NeuroImage 161 (2017) 149–170

Contents lists available at ScienceDirect

NeuroImage

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

Harmonization of multi-site diffusion tensor imaging data

Jean-Philippe Fortin [a], Drew Parker [b], Birkan Tunç [b], Takanori Watanabe [b], Mark A. Elliott [c], Kosha Ruparel [d], David R. Roalf [d], Theodore D. Satterthwaite [d], Ruben C. Gur [c,d], Raquel E. Gur [c,d], Robert T. Schultz [e], Ragini Verma [b,1], Russell T. Shinohara [a,*,1]

[a] Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, USA
[b] Center for Biomedical Image Computing and Analytics, Department of Radiology, Perelman School of Medicine, University of Pennsylvania, USA
[c] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, USA
[d] Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, USA
[e] Center for Autism Research, The Children's Hospital of Philadelphia, USA

ARTICLE INFO

Keywords:
DTI
Diffusion
Harmonization
Multi-site
ComBat
Inter-scanner

ABSTRACT

Diffusion tensor imaging (DTI) is a well-established magnetic resonance imaging (MRI) technique used for studying microstructural changes in the white matter. As with many other imaging modalities, DTI images suffer from technical between-scanner variation that hinders comparisons of images across imaging sites, scanners and over time. Using fractional anisotropy (FA) and mean diffusivity (MD) maps of 205 healthy participants acquired on two different scanners, we show that the DTI measurements are highly site-specific, highlighting the need of correcting for site effects before performing downstream statistical analyses. We first show evidence that combining DTI data from multiple sites, without harmonization, may be counter-productive and negatively impacts the inference. Then, we propose and compare several harmonization approaches for DTI data, and show that ComBat, a popular batch-effect correction tool used in genomics, performs best at modeling and removing the unwanted inter-site variability in FA and MD maps. Using age as a biological phenotype of interest, we show that ComBat both preserves biological variability and removes the unwanted variation introduced by site. Finally, we assess the different harmonization methods in the presence of different levels of confounding between site and age, in addition to test robustness to small sample size studies.

1. Introduction

Diffusion tensor imaging (DTI) is a well-established magnetic resonance imaging (MRI) technique for studying the white matter (WM) organization and tissue characteristics of the brain. Diffusion tensor imaging has been used extensively to study both brain development and pathology; see Alexander et al. (2007) for a review of DTI and several of its applications. In studies assessing white matter tissue characteristics, two commonly reported complementary scalar maps are the mean diffusivity (MD), which assesses the degree to which water diffuses at each location, and fractional anisotropy (FA), which measures the coherence of this diffusion in one particular direction. Together, MD and FA provide complementary description of white matter microstructure.

With the increasing number of publicly availably neuroimaging databases, a crucial goal is to combine large-scale imaging studies to increase the power of statistical analyses to test common biological hypothesis. For instance, for life-span studies, combining data across sites

Abbreviations: ADNI, Alzheimer's Disease NeuroImaging Initiative; AD, Axial diffusivity; CAT, Concordance at the top; ComBat, Combatting batch effects when combining batches of gene expression microarray data; CoV, Coefficient of variation; CSF, Cerebrospinal fluid; DTI, Diffusion tensor imaging; EB, Empirical Bayes; FA, Fractional anisotropy; GM, Grey matter; GS, Global scaling; IBMA, Image-based meta analysis; IPW, Inverse probability weighting; MD, Mean diffusivity; MRI, Magnetic resonance imaging; OLS, Ordinary least squares; RD, Radial diffusivity; RAVEL, Removal of artificial voxel effect by linear regression; RMSE, Root mean square error; RISH, Rotation invariant spherical harmonic; ROI, Region of interest; SVA, Surrogate variable analysis; SVD, Singular value decomposition; T1-w, T1-weighted; TBI, Traumatic brain injury; TBSS, Tract-based spatial statistics; TDC, Typically developing control; WM, White matter; WMPM, White matter parcellation map.

* Corresponding author.
E-mail address: rshi@mail.med.upenn.edu (R.T. Shinohara).

[1] Equal contribution.

http://dx.doi.org/10.1016/j.neuroimage.2017.08.047
Received 17 March 2017; Received in revised form 3 July 2017; Accepted 15 August 2017
Available online 18 August 2017

and age ranges is essential for obtaining the necessary number of participants of each age. The success of combining multi-site imaging data depends critically on the comparability of the images across sites. As with other imaging modalities, DTI images are subject to technical variability across scans, including heterogeneity in the imaging protocol, variations in the scanning parameters and differences in the scanner manufacturers (Zhu et al., 2009, 2011). Among others, the reliability of FA and MD maps have been shown to be affected by angular and spatial resolution (Zhan et al., 2010; Alexander et al., 2001; Kim et al., 2006), the number of diffusion weighting directions (Giannelli et al., 2009), the number of gradient sampling orientations (Jones, 2004), the number of b-values (Correia et al., 2009), and the b-values themselves.

In the design of multi-site studies, defining a standardized DTI protocol is a first step towards reducing inter-scanner variability. However, even in the presence of a standardized protocol, systematic differences between scanner manufacturers, field strength and other scanner characteristics will systematically affect the DTI images and induce inter-scanner variation. Image-based meta analysis (IBMA) techniques, reviewed in Salimi-Khorshidi et al. (2009), are common methods for combining results from multi-site studies with the goal of testing a statistical hypothesis. IBMA methods circumvent the need of harmonizing images across sites by performing site-specific statistical analyses and combining results afterwards. Fisher's p-value combining method and Stouffer's z-transformation test, applied to z or t-maps, are two common IBMA techniques. Fixed-effect models based on (possibly) normalized images, and mixed-effect models to model the inter- and intra-site variability, are other common techniques for the analysis of multi-site data. Indeed, meta-analysis methods have shown great promise for studies with a large number of participants at each site. For instance, the ENIGMA-DTI working group has been successfully using and validating meta-analysis techniques on such multi-site DTI data (Jahanshad et al., 2013; Kochunov et al., 2014).

