A Comparison of Publicly Available Linear MRI Stereotaxic Registration Techniques

Mahsa Dadar[1] (MS) mahsa.dadar@mail.mcgill.ca

Vladimir S. Fonov[1] (PhD) vladimir.fonov@mcgill.ca

D. Louis Collins[1] (PhD) louis.collins@mcgill.ca

Alzheimer's Disease Neuroimaging Initiative[1]

1. Image Processing Laboratory, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada.

Corresponding Author Information:

D. Louis Collins, Magnetic Resonance Imaging (MRI), Montreal Neurological Institute, 3801 University Street, Room WB315, Montréal, QC, H3A 2B4 Email: louis.collins@mcgill.ca

Tel: +1-514-398-4227

Fax: +1-514-398-2975

[1] Data used in preparation of this article were obtained from:

1) The Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.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 analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-ontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

2) The Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org

3) The Human Connectome Project (HCP; Principal Investigators: Bruce Rosen, M.D., Ph.D. Arthur W. Toga, Ph.D., Van J. Weeden, MD). HCP funding was provided by the National Institute of Dental and Craniofacial Research (NIDCR), the National Institute of Mental Health (NIMH), and the National Institute of Neurological Disorders and Stroke (NINDS). HCP data are disseminated by the Laboratory of Neuro Imaging at the University of Southern California.

4) The Pre-symptomatic Evaluation of Novel or Experimental Treatments for Alzheimer's Disease (PREVENT-AD) program (http://www.prevent-alzheimer.ca), data release 3.0 (November 30, 2016). As such, the investigators of the PREVENT-AD program contributed to the design and implementation of PREVENT- AD and/or provided data but did not participate in analysis or writing of this report. A complete listing of PREVENT-AD Research Group can be found in the PREVENT-AD database: https://preventad.loris.ca/acknowledgements/acknowledgements.php?date=[2016-11-30].

adfa, p. 1, 2011.

Abstract.

Introduction: Linear registration to a standard space is one of the major steps in processing and analyzing magnetic resonance images (MRIs) of the brain. Here we present an overview of linear stereotaxic MRI registration and compare the performance of 5 publicly available and extensively used linear registration techniques in medical image analysis.

Methods: A set of 9693 T1-weighted MR images were obtained for testing from 4 datasets: ADNI, PREVENT-AD, PPMI, and HCP, two of which have multi-center and multi-scanner data and three of which have longitudinal data. Each individual native image was linearly registered to the MNI ICBM152 average template using five versions of MRITOTAL from MINC tools, FLIRT from FSL, two versions of Elastix, spm_affreg from SPM, and ANTs linear registration techniques. Quality control (QC) images were generated from the registered volumes and viewed by an expert rater to assess the quality of the registrations. The QC image contained 60 sub-images (20 of each of axial, sagittal, and coronal views at different levels throughout the brain) overlaid with contours of the ICBM152 template, enabling the expert rater to label the registration as acceptable or unacceptable. The performance of the registration techniques was then compared across different datasets. In addition, the effect of image noise, intensity non-uniformity, age, head size, and atrophy on the performance of the techniques was investigated by comparing differences between age, scaling factor, ventricle volume, brain volume, and white matter hyperintensity (WMH) volumes between passed and failed cases for each method.

Results: The average registration failure rate among all datasets was 27.41%, 27.14%, 12.74%, 13.03%, 0.44% for the five versions of MRITOTAL techniques, 8.87% for ANTs, 11.11% for FSL, 12.35% for Elastix Affine, 24.40% for Elastix Similarity, and 30.66% for SPM. There were significant effects of signal to noise ratio, image intensity non-uniformity estimates, as well as age, head size, and atrophy related changes between passed and failed registrations.

Conclusion: Our experiments show that the Revised BestLinReg had the best performance among the evaluated registration techniques while all techniques performed worse for images with higher levels of noise and non-uniformity as well as atrophy related changes.

Keywords: MRI, linear registration, quality control

1. Introduction

Linear MR image registration, i.e. geometrically aligning two 3D images (source and target) from the same modality, different modalities, visits or subjects is a fundamental task in many aspects of medical image analysis. Image registration is used in many different areas of medicine such as multi-modality fusion, functional brain mapping, image guided surgery, and characterization of normal versus abnormal shape and variations in population studies (Maintz and Viergever, 1998). Registration of brain images to a standard stereotaxic coordinate system enables the use of anatomical priors for tissue classification and segmentation. This article reviews five publicly available linear registration techniques for MR brain images, and compares their performance in registering native un-preprocessed images to an average stereotaxic template, using a large number of subject data from 4 different studies.

A registration problem can generally be decomposed into 2 major independent components: the registration paradigm, and the optimization procedure. The registration paradigm may include landmark-based registration, segmentation-based registration, and voxel-property-based registration (Maintz and Viergever, 1998). Here we focus on voxel-wise registration methods which operate directly on the image grey intensity values, without prior data reduction by the user (as in landmark registration) or by segmentation. The standard framework for voxel-wise intensity-based registration involves optimizing a similarity metric or cost function that reflects the similarity between pairs of voxel intensities in the two images. This similarity metric provides a quantifiable measure that reflects how well the two images are aligned as the transformation parameters are changed. In case of

3D linear registrations, the transformation parameters generally include 3 translations, 3 rotations, and 3 scaling parameters in each direction. Under the assumption that the transformation parameters that optimize the similarity function would lead to the optimal registration, the registration problem is transformed into the problem of optimizing a similarity metric, which is often the cross correlation or mutual information between the two images. Registration failures may occur either when the initial assumption fails and the cost function is not ideal (i.e. returning minimum values for poor registrations) or more often when the optimization technique gets stuck in local minima and fails to find the global optimum of the cost function. To address this issue, many techniques attempt iterative multi-resolution registrations, starting by estimating an initial transformation at a lower resolution (therefore reducing the number of local optima) and refining the registration at higher resolutions (Elsen et al., 1993; Pluim et al., 2003). Another advantage of partially solving the problem at a lower resolution is that the algorithms generally require fewer computations/iterations. As a result, multi-resolution solutions also tend to reduce the computation time.

A major question concerning a computed registration transformation entails the accuracy. Since a gold standard for inter-subject registration is lacking in practice, the answer is generally non-trivial. One can identify homologous landmarks, but this is biased to the choice of landmarks and not feasible when testing thousands of datasets. One can estimate a measure of accuracy by using synthetic data, but the results might not be generalizable to practical applications, which are usually more challenging. These difficulties are generally caused by:

1) Intensity range and distribution differences between source and target images caused by differences in scanners as well as acquisition sequences, leading to various levels of noise and intensity inhomogeneity. This can also give rise to slightly different tissue contrasts.

2) Anatomical differences between source and target images, due to inter-subject variability, differences in age, surgical procedures or different atrophy patterns caused by neurodegenerative diseases.

3) Presence of pathology, such as tumors, stroke lesions, white matter hyperintensities (WMHs), infarcts, and microbleeds which can lead to drastic changes in the local intensities.

There is a widespread need to quantify registration accuracy. However, due to the lack of an absolute gold standard for inter-subject registration, such a task is impossible in practice (Maintz and Viergever, 1998). Another issue that hinders giving any statistics on a certain registration method is the incomparability of accuracy experiments done using data obtained from particular scanners and sequences since the method's implementation may be specific to that data. Finally, the inconsistency between the definition of accuracy terms between different studies also makes comparisons difficult.

