University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2015 *Paper* 43

Removing inter-subject technical variability in magnetic resonance imaging studies

Jean-Philippe Fortin* Elizabeth M. Sweeney† John Muschelli‡
Ciprian M. Crainiceanu** Russell T. Shinohara†† Alzheimer's Disease Neuroimaging Ini

*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
†Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
‡Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
**Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
††Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, rshi@upenn.edu
‡‡University of California - Los Angeles

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/upennbiostat/art43

Copyright ©2015 by the authors.

Removing inter-subject technical variability in magnetic resonance imaging studies

Jean-Philippe Fortin, Elizabeth M. Sweeney, John Muschelli, Ciprian M. Crainiceanu, Russell T. Shinohara, and Alzheimer's Disease Neuroimaging Initiative

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 (**R**emoval of **A**rtificial **V**oxel **E**ffect by **L**inear 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 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 by 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 generalizable to many imaging modalities, and shows promise for longitudinal studies. Additionally, because the choice of the control region is left to the user, RAVEL can be applied in studies of many brain disorders.

Removing inter-subject technical variability in magnetic resonance imaging studies

Jean-Philippe Fortin[1], Elizabeth M. Sweeney[1], John Muschelli[1], Ciprian M. Crainiceanu[1] and Russell T. Shinohara[2*], for the Alzheimer's Disease Neuroimaging Initiative[†]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, [2]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania

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 (**R**emoval of **A**rtificial **V**oxel **E**ffect by **L**inear 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 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 by 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 generalizable to many imaging modalities, and shows promise for longitudinal studies. Additionally, because the choice of the control region is left to the user, RAVEL can be applied in studies of many brain disorders.

*To whom correspondence should be addressed. Email: rshi@upenn.edu

[†]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

CORE Discoverable Research Archive

Hosted by The Berkeley Electronic Press

1 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.

As MRI are acquired in arbitrary units, scans are non-comparable across sites and between subjects. Therefore, intensity normalization is paramount before performing any between-subject comparisons or population-level modeling. The challenge of intensity normalization has been largely addressed in the literature [Nyúl and Udupa, 1999, Nyúl et al., 2000, Weisenfeld and Warfield, 2004, Jager et al., 2006, Madabhushi et al., 2006, Leung et al., 2010, Shinohara et al., 2011, 2014], 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, 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, the problem of batch effect correction is comparable to the problem of scan effect correction, 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

Collection of Biostatistics
Research Archive

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 therefore potentially promising for the development of biomarkers.

Although we apply RAVEL in the context of T1-w MRI of the brain, our method is generalizable to many imaging modalities. In addition, the flexibility in the choice of the control voxels makes RAVEL applicable to any disease or pathology.

2 Materials and methods

2.1 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 were also well-balanced for disease status across sites.

		Healthy	MCI	AD
	n	261	439	217
	% Female	48	36	47
	Median Age [Q_1,Q_3]	76 [72-79]	75 [70-80]	76 [71-81]
Manufacturer	% GE	42	45	47
	% Philips	11	10	14
	% Siemens	47	45	39
Field Strength	% 1.5T	85	88	88
	% 3T	15	12	12

Table 1. Summary statistics of the ADNI dataset

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

2.2 Imaging sequences and preprocessing

We considered T1-w imaging acquired on 1.5 and 3 T scanners according to the ADNI standardized protocol [Jack et al., 2008]. All 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., 2010], using the symmetric diffeomorphic image registration algorithm [Avants et al., 2008] implemented in the ANTs suite. We use non-linear registration in order to define a brain control region aligned across subjects and to find spatially coherent nuisance patterns for removal. 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 Figure 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.

2.3 RAVEL methodology

The RAVEL correction procedure adapts the linear model introduced in SVA [Leek and Storey, 2007, 2008] to intensity-normalized MRI images. The goal is to remove remaining unwanted variation in the normalized intensities by modeling the residual unwanted variation across subjects. For the optimal performance of RAVEL, we use intensities normalized with White Stripe (see Supplementary Figure S1a). We model the $m \times n$ matrix $\mathbf{V}^{WS}$ of registered and White Stripe-normalized voxel intensities, for m voxels and n subjects, as a decomposition of a biological component of interest and an unwanted component as follows:

$$\mathbf{V}^{WS} = \alpha \mathbf{1}^T + \beta \mathbf{X}^T + \gamma \mathbf{Z}^T + \mathbf{R}. \tag{1}$$

where $\alpha \mathbf{1}^T$ represents the average scan in the sample, $\beta \mathbf{X}^T$ accounts for the known clinical covariates of interest, and $\gamma \mathbf{Z}^T$ accounts for unknown, unwanted factors. We refer to $\mathbf{V}^{WS}$ as the $m \times n$ matrix of intensities, α as the $m \times 1$ vector of baseline intensities, $\mathbf{X}$ as the $n \times p$ matrix of clinical covariates, β as the $m \times p$ coefficient matrix associated with $\mathbf{X}$, $\mathbf{Z}$ as the $n \times b$ matrix of unwanted factors, γ as the $m \times b$ coefficient matrix associated with $\mathbf{Z}$, and $\mathbf{R}$ as the $m \times n$ matrix of residuals. In this model, α, β, γ and $\mathbf{Z}$ are unknown parameters that need to be estimated from the data. In the case the unwanted factors $\mathbf{Z}$ are known, the problem is reduced to simple linear regression models fit at each voxel separately.

As in RUV [Gagnon-Bartsch and Speed, 2012], we use a subset of the voxels not associated with disease to estimate the unwanted factors $\mathbf{Z}^T$. We refer to such voxels as "control voxels". An association between CSF intensities and disease status is highly unlikely [Luoma et al., 1993], and therefore CSF voxels are good candidates for inferring the unwanted component in the data. We perform a subject-specific tissue segmentation of the T1-w image and choose control voxels as voxels classified as CSF for all subjects in the study. We denote by $\mathbf{V}_c^{WS}$ the subset of the matrix of White Stripe-normalized intensities $\mathbf{V}^{\mathbf{WS}}$ confined to the control voxels. For the control voxels, Equation 1 simplifies to

$$\mathbf{V}_c^{WS} = \alpha_c \mathbf{1}^T + \gamma_c \mathbf{Z}^T + \mathbf{R}_c. \tag{2}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

because of the absence of association between the control voxels and $\mathbf{X}$. To estimate the unwanted factors $\mathbf{Z}^T$, we perform a singular value decomposition (SVD) of $\mathbf{V}_C^{WS}$ as follows

$$\mathbf{V}_c^{WS} = \mathbf{UDW}^T. \tag{3}$$

and define $\hat{\mathbf{Z}}^T$ to be the first b right-singular vectors $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_b\}$ of $\mathbf{W}$. The choice of b is discussed in the next section. Note that for $b = 1$, the estimator $\hat{\mathbf{Z}}^T$ will closely estimate the average CSF intensity for each subject. We obtain the estimates $\hat{\gamma}_i$ in Equation 1 by performing a linear regression at each voxel separately, using our estimate of $\mathbf{Z}^T$ in the equation. We define the RAVEL-corrected voxel i for subject j as

$$v_{ij}^{\text{RAVEL}} = v_{ij}^{WS} - \hat{\gamma}_i \hat{\mathbf{Z}}^T$$

where v_{ij}^{WS} is the White Stripe-normalized intensity for the i-th voxel and for the j-th subject. In summary, RAVEL aims to identify patterns of variation in the control voxels across subjects, and then assess the degree to which this variation explains the brain-wide intensity distributions. In practice, this works well if the space spanned by the unwanted factors estimated from the control voxels also spans the unwanted variation space for all voxels. A schematic of the RAVEL method is presented in Figure 1.

2.4 Estimation of the number of unwanted factors

We select the optimal number of unwanted factors b to include in Equation 1 by maximizing the discovery-validation replication rate described in section 2.7. Normalized intensities for which the top voxels associated with disease have better replication between independent experiments are more robust to technical artifacts, like site effect and differences in protocol.

Other approaches have been proposed to select b. Among others, Gagnon-Bartsch and Speed [2012] use voxels that are known to be associated with a clinical outcome to optimize b. They perform a sensitivity analysis for the parameter b, and b is chosen to optimize the number of positive control voxels that fall into the top voxels associated with the outcome. The downside of using this approach is that positive controls must be identified in advance, which is not possible for discovery studies.

Alternatively, the estimation of b could be done in an unsupervised manner by thresholding the percentage of variance explained by the first b singular vectors. This approach, which is agnostic of the outcome, can potentially provide additional safeguards against over-fitting, but could also decrease the performance of RAVEL by adding noise.

2.5 Comparison to intensity normalization methods

We compare RAVEL to two intensity normalization procedures without scan effect correction: White Stripe, as implemented in Shinohara and Muschelli [2015], and the popular histogram matching method proposed by [Nyúl and Udupa, 1999] and further refined in [Shah et al., 2011]. The histogram matching method matches the histograms of each subject to a reference population histogram using a piecewise linear transformation. We implemented the algorithm in R and we made the code available at `https://github.com/Jfortin1/RAVEL/blob/master/R/hm.R`. For better performance, we removed the background voxels before running the histogram matching algorithm. We used healthy subjects to form a reference population histogram distribution, as described in Shinohara et al. [2014].

