NeuroImage 132 (2016) 198–212

Contents lists available at ScienceDirect

NeuroImage

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

Removing inter-subject technical variability in magnetic resonance imaging studies

Jean-Philippe Fortin [a], Elizabeth M. Sweeney [a], John Muschelli [a], Ciprian M. Crainiceanu [a], Russell T. Shinohara [b,*], The Alzheimer's Disease Neuroimaging Initiative [1]

[a] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[b] Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

ARTICLE INFO

Article history:
Received 30 October 2015
Accepted 12 February 2016
Available online 23 February 2016

Keywords:
MRI
Normalization
Scan effect
Alzheimer's disease
ADNI

ABSTRACT

Magnetic resonance imaging (MRI) intensities are acquired in arbitrary units, making scans non-comparable across sites and between subjects. Intensity normalization is a first step for the improvement of comparability of the images across subjects. However, we show that unwanted inter-scan variability associated with imaging site, scanner effect, and other technical artifacts is still present after standard intensity normalization in large multi-site neuroimaging studies. We propose RAVEL (Removal of Artificial Voxel Effect by Linear regression), a tool to remove residual technical variability after intensity normalization. As proposed by SVA and RUV [Leek and Storey, 2007, 2008, Gagnon-Bartsch and Speed, 2012], two batch effect correction tools largely used in genomics, we decompose the voxel intensities of images registered to a template into a biological component and an unwanted variation component. The unwanted variation component is estimated from a control region obtained from the cerebrospinal fluid (CSF), where intensities are known to be unassociated with disease status and other clinical covariates. We perform a singular value decomposition (SVD) of the control voxels to estimate factors of unwanted variation. We then estimate the unwanted factors using linear regression for every voxel of the brain and take the residuals as the RAVEL-corrected intensities. We assess the performance of RAVEL using T1-weighted (T1-w) images from more than 900 subjects with Alzheimer's disease (AD) and mild cognitive impairment (MCI), as well as healthy controls from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. We compare RAVEL to two intensity-normalization-only methods: histogram matching and White Stripe. We show that RAVEL performs best at improving the replicability of the brain regions that are empirically found to be most associated with AD, and that these regions are significantly more present in structures impacted by AD (hippocampus, amygdala, parahippocampal gyrus, enthorinal area, and fornix stria terminals). In addition, we show that the RAVEL-corrected intensities have the best performance in distinguishing between MCI subjects and healthy subjects using the mean hippocampal intensity (AUC = 67%), a marked improvement compared to results from intensity normalization alone (AUC = 63% and 59% for histogram matching and White Stripe, respectively). RAVEL is promising for many other imaging modalities.

Published by Elsevier Inc.

Introduction

In recent years, there has been an increase in the number of multi-site neuroimaging studies, including the Human Connectome Project (HCP), the Alzheimer's Disease Neuroimaging Initiative (ADNI), and the Australian Imaging, Biomarkers and Lifestyle Flagship Study of Aging (AIBL). In structural magnetic resonance imaging (MRI) studies, larger samples of subjects yield more power to detect structural variations in different subgroups, for example, changes in the hippocampal volume associated with Alzheimer's disease (AD) and mild cognitive impairment (MCI). However, because MRI intensities are acquired in arbitrary units, it has often been found that the differences in MRI intensities between scanning parameters and studies are larger than the biological differences observed in these images. For instance, (Shinohara et al. (2014) shows that in the ADNI and AIBL studies, which have highly standardized protocols, striking differences in the raw intensities are observed between imaging sites.

Since the raw image intensities are non-comparable across sites and between subjects, intensity normalization is paramount before performing between-subject intensity comparisons at the voxel level. While intensity normalization is not as important in other applications such as morphometry and brain volumetrics (Ashburner and Friston,

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

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

http://dx.doi.org/10.1016/j.neuroimage.2016.02.036
1053-8119/Published by Elsevier Inc.

2000; Jovicich et al., 2013), it is essential for analyzing change in intensities within an MRI volume over time (Ghassemi et al., 2015a; Sweeney et al., 2016), developing intensity-based biomarkers (Chong and Lim, 2009; Meier et al., 2007; Vardhan et al., 2014) and for regression analyses at the voxel level (Hartung et al., 2014; Smith et al., 2004). The challenge of intensity normalization has been largely addressed in the literature (Jager et al., 2006; Leung et al., 2010; Madabhushi et al., 2006; Nyúl and Udupa, 1999; Nyúl et al., 2000; Shinohara et al., 2011; Shinohara et al., 2014; Weisenfeld and Warfield, 2004), with several methods reviewed in (Shah et al., 2011). Recently, a novel intensity normalization method, called White Stripe (Shinohara et al., 2014), was developed to bring raw image intensities to a biologically interpretable intensity scale. The method applies a z-score transformation to the whole brain using parameters estimated from a latent subdistribution of normal-appearing white matter (NAWM). The use of NAWM for normalization makes the method suitable for many studies of brain abnormalities, as in the case of multiple sclerosis (MS) lesions. While the method has been shown to make the white matter (WM) comparable across subjects, it was noted that residual across-subject variability was still present in the grey matter (GM).

In this work, we investigate between-scan technical variability that is left uncorrected by intensity normalization. We show that while common intensity normalization methods successfully correct for global intensity shifts associated with scanner site, substantial between-scan technical variation remains. This technical variation can be due to scanning parameters, scanner manufacturers, scanner field strength, and other factors. We refer to any post-normalization inter-scan variation that is not biological in nature as a "scan effect."