Here we have compared the performance of five different publicly available and widely used linear registration techniques to map data into stereotaxic space using multi-site and multi-scanner T1-weighted (T1w) MRI data of 9693 scans obtained from 4 different large studies. The scans contain 1.5T and 3T data from healthy individuals, subjects with mild cognitive impairment, Alzheimer's disease, and Parkinson's disease, aged between 20 and 95 years. The registration accuracy of the different linear registration techniques has been verified by manual quality control across the entire sample set to enable meaningful and reliable comparison of the performance of different techniques. In addition, passed and failed registrations for each technique are compared in terms of image signal to noise ratio, intensity non-uniformity, age, as well as the head size, ventricle and brain volume, WMH volume.

2. Methods

2.1. Data

This section describes the study and scanner information for each of the four datasets. Table 1 summarizes the acquisition parameters for each study. Table 2 shows the number of scans used from each study.

1) ADNI: The Alzheimer's Disease Neuroimaging Initiative (ADNI), is a multi-center and multi-scanner study with the aim of defining the progression of Alzheimer's disease (AD). ADNI was launched in 2003 as a public-private partnership, led by Michael W. Weiner, MD. The primary goal of ADNI was to test whether MRI and other biomarkers and clinical assessments can be combined to measure the disease progression (Mueller et al., 2005)(www.adni-info.org). ADNI data includes 1.5T and 3T scans of normal controls, individuals with mild cognitive impairment or AD patients aged 55 years or older. The data has been acquired with different models of GE Medical Systems, Philips Medical systems, and SIEMENS scanners over 59 acquisition sites.

2) PPMI: The Parkinson Progression Marker Initiative (PPMI) is a public–private partnership funded by the Michael J Fox Foundation for Parkinson's Research and funding partners (www.ppmi-info.org/fundingpartners). PPMI is an observational, multi-center and multi-scanner longitudinal study designed to identify PD biomarkers (Marek et al., 2011). PPMI data includes 1.5T and 3T scans of normal controls and *de novo* Parkinson's patients aged 30 years or older. The data has been acquired with different models of GE Medical Systems, Philips Medical systems, and SIEMENS scanners over 33 sites in 11 countries.

3) HCP: The Human Connectome Project (HCP) is a project to construct a map of structural and functional connectivity *in vivo* within and across individuals as an effort to characterize brain connectivity and function and their variability. HCP data includes young healthy adults aged between 25 and 30 years (Van Essen et al., 2012). All T1w HCP images have been scanned using a 32-channel head coil and a SIEMENS 3T scanner.

4) PREVENT-AD: The PREVENT-AD (Pre-symptomatic Evaluation of Novel or Experimental Treatments for Alzheimer's Disease, http://www.prevent-alzheimer.ca) program follows healthy individuals age 55 or older with a parental history of AD dementia (Tremblay-Mercier et al., 2014). Data used in preparation of this article were obtained from the PREVENT-AD program data release 3.0. All the T1-weighted images have been scanned using a single 3T SIEMENS MAGNETOM TrioTim syngo MR scanner (version B17).

Table 1. MRI acquisition parameters for ADNI, PPMI, HCP, and PREVENT-AD datasets.

Dataset	ADNI	PPMI	HCP	PREVENT-AD
Slice thickness (mm)	1.2	1-1.5	0.7	1
No. of slices	160-170	Min 160	210	176
Field of view (cm^2)	256×256	256×Min160	224×224	256×256
Scan matrix (cm^2)	256×256	256×Min160	224×224	256×256
Repetition time (ms)	2300-3000	5-11	2400	2300
Echo time (ms)	2.9-3.5	2-6	2.14	2.98
Pulse sequence	MPRAGE, GR	MPRAGE, SPGR	MPRAGE	MPRAGE

Table 1. Number of scans used from each of ADNI, PPMI, HCP, and PREVENT-AD datasets.

Dataset	ADNI 1.5T	ADNI 3T	PPMI 1.5T	PPMI 3T	HCP	PREVENT-AD
No. of scans	3489	3056	222	778	897	1251

2.2. Registration Methods

The image registration problem can be defined as finding a transformation that maps the target or subject image to the source or reference template image, where both images are 3D volumes with potentially different voxel sizes and dimensions. For the purposes of this paper, the reference template image is the symmetric MNI ICBM152 unbiased non-linear T1w average brain (Fonov et al., 2009, 2011) (http://nist.mni.mcgill.ca/?p=904). Registration is defined by a similarity metric (cost function) that determines the distance between the transformed target image and the source image. Table 3 shows the mathematical functions of the commonly used similarity metrics in the literature (Jenkinson et al., 2002).

Table 3. Definitions of the similarity metrics that are commonly used for linear registration (Jenkinson et al., 2002). X, Y denote source and target images represented as a set of intensities. $H(X,Y) = -\sum_{i,j} p_{ij} \log(p_{ij})$ is the entropy function, where p_{ij} represents the joint probability estimated using the joint intensity histogram. $H(X), H(Y)$ are the marginal entropy functions. Y_k is the intensity of image Y at voxels where the intensity of X is in the k^{th} intensity bin. n_k is the number of elements in Y_k. $N = \sum_k n_k$.

Cost Function	Definition
Normalized Correlation	$\frac{\sum X.Y}{\sqrt{\sum X^2}\sqrt{\sum Y^2}}$
Least Squares	$\sum (X-Y)^2$
Mutual Information	$H(X,Y) - H(X) - H(Y)$
Normalized Mutual Information	$\frac{H(X,Y)}{H(X)+H(Y)}$
Correlation Ratio	$\frac{1}{Var(Y)} \sum_k \frac{n_k}{N} Var(Y_k)$

This section reviews the registration techniques used in this study.

1) MRITOTAL: MRITOTAL is a hierarchical multi-scale 3D registration technique developed as part of the ANIMAL package (Collins et al., 1994) for the purpose of aligning a given MRI volume to an average MRI template aligned with the Talairach stereotaxic coordinate system (Talairach and Tournoux, 1988). MRITOTAL uses voxel-wise image intensity and 3D gradient magnitude as features and cross-correlation as similarity measure. The image is convolved with a 3D isotropic Gaussian kernel. The standard deviation of the kernel is used as a measure of the spatial scale and the full width at half-maximum (FWHM) of the Gaussian is used as a measure of the resolution (blurring). The registration starts at lower resolution (very blurry data) and is refined at each stage by using less blurred images. When smoothing to work at lower resolutions, values outside the field of view were assumed to be zero.