Meta-analysis techniques have several limitations, however. First, study-specific samples might not be sufficient to estimate the true biological variability in the population (Mirzaalian et al., 2016). As described by De Wit et al. (2014), adjusting for variability at the participant level is problematic in meta-analyses, since only group-level demographic and clinical information is available. Another limitation is that for a multi-site study, computing site-specific summary statistics will be affected by unbalanced data. For instance, the calculation of a variance using unbalanced datasets is highly affected by the ratio cases/controls in the sample (Linn et al., 2016b). Another limitation, for imaging studies with small sample sizes, the parameters of the z-score transformations cannot be robustly estimated, yielding suboptimal statistical inferences.

Mega-analyses, in which the imaging data are combined before performing statistical inferences, have the potential to increase power compared to meta-analyses (De Wit et al., 2014). In addition, pooling imaging data across studies has the benefit of enriching the clinical picture of the sample by increasing the variability in symptom profiles (Turner, 2014) and demographic variables. This is particularly important for age-span studies. However, pooling data across studies may increase the heterogeneity of the imaging measurements by introducing undesirable variability caused by differences in scanner protocols. Harmonization of the pooled data is therefore necessary to ensure the success of mega-analyses. The DTI harmonization technique proposed in Mirzaalian et al. (2016) is a first step towards that direction. The method is based on rotation invariant spherical harmonics (RISH) and combines the unprocessed DTI images across scanners. Unfortunately, a major drawback of the method is that it requires DTI data to have similar acquisition parameters across sites, an assumption often infeasible in multi-site observational analyses.

In this work, we adapted and compared several statistical approaches for the harmonization of DTI studies that were previously developed for other data types: Functional normalization (Fortin et al., 2014), RAVEL (Fortin et al., 2016a), Surrogate variable analysis (SVA) (Leek and Storey, 2007) and ComBat (Johnson et al., 2007), a popular batch adjustment method developed for genomics data. We also include a simple method that globally rescales the data for each site using a z-score transformation map common to all features, which we refer to as "global scaling". For the evaluation of the different harmonization techniques, we use DTI data acquired as a part of two large imaging studies ((Satterthwaite et al., 2014) and (Ghanbari et al., 2014)) with images acquired on different scanners, using different imaging protocols. The participants are teenagers, and were matched across studies for age, gender, ethnicity, and handedness.

We first analyze site-related differences in the FA, MD, radial diffusivity (RD) and axial diffusivity (AD) measurements, and show evidence of significant site effects that differ across the brain. This motivates the need for a harmonization technique that is sensitive to region-specific scanner effects. Then, we harmonize the data with several proposed harmonizations, and evaluate their performance using a comprehensive evaluation framework. We show that the ComBat is the most effective harmonization techniques as it removes unwanted variation induced by site, while preserving between-subject biological variability. ComBat is a promising harmonization technique for other imaging modalities since it does not make assumptions about the origin of the site effects.

2. Methods

2.1. Data

We consider two DTI studies from two different scanners. To investigate the effect of scanner variations on the DTI measurements, we matched the participants for age, gender, ethnicity and handedness, resulting in 105 participants retained in each study for further analysis. The characteristics of each dataset are described below.

Dataset 1 (Site 1): PNC dataset. We selected a subset of the Philadelphia Neurodevelopmental Cohort (PNC) (Satterthwaite et al., 2014), and included 105 healthy participants from 8 to 19 years old. 83 of the participants were males (22 females), and 75 participants were white (30 non-white). The DTI data were acquired on a 3T Siemens TIM Trio whole-body scanner, using a 32 channel head coil and a twice-refocused spin-echo (TRSE) single-shot EPI sequence with the following parameters: TR = 8100 ms and TE = 82 ms, b-value of 1 000 s/mm^2, 7 $b = 0$ images and 64 gradient directions. The images were acquired at 1.875 × 1.875 × 2 mm resolution. During the same session, structural T1-weighted (T1-w) MP-RAGE images were also acquired with parameters TR = 1810 ms, TE = 3.5 ms, TI = 1100 ms and FA = 9°, at 0.9375 × 0.9375 × 1 mm resolution.

Dataset 2 (Site 2): ASD dataset. The dataset contains 105 typically developing controls (TDC) from a study focusing on autism spectrum disorder (ASD) (Ghanbari et al., 2014). 83 of the participants were males (22 females), and 79 participants were white (26 non-white). The age of the participants ranges from 8 to 18 years old. The DTI data were acquired on a Siemens 3T Verio scanner, using a 32 channel head coil and a single shot spin-echo planar sequence with the following parameters: TR = 11,000 ms and TE = 76 ms, b-value of 1 000 s/mm^2, 1 $b = 0$ image and 30 gradient directions. The images were acquired at 2 mm isotropic resolution. Structural T1-w MP-RAGE images were also acquired with parameters TR = 1900 ms, TE = 2.54 ms, TI = 900 ms and FA = 9° at resolution 0.8 mm × 0.8 mm × 0.9 mm.

For benchmarking the different harmonization procedures, we use two additional subsets of the PNC database, with participants who differ from Dataset 1:

Independent Dataset 1: The dataset contains 292 additional healthy participants from the PNC with the same age range as Dataset 1 and Dataset 2 (8–18 years old).

Independent Dataset 2: The dataset contains 105 additional healthy participants from the PNC with an age range of 14–22 years old.