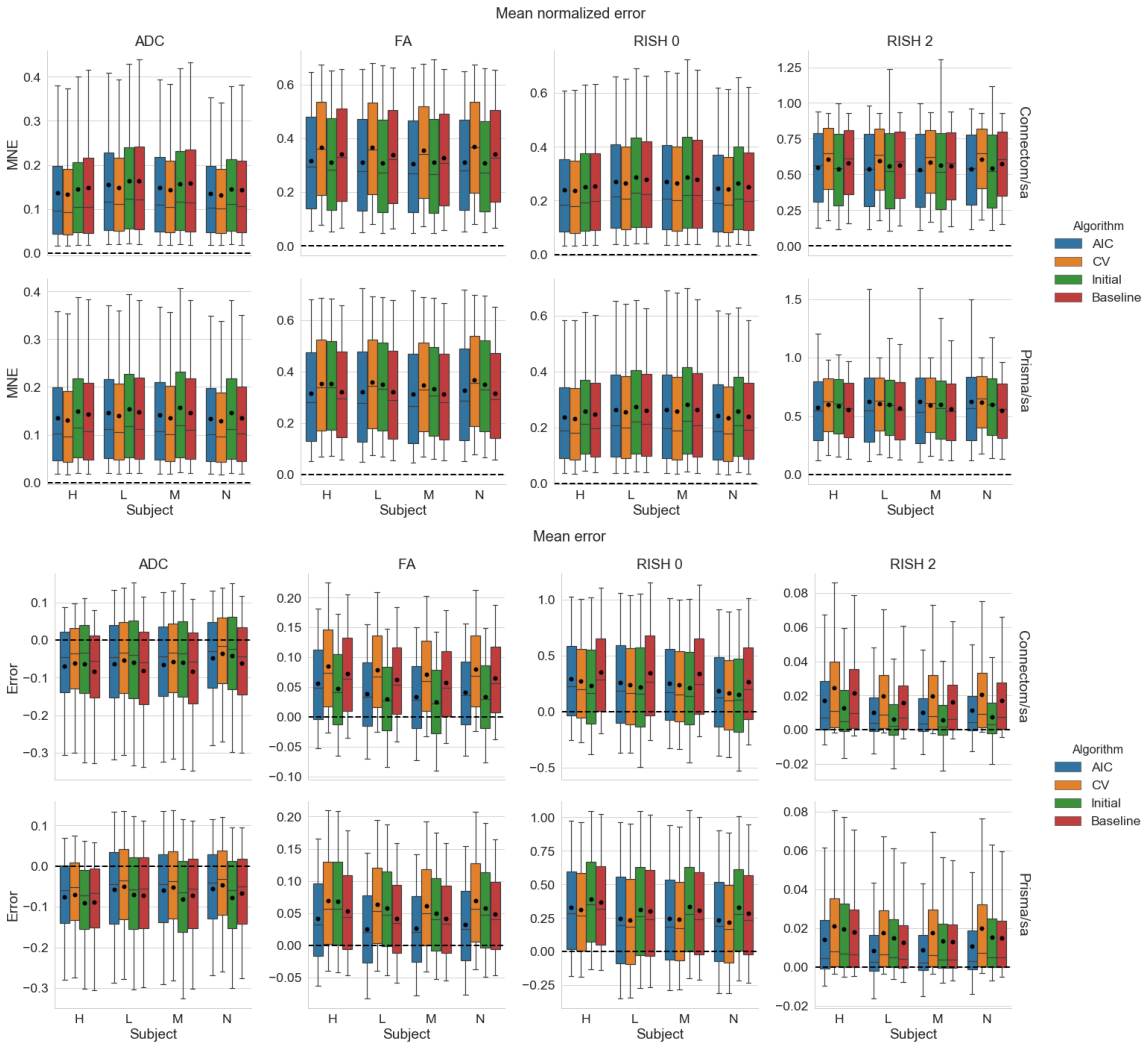

Figure 6. Boxplots of the voxelwise mean normalized error (top) and error (bottom) of each metric for the four algorithms. The black dot shows the mean error and the dashed line indicates an error of 0. The organization follows the conventions of Figure 4.

method. Notably, the anisotropy metrics (FA and RISH 2) are lower after harmonization while the difference for the isotropy metric (ADC and RISH 0) is distributed around 0.

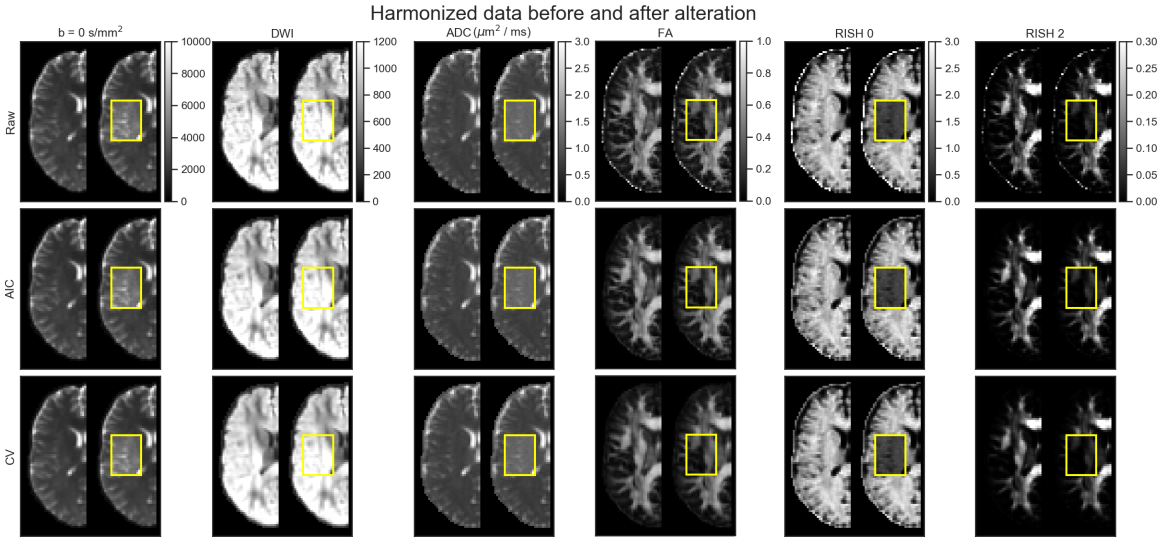

Figure 7. Examplar slice of subject 'H' on the GE scanner as original (left half) and altered (right half) metrics. Only the affected portion of the data (as shown in the yellow box) is analyzed in paired statistical testing against the same location in the original dataset. Each column shows (from left to right) a $b = 0$ s/mm^2 image, a DWI at $b = 1200$ s/mm^2, the FA, ADC, RISH 0 and RISH 2 metrics with a common colorbar per column. The top row shows the raw data, the middle row shows the data harmonized using the AIC and the bottom row shows the harmonized data using the CV. The $b = 0$ s/mm^2 image, the DWI and the ADC map increase after adding the free water compartment while the FA, RISH 0 and RISH 2 metrics are instead lower in their altered counterpart.

Figure 9 shows the relative percentage difference as boxplots for all test subjects and all scanners between the altered and normal regions. A low difference indicates that the signal removed after harmonization is the same in the baseline and altered datasets, that is the algorithm performs similarly in the presence (or not) of the simulated edema. The CV algorithm produces larger relative differences than the AIC based algorithm after harmonization between the reference and altered datasets. The larger differences are in the anisotropy metrics (FA and RISH 2) while the differences in isotropy metrics (ADC and RISH 0) are smaller on average. At this stage, it is unclear however if harmonization with the AIC regularization still contains variability attributable to the scanner or if the CV criterion is too aggressive and mistakenly removed variability due to genuine anatomical variation. Figure 10 shows the relative percentage difference that is originally present in the datasets between every pair of scanners, but before applying harmonization and alterations. This represents the amount of natural variability present in the diffusion metrics between scanners for each subjects in the region which is altered at a later stage. We do not show the signal value for the $b = 0$ s/mm^2 and $b = 1200$ s/mm^2 images since the scanners are not using the same signal scaling.