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

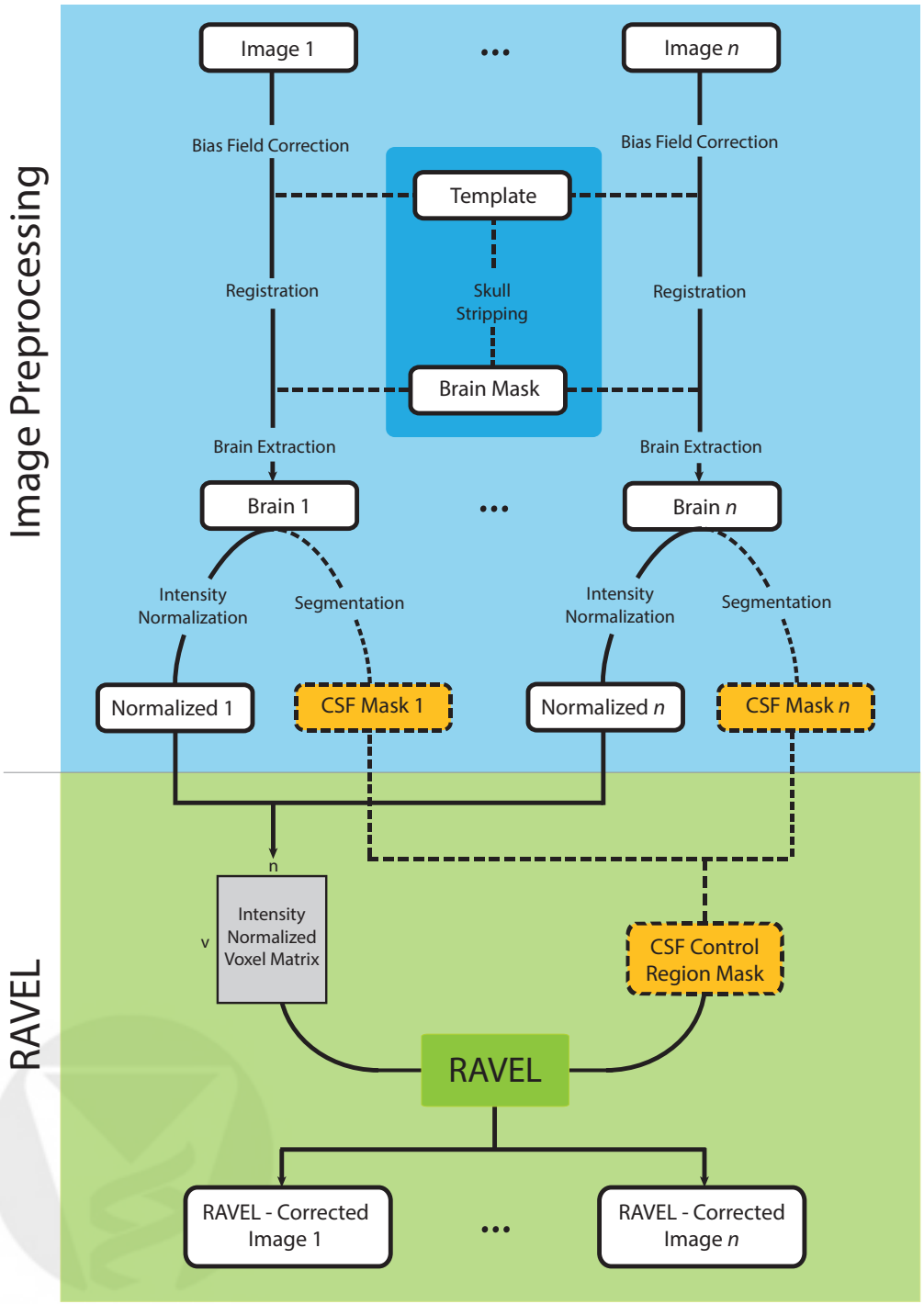


Figure 1. Schematic showing the RAVEL pipeline. The steps shown in the blue region are standard preprocessing steps that can be run in parallel. The green region shows the RAVEL algorithm.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

2.6 Identification of voxels associated with clinical covariates

Here we describe how we perform the voxel-wise analysis of the intensity distributions. For a clinical covariate x, (e.g. disease status, age, gender), we perform a simple linear regression at each voxel of the T1-w voxel intensity v on the clinical covariate x, and consider the usual t-statistic as a measure of the strength of association. We obtain a t-statistic for each of the m voxels, that is a list $\{t_1, t_2, \ldots, t_m\}$, and we rank the t-statistics in a decreasing order to get a list of rank indices $\{r_1, r_2, \ldots, r_m\}$ where r_j is such that $t_{r_j} = t_{(m-j)}$, the latter being $(m-j)$-th order statistic. For a chosen threshold q, we call the top q ranked voxels the "top voxels associated with x".

2.7 Evaluating the replicability of the top voxels associated with AD

To evaluate the replicability of the biological findings, that is the chance that an independent experiment will produce consistent results [Leek and Peng, 2015], we devised a discovery-validation cohorts scheme inspired by [Fortin et al., 2014]. The goal of the scheme is to measure replicability of the top voxels associated with the outcome of interest. If not specified otherwise, we use the disease status (AD or healthy) as the outcome of interest; we include the patients with MCI for the biomarker study described in Section 3.4 only. The discovery-validation scheme is as follows: we randomly split the full dataset into two equally sized subsets that we call discovery and validation cohorts, assigning AD and healthy patients equally between the two cohorts.

For each of the two cohorts separately, we perform a differential analysis as described in Section 2.6 to obtain two lists of ranked voxels using the differential t-statistics: $\mathbf{r}^{Dis} = \{r_1^{Dis}, r_2^{Dis}, \ldots, r_p^{Dis}\}$ and $\mathbf{r}^{Val} = \{r_1^{Val}, r_2^{Val}, \ldots, r_p^{Val}\}$, for the discovery and validation cohorts respectively. The agreement between the two lists $\mathbf{r}^{Dis}$ and $\mathbf{r}^{Val}$ serves as a measure of replicability. More specifically, we are interested in the agreement of the top-ranked voxels since those are likely more relevant and more representative of a true biological signal. For a given integer k, we look at the proportion of overlap, denoted $O(k)$, of the top k voxels from each list by

$$O(k) = \frac{|\{r_1^{Dis}, r_2^{Dis}, \ldots, r_k^{Dis}\} \cap \{r_1^{Val}, r_2^{Val}, \ldots, r_k^{Val}\}|}{k}$$

A concordance at the top (CAT) plot [Irizarry et al., 2005] is a plot showing $O(k)$ for several values of k. To quantify uncertainty of the overlap measure $O(k)$, we repeat the random discovery-validation cohort splitting one hundred times, and present the mean curve along with a 95% confidence band.

2.8 Pseudo-ROC curves and enrichment curves

In this section, we review the methodology behind pseudo-ROC curves [Bourgon, 2006] and enrichment curves. We use these curves to evaluate the performance of the different normalization and scan effect removal methods by using prior information about structural changes associated with AD. In several neuroimaging studies, prior information about a specific disease allows us to expect a set of voxels to be associated with disease. For instance, a large proportion of the hippocampus and parahippocampal voxels are known to be associated with AD and MCI. In the absence of a gold standard, these voxels can play the role of a proxy for a gold standard. We refer to these voxels as a silver standard, that is a gold standard with some contamination.

In the context of genomics, silver standards have been previously used to compare the performance of different classification methods [Bourgon, 2006] and normalization methods [Schmid et al., 2010, Fortin et al., 2014]. Bourgon [2006] show that receiver operating characteristic (ROC) curves based on a silver standard, called "pseudo-ROC curves", preserve the relative ranking of different classification methods

CORRA
Research Archive

Hosted by The Berkeley Electronic Press

with respect to ROC curves based on a gold standard. A sufficient condition for the validity of the pseudo-ROC curves ranking is that the contamination of the silver standard, with respect to the gold standard, occurs independently of the misclassification errors of the different methods compared. In the Results section, we use the t-statistics measuring the association of the voxel intensities with AD to classify voxels as either associated with AD or not. To estimate the sensitivity and specificity of each normalization method, we use voxels from 5 regions known to be associated with AD from an extensive search of the literature (see Table 2) as a silver standard.

Brain region	References
Hippocampus	Fox et al. [1996], Mori et al. [1997], Jack et al. [1999] Visser et al. [1999], Jack et al. [2000], Xu et al. [2000] Callen et al. [2001], Du et al. [2001], Bottino et al. [2002] Chételat et al. [2002], Pennanen et al. [2004], Wolf et al. [2004] Chételat et al. [2005], Ridha et al. [2006], Farrow et al. [2007] Whitwell et al. [2007], Poulin et al. [2011]
Amygdala	Scott et al. [1991, 1992], Vereecken et al. [1994] Mori et al. [1997], Callen et al. [2001], Bottino et al. [2002] Horínek et al. [2006], Farrow et al. [2007], Whitwell et al. [2007] Poulin et al. [2011], Miller et al. [2015]
Parahippocampal gyrus	Mori et al. [1997], Visser et al. [1999], Callen et al. [2001] Bottino et al. [2002], Chételat et al. [2005], Khan et al. [2014]
Enthorinal region	Gómez-Isla et al. [1996], Xu et al. [2000], Du et al. [2001] Pennanen et al. [2004], Whitwell et al. [2007] Braak and Del Tredici [2012], Khan et al. [2014]
Fornix and S. Terminalis	Callen et al. [2001], Mielke et al. [2009], Liu et al. [2011]

Table 2. Brain regions previously reported to undergo a structural change in the progression of AD.

A second approach for the benchmarking of different normalization/scan effect correction methods is to count the number of candidate voxels that fall into the list of the top k voxels associated with disease. We refer to the curve that depicts the counts for different values of k as the "enrichment curve".

3 Results

We compared RAVEL to three normalization strategies: raw image intensities (no normalization), White Stripe [Shinohara et al., 2014], and histogram matching [Shah et al., 2011]. We recall that RAVEL correction was performed on the White Stripe-normalized intensities for better performance. (see Supplementary Figure S1a).

3.1 RAVEL reduces inter-subject variability

We used a subset of the CSF intensities as control voxels to estimate factors of unwanted variation in the RAVEL model. We obtained 9869 CSF control voxels; we recall that a voxel is qualified as a CSF control if it

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

is classified as CSF for all subjects. As expected, the CSF control voxels were located primarily in the center of the ventricles (Figure 2a). Maximizing the discovery-validation replication rate explained in the Methods section, we only kept the first singular vector as the unwanted factor term $\mathbf{Z}$ in Equation 1, corresponding to $b = 1$ (see Supplementary Figure S1b). Unsurprisingly, the singular vector is highly correlated with the mean CSF intensity for each subject (correlation of 95.7%).