The initial BestLinReg algorithm is a 5-stage hierarchical technique based on MRITOTAL that was developed by Robbins et al. as part of the MINC tools for cortical surface analysis (Robbins,

2004; Robbins et al., 2004). Similar to MRITOTAL, it starts the optimization with highly blurred images (σ_1=16 mm) in the first stage and a sampling step size ($SampS_1$) of 8 mm and a simplex optimization algorithm with a simplex size ($SimpS_1$) of 32 mm. The tolerance parameter (Tol) for the cost function for the initial stage is set to 0.01. In the second and third stages, less blurred images (σ_2=8 mm, σ_3=4 mm) are used as well as smaller step sizes and simplex size ($SampS_{2\text{-}3}$=4 mm, $SimpS_2$=16, $SimpS_3$=8) and higher tolerance (Tol=0.004). In the last two stages, it uses the gradient magnitude of the blurred image with different levels of blurring (σ_4=8 mm, σ_5=4 mm) and the same sampling step size ($SampS_{4\text{-}5}$=4 mm) and tolerance (Tol=0.004) and smaller simplex sizes ($SimpS_4$=4 mm, $SimpS_5$=2 mm). The Revised BestLinReg is another version of BestLinReg with different parameter configurations that has been developed as part of the Cortical Thickness and Surface Analysis (CIVET 2.1) pipeline (Lepage et al., 2017). Revised BestLinReg only estimates an initial translation by calculating the centers of mass in the images in the first stage (3 translations). It then runs the second and third stages of the registration with 6 and 7 parameters (3 translations, 3 rotations plus 1 scaling parameter) and the last two stages with full 9-parameter registrations. The optimization parameters are also modified to adapt to these changes ($\sigma_{2\text{-}3}$=8 mm, σ_4=4 mm, σ_5=2 mm, $SimpS_{2\text{-}3}$=16 mm, $SimpS_4$=8 mm, $SimpS_5$=4 mm, $Tol_{2\text{-}4}$=0.0001, Tol_5=0.0005). The source code for all versions are available at https://github.com/bic-mni. Figure 1 summarizes the registration steps in each version. In this experiment, both normalized mutual information (MI) and cross correlation (XCorr) cost functions were tested for BestLinReg. For revised BestLinReg, only normalized mutual information was used.

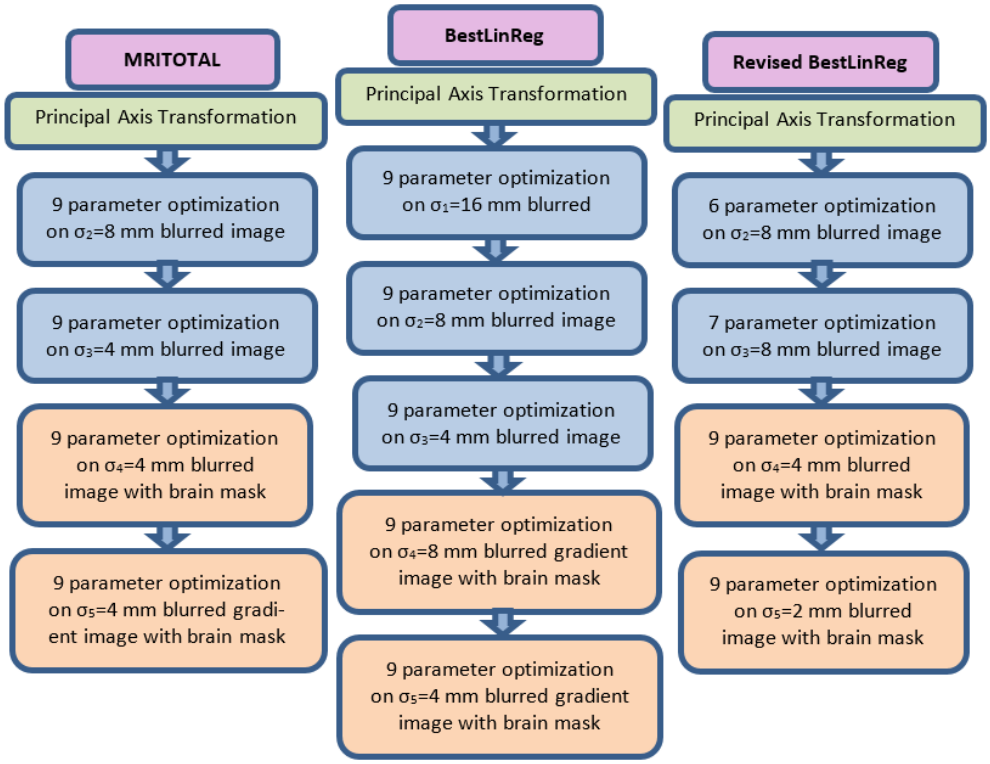

Fig. 1. - Registration steps for MRITOTAL, BestLinReg and Revised BestLinReg methods.

2) FSL FLIRT: FMRIB's Linear Image Registration Tool (FLIRT) is a multi-start, multi-resolution registration tool (Jenkinson et al., 2002, 2012). The registration starts with a large-scale search at 8 mm resolution (e.g. applying a set of initial rotations), followed by a series of multi-start local optimizations at 4 mm resolution, based on the best candidates of the previous stage. The registration is refined in the last stage using a sequence of local optimizations at 2 mm and 1mm resolutions. In addition to the multi-resolution approach, FLIRT uses modified cost functions, where the voxels at the edge of the common overlapping field of view are down-weighted. Fuzzy binning techniques for histogram estimation are also used in an attempt to reduce the number of local minima. The available cost functions include normalized cross correlation, mutual information, and correlation ratio. In this experiment, the default cost function (correlation ratio) was used.

3) Elastix: Elastix is a registration tool (Klein et al., 2010) built on top of Insight Toolkit (ITK) (Johnson et al., 2015; Yoo et al., 2002). Elastix has a parametric and modular framework,

where the user can configure different components of the registration in a parameter text file. The following linear transformation models are supported by Elastix: 3-parameter translation, 6-parameter rigid registration (3 translations and 3 rotations), 7-parameter similarity (rigid plus isotopic scaling), and 12-parameter affine (3 translations, 3 rotations, 3 scales, and 3 sheers). The available cost functions include mean squared difference, normalized correlation, (normalized) mutual information, multi-feature α-mutual information, κ-statistic, and bending energy penalty term. The user can also linearly combine various cost functions. The available optimizers include gradient descent, quasi-Newton, nonlinear conjugate gradient descent, Kiefer-Wolfowitz, Robbins-Monro, adaptive stochastic gradient descent, and evolutionary strategy. For sampling, Elastix supports the use of all voxels, a subset of voxels selected on a uniform grid, and random sampling of voxels on and off the voxel grid (at non-voxel locations). For computing the cost function, Elastix supports several interpolation techniques including nearest neighbour, linear and Nth-order B-Spline interpolation. In this experiment, Mattes mutual information, adaptive stochastic gradient descent optimizer and B-spline interpolation were used in the image pyramid schedule with 3 levels of resolution (downsampling at 8, 4, and 2 voxels respectively). This configuration was selected since it has been widely used and works well for both mono-modality and multi-modality registrations (http://elastix.bigr.nl/wiki/index.php/Parameter_file_database).

4) SPM: spm_affreg is an affine registration tool from Statistical Parametric Mapping software package (SPM12) (Ashburner et al., 1997; Ogden, 1997; Penny et al., 2011) which performs affine registration using a least squares technique. A maximum a posteriori Bayesian approach is adopted, where the spatial transformation is estimated using prior knowledge of the normal variability of brain size, orientation and position in the scanner. The a priori distribution of the parameters have been previously determined from affine transformations estimated from T1w brain images of 51 normal adults (Ashburner et al., 1997). The optimization is performed by iteratively solving a linear approximation of the sum of squared differences between the two images using Taylor's the-

orem. Images are resampled at the desired coordinates using trilinear interpolation of the voxel lattice. In this experiment, the default settings were used which include inter-subject registration regularisation, and 5 mm spacing between sample points. In addition, before running spm_affreg, the images were smoothed by applying a three-dimensional Gaussian filter with FWHM of 12 mm as common practice. Note that the SPM12 image processing pipeline uses a different affine alignment strategy by default.