Figure 11 shows boxplots of the effect size as computed by a paired t-test after harmo-

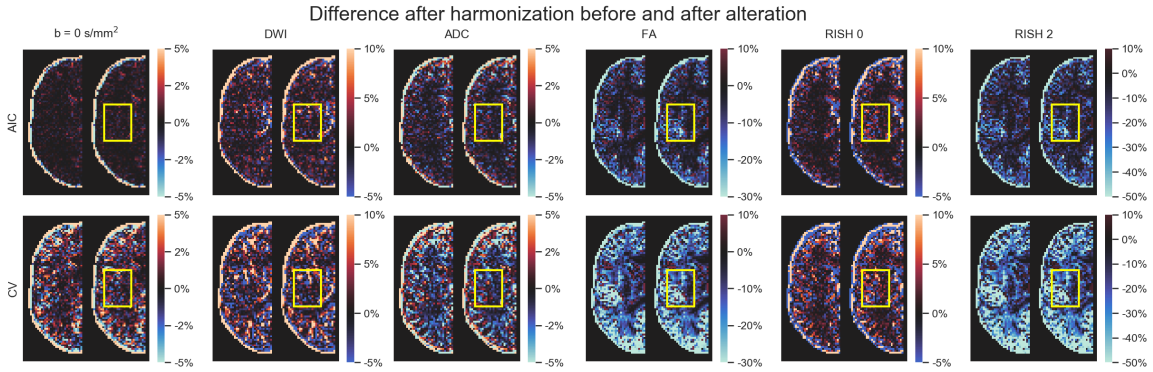

Figure 8. Examplar slice of subject 'H' on the GE scanner as original (left half) and altered (right half) metrics with the yellow box indicating the altered region specifically. Each column shows (from left to right) a $b = 0$ s/mm^2 image, a DWI at $b = 1200$ s/mm^2, the ADC, FA, RISH 0 and RISH 2 metrics with a common colorbar per column as in Figure 7. The top row (resp. the bottom row) shows the relative percentage difference between the harmonized data using the AIC (resp. the CV) and the raw data. If the affected region is similar in both images, this means that the harmonization algorithm did not remove the artificial alterations that were introduced and only removed variability attributable to the scanner equally in both cases.

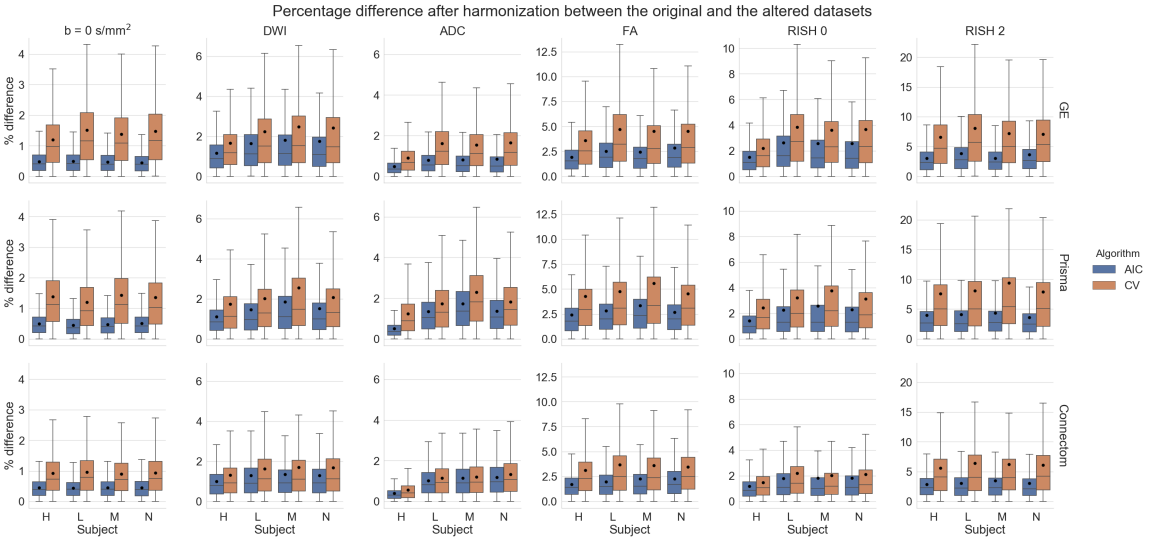

Figure 9. Boxplots of the percentage difference between the harmonized datasets with and without alteration for all subjects for the AIC and CV criterion in the altered region only. The top row shows the difference for the GE scanner, the middle row for the Prisma scanner and the bottom row for the Connectom scanner. A value close to 0 indicates that the harmonization procedure removed a similar amount of the signal in the reference datasets and in the altered datasets.

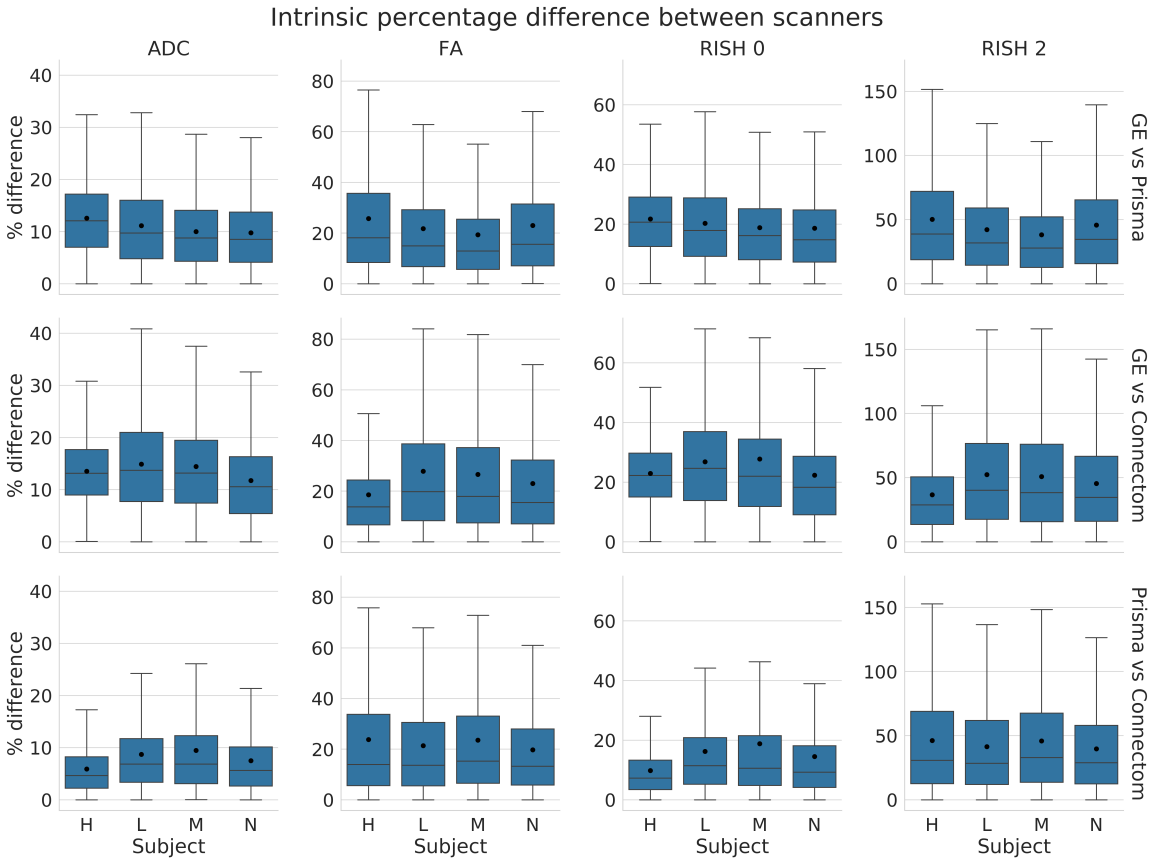

Figure 10. Boxplots of the percentage difference (before harmonization) between datasets acquired on different scanners in the selected region before alterations. The top row shows the difference between the GE scanner and Prisma scanner, the middle row between the GE and Connectom scanner and the bottom row between the Prisma and the Connectom scanner. In this case, the percentage difference is computed as $100 \times \frac{|\text{scanner1 - scanner2}|}{(\text{scanner1 + scanner2})/2}$, similarly to Equation (6).

