



Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration.

Arno Klein, Jesper Andersson, Babak A. Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E. Christensen, D. Louis Collins, Pierre Hellier, Joo Hyun Song, et al.

► To cite this version:

Arno Klein, Jesper Andersson, Babak A. Ardekani, John Ashburner, Brian Avants, et al.. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration.. *NeuroImage*, 2009, 46 (3), pp.786-802. 10.1016/j.neuroimage.2008.12.037 . inserm-00360790v2

HAL Id: inserm-00360790

<https://inserm.hal.science/inserm-00360790v2>

Submitted on 2 Mar 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration

Arno Klein ^{a,*}, Jesper Andersson ^b, Babak A. Ardekani ^{c,d},
John Ashburner ^e, Brian Avants ^f, Ming-Chang Chiang ^g,
Gary E. Christensen ^h, D. Louis Collins ⁱ, James Gee ^f,
Pierre Hellier ^{j,k}, Joo Hyun Song ^h, Mark Jenkinson ^b,
Claude Lepage ⁱ, Daniel Rueckert ^m, Paul Thompson ^g,
Tom Vercauteren ^{n,ℓ}, Roger P. Woods ^o, J. John Mann ^a,
Ramin V. Parsey ^a

^a*New York State Psychiatric Institute, Columbia University, NY, NY 10032, USA*

^b*FMRIB Centre, University of Oxford, Department of Clinical Neurology, John Radcliffe Hospital, Oxford OX3 9DU, UK*

^c*Nathan Kline Institute, Orangeburg, NY, 10962, USA*

^d*New York University School of Medicine, NY, NY 10016, USA*

^e*Functional Imaging Laboratory, Wellcome Trust Centre for Neuroimaging, London WC1N 3BG, UK*

^f*Penn Image Computing and Science Laboratory, Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104-2644, USA*

^g*Laboratory of Neuro Imaging, UCLA School of Medicine, Los Angeles, CA 90095-7332, USA*

^h*Dept. of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA*

ⁱ*McConnell Brain Imaging Center, Montreal Neurological Institute, Montreal, QC H3A 2B4, Canada*

^j*INRIA Rennes, Bretagne Atlantique Research Centre, Campus universitaire de Beaulieu, 35042 Rennes Cedex, France*

^k*INSERM, Visages U746, IRISA, Campus de Beaulieu, Rennes, France*

^ℓ*INRIA Sophia Antipolis - Méditerranée, 06902 Sophia Antipolis, France*

^m*Visual Information Processing, Department of Computing, Imperial College, London SW7 2BZ, UK*

ⁿ*Mauna Kea Technologies, 75010 Paris, France*

^o*Department of Neurology, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA*

Abstract

All fields of neuroscience that employ brain imaging need to communicate their results with reference to anatomical regions. In particular, comparative morphometry and group analysis of functional and physiological data require coregistration of brains to establish correspondences across brain structures. It is well established that linear registration of one brain to another is inadequate for aligning brain structures, so numerous algorithms have emerged to nonlinearly register brains to one another. This study is the largest evaluation of nonlinear deformation algorithms applied to brain image registration ever conducted. Fourteen algorithms from laboratories around the world are evaluated using 8 different error measures. More than 45,000 registrations between 80 manually labeled brains were performed by algorithms including: AIR, ANIMAL, ART, Diffeomorphic Demons, FNIRT, IRTK, JRD-fluid, ROMEO, SICLE, SyN, and four different SPM5 algorithms (“SPM2-type” and regular Normalization, Unified Segmentation, and the DARTEL Toolbox). All of these registrations were preceded by linear registration between the same image pairs using FLIRT. One of the most significant findings of this study is that the relative performances of the registration methods under comparison appear to be little affected by the choice of subject population, labeling protocol, and type of overlap measure. This is important because it suggests that the findings are generalizable to new subject populations that are labeled or evaluated using different labeling protocols. Furthermore, we ranked the 14 methods according to three completely independent analyses (permutation tests, one-way ANOVA tests, and indifference-zone ranking) and derived three almost identical top rankings of the methods. ART, SyN, IRTK, and SPM’s DARTEL Toolbox gave the best results according to overlap and distance measures, with ART and SyN delivering the most consistently high accuracy across subjects and label sets. Updates will be published on the <http://www.mindboggle.info/papers/> website.

* Corresponding author

Email address: arno@binarybottle.com (Arno Klein).