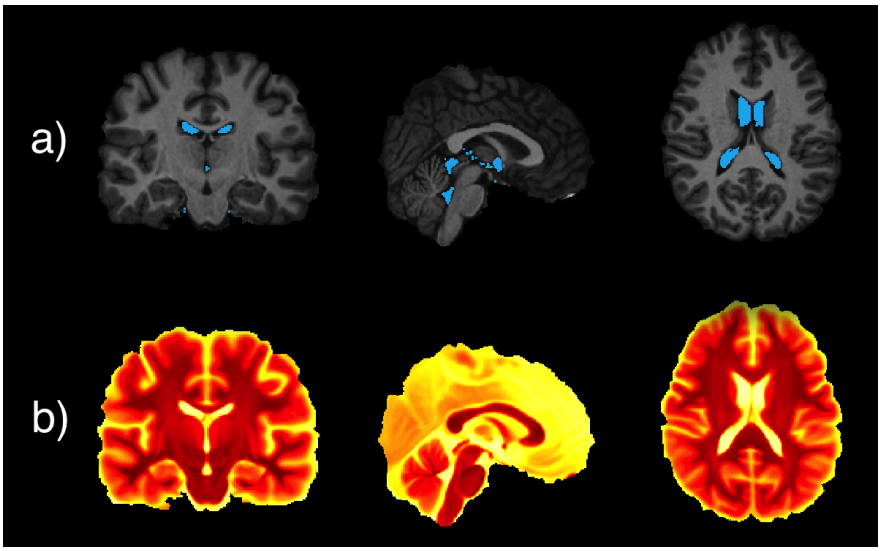


Figure 2. Estimation of technical variability using CSF control voxels. (a) The voxels selected in the RAVEL model as control voxels for CSF are shown in blue overlaid on the template; the control voxels were selected as voxels classified as CSF for every subject. (b) Heatmap of the RAVEL coefficient $\hat{\gamma}$ from Equation 1 depicted on the template, using $b = 1$ in Equation 1. The coefficient depends on the brain tissue, with a high coefficient for voxels in CSF (yellow regions), a moderate coefficient in GM (orange and lighter red) and a low coefficient for WM (darker red).

In Figure 2b, we depict the coefficient $\hat{\gamma}$ at each voxel. We notice that the distribution of $\hat{\gamma}$ varies across brain tissues, for instance darker red in WM and yellow in CSF. This shows that the method allows an unsupervised tissue-specific normalization. This prevents over-normalization in situations where the technical variation of the CSF intensities is not representative of the variation of other tissues.

In Figure 3, we show the histograms of intensities before and after RAVEL correction. The first row shows the unnormalized image histograms and the second row shows the histograms for the images normalized with White Stripe. The last row depicts the histograms for the White Stripe-normalized images with RAVEL correction. In accordance with the findings of Shinohara et al. [2014], the White Stripe-normalized images show good comparability of the WM across subjects. This can be seen by the similar WM densities centered around zero (Figure 3 second row, third column). For GM, the White Stripe densities are less clustered and show more variability, which is even more exaggerated for the CSF intensities. This shows that scaling and centering using a NAWM stripe is not enough to make GM and CSF intensities comparable across subjects. This can be explained by differential WM to GM and WM to CSF contrast ratios across images and protocols. In the third row, one can see that RAVEL substantially corrects for the extra variability in CSF and GM intensities that is not accounted for by intensity normalization. RAVEL also preserves the comparability of the WM intensities. The histograms for each tissue class cluster together well and show similar characteristics (mean, scale and range).

The main source of variation in the unnormalized images is from scanning site; on average, 67.8% of the variation in the intensities is explained by scanning site (R^2 averaged across voxels). Interestingly, we

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

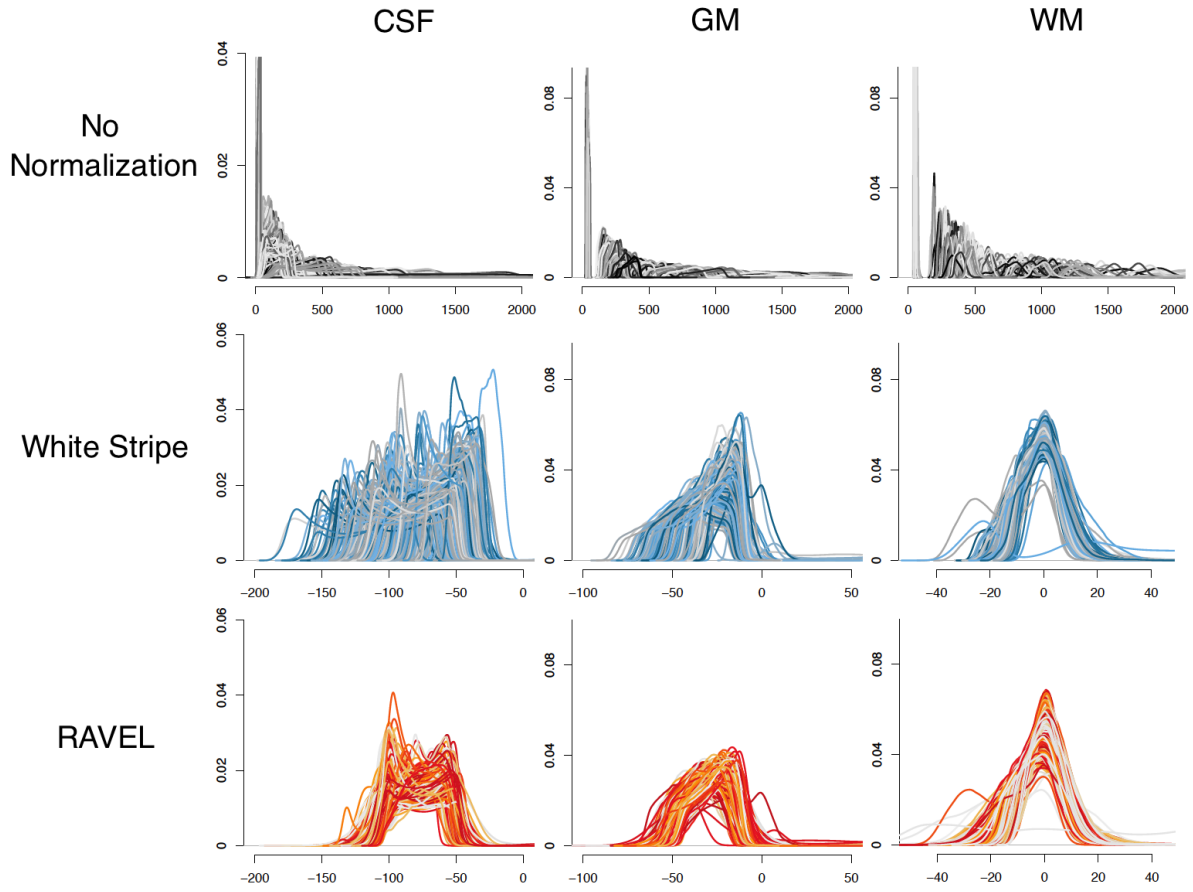


Figure 3. Effect of RAVEL on the histograms of intensities. Rows correspond to different preprocessing steps, and columns to different brain tissues. Each curve represents the corresponding histogram of intensities for one subject.

observed much less variation explained by scanning site for both intensity-normalized datasets (18% for both White Stripe and histogram matching) and for RAVEL (18%). We randomly permuted the scanning site variable 100 times and obtained a null distribution of the average R^2 with range of $[16.1\%, 16.5\%]$. This implies that after intensity normalization alone, the variability between different sites is close to the within-site variability. However, as shown in Figure 3, RAVEL removes additional technical variability in comparison to intensity normalization alone.

3.2 RAVEL improves replicability of large MRI studies

We and others have shown in the study of large epigenetic data that the ability to reduce technical variation does not necessarily lead to a better detection of features associated with the outcome of interest [Fortin et al., 2014, Dedeurwaerder et al., 2014]. A good normalization method should both reduce technical variability and enhance the replicability and robustness of biological findings. Here, we evaluate the

Collection of Biostatistics Research Archive

performance of RAVEL in terms of estimating brain regions associated with AD.

We randomly split the ADNI dataset into discovery and validation cohorts one hundred times, and we present in Figure 4b the mean CAT curves with 95% confidence bands. As expected, the unnormalized data, that is the raw images intensities, show very poor replication of the results (maximum of 0.17), while RAVEL improves replication of the findings substantially (up to 0.65 overlap in the top findings) upon intensity normalization methods alone.

The replicated voxels fall into regions that are known to be associated with AD. In Figure 4a, we show voxels associated with AD that were replicated among the top 50,000 voxels for all random splittings. No normalization led to zero voxels replicated across splittings. This is not surprising since raw image intensities are expressed in arbitrary units. White Stripe replicated 1541 voxels, while histogram matching and RAVEL replicated 3758 and 4897 voxels respectively (Figure 4c). In addition, RAVEL is the most powerful method for finding replicated voxels in the hippocampus and amygdala, two structures known to be associated with AD. The number of replicated voxels for the hippocampus are the following: 0 for no normalization, 396 for White Stripe, 1693 for histogram matching and 2405 for RAVEL. For the amygdala, we obtained the following counts: 0 for no normalization, 323 for White Stripe, 368 for histogram matching and 518 for RAVEL. The validity of those regions is discussed in the next section.