nization towards a common space for all scanners. Tests were conducted for every subject between each scanner in addition to the altered versions of the datasets as previously described in Section 3.7. For the ADC metric, both methods yield a lower effect size on average than the raw, unprocessed data and preserve the effect size due to the alterations as shown in the middle row. The RISH 0 metric shows similar behavior with the CV based method producing an average effect size slightly higher than the raw datasets. Now looking at the anisotropy metrics (FA and RISH 2), the effect size is reduced or equal on average in most cases (except for subject 'H' when only one scan is altered) when scans are harmonized with the AIC algorithm. The CV based algorithm shows a higher effect size for harmonization between scans and a lower effect size when both scans are altered. As we only report the absolute value of the effect size, this is due to both a lower group mean and group standard deviation than the raw datasets. This difference in group means and standard deviations prevents a direct comparison of results between each rows, which can not be

directly compared as they are unlikely to share a common numerator or denominator. The harmonization process is likely only removing scanner effects present in each dataset as the middle row (where only one of the compared dataset is affected) shows similar reductions in effect size, but is still on the same magnitude as the raw datasets since the alteration is preserved.

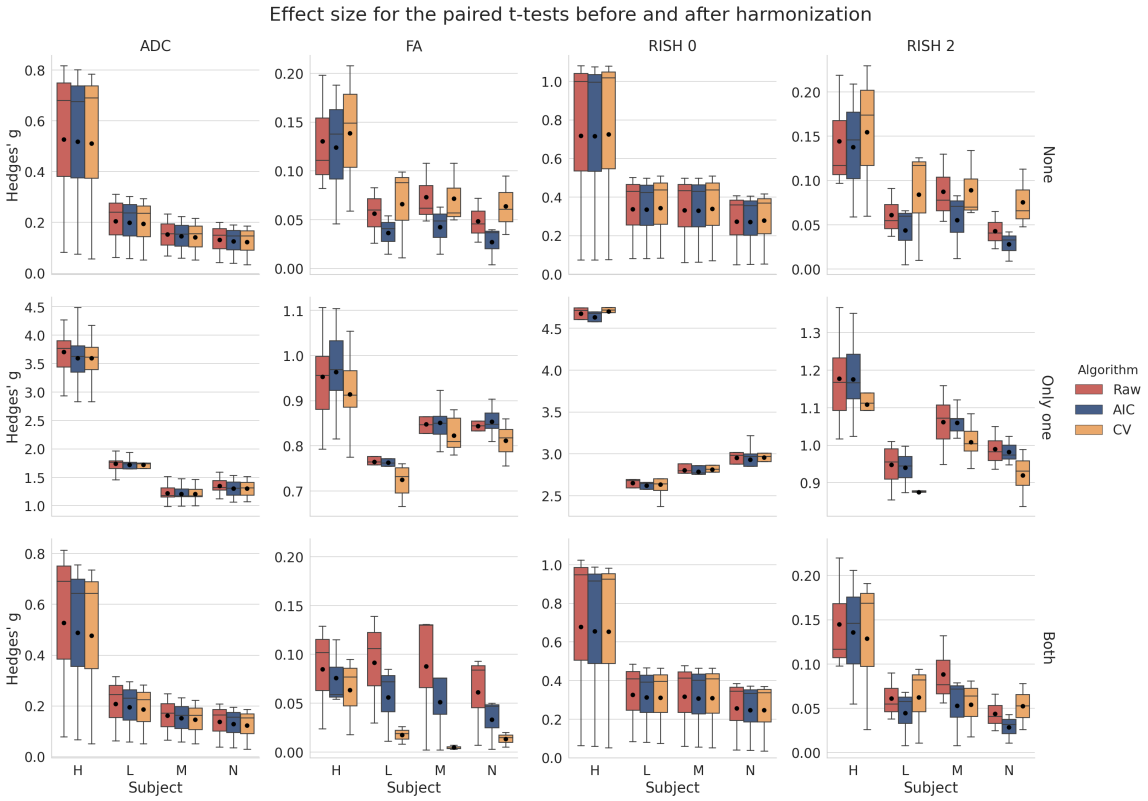

Figure 11. Boxplots of Hedges' g effect size for each metric with the mean value as the black dot. The raw data is shown in red (no harmonization), the data harmonized with the AIC in blue and finally the data harmonized with the CV in orange, similarly to the previous figures. The top row shows the effect size when both datasets are in their original version (**None** of the datasets are altered), the middle row when **only one** of the dataset is altered and the bottom row when **both** datasets are altered as indicated on the right of each row. The top and bottom row are only affected by scanner effects. The middle row shows larger effects size due to one of the compared dataset being altered in addition to the scanner effects.

Figure 12 shows the effect size, with a 95% confidence interval (CI), for the paired t-test between the original and altered datasets on each scanner. While Figure 11 showed the general trend for all results, we instead now focus on the effect size attributable solely to the alterations we previously induced. Results show that the ADC and RISH 0 metrics have the smallest CI, showing the lowest variability in the 3000 voxels in the altered region. All CI are overlapping and therefore have a 95% chance of containing the true mean effect size for every case. The FA and RISH 2 metrics have both larger CI, showing larger variability

in their sample values, but are overlapping with the raw datasets CI in most cases. Only the CV based harmonization method CI is outside the raw datasets CI for two cases. This shows that the effect size is likely preserved after applying the harmonization algorithm in most cases since the only source of variability is the effects we induced in that region to create the altered datasets. The individual effect sizes, p-values and other intermediary statistics for every tested combination that generated the boxplots shown in Figure 11 are available as Supplementary materials.