5) ANTs: ANTs linear registration also uses a multi-resolution hierarchical method, starting by aligning the centers (3 translations), aligning the orientations (3 translations + 3 rotations), accounting for the scaling factors (3 translations + 3 rotations + 1 scaling), and finally, a fully affine transformation (Avants et al., 2011, 2014). The similarity metric can be defined separately for each step. In this experiment, we have used Mattes mutual information metric for all steps since it has been shown to produce the best results for affine registration (Avants et al., 2011). The default stochastic gradient descent is used for optimizing the cost function. The optimization stops either when the slope of change in the energy function is negative or very small or when the maximum number of iterations is reached (Avants et al., 2014). Other parameters were selected based on guidelines from ANTs documentation.

The selected images from each dataset were registered to the MNI ICBM152 average template using the abovementioned methods and settings with a 9-parameter registration for MRITOTAL and FSL, a 7-parameter and a 12-parameter registration for Elastix since it does not support 9-parameter registrations, and a 12-parameter affine registration for SPM and ANTs since they do not support 9-parameter registrations. The scripts containing the details and parameters used for all experiments are available at https://bitbucket.org/bicnist/bic-nist-registration.

2.3. Quality Control

If a technique failed to produce an output, the outcome of the registration was considered as a failure. For the rest of the registrations that produced an output, the obtained transformations were used to transform the images from the native space to template space, resampling it in the template voxel space. To create a QC image, 60 images were extracted from axial, sagittal, and coronal slices (20 each) throughout the resampled volume and the contours of the ICBM152 template were overlaid on each slice image. The slices were selected to cover the brain from bottom to top (axial), left to right (sagittal), and back to front (coronal) for each brain. These 60 images were then concatenated into a single large composite image that was viewed by the human expert to assess the registration and label the outcome as acceptable or failure. Figure 2 shows an example of a QC image for a passed registration.

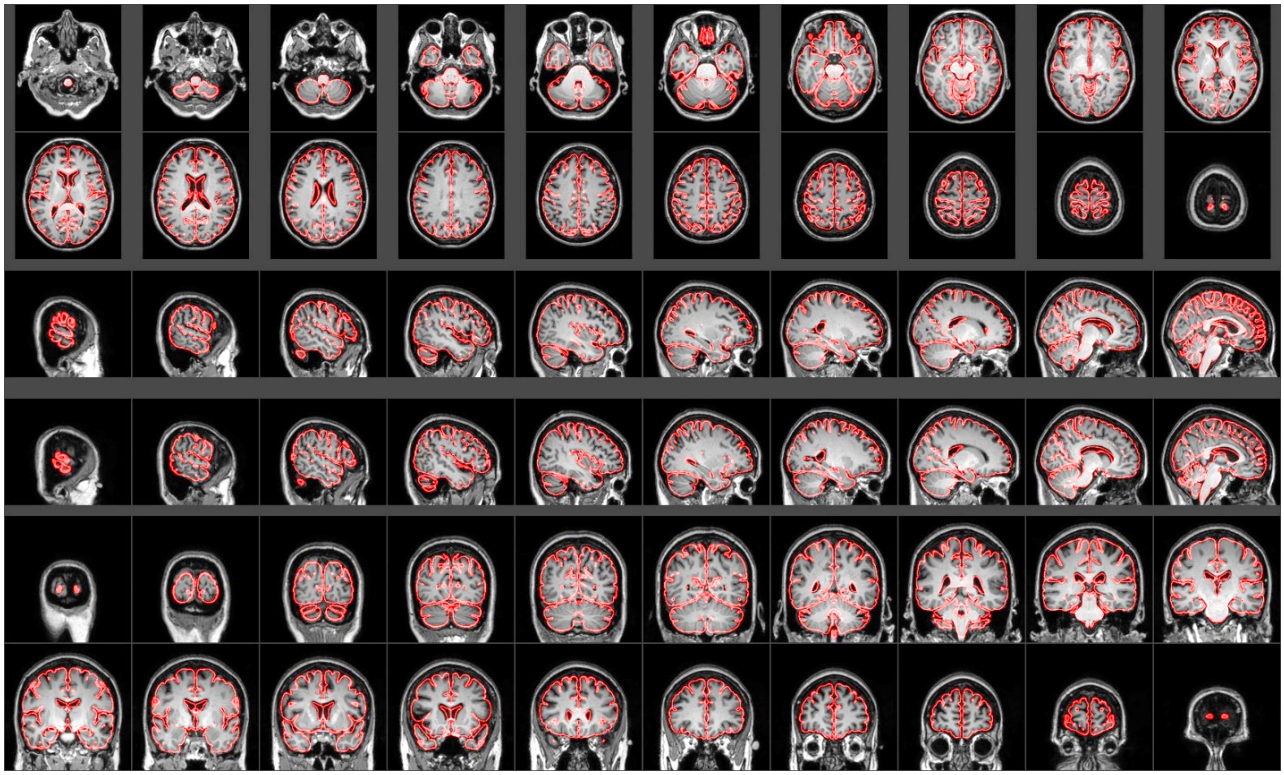

Fig. 1. - Sample image created for assessing the quality of the registrations. Axial, sagittal, and coronal slices showing contours of the average template brain overlaid on the registered image of a single subject in the template space.

The human expert started by assessing the alignment of the images on sagittal and then coronal views. If these images were well aligned, axial slices were assessed to evaluate whether rotation in the axial plane (generally where the highest variability was found) had been correctly estimated. The anatomical landmarks that were used to assess the alignment included the outline of the brain, central sulcus, cingulate sulcus, and parieto-occipital fissure. Since the ventricles are larger in aging and AD brains, the outline of the ventricles was not used as a landmark in the QC. The QC procedure took approximately 30 hours per method for the entire dataset. The human rater was blind to both the registration technique and the dataset information.

The intra-rater Dice similarity and accuracy were 0.96 and 93%, respectively, assessed by manually assessing 1000 randomly selected images a second time. Figure 3 shows examples of six registrations that were labeled as unacceptable by the human rater due to incorrect estimates of translation (Fig. 3.a), translation and scaling (Fig. 3.b), scaling in all directions (Fig. 3.c), scaling in axial plane (Fig. 3.d, Fig. 3.e), and rotation (Fig. 3.f). Figure 4 shows one axial, sagittal, and coronal slice from each image in Figure 3 in greater detail.