In summary, White Stripe and histogram matching, by correcting for inter-subject variability in the white matter, substantially increased the number of replicated voxels associated with AD in comparison to no normalization. RAVEL led to a 3-fold increase in the number of replicated voxels with respect to White Stripe. This was achieved by additionally modeling brain-wide unwanted variability using a CSF control region. This is consistent with the idea that while CSF is not interesting on its own with respect to disease, it can be used powerfully to distinguish signal from noise in the entire brain.

3.3 RAVEL uncovers known regions associated with AD

The discovery-validation scheme discussed above allowed us to evaluate the replicability of the top voxels associated with AD. In the current section, we aim to evaluate the validity of the results by comparing the top voxels to brain regions known to undergo a structural change in the progression of AD. Those structural changes include, among others, GM and WM atrophy, neuronal loss, amyloid senile plaques, loss of fiber tract integrity and tau lesions. In the context of AD, these changes have been described in the hippocampal formation and several parahippocampal structures. The list includes, but is not limited to, the hippocampus, the amygdala, the enthorinal cortex, the fornix, the stria terminalis and the parahippocampal gyrus. Table 2 lists several studies that have reported structural changes in these regions.

Using the template parcellation map [Oishi et al., 2010], we considered 67,983 voxels that are part of the regions listed in Table 2. These voxels represent 3.5% of the template and are potential candidates for association with AD. We use these voxels as a silver standard to evaluate the performance of the different normalization methods and RAVEL. For different values of k, we count the number of the top k voxels associated with AD that are in the silver standard, which are said to be enriched for the truth. The enrichment curves, depicted in Figure 5a (solid lines), show the number of enriched voxels for different values of k, for each normalization method. The dotted line at the bottom represents the number of voxels expected by chance only ($y = 0.035$k). To account for variability in the enrichment curves, we nonparametrically bootstrapped with replacement by subject to recalculate the top voxels associated with AD and recompute the curves. The shaded regions of Figure 5 represent bootstrapped 95% confidence bands. We observe that RAVEL discovers significantly more voxels that are truly associated with AD than the competing methods. The top voxels associated with RAVEL are also more stable than other methods, as measured by the width of the 95% confidence bands. Notably, RAVEL offers a substantial improvement with respect to intensity normalization with White Stripe alone.

Next, we obtained pseudo-ROC curves to measure the specificity and sensitivity of RAVEL for detecting

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

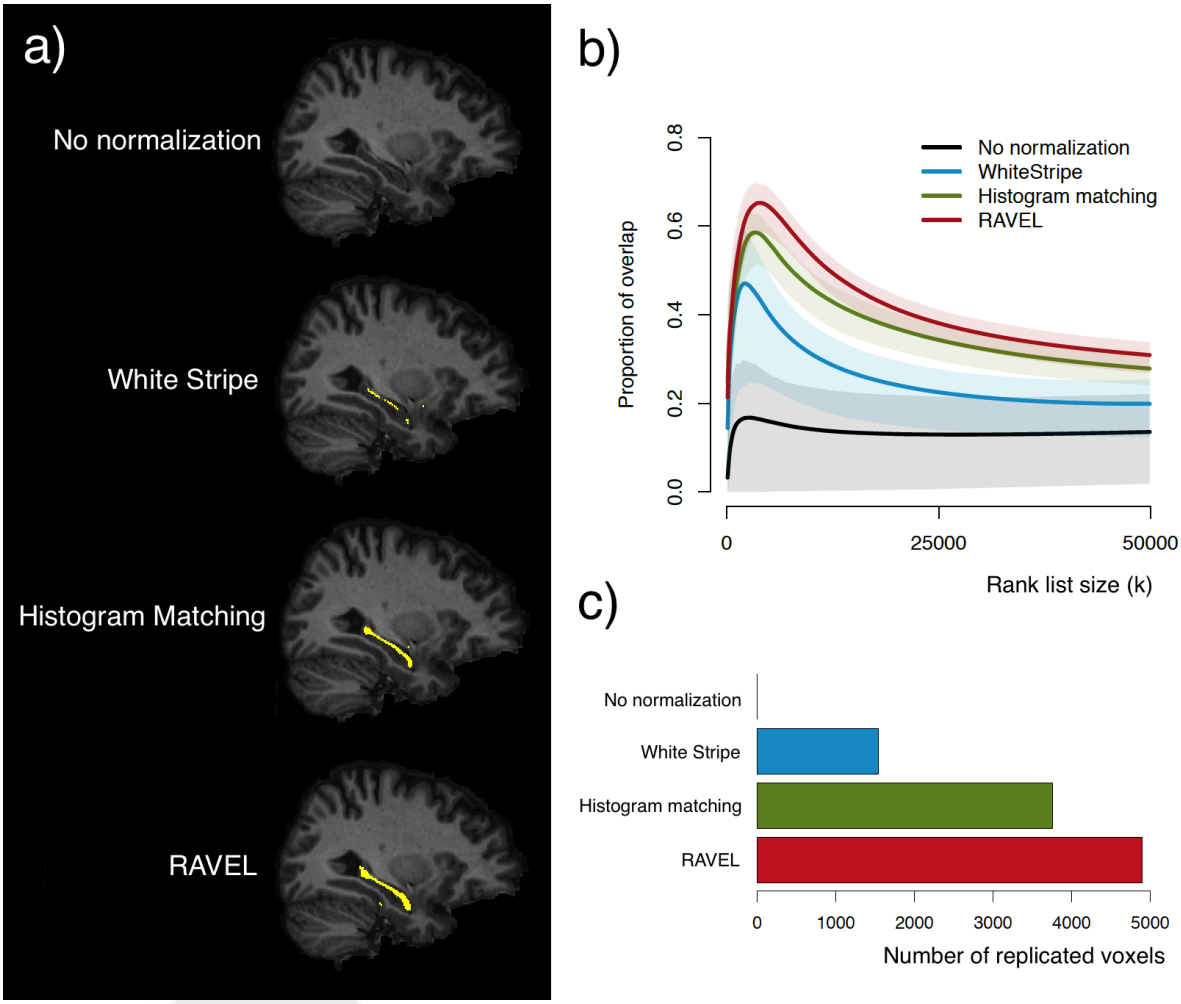


Figure 4. RAVEL improves replicability of voxels associated with AD (a) In template space, we depict in yellow the voxels that are replicated across all random splittings, from the list of the top 50,000 associated with AD. (b) Mean CAT curves for association with AD with 95% confidence bands. (c) Number of voxels replicated for each method in (a). RAVEL shows excellent performance at replicating the discovery of regions of the brain associated with AD.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

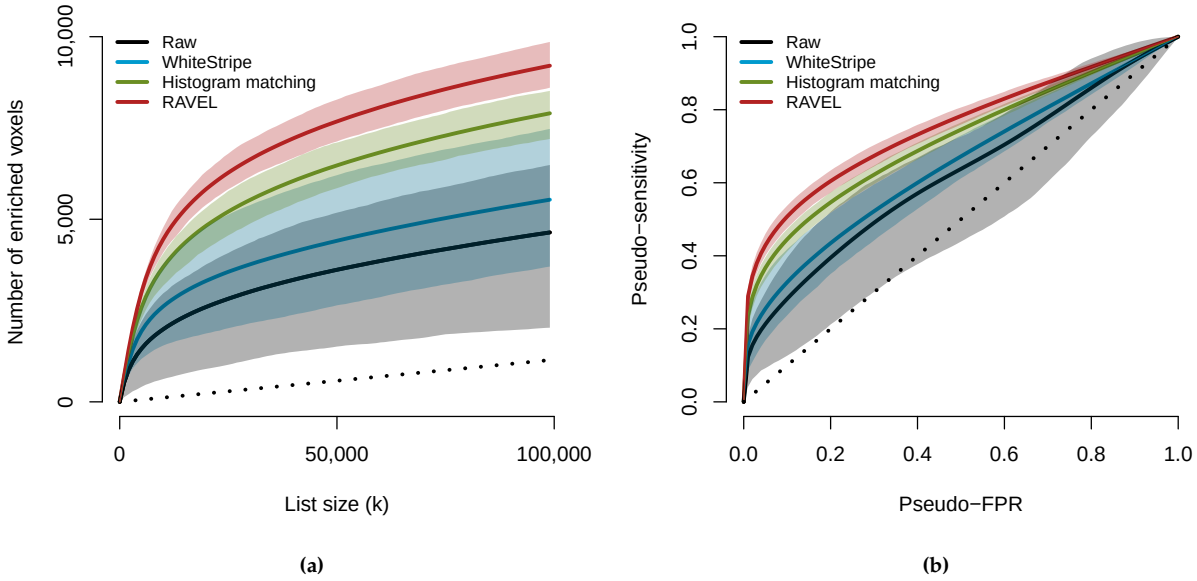


Figure 5. The top voxels associated with AD are enriched for the hippocampus and parahippocampal regions (a) For the top k voxels associated with AD (x-axis), the solid lines display the number of voxels out of the k voxels falling into five structures known to be associated with the progression of AD: the hippocampus, amygdala, enthorinal cortex, fornix and stria terminalis and parahippocampal gyrus. The dotted line represents the number of voxels expected by chance only. The shaded areas represent 95% confidence bands computed using 100 bootstrapped samples. (b) From the t-statistics measuring the association of the voxel intensities with AD, we present the pseudo-ROC curves for classifying a voxel as a member of the five regions described in (a). RAVEL shows significantly better sensitivity and specificity than the other methods for detecting hippocampus and parahippocampal changes associated with AD.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

a true association between voxel intensities and AD. In Figure 5b, we present the pseudo-ROC curves for classifying voxels as associated with AD or not, using the differential analysis (voxel-wise) t-statistics as a measure of association. The voxels from the regions listed in Table 2 are used as a silver standard. As with the enrichment curves, we present bootstrapped 95% confidence bands. RAVEL outperforms histogram matching, White Stripe and raw image intensities for the full range of specificity.

In Supplementary Figure S2, we show in template space the negative log p-value at each voxel for association between the intensities and AD status.