URL: <http://www.binarybottle.com> (Arno Klein).

1 Introduction

Brain mapping – mapping the structures, physiology, functions, and connectivity of brains in individuals and in different populations – is possible due to a diverse but often disconnected array of brain imaging technologies and analysis methods. To make the best use of brain image data, researchers have attempted for over 40 years to establish a common reference frame such as a three-dimensional coordinate or labeling system to consistently and accurately communicate the spatial relationships within the data (Talairach and Szikla, 1967; Talairach and Tournoux, 1988; Drury et al., 1996; Fischl et al., 1999; Clouchoux et al., 2005). A common reference frame helps us to:

1. communicate and compare data
(across subjects, time, conditions, and image types),
2. classify data
(by meaningful spatial positions or extent), and
3. find patterns in data
(to infer structural or functional relationships).

These three benefits are contingent on one serious premise: positions and sizes in one brain must correspond to positions and sizes in another brain to make comparisons.

This premise almost universally does not hold when brain image data are compared across individuals. The noise that this introduces is often accepted by researchers who generally assume that if they have found corresponding features across two brains, the intervening points between those features correspond to one another as well. Brains are so variable in shape that there simply may not exist a point-to-point correspondence across any two brains, or even in the same brain over time.

Explicit manual labeling of brain regions is the preferred approach for establishing anatomical correspondence, but it is too prohibitive in terms of time and resources, particularly in cases where neuroanatomists are not available, in intraoperative or other time-sensitive scenarios, and in high-throughput environments that need to process dozens to thousands of brain images.¹

¹ To indicate the level of investment required to manually label brain anatomy, the Center for Morphometric Analysis (CMA) at the Massachusetts General Hospital (MGH) expects at least one month of training to train new technicians to the point of acceptable inter-rater reliability using their Cardviews (Caviness et al., 1996) labeling protocol and software; once trained, it takes hours to weeks to manually label a single brain. For 12 of the brains used in this study, a trained assistant took two weeks to label each brain. At this rate, performing a modest imaging study with 20 subjects and 20 controls would require 20 months devoted strictly to labeling. Manual labeling also suffers from inconsistencies within and across human labelers

Automatically determining anatomical correspondence is almost universally done by registering brains to one another or to a template. There has been a proliferation of different approaches to perform image registration that demands a comparison to guide choices regarding algorithms, software implementation, setup and parameters, and data preprocessing options. To better enable individuals to make these choices, the Valmet software tool (<http://www.ia.unc.edu/public/valmet/>) (Gerig et al., 2001) and the Non-rigid Image Registration Evaluation Project (NIREP) (<http://www.nirep.org>) were developed. The Windows-based Valmet was in 2001 the first publicly available software tool for measuring (as well as visualizing) the differences between corresponding image segmentations, but has received only one minor update since 2001 (in 2004). It uses several algorithms to compare segmentations: overlap ratio, Hausdorff distance, surface distance, and probabilistic overlap. The NIREP project “has been started to develop, establish, maintain, and endorse a standardized set of relevant benchmarks and metrics for performance evaluation of nonrigid image registration algorithms.” The initial phase of the project will include 16 manually labeled brain images (32 labeled regions in 8 men and 8 women) and four evaluation metrics: 1. relative overlap (equivalent to the “union overlap” defined in the Materials and methods section), 2. variance of the registered intensity images for an image population, 3. inverse consistency error between a forward and reverse transformation between two images, and 4. transitivity (how well all the pairwise registrations of the image population satisfy the transitivity property).

In this study we set out to evaluate what we believe are the most important nonlinear deformation algorithms that have been implemented in fully automated software programs and applied to human brain image registration. We measure accuracy at the scale of gross morphological structures (gyri, sulci, and subcortical regions) acquired by magnetic resonance imaging (MRI). There have been two significant prior studies that compared more than three nonlinear deformation algorithms for evaluating whole-brain registration.

The first was communicated in a series of publications by Hellier et al. (Hellier et al., 2001a, 2002, 2003); they compared five different fully automated nonlinear brain image registration software programs using the same set of quantitative measures. These included global measures comparing 17 deformed MRI source images and one target image: average brain volume, gray matter overlap, white matter overlap, and correlation of a measure of curvature, and local measures of distance and shape between corresponding principal sulci. Our study includes a version of each of the five methods and is different primarily because (1) all tests were conducted by a single individual (the first author) who had not authored any of the software packages, but received

(Caviness et al., 1996; Fiez et al., 2000; Towle et al., 2003)

guidance from the principal architects of the respective algorithms, (2) its focus is on manually labeled anatomical regions, and (3) each and every brain was used as a source and as a target for registration rather than selecting a single target.