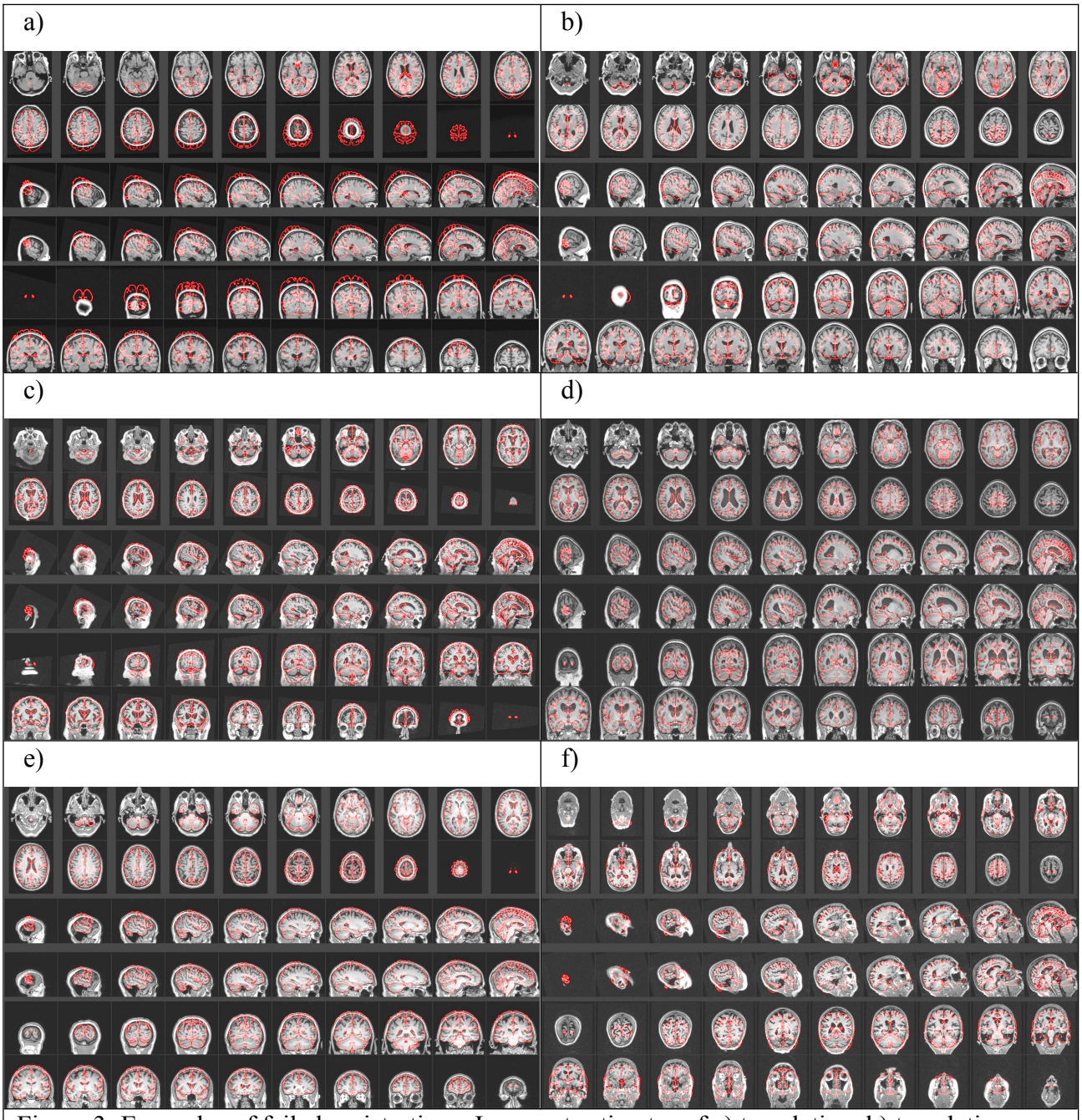

Figure 3: Examples of failed registrations. Incorrect estimates of a) translation, b) translation and scaling, c) scaling in all directions, d, e) scaling in axial plane, and f) rotation.

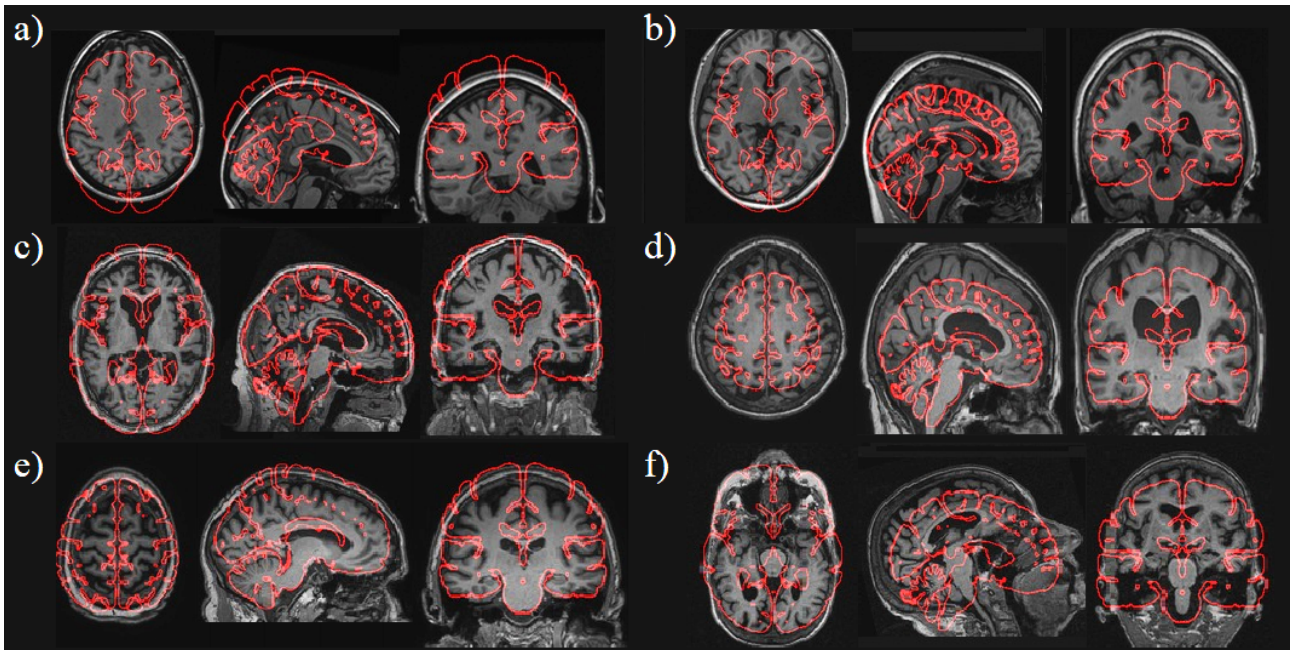

Figure 4: Examples of one axial, sagittal, and coronal slices of failed registrations in more detail. Incorrect estimates of a) translation, b) translation and scaling, c) scaling in all directions, d, e) scaling in axial plane, and f) rotation.

2.4. Effect of noise and image intensity non-uniformity on registration quality

The effects of signal to noise ratio (SNR) and image intensity non-uniformity on registration quality were investigated using permutation tests (N=10,000) between the estimates for passed and failed cases. An estimate of the SNR was obtained using a robust Rician noise estimation technique (Coupé et al., 2010). As a surrogate of the amount of intensity inhomogeneity, an estimate of the standard deviation (STD) of intensity non-uniformity was obtained from the N3 non-uniformity correction method (Sled et al., 1998).