3.4 RAVEL-corrected intensities improve prediction of AD and MCI

We investigated the potential use of T1-w RAVEL-corrected intensities as biomarkers for disease identification and progression. We first compared the average hippocampal intensity between AD patients and healthy controls. We used the template parcellation map to identify the 9847 voxels labelled as hippocampus. Using the mean intensity of the hippocampus as a score, we classified each subject as either having AD or being healthy, thresholding the scores at different levels. The corresponding ROC curves are presented in Figure 6a. We obtained an area under the curve (AUC) of 81.7% for RAVEL (95% CI [77.6, 85.4]), as opposed to 74.9% for histogram matching ([70.4, 79.2]), 64.4% for White Stripe ([58.9, 69.0]), and 57.0% for no normalization ([52.1, 62.0]). We obtained the 95% confidence intervals by bootstrapping the samples with replacement 1000 times. Similarly, we used the average hippocampal intensity to distinguish between MCI patients and healthy controls; the corresponding ROC curves are presented in Figure 6b. We obtained an AUC of 67.3% for RAVEL (95% CI [63.1, 71.3]), as opposed to 63.4% for histogram matching ([59.6, 67.7]), 59.0% for White Stripe ([54.8, 63.4]), and 52.9% for no normalization ([48.4, 57.3]). This shows that RAVEL-corrected intensities are more representative of true biological variation than intensity-normalized intensities alone, indicating that the development of biomarkers using MRI studies in many neurological and psychiatric disorders could benefit from the RAVEL scan effect correction tool.

4 Discussion

In this work, we have presented the scan effect correction tool RAVEL, to correct for inter-scan unwanted variability in MRI studies that is present after intensity normalization. We have shown that RAVEL, applied after normalizing the intensities with White Stripe [Shinohara et al., 2014], substantially improves the replicability of the regions of the brain found to be the most associated with AD. RAVEL, inspired by the batch effect correction tools SVA and RUV Leek and Storey [2007, 2008], Gagnon-Bartsch and Speed [2012], infers the unwanted variation in the images by using regions of the brain that are not associated with disease. After registering all images to a common template, we used voxels that were labelled as CSF for all images as control voxels. We used a linear regression model at each voxel to regress out the variation in the intensities explained by variation in the control CSF voxels intensities. We used an SVD to reduce the dimensionality of the control voxels, and selected the number of components to include in the regression models by maximizing the replication rate of biological findings between independent subsets of the data.

We have shown that while common intensity normalizations remove a large part of the unwanted site effects for T1-w imaging, significant unwanted variation remains uncorrected. We encapsulated this post-normalization residual variability using the term *scan effect*. We have shown that the scan effect correction tool RAVEL successfully improves the comparability of the images in a large subset of the ADNI database by removing this extra variability. We measured the performance of RAVEL and other methods by estimating the replicability of the top voxels associated with AD in independent subsets of the ADNI dataset. To do so, we randomly divided the ADNI dataset into discovery and validation cohorts several times, and computed the top-replicated voxels for each random split. We have also shown that the top voxels associated with AD in our analysis and replicated in the discovery-validation division are more enriched for

Research Archive

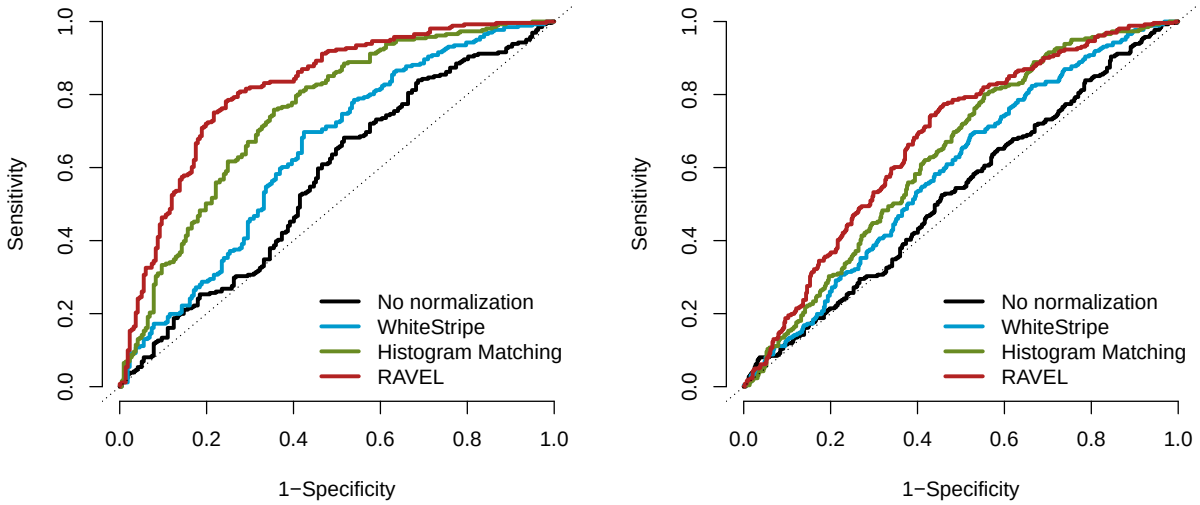


Figure 6. RAVEL improves the prediction of AD and MCI. (a) The mean hippocampus intensity was used to predict AD. The AUC is 81.7 % for RAVEL, 74.9% for histogram matching, 64.4% for White Stripe and 57.0% for no normalization, with 95% CIs [77.6, 85.4], [70.4, 79.2], [58.9, 69.0] and [52.1, 62.0] respectively. (b) The mean hippocampus intensity was used to predict MCI. The AUC is 67.3% for RAVEL, 63.4% for histogram matching, 59.0% for White Stripe and 52.9% for no normalization with 95% CIs [63.1, 71.3], [59.6, 67.7], [54.8, 63.4] and [48.4, 57.3] respectively.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

brain regions known to be associated with AD than those found using intensity-normalized data only. This shows that RAVEL is a potent method for improving the discovery of brain regions associated with disease. Finally, we have also shown that the RAVEL correction improves the prediction of AD and MCI compared to healthy controls, using the mean hippocampal intensity as a predictor. This suggests that RAVEL is a promising method that may facilitate the development of biomarkers using MRI intensities. Furthermore, with the recent emphasis on multivariate pattern analysis for biomarker development [Davatzikos et al., 2005, De Martino et al., 2008, Vemuri et al., 2008, Craddock et al., 2009, Davatzikos et al., 2011, Gaonkar and Davatzikos, 2013], RAVEL promises to produce more generalizable biomarkers that are less susceptible to biases associated with scanner and site imbalances.

The idea of using a control region of the brain which is not associated with disease is not new. In [Pujol et al., 1992, Bakshi et al., 2002, Tjoa et al., 2005, Brass et al., 2006, Neema et al., 2009], the regions of interest were divided by the mean signal intensity of a CSF region to correct for potential inter-subject variation. Shinohara et al. [2014] used a NAWM stripe to estimate a scaling and shifting parameter in their z-score normalization method. In Mejia et al. [2015], in the context of estimating quantitative T_1 maps (qT_1) from conventional MRI, the authors proposed an adaptation of the z-score normalization method by using a combination of NAWM and cerebellar gray matter (CBGM), where the NAWM was used for the scaling parameter and the CBGM was used for the shifting parameter. In [Ghassemi et al., 2015], the authors used the median GM intensity for the shifting parameter, and the difference between the median intraconal orbital fat intensity and the median GM intensity for the scaling parameter. In [Sweeney et al., 2013], the authors use the whole brain to estimate the two parameters. We note that the different versions of the z-score transformation used in [Shinohara et al., 2014, Sweeney et al., 2013, Mejia et al., 2015, Ghassemi et al., 2015] only leave room for the choice of two control regions at maximum, corresponding to the mean and scale parameters. While this improves comparability between subjects in comparison to the unnormalized intensities, as shown in Figure 4b, we have shown that RAVEL improves dramatically upon a z-score transformation only.

There are several limitations to our method. If control regions are misspecified, i.e. the region does not carry any information about the technical variability across subjects, or worse yet, if the control regions are inadvertently associated with the outcome of interest, the RAVEL correction may remove biological signals of interest. In both cases, however, cross-validation using the concordance curves from the discovery-validation scheme introduced in Section 2.7, allows the user to estimate directly the performance of RAVEL on their dataset.

Another limitation is the use of nonlinear registration to align voxels across subjects. The registration step is necessary to apply the voxel-wise linear models from Equation 1. Because patients with AD and MCI have different volumes of WM, GM and CSF in comparison with healthy controls, misregistration error might be associated with the outcome of interest. However, this is a problem inherent to any cross-subject voxel analysis, and remains an active subject of research in image analysis. While voxels that are associated with disease can be a consequence of differential misregistration, this does not change the results of the present work, as misregistered voxels should be detected by intensity normalization method, after scan effect correction. It may also be possible to approximate RAVEL corrections using mean values in reference regions; indeed, in the ADNI the mean T1-w intensity in CSF after White Stripe correction was highly correlated with the first RAVEL factor. Thus, in the case of the well-controlled ADNI protocol, adjusting by regression on the mean in CSF would yield similar results. In cases where there is more heterogeneity in acquisitions, and in imaging modalities that are more difficult to calibrate, additional RAVEL factors are likely and using the mean in the reference region may not perform well.

A first extension of the presented methodology is to precede the RAVEL correction tool by a variant of the White Stripe intensity normalization method. For instance, as used in [Sweeney et al., 2013], a whole-brain z-transformation might be used instead, where the mean and scaling parameters are estimating using all brain intensities. Subsequently, the RAVEL correction model can be applied using additional control regions, and mask erosion could be performed to improve the homogeneity of the selected control regions.