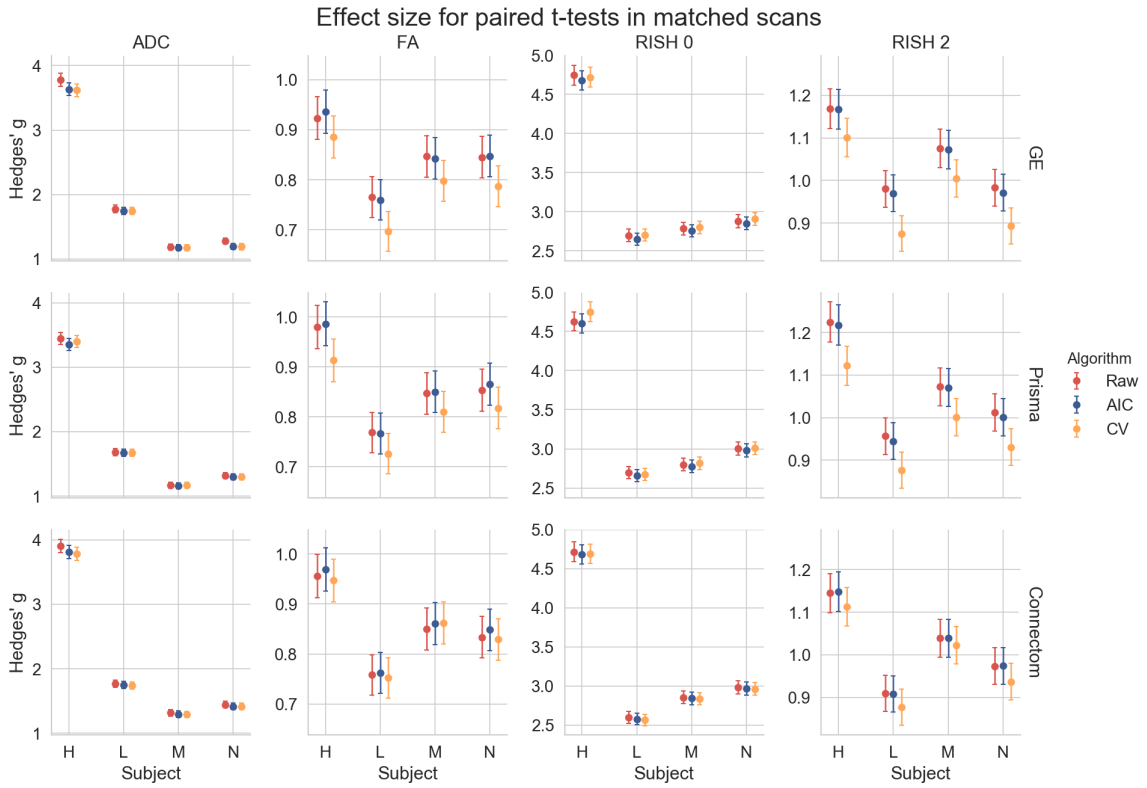

Figure 12. Hedges' g effect size for each metric between the original and altered datasets on the same scanner with a 95% CI. The top row shows the effect size between the original and altered dataset on the GE scanner, the middle row for the Prisma scanner and the bottom row for the Connectom scanner. Most of the CI are overlapping except for the CV in the cases of subject 'L' on the GE scanner and subject 'H' on the Prisma scanner. This effect size is only due to the alterations performed in the experiments and is free of any other source of variability, such as registration error or scanner effects.

5 Discussion

5.1 Reducing variability across scanners

We have presented a new algorithm based on dictionary learning to harmonize datasets acquired on different scanners using the benchmark database from the CDMRI 2017 harmonization challenge (Tax et al., 2019). The flexibility of the method lies in its

ability to pool datasets from any scanner, without the need of paired datasets or spatial correspondence, by adapting the regularization parameter λ_i automatically to each subset of training examples in Equation (1), ensuring that the relevant information to reconstruct the data is encoded in the dictionary $\boldsymbol{D}$. Only features deemed important to the reconstruction are stored as the ℓ_1 norm on the coefficients $\boldsymbol{\alpha}$ encourages a sparse reconstruction and forces most of the coefficients to zero (Candès, Wakin, & Boyd, 2008; Daubechies, Devore, Fornasier, & Güntürk, 2010; St-Jean et al., 2016). In the reconstruction step, a new value of λ_i is automatically selected for each reconstructed patch, ensuring that the regularization is tuned uniquely so that the reconstruction matches the original patch, but using only features found in the target scanner. This is of course at the cost of additional computations since a least-square problem needs to be solved for each candidate value λ_i, but convex and efficient numerical routines reusing the previous solution as a starting point can be used to alleviate computational issues (Friedman et al., 2010). To the best of our knowledge, this is the first case where an automatic search of the regularization parameter has been used in both stages of the optimization.

For the reconstruction step, we introduced two alternatives to compute λ_i through the AIC or CV using held out parts of the signal. While other choices are possible, such as the Bayesian information criterion (Schwarz, 1978), we chose here the AIC for simplicity and because it is in fact equivalent to leave one out CV in the asymptotic limit (Stone, 1977). Cross-validation was done with a classical approach as done in statistics, predicting the signal on parts of the current reconstructed patch as opposed to e.g., reconstructing a completely separate patch with the same value of λ_i as may be done in machine learning. This could explain why the AIC based method performed better than the CV criterion for the anisotropy metrics in the SA protocol since the held out data, which is selected at random for every case, may sometimes unbalance the angular part of the signal because of the random splitting process used during CV. The AIC would not be affected as it can access the whole data for prediction but instead penalizes reconstructions that do not substantially reduce the mean ℓ_2 error and are using too many coefficients—a likely situation of overfitting. This also makes the AIC faster to compute since there is no need to refit the whole model from the beginning unlike the CV. While we used 3-fold cross-validation in this work to limit computations, better results may be obtained by increasing the number of folds held out in total as additional data would be available at each step. However, it is important to keep in mind that the whole model needs to be fitted K-times for K-fold cross-validation, which may be prohibitive from a computational standpoint if many datasets are to be harmonized.

One major advantage of the harmonization approach we presented is its ability to process raw datasets without the requirement of paired samples or spatial alignment during training. In our experiments, the data was given at random for the training phase and we mixed patches from all subjects and all scanners altogether in the additional experiments we described in Section 3.5, preventing overfitting to a particular scanner in the process. Other approaches instead go through an alternative representation such as the SH basis (Blumberg et al., 2019; Cetin Karayumak et al., 2019; Mirzaalian et al., 2016) or harmonize only the extracted scalar maps from diffusion MRI instead (Alexander et al., 2017; Fortin et al., 2017). In the latter cases, it is not clear if the mapping developed for a particular scalar map is in fact similar between metrics as scanner effects may behave differently,

e.g., isotropy metrics may be subject to global effects while anisotropy metrics may exhibit orientational bias due to low SNR in some given gradient directions. We also observed in our experiments that the error for the ADC and RISH 0 metrics were similar for most methods while the error was larger for the FA and RISH 2 metrics for the baseline method, which are orientation dependent. This shows that the "optimal" mapping function could likely be task dependent if one wants to harmonize directly the scalar maps between scanners, which could complicate interpretation between studies that are not using a matched number of b-values or gradient orientation. In the original CDMRI challenge (Tax et al., 2019), the best performing algorithm for some cases of the anisotropy metrics was the baseline algorithm. This was attributed to the blurring resulting of the SH basis interpolation in the angular domain with a trilinear interpolation when the spatial resolution of the datasets is not matching. These results were obtained by applying the harmonization on the GE scanner datasets only while leaving the target scanners (Prisma and Connectom) datasets intact. This task consists in matching the distribution from a source scanner to a target scanner, but without harmonizing the target scanner. This blurring introduced by interpolation could also explain why the baseline method outperforms some of the compared algorithms for the KL divergence in Figures 3 and 5 as this SH interpolation step was not included in the AIC or CV algorithms of this manuscript.

In the additional experiments, we introduced the idea of creating a neutral scanner-space instead of mapping the datasets towards a single target scanner. We also harmonized datasets that had been altered towards that common space and showed that the induced effect sizes are preserved while at the same time preserving normal anatomical variability. This approach has the benefit of removing variability attributable to multiple scanners, instead of trying to force a source scanner to mimic variability that is solely attributable to a target scanner. It is also important to mention here that a good harmonization method should remove unwanted variability due to instrumentation, all the while preserving genuine anatomical effects as also pointed out previously by Fortin et al. (2017). While this statement may seem obvious, success of harmonization towards a common space is much more difficult to quantify than harmonization between scanners since we can not look at difference maps between harmonized datasets anymore. As a thought experiment, a harmonization method that would map all datasets towards a constant value would show no difference between the harmonized datasets themselves, therefore entirely removing all variability. It would however commit very large errors when compared against the original version of those same datasets. From Figure 7, we see that the harmonized datasets are similar to their original version, but Figures 8 and 9 show that the CV based algorithm has larger relative differences with the data before harmonization. It is however not obvious if the CV based algorithm is removing too much variability by underfitting the data or if the AIC based method is not removing enough, overfitting the data. Figure 12 suggests that the CV criterion might underfit the data due to the lower effect size, but this could be due to using only 3 fold CV in our experiments to limit computation time. Results might be improved by using more folds as the AIC approximates the CV as we have previously mentioned. Our recommendation is therefore to use the AIC criterion on large cohort where computation resources are limited, but improvements could be possible by increasing the number of folds for CV or even using a separate test set to build the dictionary if enough data is available to do so.