2.5. Effect of age, head size, atrophy and WMH load on registration quality

The effects of age, head size, atrophy and WMH load on registration quality were also investigated. While age is available directly from the different imaging databases, estimates of head size, brain atrophy and WMH load are required. Surrogates of these values were obtained by processing each MRI volume through our standard pipeline

(https://github.com/BIC-MNI/bic-pipelines/blob/master/bin/standard_pipeline.pl) (Aubert-Broche et al., 2013). Image denoising (Coupe et al., 2008), intensity non-uniformity correction (Sled et al., 1998), and image intensity normalization into range (0-100) were performed. After preprocessing, all images were first linearly (using a 9-parameter rigid registration) and then nonlinearly registered to an average template as part of the ANIMAL software (Collins and Evans, 1997; Collins et al., 1994). The scaling parameter (here referred to as scale factor) used to scale individual scans to the standard template was used as a surrogate of head size. The brain tissue as well as the ventricles were segmented as part of the ANIMAL software. Normalized brain volume and ventricle volume were used as surrogates of brain atrophy in the permutation tests below.

The WMH load of subjects in the ADNI, PPMI, and PREVENT-AD datasets was estimated by segmenting the WMH lesions using a previously validated fully automated technique (Dadar et al., 2017a, 2017b). The WMH segmentation tool is based on a Random Forests classifier that is trained to detect WMHs in multi-center and multi-scanner datasets, using either T1-weighted and FLAIR or T1-weighterd and T2-weighted/PD images. Since HCP subjects were young healthy individuals and visual assessment showed that they did not have significant amounts of WMHs, they were excluded from the WMH comparisons.

The quality of all segmentations was verified by visual assessment and failed segmentation cases were discarded (N=48). The effect of age, structure volumes and WMH loads on registration quality was evaluated using permutation tests (N=10,000) between the estimates for passed and failed registrations.

3. Results

Overall performance.

Table 4 compares the performance of different techniques in terms of percentage of registration failures across different datasets. Performance ranged from 53.83% success in ADNI 1.5T with

MRITOTAL to 100% success with PPMI 1.5T and Revised BestLinReg. The Revised BestLinReg method had the best overall performance across all datasets (failure rate= 0.44%), followed by FSL (failure rate= 11.11%), BestLinReg MI (failure rate= 12.74%), BestLinReg XCorr (failure rate= 13.03%), ANTs (failure rate= 8.87%), Elastix Similarity (failure rate= 24.40%), Elastix Affine (failure rate= 12.35%), MRITOTAL (failure rate= 27.14%), MRITOTAL ICBM (failure rate= 27.41%), and SPM (failure rate= 30.66%). ANTs failed to converge within the two-hour limit and did not produce any outputs for 98.55% of the scans from the HCP dataset.

Table 4. Registration error rates. Comparison between the performance of different linear registration techniques. Data are the percentage of registration failures assessed by a human expert (i.e., a smaller number shows better performance), across the different datasets.

Dataset	ADNI 1.5T	ADNI 3T	PPMI 1.5T	PPMI 3T	HCP 3T	PREVENT-AD 3T
MRITOTAL	46.17	35.01	22.97	27.51	7.92	24.86
MRITOTAL ICBM	45.03	37.40	18.02	26.99	13.27	22.14
BestLinReg MI	15.36	16.30	7.66	15.68	8.92	12.55
BestLinReg XCorr	9.03	8.48	14.41	16.20	18.17	11.91
Revised BestLinReg	0.46	0.69	0.00	0.90	0.11	0.48
FSL	13.01	12.24	14.41	18.38	4.24	4.40
Elastix Similarity	33.99	28.93	13.96	18.25	28.09	23.18
Elastix Affine	18.40	11.75	11.26	11.31	6.35	2.08
SPM	28.00	25.22	36.94	43.32	49.94	21.74
ANTs	5.33	6.41	35.59	22.37	9.36	11.19

Table 5 shows the percentage of failures that were common between each two methods.

Table 5. Registration failures common between different methods. Data are the Dice Kappa percentages of cases that failed for each two registration methods.

Method	MRITOTAL	MRITOTAL ICBM	BestLinReg MI	BestLinReg XCorr	Revised BestLinReg	FSL	Elastix Similarity	Elastix Affine	SPM	ANTs
MRITOTAL	-	60.20	21.70	16.97	1.07	19.39	36.35	22.32	45.97	12.23
MRITOTAL ICBM	-	-	21.34	16.58	0.94	19.84	34.89	20.85	45.50	14.78
BestLinReg MI	-	-	-	25.02	3.15	24.66	21.17	21.56	24.67	13.10
BestLinReg XCorr	-	-	-	-	1.28	20.37	16.35	15.53	17.66	19.68
Revised BestLinReg	-	-	-	-	-	4.88	1.41	5.13	1.42	2.57
FSL	-	-	-	-	-	-	18.40	35.41	23.85	15.61
Elastix Similarity	-	-	-	-	-	-	-	24.76	41.43	20.25
Elastix Affine	-	-	-	-	-	-	-	-	30.37	17.51
SPM	-	-	-	-	-	-	-	-	-	26.69
ANTs	-	-	-	-	-	-	-	-	-	-

Factors affecting performance.

MRITOTAL, MRITOTAL ICBM, and Elastix performed significantly better on 3T scans in the ADNI dataset ($p<0.001$, unpaired t-test). The differences for 1.5T vs 3T for other methods and for the PPMI dataset were not significant.

Table 6 shows the p values of the permutation tests comparing SNR and the average and standard deviation of image intensity non-uniformity estimates between the passed and failed registrations. The amount of image non-uniformity was associated with registration success for all methods except Elastix Similarity. The SNR level was associated with success for half the methods tested: MRITOTAL, MRITOTAL ICBM, BestLinReg MI, Elastix Similarity and SPM.

Table 6. Effect of SNR and NU on registration QC. p values of permutation tests comparing SNR and intensity non-uniformity measures. SNR= Signal to Noise Ratio. NU= Intensity Non-uniformity.

Method	SNR	NU STD
MRITOTAL	<0.001	<0.001
MRITOTAL ICBM	<0.001	<0.001
BestLinReg MI	0.015	<0.001
BestLinReg XCorr	0.121	<0.001
Revised BestLinReg	0.420	0.014
FSL	0.198	<0.001
Elastix Similarity	<0.001	0.158
Elastix Affine	0.040	<0.001
SPM	<0.001	<0.001
ANTs	0.159	<0.001

Table 7 shows the p values of the permutation tests comparing age and different measures of atrophy related changes between passed and failed registrations. Interestingly, Age affects registration success for all techniques except Revised BestLinReg (albeit the p value is marginally significant before correction for multiple comparisons). Larger ventricle size is associated with registration failures for all methods except Elastix Similarity and SPM. Brain size adversely affects registration success for all methods except Elastix. Head size is not associated with registration failures only for Elastix Similarity and SPM. Finally, the WMH load is associated with registration failure or all methods except BestLinReg MI, BestLinReg XCorr and Revised BestLinReg.

Table 7. Effect of Age, atrophy, brain size and WMH load on registration quality. p values of permutation tests comparing age and measures of atrophy related changes between passed and failed registrations. WMH= White Matter Hyperintensity.