Collection of Biostatistics
Research Archive

Although we have shown the performance of RAVEL in the context of T1-w MRI of the brain, RAVEL is a promising scan effect correction tool for many imaging modalities, such as quantitative images, maps derived from diffusion tensor imaging (DTI), functional imaging and many other modalities. Furthermore, the choice of the control regions, left to the user, makes the method applicable to virtually any disease and pathology. The RAVEL software can be found at https://github.com/Jfortin1/RAVEL.

Abbreviations

AD: Alzheimer's disease; ADNI: Alzheimer's Disease Neuroimaging Initiative; ANTs: Advanced Normalization Tools; AUC: area under the curve; BET: Brain Extraction Tool; CAT: concordance at the top; CBGM: cerebellar gray matter; CSF: cerebrospinal fluid; DTI: diffusion tensor imaging; FAST: FMRIB's Automated Segmentation Tool; FMRIB: Oxford Centre for Functional MRI of the Brain; FSL: FMRIB Software Library; GM: grey matter; MCI: mild cognitive impairment; MRI: magnetic resonance imaging; NAWM: normal-appearing white matter; NIFTI: The Neuroimaging Informatics Technology Initiative; RAVEL: Removal of Artificial Voxel Effect by Linear regression; ROC: receiver operating characteristic; RUV: removing unwanted variation; SVA: surrogate variable analysis; SVD: singular value decomposition; T1-w: T1-weighted; WM: white matter; WMPM: white matter parcellation map.

Competing interests

The authors declare that they have no competing interests.

Authors contributions

JPF and RTS developed the method. JPF analyzed the data and wrote the software. RTS supervised the study. JPF and RTS wrote the manuscript with comments from EMS, JM and CMC. All authors read and approved the final manuscript.

Funding

The research of Shinohara and Crainiceanu was supported by Award Numbers R01NS085211 from the National Institute of Neurological Disorders and Stroke. RTS is partially supported by R01EB017255 from the National Institute for Biomedical Imaging and Bioengineering.

Acknowledgements

We would like to thank Paul Yushkevich and Sandhitsu Das for insightful discussions concerning biomarkers in AD.

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60million, 5-year public private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California San Francisco. ADNI is the result of the efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

References

B B Avants, C L Epstein, M Grossman, and J C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal*, 12 (1):26–41, Feb 2008. doi:10.1016/j.media.2007.06.004.

Brian B Avants, Benjamin m Kandel, Jeff T Duda, and Philip A Cook. Antsr: Ants in r. 2015. URL https://github.com/stnava/ANTsR.

Rohit Bakshi, Ralph H B Benedict, Robert A Bermel, Shelton D Caruthers, Srinivas R Puli, Christopher W Tjoa, Andrew J Fabiano, and Lawrence Jacobs. T2 hypointensity in the deep gray matter of patients with multiple sclerosis: a quantitative magnetic resonance imaging study. *Arch Neurol*, 59(1):62–8, Jan 2002.

Cássio M C Bottino, Cláudio C Castro, Regina L E Gomes, Carlos A Buchpiguel, Renato L Marchetti, and Mário R Louzã Neto. Volumetric mri measurements can differentiate alzheimer's disease, mild cognitive impairment, and normal aging. *Int Psychogeriatr*, 14(1):59–72, Mar 2002.

Richard Walter Bourgon. *Chromatin immunoprecipitation and high-density tiling microarrays: a generative model, methods for analysis, and methodology assessment in the absence of a "gold standard"*. PhD thesis, University of California, Berkeley, 2006.

Heiko Braak and Kelly Del Tredici. Alzheimer's disease: pathogenesis and prevention. *Alzheimers Dement*, 8(3):227–33, May 2012. doi:10.1016/j.jalz.2012.01.011.

S D Brass, R H B Benedict, B Weinstock-Guttman, F Munschauer, and R Bakshi. Cognitive impairment is associated with subcortical magnetic resonance imaging grey matter t2 hypointensity in multiple sclerosis. *Mult Scler*, 12(4):437–44, Aug 2006.

D J Callen, S E Black, F Gao, C B Caldwell, and J P Szalai. Beyond the hippocampus: Mri volumetry confirms widespread limbic atrophy in ad. *Neurology*, 57(9):1669–74, Nov 2001.

G Chételat, B Landeau, F Eustache, F Mézenge, F Viader, V de la Sayette, B Desgranges, and J-C Baron. Using voxel-based morphometry to map the structural changes associated with rapid conversion in mci: a longitudinal mri study. *Neuroimage*, 27(4):934–46, Oct 2005. doi:10.1016/j.neuroimage.2005.05.015.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Gaël Chételat, Béatrice Desgranges, Vincent De La Sayette, Fausto Viader, Francis Eustache, and Jean-Claude Baron. Mapping gray matter loss with voxel-based morphometry in mild cognitive impairment. *Neuroreport*, 13(15):1939–43, Oct 2002.

R Cameron Craddock, Paul E Holtzheimer, 3rd, Xiaoping P Hu, and Helen S Mayberg. Disease state prediction from resting state functional connectivity. *Magn Reson Med*, 62(6):1619–28, Dec 2009. doi:10.1002/mrm.22159.

C Davatzikos, K Ruparel, Y Fan, D G Shen, M Acharyya, J W Loughead, R C Gur, and D D Langleben. Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage*, 28(3):663–8, Nov 2005. doi:10.1016/j.neuroimage.2005.08.009.

Christos Davatzikos, Priyanka Bhatt, Leslie M Shaw, Kayhan N Batmanghelich, and John Q Trojanowski. Prediction of mci to ad conversion, via mri, csf biomarkers, and pattern classification. *Neurobiol Aging*, 32 (12):2322.e19–27, Dec 2011. doi:10.1016/j.neurobiolaging.2010.05.023.

Federico De Martino, Giancarlo Valente, Noël Staeren, John Ashburner, Rainer Goebel, and Elia Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns. *Neuroimage*, 43(1):44–58, Oct 2008.

Sarah Dedeurwaerder, Matthieu Defrance, Martin Bizet, Emilie Calonne, Gianluca Bontempi, and François Fuks. A comprehensive overview of infinium humanmethylation450 data processing. *Brief Bioinform*, 15 (6):929–41, Nov 2014. doi:10.1093/bib/bbt054.

A T Du, N Schuff, D Amend, M P Laakso, Y Y Hsu, W J Jagust, K Yaffe, J H Kramer, B Reed, D Norman, H C Chui, and M W Weiner. Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and alzheimer's disease. *J Neurol Neurosurg Psychiatry*, 71(4):441–7, Oct 2001.

Tom F D Farrow, Subha N Thiyagesh, Iain D Wilkinson, Randolph W Parks, Leanne Ingram, and Peter W R Woodruff. Fronto-temporal-lobe atrophy in early-stage alzheimer's disease identified using an improved detection methodology. *Psychiatry Res*, 155(1):11–9, May 2007. doi:10.1016/j.pscychresns.2006.12.013.

Jean-Philippe Fortin, Aurelie Labbe, Mathieu Lemire, Brent Zanke, Thomas Hudson, Elana Fertig, Celia Greenwood, and Kasper D Hansen. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology*, 15(11):503, 2014. doi:10.1186/s13059-014-0503-2.

N C Fox, E K Warrington, P A Freeborough, P Hartikainen, A M Kennedy, J M Stevens, and M N Rossor. Presymptomatic hippocampal atrophy in alzheimer's disease. a longitudinal mri study. *Brain*, 119 (Pt 6): 2001–7, Dec 1996.

J A Gagnon-Bartsch and T P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012. doi:10.1093/biostatistics/kxr034.

Bilwaj Gaonkar and Christos Davatzikos. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. *Neuroimage*, 78:270–83, Sep 2013. doi:10.1016/j.neuroimage.2013.03.066.

Rezwan Ghassemi, Robert Brown, Sridar Narayanan, Brenda Banwell, Kunio Nakamura, and Douglas L Arnold. Normalization of white matter intensity on t1-weighted images of patients with acquired central nervous system demyelination. 25(2):184–90, 2015. doi:10.1111/jon.12129.

T Gómez-Isla, J L Price, D W McKeel, Jr, J C Morris, J H Growdon, and B T Hyman. Profound loss of layer ii entorhinal cortex neurons occurs in very mild alzheimer's disease. *J Neurosci*, 16(14):4491–500, Jul 1996.

D Horínek, P Petrovický, J Hort, J Krásenský, J Brabec, M Bojar, M Vanecková, and Z Seidl. Amygdalar volume and psychiatric symptoms in alzheimer's disease: an mri analysis. *Acta Neurol Scand*, 113(1): 40–5, Jan 2006. doi:10.1111/j.1600-0404.2006.00540.x.

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Rafael A Irizarry, Daniel Warren, Forrest Spencer, Irene F Kim, Shyam Biswal, Bryan C Frank, Edward Gabrielson, Joe G N Garcia, Joel Geoghegan, Gregory Germino, Constance Griffin, Sara C Hilmer, Eric Hoffman, Anne E Jedlicka, Ernest Kawasaki, Francisco Martínez-Murillo, Laura Morsberger, Hannah Lee, David Petersen, John Quackenbush, Alan Scott, Michael Wilson, Yanqin Yang, Shui Qing Ye, and Wayne Yu. Multiple-laboratory comparison of microarray platforms. *Nature Methods*, 2(5):345–50, 2005. doi:10.1038/nmeth756.

C R Jack, Jr, R C Petersen, Y C Xu, P C O'Brien, G E Smith, R J Ivnik, B F Boeve, S C Waring, E G Tangalos, and E Kokmen. Prediction of ad with mri-based hippocampal volume in mild cognitive impairment. *Neurology*, 52(7):1397–403, Apr 1999.

C R Jack, Jr, R C Petersen, Y Xu, P C O'Brien, G E Smith, R J Ivnik, B F Boeve, E G Tangalos, and E Kokmen. Rates of hippocampal atrophy correlate with change in clinical status in aging and ad. *Neurology*, 55(4): 484–89, Aug 2000.