5.2 Dependence of isotropy and anisotropy metrics on scanning parameters

While it is usually advocated that protocols should use similar scanning parameters as much as possible to ensure reproducibility, this is not always easily feasible depending on the sequences readily available from a given vendor and differences in their implementations. Subtle changes in TE and TR influence the measured signal as shown in Figure 13 by changing the relative T2 and T1 weighting of the measured diffusion signal, respectively. While dMRI local models are usually applied on a per voxel basis, changes in these weightings will yield different values of the diffusion metrics, which makes comparisons between scans more difficult as the weighting depends on the different (and unknown) values of T1 and T2 of each voxel (Brown, Haacke, Cheng, Thompson, & Venkatesan, 2014, Chap. 8). Even if these changes are global for all voxels, matched b-values are not sufficient to ensure that the diffusion time is identical between scans as changes in TE influence diffusion metrics such as increased FA (Qin et al., 2009), but this effect may only manifest itself at either long or very short diffusion times in the human brain (Clark, Hedehus, & Moseley, 2001; Kim, Chi-Fishman, Barnett, & Pierpaoli, 2005). Proper care should be taken to match the diffusion time beyond the well-known b-value, which may not always be the case if different sequences are used e.g., PGSE on the Siemens scanners and TRSE on the GE scanner as used in this manuscript. Additional effects due to gradients and b-values spatial distortions (Bammer et al., 2003) could also adversely affect the diffusion metrics, especially on the Connectom scanner as it uses strong gradients of 300 mT/m (Tax et al., 2019). Isotropy metrics are not necessarily free of these confounds as gradients nonlinearity create a spatially dependent error on the b-values (Paquette, Eichner, & Anwander, 2019). This could explain the larger mean error for the CV and baseline methods on the Connectom scanner harmonization task, especially for the anisotropy metrics. While correcting for these effects is not straightforward, gradient timings should be reported in addition to the usual parameters (e.g., TE, TR, b-values and number of gradient directions) in studies to ease subsequent harmonization. Accounting for these differences during analysis could be done e.g., by using a (possibly mono-exponential) model including diffusion time and predicting the diffusion metrics of interest at common scanning parameters values between the acquisitions to harmonize.

5.3 Limitations

Limitations of harmonization. As Burnham and Anderson (2004) stated, "in a very important sense, we are not trying to model the data; instead, we are trying to model the information in the data". This is indeed the approach taken in the challenge by the participants, the four other entries relying on deep learning and neural networks for the most part with all methods (including ours) optimizing a loss function which considered the difference between the original and the harmonized dataset. With the rapid rise of the next generation of deep learning methods such as generative adversarial networks (GAN) and extensions (Goodfellow et al., 2014), it is now possible to instead model implicitly the *distribution* of the data. This allows generation of datasets from a completely different imaging modality such as synthesizing target CT datasets from source MRI datasets (Wolterink et al., 2017). However, if proper care is not taken to sample truthfully the distribution of the data (e.g., not including enough tumor samples in a harmonization task

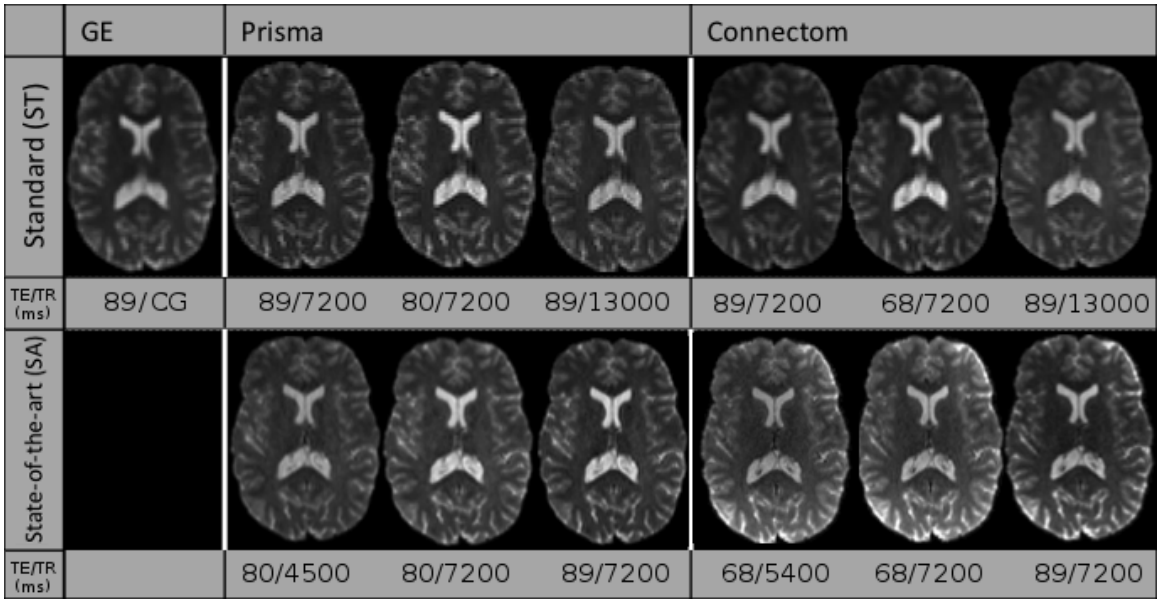

Figure 13. Example $b = 0$ s/mm^2 images for the standard protocol (top row) and the state-of-the-art protocol (bottom row) for a single subject acquired on the three scanners at different combinations of TE and TR. Note that the $b = 0$ s/mm^2 image for the GE scanner was only acquired at a single TE with a cardiac gated (CG) TR. The figure is adapted from Tax et al. (2019), available under the CC-BY 4.0 license.

between datasets with pathological data), this can lead to severe issues. Cohen, Luck, and Honari (2018) recently showed that in such a case, GAN based methods could try to remove the pathology in the data to match the distribution of healthy subjects that the method previously learned, precluding potential applications to new datasets or pathological cases not represented "well enough" in the training set. The same concept would likely apply to systematic artifacts; if every dataset from a target scanner is corrupted by e.g., a table vibration artifact, it may very well be possible that a harmonization algorithm will try to imprint this artifact to the source datasets to match the target datasets. The same remark would apply to our harmonization algorithm; if systematic artifacts are in the data, the learned dictionary may very well try to reconstruct these systematic artifacts. However, when rebuilding the source dataset using this corrupted target dictionary, we expect that the artifact would be mitigated since it would not appear in the source dataset and hence should not be reconstructed by Equation (1) as it would penalize the ℓ_2 norm part of the cost function. This remark also applies to normal variability of the subjects; if the training datasets are too heterogeneous (e.g., young and healthy subjects mixed in with an older population affected by a neurological trait of interest), harmonization algorithms may mistakenly identify (and subsequently remove) information attributable to biological differences between subjects rather than scanner variability. It is therefore implicitly assumed in our algorithm that the datasets to harmonize are representative and well matched (e.g., age, gender) when removing scanner-only differences as other sources of expected variability can be alternatively included in the statistical testing step of the study at hand. While offering

a promising avenue, care must be taken when analyzing harmonization methods to ensure that they still faithfully represent the data as optimal values of the cost functions themselves or "good" reconstruction of the diffusion metrics only may not ensure this fact (Rohlfing, 2012).