Method	Age	Ventricle	Brain	Scale	WMH
MRITOTAL	<0.001	<0.001	<0.001	<0.001	<0.001
MRITOTAL ICBM	<0.001	<0.001	<0.001	<0.001	<0.001
BestLinReg MI	<0.001	<0.001	<0.001	<0.001	0.240
BestLinReg XCorr	0.006	0.003	<0.001	0.004	0.162
Revised BestLinReg	0.026	<0.001	0.004	<0.001	0.321
FSL	<0.001	<0.001	<0.001	<0.001	0.001
Elastix Similarity	<0.001	0.308	0.936	0.303	<0.001
Elastix Affine	<0.001	<0.001	0.472	<0.001	<0.001
SPM	<0.001	0.502	<0.001	0.501	<0.001
ANTs	<0.001	<0.001	<0.001	<0.001	<0.001

4. Discussion

In brain imaging, linear stereotaxic registration aims to align the subject's brain into a standardized space to allow for more comprehensive comparisons of the anatomy and pathologies at the population level. Such a mapping generally corrects for location, orientation, and overall size of the brain (3 translation, 3 rotation, and 3 scaling parameters in 3D transformations). Choosing a registration technique among the various tools that are publicly available and widely used is difficult, since there is no single technique that can handle every brain registration task (registering different image modalities, acquisition sequences, inter/intra subject registration). Moreover, comparisons between different techniques should be driven by evaluations on the same datasets, which is generally not the case. The experiments in this paper were designed to compare five commonly used publicly available registration tools based on their performance in registering un-preprocessed native T1-weighted MRIs of brains aged between 25-95 years to an average template of young healthy brain (the MNI ICBM 152 unbiased non-linear average).

In evaluating registration performance, many comparison studies use synthetically generated data that is created by applying a set of transformations to the original images to assess the quality of linear registration techniques (Jenkinson et al., 2002). This greatly simplifies the problem since it ensures that there would be a perfect match between the source and target images, which is generally not the case. Here instead, we register native images to an average template, a task necessary in any population study, and also needed for many preprocessing and segmentation techniques. The experiments here enable meaningful comparisons between different registration techniques, since they have been applied to 1.5T and 3T data from various datasets, two of which contain multi-site, multi-scanner data. Furthermore, including subjects with a wide age range (25-95 years) and patients with neurodegenerative diseases from the ADNI and PPMI cohorts enables evaluation of the techniques in the presence of brain changes such as AD- and PD-related atrophy and vascular dis-

ease indicated by white matter abnormalities. Indeed, our experiments showed that the brain changes caused by aging, atrophy, and WMHs significantly reduce the accuracy of the registrations.

Our experiments showed that the revised BestLinReg technique had the best performance among all registration techniques and datasets, with only 51 registration failures out of 9693 registrations. The MRITOTAL and the standard BestLinReg techniques tend to slightly underestimate the scaling parameters when using the cross-correlation similarity metric. BestLinReg with mutual information tended to overestimate them. Elastix single scale registrations tended to align the template in the coronal plane, but were not able to correct for the front-to-back or top-to-bottom differences in the brains.

The SNR does not appear to have a significant adverse effect on the performance of BestLinReg XCorr, Revised BestLinReg, FSL or ANTS techniques. This is likely due to the internal blurring used in these methods, which generally reduces the effects of noise. Additionally, intensity nonuniformity seems to adversely affect Revised BestLinReg and Elastix methods less (albeit this effect is no longer significant when corrected for multiple comparisons).

Older Age adversely affects registration success for all techniques (except for Revised BestLinReg, but this no longer holds when correcting for multiple comparisons). This is likely due to the morphological changes that are associated with aging, i.e. larger ventricles, grey matter and white matter atrophy, and white matter hyperintensities. This is further validated by the fact that ventricle and brain size (both reflecting atrophy) as well as higher white matter hyperintensity load also seem to adversely affect registration success. Specifically, in MRITOTAL techniques, the last steps are driven by gradient magnitude, and larger ventricles will have more energy in the objective function, possibly biasing the transform. Since the target image (MNI ICBM 152 template) is generated based on healthy young individuals, registration of older brains with different intensity distributions proves more challenging. This supports the fact that older subjects need to be registered to

an age-specific or population-specific (e.g. Alzheimer's disease, Parkinson's disease) template for analyzing datasets to reduce the registration failure caused by these effects. Various groups have attempted to create age-specific or population specific (e.g. Alzheimer's disease population) templates (Dickie et al., 2016; Fillmore et al., 2015), including our group (Fonov et al., 2009, 2011). All templates are publicly available at (http://nist.mni.mcgill.ca/?page_id=714, https://datashare.is.ed.ac.uk/handle/10283/1957).

The head size (estimated by scaling factor) is not associated with registration failures only for BestLinReg XCorr, Elastix, and SPM. MRITOTAL, MRITOTAL ICBM, BestLinReg MI, Revised BestLinReg, and FSL techniques seem to have more registration failures for larger head sizes, whereas ANTs seems to have more registration failures for smaller head sizes.

The larger number of available configurations provides the users with the opportunity to optimize the registrations based on the specific data set and task of interest. As an example, when dealing with source and target images that have very different tissue contrasts, using similarity metrics such as mutual information generally works better than least squares or correlation metrics. Similarly, if available, one can choose different optimizers based on time and computational power constraints. Another important registration parameter that is not always supported is the type of linear transformation. For example, Elastix, SPM, and ANTs do not support a 9-parameter registration. Therefore, one has to either opt for a suboptimal 7-parameter transformation which assumes the same amount of scaling in all directions, or a 12-parameter transformation with shearing which warps the shape of geometric figures.

Our study is not without limitations. Not all the available options were tested for different registration techniques. For FSL FLIRT, and SPM the default options were used. For Elastix, and ANTs the widely used mutual information and stochastic gradient descent optimizer options were selected. This might lead to suboptimal results. However, since one rarely attempts all the available configurations (especially for Elastix, which provides 10s of different possible combinations), we tested

only the most commonly used options. Furthermore, HCP dataset had been defaced prior to the registrations in order to ensure anonymity, while the other datasets had not (ADNI, PPMI, and PREVENTAD). This face removal step might affect the performance of some registration techniques more than the others. SPM and BestLinReg XCorr had a poorer performance on HCP dataset compared with other datasets (Table 4). However, since HCP also had a higher level of intensity nonuniformity, it's not possible to speculate whether this lower performance was caused by the defacing.

Preprocessing the images can have a significant effect in improving the performance of most registration techniques. As was shown in our experiments, the signal to noise ratio and image intensity non-uniformity significantly affected the quality of the registrations for all techniques. To make the comparisons fair, no preprocessing was performed on the native images here before the registrations. This decision was also made based on the fact that many preprocessing techniques need an initial registration, and a registration method that would be dependent on preprocessing cannot be used in pipelines that use such preprocessing techniques.

Linear brain MRI registration to an average template is an ill-posed problem because the shape and cortical topology of the brain varies strongly from one individual to another, especially in existence of brain atrophy. Therefore, intensity-based registration algorithms are expected to fail at least on some pathological cases. Variabilities in data from different scanner models and image acquisition sequences further adds to the complexity of this problem. So far, no general technique is able to accurately register any two images all of the time. This comparison provides some insight into the performance of several publicly available registration tools and facilitates the choice of a registration technique for a specific application.

Acknowledgement:

ADNI data is from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles.