The second is a recent paper (Yassa and Stark, 2009) that compares nonlinear registration methods applied to regions in the medial temporal lobe; six of the methods are fully automated and two are semi-automated (requiring manual identification of landmarks). They apply these methods either to manually labeled brain regions, to weighted masks for these regions, or to the original unlabeled brains, as in our study. The four methods that they applied to unlabeled brains (and evaluated on regions in the medial temporal lobe) are the Talairach piecewise linear approach and three SPM programs (included in our study). Registering labeled regions obviously requires that the regions be labeled; their ROI-AL approach ‘labels to register’ rather than ‘registers to label’ or ‘registers without labels.’ They used two evaluation measures on pairs of images (20 MRI volumes total): an overlap measure (equivalent to the “target overlap” defined in the Materials and methods section) and a measure of blur in a group average of coregistered images. What sets our study apart from both of these prior studies is the unparalleled scale and thoroughness of the endeavor:

What sets our study apart from both of these prior studies is the unparalleled scale and thoroughness of the endeavor:

- over 14 nonlinear algorithms
- each algorithm applied at least 2,168 times (over 45,000 registrations total)
- 80 manually labeled brain images
- 4 different whole-brain labeling protocols (56 to 128 labeled regions)
- 8 different evaluation measures
- 3 independent analysis methods

This study evaluates 15 registration algorithms, one linear (FLIRT) and 14 nonlinear: AIR, ANIMAL, ART, Diffeomorphic Demons, FNIRT, IRTK, JRD-fluid, ROMEO, SICLE, SyN, and four different SPM5 algorithms (“SPM2-type” and regular Normalization, Unified Segmentation, and the DARTEL Toolbox; DARTEL was also run in a pairwise manner and all four SPM algorithms were run with and without removal of skulls from the images). The linear algorithm was included as an initialization step to establish a baseline prior to applying the nonlinear algorithms. Comparisons among the algorithms and their requirements are presented in Table 1 and in the Appendix, software commands are in Supplementary section 7, and brief descriptions are in Supplementary section 8. Many of them are in common use for registering structural MRIs to each other or to templates for neuromorphometric research or as an intermediary to compare functional or physiological data (Gholipour

et al., 2007), but some of them exist only as pre-release code made available by their respective authors for this study. See the “Algorithms excluded from the study” section in the Discussion for algorithms excluded from the study. Additional materials and updated information will be made publicly available via the website <http://www.mindboggle.info/papers/>.

2 Materials and methods

In this section, we first briefly describe the acquisition and preparation of the brain image and label data. Then we outline the preprocessing (brain extraction and formatting), linear registration, and nonlinear registration stages applied to the data, our evaluation measures, and our analysis methods. The first author performed these latter steps on an OSX system (Mac Pro 2-Quad-Core (8-processor) Intel Xeon, 3 GHz, 6 GB RAM) with a 10.4 operating system, except where noted (see Supplementary section 7). Custom Python (<http://www.python.org>) and Matlab (<http://www.mathworks.com>) software programs performed the preprocessing steps, called the different programs to process thousands of pairs of images, computed the results for evaluation, and produced the visualizations in the Results section.