To correct for scan effects, we propose Removal of Artificial Voxel Effect by Linear regression (RAVEL). RAVEL is a tool for removing unwanted variation present after intensity normalization. RAVEL is inspired by the batch effect correction tools SVA (Leek and Storey, 2007; Leek and Storey, 2008) and RUV (Gagnon-Bartsch and Speed, 2012) used broadly in genomics. In the analysis of gene expression and other genomic data, residual noise after intensity normalization is referred to as batch effects, because experiments are often performed in batches run on different dates. If not accounted for, batch effects have been shown to lead to spurious associations (Leek et al., 2010). To make a parallel with brain-imaging studies, batch effects are comparable to scan effects, where a single scan plays the role of a batch.

We use the linear model introduced in (Leek and Storey, 2007) to decompose the variation of the normalized intensities into a biological component of interest (variation associated with clinical covariates) and an unknown, unwanted variation component to be estimated from the data. The unwanted variation component encapsulates both technical variation and biological variation that is not of interest in the study. We register the different scans to a common template to allow the use of voxel-wise linear models, and estimate the unwanted variation component from regions of the brain that are not expected to be associated with the clinical covariates of interest. This follows the methodology of the RUV batch effect correction tool (Gagnon-Bartsch and Speed, 2012) which was later discussed in Leek (2014) for RNA sequencing. Unlike intensity-normalization methods, RAVEL utilizes all images in the study to leverage information about unwanted variability. Here, we use voxels that are consistently labelled as cerebrospinal fluid (CSF) across subjects as a control region; these voxels are not expected to be associated with disease (Luoma et al., 1993).

We evaluate the performance of RAVEL using a large subset of the ADNI database consisting of more than 900 subjects. We demonstrate our method by using the T1-weighted (T1-w) images from subjects with AD and MCI, as well as healthy controls. We follow the work of Fortin et al. (2014) to benchmark RAVEL against two intensity normalization procedures without any scan effect correction: the popular histogram matching algorithm and White Stripe. We focus on showing that RAVEL improves the replicability of the biological findings. Critically, we show that a reduction of technical variation does not result in removing biological variability. Namely, making intensity densities more similar does not necessarily improve sensitivity to biological changes; on the contrary, overmatching of distributions can result in the removal of biologically relevant signal. To show improvement in terms of biological findings, we first demonstrate that the top voxels associated with AD in the RAVEL-corrected dataset are more replicable across independent subsets of subjects. We measure the replicability of the results by randomly splitting the ADNI dataset into discovery and validation cohorts multiple times. Then, we show that the top voxels associated with AD after RAVEL correction are more enriched for brain regions known to undergo structural changes in AD. Finally, we show that the average hippocampal intensity after RAVEL correction performs better than intensity-normalized-only images in discriminating between AD patients and healthy controls, and between MCI patients and healthy controls. This shows that RAVEL-corrected T1-w intensities are more biologically meaningful than intensity-normalized-only images for group comparisons, and also potentially promising for the development of biomarkers.

Materials and methods

Study population

Our dataset consists of a subset of 917 subjects downloaded from the ADNI database (adni.loni.usc.edu). For each subject, we selected a study visit at random. We obtained 506, 184, and 227 subjects from the ADNI, ADNI-2, and ADNI-GO phases, respectively. We present summary statistics of the study population in Table 1. The selected scans were acquired at 83 different imaging sites, with a median number of 10 patients per site. The scans are also well-balanced for disease status across sites. The different scanning parameters are presented in Table A.1.

Imaging sequences and preprocessing

We considered T1-w imaging acquired on T1.5 and T3 scanners according to the ADNI standardized protocol (Jack et al., 2008). The analysis was performed in R (R Core Team, 2014), using the packages oro.nifti (Whitcher et al., 2011), fslr (Muschelli et al., 2015), ANTsR (Avants et al., 2015), and WhiteStripe (Shinohara and Muschelli, 2015).

We applied the N4 inhomogeneity correction algorithm (Tustison et al., 2010) to each image. We nonlinearly registered all T1-w images to a high-resolution T1-w image atlas (Oishi et al., 2009), using the symmetric diffeomorphic image registration algorithm (Avants et al., 2008) implemented in the ANTs suite. We used non-linear registration in order to define a brain control region aligned across subjects and to find spatially coherent nuisance patterns for removal. Compared to the population-level atlases, the advantage of using a single-subject atlas is that it contains sharp definitions of anatomical structures, many of which are highly variable across individuals and cannot be easily delineated in population atlases. We emphasize that all of the techniques proposed here can be applied directly to data in either multi- or single-subject template spaces. To remove extra-cerebral tissue from each scan, we first created a brain mask on the template using the skull-stripping algorithm FSL BET (Smith, 2002) using the fslr package and subsequently applied this resulting brain mask to all N4-corrected and registered images. The preprocessing pipeline is summarized at the top of Fig. 1.

In addition to the template brain segmentation, we performed a 3-class tissue segmentation by running the FSL FAST segmentation algorithm (Zhang et al., 2001) on the N4-corrected, registered, and skull-stripped images, for each subject separately.

RAVEL methodology

The RAVEL correction procedure adapts the linear model introduced in SVA (Leek and Storey, 2007; Leek and Storey, 2008) to intensity-