PPMI data was obtained from the Parkinsons Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI is sponsored and partially funded by the Michael J Fox Foundation for Parkinsons Research and funding partners, including AbbVie, Avid Radiopharmaceuticals, Biogen, Bristol-Myers Squibb, Covance, GE Healthcare, Genentech, GlaxoSmithKline (GSK), Eli Lilly and Company, Lundbeck, Merck, Meso Scale Discovery (MSD), Pfizer, Piramal Imaging, Roche, Servier, and UCB (www.ppmi-info.org/fundingpartners).

HCP data was obtained from the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

PREVENT-AD data were obtained from the Pre-symptomatic Evaluation of Novel or Experimental Treatments for Alzheimer's Disease (PREVENT-AD, http://www.prevent-alzheimer.ca) program data release 3.0 (2016-11-30). Data collection and sharing for this project were supported by its sponsors, McGill University, the Fonds de Research du Québec – Santé, the Douglas Hospital Research Centre and Foundation, the Government of Canada, the Canadian Foundation for Innovation, the Levesque Foundation, and an unrestricted gift from Pfizer Canada. Private sector contributions are facilitated by the Development Office of the McGill University Faculty of Medicine and by the Douglas Hospital Research Centre Foundation (http://www.douglas.qc.ca/).

We also wish to thank the *Famille Louise & André Charron* for financial support.

5. References:

Ashburner, J., Neelin, P., Collins, D.L., Evans, A., and Friston, K. (1997). Incorporating prior knowledge into image registration. Neuroimage *6*, 344–352.

Aubert-Broche, B., Fonov, V.S., García-Lorenzo, D., Mouiha, A., Guizard, N., Coupé, P., Eskildsen, S.F., and Collins, D.L. (2013). A new method for structural volume analysis of longitudinal brain MRI data and its application in studying the growth trajectories of anatomical brain structures in childhood. NeuroImage *82*, 393–402.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., and Gee, J.C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage *54*, 2033–2044.

Avants, B.B., Tustison, N.J., Stauffer, M., Song, G., Wu, B., and Gee, J.C. (2014). The Insight ToolKit image registration framework. Front. Neuroinformatics *8*.

Collins, D.L., and Evans, A.C. (1997). Animal: validation and applications of nonlinear registration-based segmentation. Int. J. Pattern Recognit. Artif. Intell. *11*, 1271–1294.

Collins, D.L., Neelin, P., Peters, T.M., and Evans, A.C. (1994). Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J. Comput. Assist. Tomogr. *18*, 192–205.

Coupe, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., and Barillot, C. (2008). An Optimized Blockwise Nonlocal Means Denoising Filter for 3-D Magnetic Resonance Images. IEEE Trans. Med. Imaging *27*, 425–441.

Coupé, P., Manjón, J.V., Gedamu, E., Arnold, D., Robles, M., and Collins, D.L. (2010). Robust Rician noise estimation for MR images. Med. Image Anal. *14*, 483–493.

Dadar, M., Maranzano, J., Misquitta, K., Anor, C.J., Fonov, V.S., Tartaglia, M.C., Carmichael, O.T., Decarli, C., Collins, D.L., and Alzheimer's Disease Neuroimaging Initiative (2017a). Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging. NeuroImage *157*, 233–249.

Dadar, M., Pascoal, T., Manitsirikul, S., Misquitta, K., Tartaglia, C., Brietner, J., Rosa-Neto, P., Carmichael, O., DeCarli, C., and Collins, D.L. (2017b). Validation of a Regression Technique for Segmentation of White Matter Hyperintensities in Alzheimer's Disease. IEEE Trans. Med. Imaging.

Dickie, D.A., Rodriguez, D., Danso, S., Deary, I.J., Job, D.E., Bastin, M.E., Pernet, C., Shenkin, S.D., Wardlaw, J., and Robson, A. (2016). Brain Imaging of Normal Subjects (BRAINS) age-specific MRI atlases from young adults to the very elderly (v1. 0).

Elsen, P.A. van den, Pol, E.J.D., and Viergever, M.A. (1993). Medical image matching-a review with classification. IEEE Eng. Med. Biol. Mag. *12*, 26–39.

Fillmore, P.T., Phillips-Meek, M.C., and Richards, J.E. (2015). Age-specific MRI brain and head templates for healthy adults from 20 through 89 years of age. Front. Aging Neurosci. *7*.

Fonov, V., Evans, A., McKinstry, R., Almli, C., and Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage *47*, S102.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., and Collins, D.L. (2011). Unbiased average age-appropriate atlases for pediatric studies. NeuroImage *54*, 313–327.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. NeuroImage *17*, 825–841.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., and Smith, S.M. (2012). Fsl. Neuroimage *62*, 782–790.

Johnson, H.J., McCormick, M.M., and Ibanez, L. (2015). The ITK Software Guide: Design and Functionality. Publ. Kitware Inc ISBN 9781–930934.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., and Pluim, J.P. (2010). Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging *29*, 196–205.

Lepage, C., Lewis, L., Jeon, S., Bermudez, P., Khalili-Mahani, N., Omidyeganeh, M., Zijdenbos, A., Vincent, R., Adalat, R., and Evans, A. (2017). Human MR Evaluation of Cortical Thickness Using CIVET 2.1. In Annual Meeting of the Organization for Human Brain Mapping, p.

Maintz, J.B.A., and Viergever, M.A. (1998). A survey of medical image registration. Med. Image Anal. *2*, 1–36.

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., et al. (2011). The Parkinson Progression Marker Initiative (PPMI). Prog. Neurobiol. *95*, 629–635.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., and Beckett, L. (2005). The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. N. Am. *15*, 869–877.

Ogden, R.W. (1997). Non-linear elastic deformations (Courier Corporation).

Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., and Nichols, T.E. (2011). Statistical parametric mapping: the analysis of functional brain images (Academic press).

Pluim, J.P.W., Maintz, J.B.A., and Viergever, M.A. (2003). Mutual-information-based registration of medical images: a survey. IEEE Trans. Med. Imaging *22*, 986–1004.

Robbins, S.M. (2004). Anatomical standardization of the human brain in euclidean 3-space and on the cortical 2-manifold (McGill University).

Robbins, S., Evans, A.C., Collins, D.L., and Whitesides, S. (2004). Tuning and comparing spatial normalization methods. Med. Image Anal. *8*, 311–323.

Sled, J.G., Zijdenbos, A.P., and Evans, A.C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. Med. Imaging IEEE Trans. On *17*, 87–97.

Talairach, J., and Tournoux, P. (1988). Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging.

Tremblay-Mercier, J., Madjar, C., Etienne, P., Poirier, J., and Breitner, J. (2014). A PROGRAM OF PRE-SYMPTOMATIC EVALUATION OF EXPERIMENTAL OR NOVEL TREATMENTS FOR ALZHEIMER'S DISEASE (PREVENT-AD): DESIGN, METHODS, AND PERSPECTIVES. Alzheimers Dement. J. Alzheimers Assoc. *10*, P808.

Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E.J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., et al. (2012). The Human Connectome Project: A data acquisition perspective. NeuroImage *62*, 2222–2231.

Yoo, T.S., Ackerman, M.J., Lorensen, W.E., Schroeder, W., Chalana, V., Aylward, S., Metaxas, D., and Whitaker, R. (2002). Engineering and algorithm design for an image processing API: a technical report on ITK-the insight toolkit. Stud. Health Technol. Inform. 586–592.