Limitations of our algorithm and possible improvements. Our additional experiments with simulated free water have shown how harmonization can, to a certain extent, account for data abnormalities not part of the training set. However, the presence of CSF and the boundary between gray matter and CSF (or a linear combination of those elements) may yield enough information for the reconstruction to encode these features in the dictionary. This can provide new elements that are not used for the reconstruction of normal white matter but may be useful for the altered data in the experiments. It is not necessarily true that this property would also be valid for other neurological disorders such as tumors, especially if their features are not well represented in the training data as we have mentioned previously in Section 5.3. Another aspect that we did not explicitly cover is multishell data i.e., datasets acquired with multiple b-values, which was in fact part of the following CDMRI challenge (Ning et al., 2019). Nevertheless, our method can still be used on such datasets, but would not be aware of the relationship between DWIs beyond the angular domain. Other approaches to build the dictionary could be used to inform the algorithm and potentially increase performance on such datasets by explicitly modeling the spatial and angular relationship (Schwab et al., 2018) or using an adaptive weighting considering the b-values in the angular domain (Duits, St-Onge, Portegies, & Smets, 2019) amongst other possible strategies. This weighting strategy could be used for repeated acquisitions or if multishell datasets without an equal repartition of the data across shells needs to be harmonized instead of the strictly angular criterion we used in this manuscript. Note however that redefining the extraction step would only affect the initial creation of the patches as defined in Appendix A, leaving Equation (1) unchanged. Modeling explicitly the angular part of the signal could also be used to sample new gradients directions directly, an aspect we covered in the original CDMRI challenge by using the spherical harmonics basis (Descoteaux et al., 2007). Correction for the nature of the noise distribution could also be subsequently included as a processing step before harmonization since reconstruction algorithms vary by scanner vendor (Dietrich et al., 2008; St-Jean et al., 2018), leading to differences between scans due to changes in the noise floor level (Sakaie et al., 2018). Improvements could also potentially be achieved by considering the group structure shared by overlapping patches when optimizing Equation (1) (Simon, Friedman, Hastie, & Tibshirani, 2013). While this structure would need to be explicitly specified, optimizing jointly groups of variables has recently led to massive improvements in other applications of diffusion MRI such as reduction of false positive connections in tractography (Schiavi et al., 2019). In the end, the aim of harmonization procedures is to reduce variability arising from non-biological effects of interest in the application at hand. Future benefits for this class of methods should therefore be evaluated on the end result of a study, rather than using proxy metrics of the diffusion signal for evaluation as is commonly done. In the current work, registration errors or misalignment between subjects may influence negatively the evaluation of the algorithms as previously outlined in the CDMRI challenge (Tax et al., 2019), even though *a priori* alignment is not an assumption of the presented harmonization algorithm. Further validation of the proposed harmonization

algorithm is therefore planned on a large-scale retrospective multicenter study to evaluate the effect of harmonization on clinical outcomes.

6 Conclusions

In this paper, we have developed and evaluated a new harmonization algorithm to reduce intra and inter scanner differences. Using the public database from the CDMRI 2017 harmonization challenge, we have shown how a mapping to reduce variability attributable to the scanning protocol can be constructed automatically through dictionary learning using datasets acquired on different scanners. These datasets do not require to be matched or spatially registered, making the algorithm applicable in retrospective multicenter studies. The harmonization can also be done for different spatial resolutions through careful matching of the ratio between the spatial patch size used to build the dictionary and the spatial resolution of the target scanner. We also introduced the concept of mapping datasets towards an arbitrary scanner-space and used the proposed algorithm to reconstruct altered versions of the test datasets corrupted by a free water compartment, even if such data was not part of the training datasets. Results have shown that the effect size due to alterations is preserved after harmonization, while removing variability attributable to scanner effects in the datasets. We also provided recommendation when harmonizing protocols, such as reporting the gradient timings to inform subsequent harmonization algorithms which could exploit these values across studies. As perfect matching of scanner parameters is difficult to do in practice due to differences in vendor implementations, an alternative approach could be to account for these differences through models of diffusion using these additional parameters. Nevertheless, as the algorithm is freely available, this could help multicenter studies in pooling their data while removing scanner specific confounds and increase statistical power in the process.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi:10.1109/TAC.1974.1100705

Alexander, D. C., Zikic, D., Ghosh, A., Tanno, R., Wottschel, V., Zhang, J., ... Criminisi, A. (2017). Image quality transfer and applications in diffusion MRI. *NeuroImage*, *152*, 283–298. doi:10.1016/j.neuroimage.2017.02.089

Assemlal, H.-E., Tschumperlé, D., Brun, L., & Siddiqi, K. (2011). Recent advances in diffusion MRI modeling: Angular and radial reconstruction. *Medical Image Analysis*, *15*(4), 369–396. doi:10.1016/j.media.2011.02.002

Bammer, R., Markl, M., Barnett, A., Acar, B., Alley, M., Pelc, N., ... Moseley, M. (2003). Analysis and generalized correction of the effect of spatial gradient field distortions in diffusion-weighted imaging. *Magnetic Resonance in Medicine*, *50*(3), 560–569. doi:10.1002/mrm.10545

Basser, P., Mattiello, J., & LeBihan, D. (1994). MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, *66*(1), 259–267. doi:10.1016/S0006-3495(94)80775-1

Basser, P. J., & Pierpaoli, C. [Carlo]. (1996). Microstructural and Physiological Features of Tissues Elucidated by Quantitative-Diffusion-Tensor MRI. *Journal of Magnetic Resonance, Series B*, *111*(3), 209–219. doi:10.1006/jmrb.1996.0086

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300. doi:10.2307/2346101

Blumberg, S. B., Palombo, M., Khoo, C. S., Tax, C. M. W., Tanno, R., & Alexander, D. C. (2019). Multi-stage Prediction Networks for Data Harmonization. In *Lecture notes in computer science* (Vol. 11767 LNCS, pp. 411–419). doi:10.1007/978-3-030-32251-9_45

Brown, R. W., Haacke, E. M., Cheng, Y.-C. N., Thompson, M. R., & Venkatesan, R. (2014). *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference. *Sociological Methods & Research*, *33*(2), 261–304. doi:10.1177/0049124104268644

Candès, E. J., Wakin, M. B., & Boyd, S. P. (2008). Enhancing Sparsity by Reweighted l1 Minimization. *Journal of Fourier Analysis and Applications*, *14*(5-6), 877–905. doi:10.1007/s00041-008-9045-x

Cetin Karayumak, S., Bouix, S., Ning, L., James, A., Crow, T., Shenton, M., ... Rathi, Y. (2019). Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. *NeuroImage*, *184*(May 2018), 180–200. doi:10.1016/j.neuroimage.2018.08.073

Clark, C. A., Hedehus, M., & Moseley, M. E. (2001). Diffusion time dependence of the apparent diffusion tensor in healthy human brain and white matter disease. *Magnetic Resonance in Medicine*, *45*(6), 1126–1129. doi:10.1002/mrm.1149

Cohen, J. P., Luck, M., & Honari, S. (2018). Distribution Matching Losses Can Hallucinate Features in Medical Image Translation. In *Medical image computing and computer assisted intervention – miccai 2018* (Vol. 57, pp. 529–536). doi:10.1007/978-3-030-00928-1_60

Daubechies, I., Devore, R., Fornasier, M., & Güntürk, C. S. (2010). Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, *63*, 1–38. doi:10.1002/cpa.20303

De Luca, A., Franklin, S., Lucci, C., Hendrikse, J., Froeling, M., & Leemans, A. (2019). Investigation of the dependence of free water and pseudo-diffusion MRI estimates on the cardiac cycle. In *International society for magnetic resonance in medicine (ismrm)* (p. 344).