Clifford R Jack, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

F Jager, Y Deuerling-Zheng, B Frericks, F Wacker, and H Hornegger. A new method for mri intensity standardization with application to lesion detection in the brain. *Vision Modeling and Visualization*, pages 269–276, 2006.

Usman A Khan, Li Liu, Frank A Provenzano, Diego E Berman, Caterina P Profaci, Richard Sloan, Richard Mayeux, Karen E Duff, and Scott A Small. Molecular drivers and cortical spread of lateral entorhinal cortex dysfunction in preclinical alzheimer's disease. *Nat Neurosci*, 17(2):304–11, Feb 2014. doi:10.1038/nn.3606.

Jeffrey T Leek. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res*, 42(21), Dec 2014.

Jeffrey T Leek and Roger D Peng. Opinion: Reproducible research can still be wrong: adopting a prevention approach. *Proc Natl Acad Sci U S A*, 112(6):1645–6, Feb 2015. doi:10.1073/pnas.1421412111.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):1724–1735, 2007. doi:10.1371/journal.pgen.0030161.

Jeffrey T Leek and John D Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008. doi:10.1073/pnas.0808709105.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010. doi:10.1038/nrg2825.

Kelvin K Leung, Matthew J Clarkson, Jonathan W Bartlett, Shona Clegg, Clifford R Jack, Jr, Michael W Weiner, Nick C Fox, Sébastien Ourselin, and Alzheimer's Disease Neuroimaging Initiative. Robust atrophy rate measurement in alzheimer's disease using multi-site serial mri: tissue-specific intensity normalization and parameter selection. *Neuroimage*, 50(2):516–23, Apr 2010. doi:10.1016/j.neuroimage.2009.12.059.

Yawu Liu, Gabriela Spulber, Kimmo K Lehtimäki, Mervi Könönen, Ilona Hallikainen, Heidi Gröhn, Miia Kivipelto, Merja Hallikainen, Ritva Vanninen, and Hilkka Soininen. Diffusion tensor imaging and tract-based spatial statistics in alzheimer's disease and mild cognitive impairment. *Neurobiol Aging*, 32(9): 1558–71, 2011. doi:10.1016/j.neurobiolaging.2009.10.006.

K Luoma, R Raininko, P Nummi, and R Luukkonen. Is the signal intensity of cerebrospinal fluid constant? intensity measurements with high and low field magnetic resonance imagers. *Magn Reson Imaging*, 11(4): 549–55, 1993.

Collection of Biostatistics
Research Archive

Anant Madabhushi, Jayaram K Udupa, and Gul Moonis. Comparing mr image intensity standardization against tissue characterizability of magnetization transfer ratio imaging. *J Magn Reson Imaging*, 24(3): 667–75, Sep 2006. doi:10.1002/jmri.20658.

Amanda Mejia, Elizabeth M Sweeney, Blake Dewey, Govind Nair, Pascal Sati, Colin Shea, Daniel S Reich, , and Russell T Shinohara. Statistical estimation of t1 relaxation time using conventional magnetic resonance imaging. *UPenn Biostatistics Working Papers*, Working Paper 37, 2015.

M M Mielke, N A Kozauer, K C G Chan, M George, J Toroney, M Zerrate, K Bandeen-Roche, M-C Wang, P Vanzijl, J J Pekar, S Mori, C G Lyketsos, and M Albert. Regionally-specific diffusion tensor imaging in mild cognitive impairment and alzheimer's disease. *Neuroimage*, 46(1):47–55, May 2009. doi:10.1016/j.neuroimage.2009.01.054.

Michael I Miller, Laurent Younes, J Tilak Ratnanather, Timothy Brown, Huong Trinh, David S Lee, Daniel Tward, Pamela B Mahon, Susumu Mori, Marilyn Albert, and BIOCARD Research Team. Amygdalar atrophy in symptomatic alzheimer's disease based on diffeomorphometry: the biocard cohort. *Neurobiol Aging*, 36 Suppl 1:S3–S10, Jan 2015.

E Mori, Y Yoneda, H Yamashita, N Hirono, M Ikeda, and A Yamadori. Medial temporal structures relate to memory impairment in alzheimer's disease: an mri volumetric study. *J Neurol Neurosurg Psychiatry*, 63 (2):214–21, Aug 1997.

John Muschelli, Elizabeth M Sweeney, Martin A Lindquist, and Ciprian M Crainiceanu. fslr: Connecting the fsl software with r. *The R Journal*, 7(1):163–175, Feb 2015.

Mohit Neema, Ashish Arora, Brian C Healy, Zachary D Guss, Steven D Brass, Yang Duan, Guy J Buckle, Bonnie I Glanz, Lynn Stazzone, Samia J Khoury, Howard L Weiner, Charles R G Guttmann, and Rohit Bakshi. Deep gray matter involvement on brain mri scans is associated with clinical progression in multiple sclerosis. *J Neuroimaging*, 19(1):3–8, Jan 2009. doi:10.1111/j.1552-6569.2008.00296.x.

L G Nyúl and J K Udupa. On standardizing the mr image intensity scale. *Magn Reson Med*, 42(6):1072–81, Dec 1999.

L G Nyúl, J K Udupa, and X Zhang. New variants of a method of mri scale standardization. *IEEE Trans Med Imaging*, 19(2):143–50, Feb 2000. doi:10.1109/42.836373.

Kenichi Oishi, AndreiaV Faria, and Susumu Mori. Jhu-mni-ss atlas. 05 2010.

Corina Pennanen, Miia Kivipelto, Susanna Tuomainen, Päivi Hartikainen, Tuomo Hänninen, Mikko P Laakso, Merja Hallikainen, Matti Vanhanen, Aulikki Nissinen, Eeva-Liisa Helkala, Pauli Vainio, Ritva Vanninen, Kaarina Partanen, and Hilkka Soininen. Hippocampus and entorhinal cortex in mild cognitive impairment and early ad. *Neurobiol Aging*, 25(3):303–10, Mar 2004. doi:10.1016/S0197-4580(03)00084-8.

Stéphane P Poulin, Rebecca Dautoff, John C Morris, Lisa Feldman Barrett, Bradford C Dickerson, and Alzheimer's Disease Neuroimaging Initiative. Amygdala atrophy is prominent in early alzheimer's disease and relates to symptom severity. *Psychiatry Res*, 194(1):7–13, Oct 2011. doi:10.1016/j.pscychresns.2011.06.014.

Jesús Pujol, Carme Junqué, Pere Vendrell, Josep M Grau, Josep L Martí-Vilalta, Carme Olivé, and Jaume Gili. Biological significance of iron-related magnetic resonance imaging changes in the brain. *Archives of neurology*, 49(7):711–717, 1992.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

Basil H Ridha, Josephine Barnes, Jonathan W Bartlett, Alison Godbolt, Tracey Pepple, Martin N Rossor, and Nick C Fox. Tracking atrophy progression in familial alzheimer's disease: a serial mri study. *Lancet Neurol*, 5(10):828–34, Oct 2006. doi:10.1016/S1474-4422(06)70550-6.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Ramona Schmid, Patrick Baum, Carina Ittrich, Katrin Fundel-Clemens, Wolfgang Huber, Benedikt Brors, Roland Eils, Andreas Weith, Detlev Mennerich, and Karsten Quast. Comparison of normalization methods for illumina beadchip humanht-12 v3. *BMC Genomics*, 11:349, 2010. doi:10.1186/1471-2164-11-349.

S A Scott, S T DeKosky, and S W Scheff. Volumetric atrophy of the amygdala in alzheimer's disease: quantitative serial reconstruction. *Neurology*, 41(3):351–6, Mar 1991.

S A Scott, S T DeKosky, D L Sparks, C A Knox, and S W Scheff. Amygdala cell loss and atrophy in alzheimer's disease. *Ann Neurol*, 32(4):555–63, Oct 1992. doi:10.1002/ana.410320412.

Mohak Shah, Yiming Xiao, Nagesh Subbanna, Simon Francis, Douglas L Arnold, D Louis Collins, and Tal Arbel. Evaluating intensity normalization on mris of human brain with multiple sclerosis. *Med Image Anal*, 15(2):267–82, Apr 2011. doi:10.1016/j.media.2010.12.003.

Russell T Shinohara, Ciprian M Crainiceanu, Brian S Caffo, María Inés Gaitán, and Daniel S Reich. Population-wide principal component-based quantification of blood-brain-barrier dynamics in multiple sclerosis. *Neuroimage*, 57(4):1430–46, Aug 2011. doi:10.1016/j.neuroimage.2011.05.038.

Russell T Shinohara, Elizabeth M Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J Mateen, Peter A Calabresi, Samson Jarso, Dzung L Pham, Daniel S Reich, Ciprian M Crainiceanu, Australian Imaging Biomarkers Lifestyle Flagship Study of Ageing, and Alzheimer's Disease Neuroimaging Initiative. Statistical normalization techniques for magnetic resonance imaging. *Neuroimage Clin*, 6:9–19, 2014. doi:10.1016/j.nicl.2014.08.008.

Taki Shinohara and John Muschelli. Whitestripe: White matter normalization for magnetic resonance images using whitestripe. 2015. URL `https://cran.r-project.org/web/packages/WhiteStripe/index.html`.

J G Sled, A P Zijdenbos, and A C Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Trans Med Imaging*, 17(1):87–97, Feb 1998. doi:10.1109/42.668698.

Stephen M Smith. Fast robust automated brain extraction. *Hum Brain Mapp*, 17(3):143–55, Nov 2002. doi:10.1002/hbm.10062.

Elizabeth M Sweeney, Russell T Shinohara, Navid Shiee, Farrah J Mateen, Avni A Chudgar, Jennifer L Cuzzocreo, Peter A Calabresi, Dzung L Pham, Daniel S Reich, and Ciprian M Crainiceanu. Oasis is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in mri. *Neuroimage Clin*, 2:402–13, 2013.