2.1 Data preparation: images, labels, brain extraction, and formatting

Algorithm	Deformation	$\simeq$ dof	Similarity	Regularization
FLIRT	Linear, rigid-body	9, 6	norm. CR	
AIR	5th-order polynomial warps	168	MSD (opt. intensity scaling)	Incremental increase of polynomial order; MRes: sparse-to-fine voxel sampling
ANIMAL	Local translations	69K	CC	MRes, local Gaussian smoothing; stiffness parameter weights mean deformation vector at each node
ART	Non-parametric, homeomorphic	7M	norm. CC	MRes median and low-pass Gaussian filtering
Diffeomorphic Demons	Non-parametric, diffeomorphic displacement field	21M	SSD	MRes: Gaussian smoothing
FNIRT	Cubic B-splines	30K	SSD	Membrane energy* MRes: down- to up-sampling; number of basis components
IRTK	Cubic B-splines	1.4M	norm. MI	None used in the study; MRes: control mesh and image
JRD-fluid	Viscous fluid: variational calculus (diffeomorphic)	2M	Jensen-Rényi divergence	Compressible viscous fluid governed by the Navier-Stokes equation for conservation of momentum; MRes
ROMEO	Local affine (12 dof)	2M	Displaced frame difference	First-order explicit regularization method, brightness constancy constraint MRes: adaptive multigrid (octree subdivision), Gaussian smoothing
SICLE	3-D Fourier series (diffeomorphic)	8K	SSD	Small-deformation linear elasticity, inverse consistency MRes: number of basis components
SyN	Bi-directional diffeomorphism	28M	CC	MRes Gaussian smoothing of the velocity field, transformation symmetry
SPM5:				
“SPM2-type” Normalization	Discrete cosine transforms	1K	MSD	Bending energy, basis cutoff
Normalization	Discrete cosine transforms	1K	MSD	Bending energy, basis cutoff
Unified Segmentation	Discrete cosine transforms	1K	Generative segmentation model	Bending energy, basis cutoff
DARTEL Toolbox	Finite difference model of a velocity field (constant over time, diffeomorphic)	6.4M	Multinomial model (“congealing”)	Linear-elasticity; MRes: full-multigrid (recursive)

Table 1

Deformation model, approximate number of degrees of freedom (dof), similarity measure, and regularization method for each of the algorithms evaluated in this study. The dof is estimated based on the parameters and data used in the study; approximate equations, where available, are given in each algorithm’s description in the Supplementary section 8. Software requirements, input, and run time for the algorithms are in the Appendix. *Since this study was conducted, FNIRT uses bending energy as its default regularization method. MRes=multiresolution; norm=normalized; MSD=mean squared difference; SSD=sum of squared differences; CC=cross-correlation; CR=correlation ratio; MI=mutual information