Descoteaux, M., Angelino, E., Fitzgibbons, S., & Deriche, R. (2007). Regularized, fast, and robust analytical Q-ball imaging. *Magnetic Resonance in Medicine*, *58*(3), 497–510. doi:10.1002/mrm.21277

Dietrich, O., Raya, J. G., Reeder, S. B., Ingrisch, M., Reiser, M. F., & Schoenberg, S. O. (2008). Influence of multichannel combination, parallel imaging and other reconstruction techniques on MRI noise characteristics. *Magnetic Resonance Imaging*, *26*(6), 754–762. doi:10.1016/j.mri.2008.02.001

Duchesne, S., Chouinard, I., Potvin, O., Fonov, V. S., Khademi, A., Bartha, R., ... Black, S. E. (2019). The Canadian Dementia Imaging Protocol: Harmonizing National Cohorts. *Journal of Magnetic Resonance Imaging*, *49*(2), 456–465. doi:10.1002/jmri.26197

Duits, R., St-Onge, E., Portegies, J., & Smets, B. (2019). Total Variation and Mean Curvature PDEs on the Space of Positions and Orientations. In A. M. Bruckstein, B. M. ter

Haar Romeny, A. M. Bronstein, & M. M. Bronstein (Eds.), *Lecture notes in computer science* (Vol. 11603, pp. 211–223). doi:10.1007/978-3-030-22368-7_17

Elad, M., & Aharon, M. (2006). Image Denoising Via Sparse and Redundant Representations Over Learned Dictionaries. *IEEE Transactions on Image Processing*, *15*(12), 3736–3745. doi:10.1109/TIP.2006.881969

Federau, C., Hagmann, P., Maeder, P., Müller, M., Meuli, R., Stuber, M., & O'Brien, K. (2013). Dependence of Brain Intravoxel Incoherent Motion Perfusion Parameters on the Cardiac Cycle. *PLoS ONE*, *8*(8), e72856. doi:10.1371/journal.pone.0072856

Feinberg, D. A., Moeller, S., Smith, S. M., Auerbach, E., Ramanna, S., Glasser, M. F., ... Yacoub, E. (2010). Multiplexed echo planar imaging for sub-second whole brain fmri and fast diffusion imaging. *PLoS ONE*, *5*(12). doi:10.1371/journal.pone.0015710

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., ... Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, *161*, 149–170. doi:10.1016/j.neuroimage.2017.08.047

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, *33*(1), 1–22.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative Adversarial Networks. In *Proceedings of the international conference on neural information processing systems (nips 2014)* (pp. 2672–2680).

Gramfort, A., Poupon, C., & Descoteaux, M. (2014). Denoising and fast diffusion imaging with physically constrained sparse dictionary learning. *Medical image analysis*, *18*(1), 36–49. doi:10.1016/j.media.2013.08.006

Gronenschild, E. H. B. M., Habets, P., Jacobs, H. I. L., Mengelers, R., Rozendaal, N., van Os, J., & Marcelis, M. (2012). The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements. *PLoS ONE*, *7*(6), e38234. doi:10.1371/journal.pone.0038234

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational Statistics*, *6*(2), 107–128. doi:10.3102/10769986006002107

Hofman, A., Brusselle, G. G. O., Murad, S. D., van Duijn, C. M., Franco, O. H., Goedegebure, A., ... Vernooij, M. W. (2015). The Rotterdam Study: 2016 objectives and design update. *European Journal of Epidemiology*, *30*(8), 661–708. doi:10.1007/s10654-015-0082-x

Hofman, A., Grobbee, D. E., De Jong, P. T. V. M., & Van den Ouweland, F. A. (1991). Determinants of disease and disability in the elderly: The Rotterdam elderly study. *European Journal of Epidemiology*, *7*(4), 403–422. doi:10.1007/BF00145007

Huisman, T. A. G. M., Loenneker, T., Barta, G., Bellemann, M. E., Hennig, J., Fischer, J. E., & Il'yasov, K. A. (2006). Quantitative diffusion tensor MR imaging of the brain: field strength related variance of apparent diffusion coefficient (ADC) and fractional anisotropy (FA) scalars. *European Radiology*, *16*(8), 1651–1658. doi:10.1007/s00330-006-0175-8

St-Jean, S., Coupé, P., & Descoteaux, M. (2016). Non Local Spatial and Angular Matching: Enabling higher spatial resolution diffusion MRI datasets through adaptive denoising. *Medical Image Analysis*, *32*(2016), 115–130. doi:10.1016/j.media.2016.02.010

St-Jean, S., De Luca, A., Viergever, M. A., & Leemans, A. (2018). Automatic, Fast and Robust Characterization of Noise Distributions for Diffusion MRI. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, & G. Fichtinger (Eds.), *Medical image computing and computer assisted intervention – MICCAI 2018* (pp. 304–312). doi:10.1007/978-3-030-00928-1_35

St-Jean, S., Viergever, M. A., & Leemans, A. (2017). A unified framework for upsampling and denoising of diffusion MRI data. *25th Annual Meeting of ISMRM*, 3533.

St-Jean, S., Viergever, M. A., & Leemans, A. (2019). samuelstjean/harmonization: Harmonization of diffusion MRI datasets with adaptive dictionary learning. doi:10.5281/zenodo.3385925

Johansen-Berg, H., & Behrens, T. E. J. (2009). *Diffusion MRI*. doi:10.1016/B978-0-12-374709-9.X0001-6

Jones, D. K. [Derek K.]. (2011). *Diffusion MRI: Theory, Methods, and Applications*. Cambridge: Cambridge University Press.

Kim, S., Chi-Fishman, G., Barnett, A. S., & Pierpaoli, C. (2005). Dependence on diffusion time of apparent diffusion tensor of ex vivo calf tongue and heart. *Magnetic Resonance in Medicine*, *54*(6), 1387–1396. doi:10.1002/mrm.20676

Kristo, G., Leemans, A., Raemaekers, M., Rutten, G.-J., de Gelder, B., & Ramsey, N. F. (2013). Reliability of two clinically relevant fiber pathways reconstructed with constrained spherical deconvolution. *Magnetic Resonance in Medicine*, *70*(6), 1544–1556. doi:10.1002/mrm.24602

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. doi:10.1214/aoms/1177729694

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(NOV), 1–12. doi:10.3389/fpsyg.2013.00863

Larkman, D. J., Hajnal, J. V., Herlihy, A. H., Coutts, G. A., Young, I. R., & Ehnholm, G. (2001). Use of multicoil arrays for separation of signal from multiple slices simultaneously excited. *Journal of Magnetic Resonance Imaging*, *13*(2), 313–317. doi:10.1002/1522-2586(200102)13:2<313::AID-JMRI1045>3.0.CO;2-W

Lustig, M., Donoho, D., & Pauly, J. M. (2007). Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, *58*(6), 1182–1195. doi:10.1002/mrm.21391

Magnotta, V. A., Matsui, J. T., Liu, D., Johnson, H. J., Long, J. D., Bolster, B. D., ... Paulsen, J. S. (2012). MultiCenter Reliability of Diffusion Tensor Imaging. *Brain Connectivity*, *2*(6), 345–355. doi:10.1089/brain.2012.0112

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2010). Online Learning for Matrix Factorization and Sparse Coding. *Journal of Machine Learning Research*, *11*, 19–60.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., & Zisserman, A. (2009). Non-local sparse models for image restoration. In *2009 ieee 12th international conference on computer vision* (Vol. 2, pp. 2272–2279). doi:10.1109/ICCV.2009.5459452

Merlet, S., Caruyer, E., Ghosh, A., & Deriche, R. (2013). A computational diffusion MRI and parametric dictionary learning framework for modeling the diffusion signal and its features. *Medical Image Analysis*, *17*(7), 830–843. doi:10.1016/j.media.2013.04.011

Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., ... Rathi, Y. (2016). Inter-site and inter-scanner diffusion MRI data harmonization. *NeuroImage*, *135*, 311–323. doi:10.1016/j.neuroimage.2016.04.041

Ning, L., Bonet-Carne, E., Grussu, F., Sepehrband, F., Kaden, E., Veraart, J., ... Tax, C. W. M. (2019). Muti-shell Diffusion MRI Harmonisation and Enhancement Challenge (MUSHAC): Progress and Results. In E. Kaden, F. Grussu, L. Ning, C. M. W. Tax, & J. Veraart (Eds.), (pp. 217–224). doi:10.1007/978-3-030-05831-9_18

Paquette, M., Eichner, C., & Anwander, A. (2019). Gradient Non-Linearity Correction for Spherical Mean Diffusion Imaging. In *International symposium on magnetic resonance in medicine (ismrm'19)* (p. 550).