C W Tjoa, R H B Benedict, B Weinstock-Guttman, A J Fabiano, and R Bakshi. Mri t2 hypointensity of the dentate nucleus is related to ambulatory impairment in multiple sclerosis. *J Neurol Sci*, 234(1-2):17–24, Jul 2005. doi:10.1016/j.jns.2005.02.009.

Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE Trans Med Imaging*, 29(6):1310–20, Jun 2010. doi:10.1109/TMI.2010.2046908.

Prashanthi Vemuri, Jeffrey L Gunter, Matthew L Senjem, Jennifer L Whitwell, Kejal Kantarci, David S Knopman, Bradley F Boeve, Ronald C Petersen, and Clifford R Jack, Jr. Alzheimer's disease diagnosis in individual subjects using structural mr images: validation studies. *Neuroimage*, 39(3):1186–97, Feb 2008. doi:10.1016/j.neuroimage.2007.09.073.

T H Vereecken, O J Vogels, and R Nieuwenhuys. Neuron loss and shrinkage in the amygdala in alzheimer's disease. *Neurobiol Aging*, 15(1):45–54, 1994.

P J Visser, P Scheltens, F R Verhey, B Schmand, L J Launer, J Jolles, and C Jonker. Medial temporal lobe atrophy and memory dysfunction as predictors for dementia in subjects with mild cognitive impairment. *J Neurol*, 246(6):477–85, Jun 1999.

Collection of Biostatistics Research Archive

N L Weisenfeld and S K Warfield. Normalization of joint image-intensity statistics in mri using the kullback–leibler divergence. *Biomedical Imaging: Nano to Macro, 2004 IEEE International Symposium on (101–104IEEE)*, 2004.

Brandon Whitcher, Volker J. Schmid, and Andrew Thornton. Working with the DICOM and NIfTI data standards in R. *Journal of Statistical Software*, 44(6):1–28, 2011. URL http://www.jstatsoft.org/v44/i06/.

Jennifer L Whitwell, Scott A Przybelski, Stephen D Weigand, David S Knopman, Bradley F Boeve, Ronald C Petersen, and Clifford R Jack, Jr. 3d maps from multiple mri illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to alzheimer's disease. *Brain*, 130(Pt 7):1777–86, Jul 2007. doi:10.1093/brain/awm112.

Henrike Wolf, Anke Hensel, Frithjof Kruggel, Steffi G Riedel-Heller, Thomas Arendt, Lars-Olof Wahlund, and Hermann-Josef Gertz. Structural correlates of mild cognitive impairment. *Neurobiol Aging*, 25(7): 913–24, Aug 2004. doi:10.1016/j.neurobiolaging.2003.08.006.

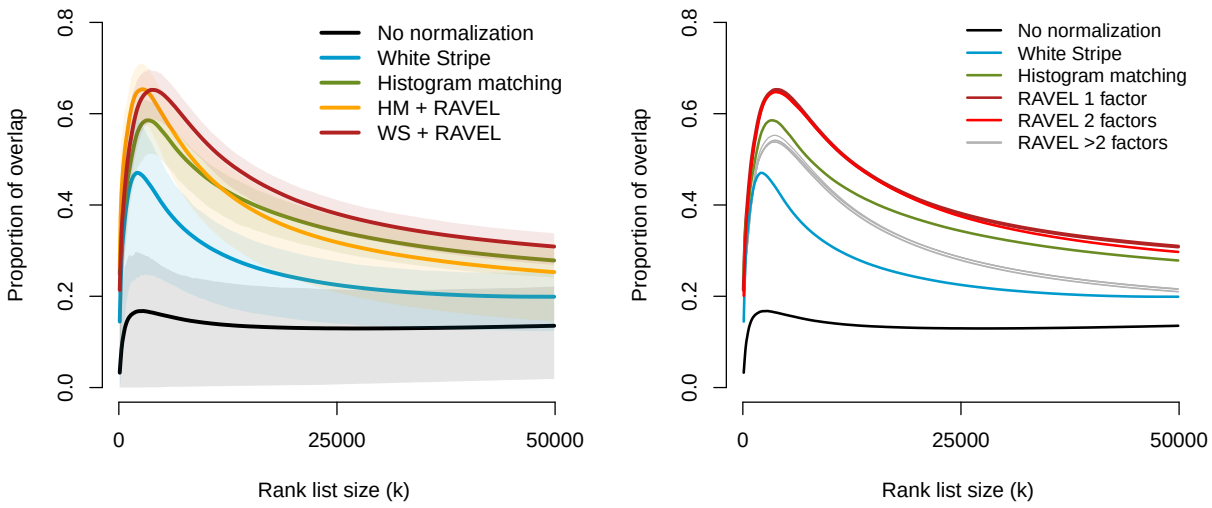
Y Xu, C R Jack, Jr, P C O'Brien, E Kokmen, G E Smith, R J Ivnik, B F Boeve, R G Tangalos, and R C Petersen. Usefulness of mri measures of entorhinal cortex versus hippocampus in ad. *Neurology*, 54(9):1760–7, May 2000.

Y Zhang, M Brady, and S Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging*, 20(1):45–57, Jan 2001. doi:10.1109/42.906424.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

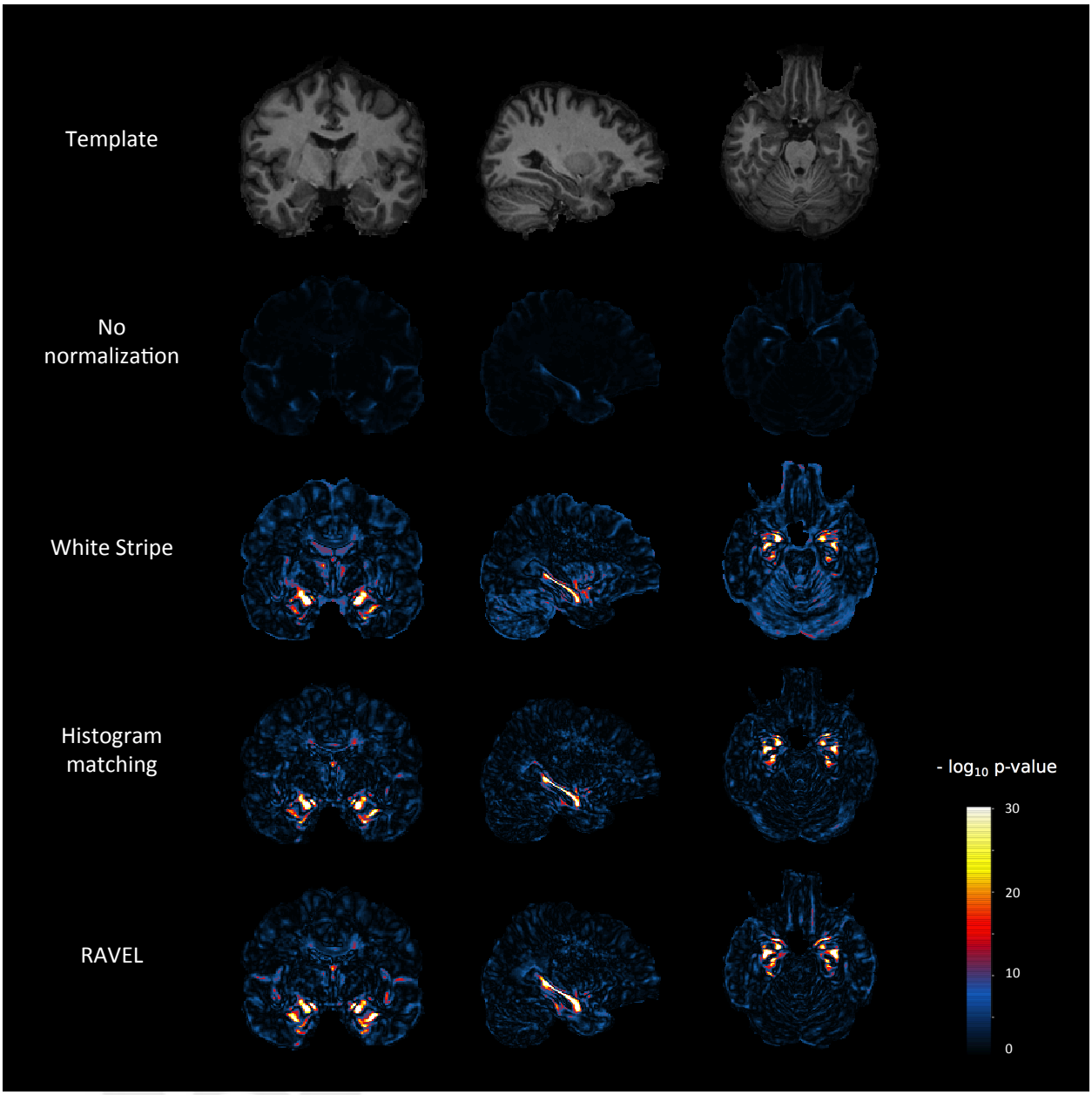
Hosted by The Berkeley Electronic Press

Appendix



Supplementary Figure S1. CAT plots (a) Like Figure 4b, but distinguish between RAVEL run on intensities normalized by White Stripe (default) and RAVEL run on intensities normalized by histogram matching. (b) Like Figure 4b, but for different numbers of unwanted factors in the RAVEL model. The pink line is for RAVEL with 2 factors, and the grey lines represent RAVEL with 3 to 15 factors. We can observe that the choice of 1 or 2 factors in the RAVEL model optimizes the replication of the voxels associated with AD.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive



Supplementary Figure S2. Voxel-level p-value maps from AD vs. healthy patient differential analysis At each voxel, we computed a t-statistic for testing a difference in intensities between AD and healthy patients. For each normalization method, we report the negative log p-values from the t-test. We include at the top of the figure the template for anatomical reference.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press