Pasternak, O., Sochen, N., Gur, Y., Intrator, N., & Assaf, Y. (2009). Free water elimination and mapping from diffusion MRI. *Magnetic Resonance in Medicine*, *62*(3), 717–730. doi:10.1002/mrm.22055

Pierpaoli, C., & Jones, D. K. [D K]. (2004). Removing CSF Contamination in Brain DT-MRIs by Using a Two-Compartment Tensor Model. In *Proceedings of the 12th annual meeting of ismrm, kyoto* (p. 1215).

Pinto, M. S., Paolella, R., Billiet, T., Van Dyck, P., Guns, P.-J., Jeurissen, B., ... Sijbers, J. (2020). Harmonization of Brain Diffusion MRI: Concepts and Methods. *Frontiers in Neuroscience*, *14*(May). doi:10.3389/fnins.2020.00396

Pohl, K. M., Sullivan, E. V., Rohlfing, T., Chu, W., Kwon, D., Nichols, B. N., ... Pfefferbaum, A. (2016). Harmonizing DTI measurements across scanners to examine the development of white matter microstructure in 803 adolescents of the NCANDA study. *NeuroImage*, *130*, 194–213. doi:10.1016/j.neuroimage.2016.01.061

Qin, W., Shui Yu, C., Zhang, F., Du, X. Y., Jiang, H., Xia Yan, Y., & Cheng Li, K. (2009). Effects of echo time on diffusion quantification of brain white matter at 1.5T and 3.0T. *Magnetic Resonance in Medicine*, *61*(4), 755–760. doi:10.1002/mrm.21920

Rohlfing, T. (2012). Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable. *IEEE Transactions on Medical Imaging*, *31*(2), 153–163. doi:10.1109/TMI.2011.2163944

Rueda, A., Malpica, N., & Romero, E. (2013). Single-image super-resolution of brain MR images using overcomplete dictionaries. *Medical Image Analysis*, *17*(1), 113–132. doi:10.1016/j.media.2012.09.003

Sakaie, K., Zhou, X., Lin, J., Debbins, J., Lowe, M., & Fox, R. J. (2018). Technical Note: Retrospective reduction in systematic differences across scanner changes by accounting for noise floor effects in diffusion tensor imaging. *Medical Physics*, *45*(9), 4171–4178. doi:10.1002/mp.13088

Schiavi, S., Barakovic, M., Pineda, M. O., Descoteaux, M., Thiran, J.-P., & Daducci, A. (2019). Reducing false positives in tractography with microstructural and anatomical priors. *bioRxiv*. doi:10.1101/608349

Schwab, E., Vidal, R., & Charon, N. (2018). Joint spatial-angular sparse coding for dMRI with separable dictionaries. *Medical Image Analysis*, *48*, 25–42. doi:10.1016/j.media.2018.05.002

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, *22*(2), 231–245. doi:10.1080/10618600.2012.681250

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 44–47. doi:10.1111/j.2517-6161.1977.tb01603.x

Student. (1908). The Probable Error of a Mean. *Biometrika*, *6*(1), 1–25.

Tax, C. M., Grussu, F., Kaden, E., Ning, L., Rudrapatna, U., John Evans, C., ... 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

Tibshirani, R. J., & Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, *40*(2), 1198–1232. doi:10.1214/12-AOS1003

Tournier, J.-D. (2019). Diffusion MRI in the brain – Theory and concepts. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *112-113*, 1–16. doi:10.1016/j.pnmrs.2019.03.001

Tournier, J.-D., Mori, S., & Leemans, A. (2011). Diffusion tensor imaging and beyond. *Magnetic Resonance in Medicine*, *65*(6), 1532–1556. doi:10.1002/mrm.22924

Vemuri, B. C., Sun, J., Banerjee, M., Pan, Z., Turner, S. M., Fuller, D. D., ... Entezari, A. (2019). A geometric framework for ensemble average propagator reconstruction from diffusion MRI. *Medical Image Analysis*, *57*, 89–105. doi:10.1016/j.media.2019.06.012

Vollmar, C., O'Muircheartaigh, J., Barker, G. J., Symms, M. R., Thompson, P., Kumari, V., ... Koepp, M. J. (2010). Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0 T scanners. *NeuroImage*, *51*(4), 1384–1394. doi:10.1016/j.neuroimage.2010.03.046

Wolterink, J. M., Dinkla, A. M., Savenije, M. H. F., Seevinck, P. R., van den Berg, C. A. T., & Išgum, I. (2017). Deep MR to CT Synthesis Using Unpaired Data. In *Lecture notes in computer science* (Vol. 10557 LNCS, pp. 14–23). doi:10.1007/978-3-319-68127-6_2

Yang, J., Wang, Z., Lin, Z., Cohen, S., & Huang, T. (2012). Coupled Dictionary Training for Image Super-Resolution. *IEEE Transactions on Image Processing*, *21*(8), 3467–3478. doi:10.1109/TIP.2012.2192127

Yang, J., Wright, J., Huang, T., & Ma, Y. (2010). Image Super-Resolution Via Sparse Representation. *IEEE Transactions on Image Processing*, *19*(11), 2861–2873. doi:10.1109/TIP.2010.2050625

Zhu, A. H., Moyer, D. C., Nir, T. M., Thompson, P. M., & Jahanshad, N. (2019). Challenges and Opportunities in dMRI Data Harmonization. In E. Kaden, F. Grussu, L. Ning, C. M. W. Tax, & J. Veraart (Eds.), *Computational diffusion mri miccai 2019* (pp. 157–172). doi:10.1007/978-3-030-05831-9_13

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, *35*(5), 2173–2192. doi:10.1214/009053607000000127

Appendix

The harmonization algorithm

This appendix outlines the harmonization algorithm in two separate parts. Algorithm 1 first shows how to build a target dictionary as depicted in the top part of Figure 1. The bottom part of the diagram shows how to rebuild a dataset given the dictionary and is detailed in Algorithm 2. Our implementation is also freely available at https://github.com/samuelstjean/harmonization (St-Jean et al., 2019).

Algorithm 1: The proposed harmonization algorithm - building a target dictionary.

Data: Datasets, patch size, angular neighbor
Result: Dictionary $\boldsymbol{D}$

Step 1 : *Extracting patches from all datasets*;
foreach *Datasets* **do**
  Find the closest angular neighbors;
  Create a 4D block with a $b = 0$ s/mm^2 image and the angular neighbors;
  Extract all 3D patches and store the result in an array Ω;
end
Step 2 : *Build the target dictionary*;
while *Number of max iterations not reached* **do**
  Randomly pick patches from Ω;
  Solve Equation (1) for $\boldsymbol{\alpha}$ with $\boldsymbol{D}$ fixed;
  Solve Equation (1) for $\boldsymbol{D}$ with $\boldsymbol{\alpha}$ fixed using e.g., Mairal, Bach, Ponce, and Sapiro (2010, Algorithm 2);
end

Algorithm 2: The proposed harmonization algorithm - reconstruction of the harmonized data.

Data: Dataset, dictionary
Result: Harmonized dataset

Step 1 : *Extracting patches from the dataset to harmonize*;
foreach *Dataset* **do**
  Find the closest angular neighbors;
  Create a 4D block with a $b = 0$ s/mm^2 image and the angular neighbors;
  Extract all overlapping 3D patches and store the result as Ω;
end

if *Matching across spatial resolution* **then**
  Downsample $\boldsymbol{D}$ into $\boldsymbol{D}_{\text{small}}$ spatially before reconstruction;
else
  $\boldsymbol{D}_{\text{small}} = \boldsymbol{D}$;
end

Step 2 : *Find the harmonized patch*;
foreach *patches* $\in \Omega$ **do**
  Find the coefficients $\boldsymbol{\alpha}$ by solving Equation (1) for $\boldsymbol{D}_{\text{small}}$ fixed;
  Find the harmonized representation $\boldsymbol{X} = \boldsymbol{D}\boldsymbol{\alpha}$;
end
foreach *patches* $\in \Omega$ **do**
  Put back each patch at its spatial location and average overlapping parts;
end