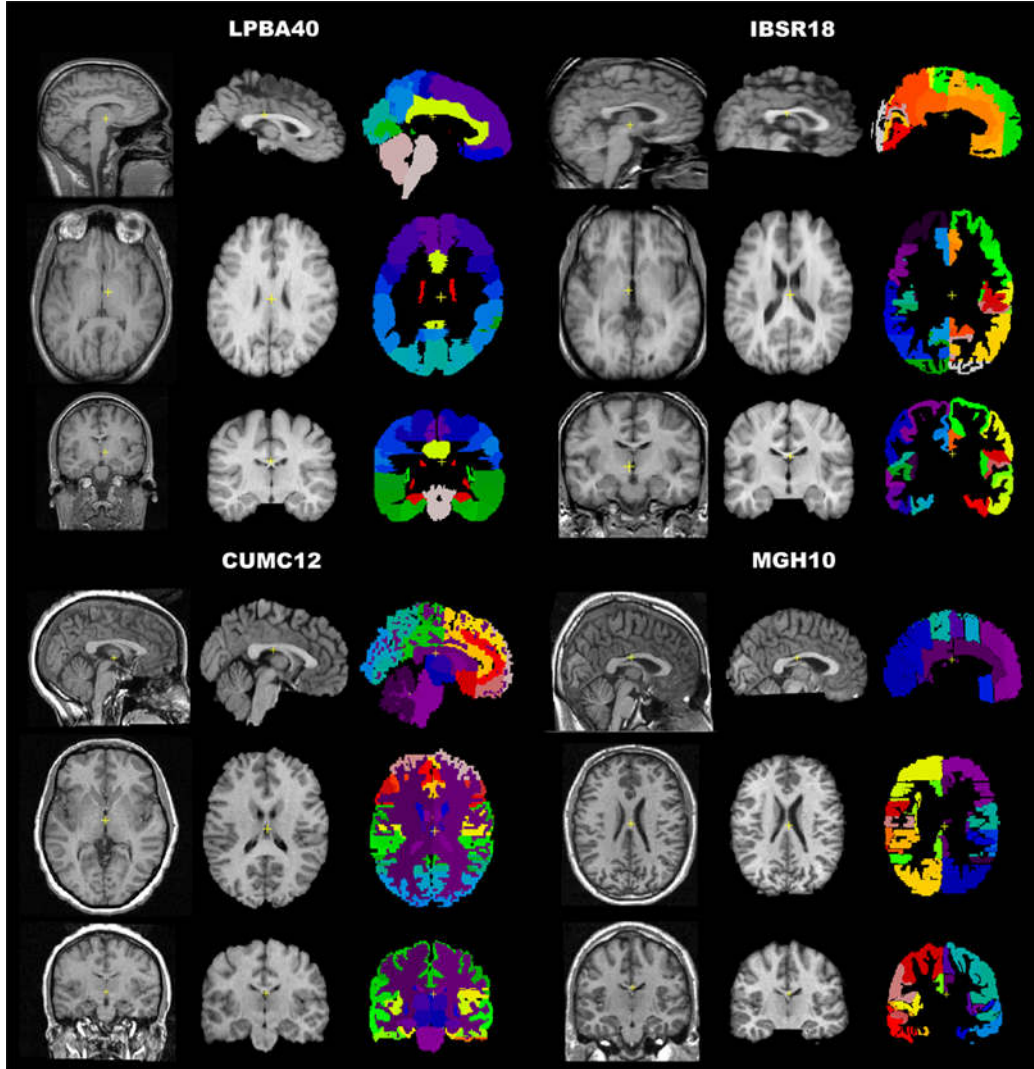


Fig. 1. Brain image data. The study used four different image datasets with a total of 80 brains. The datasets contain different numbers of subjects (n) and different numbers of labeled anatomical regions (r) derived from different labeling protocols: LPBA40 (LONI Probabilistic Brain Atlas: $n=40$, $r=56$), IBSR18 (Internet Brain Segmentation Repository: $n=18$, $r=84$), CUMC12 (Columbia University Medical Center: $n=12$, $r=128$), and MGH10 (Massachusetts General Hospital: $n=10$, $r=74$). A sample brain from each dataset is shown. For each brain, there are three columns (left to right): original T1-weighted MRI, extracted brain registered to nonlinear MNI152 space, and manual labels registered to nonlinear MNI152 space (used to extract the brain). Within each column the three rows (top to bottom) correspond to sagittal (front facing right), horizontal (front facing top, right on right side), and coronal (right on right side) views. The LPBA40 brains had already been extracted and registered to MNI (MNI305 vs. MNI152) space (Shattuck et al., 2008). The scale, position, and contrast of the MR images have been altered for the figure. The colors for the manual labels do not correspond across datasets.

2.1.1 Image acquisition and manual labels

Brain image data (T1-weighted MRIs and corresponding manual labels) for 80 normal subjects were acquired from four different sources (see Fig. 1 and Table 2, and Caveats section in the Discussion regarding label reliability):

LPBA40: 40 brain images and their labels used to construct the LONI Probabilistic Brain Atlas (LPBA40) at the Laboratory of Neuro Imaging (LONI) at UCLA (Shattuck et al., 2008) are available online (<http://www.loni.ucla.edu/Atlases/LPBA40>). They were preprocessed according to existing LONI protocols to produce skull-stripped brain volumes. These volumes were aligned to the MNI305 atlas (Evans et al., 1993) using rigid-body transformation to correct for head tilt and reduce bias in the manual labeling process. This produced a transform from native space to labeling space and an associated inverse transform. In each of the 40 subjects, 56 structures were manually labeled according to custom protocols (<http://www.loni.ucla.edu/Protocols/LPBA40>) using BrainSuite software (<http://brainsuite.usc.edu/>). Brain masks were constructed from the manual labels and projected back to the native (labeling) space to produce brain-only MRI volumes. These volumes were then corrected for non-uniformity using BrainSuite’s Bias Field Corrector. Sulci were used as boundaries; white matter voxels that occurred between the boundaries of sulci and their surrounding gray matter were included in the structure. This is the only dataset where white matter is included with gray-matter regions.

After all of the registrations were conducted, we found errors in two of the LPBA40 subjects, particularly with the right putamen. We brought this to LONI’s notice and it is being corrected for future downloads. The impact of these errors on the present study appears to be negligible, as may be seen in Figs. 7 and 13, where there appears to be little difference between the average values for the left and right putamen.

IBSR18: 18 brain images acquired at different laboratories are available through the Internet Brain Segmentation Repository (<http://www.cma.mgh.harvard.edu/ibsr/>) as IBSR v2.0. The T1-weighted images have been rotated to be in Talairach alignment (Talairach and Tournoux, 1988) and have been processed by the CMA (Center for Morphometric Analysis, Massachusetts General Hospital (MGH) in Boston) ‘autoseg’ bias field correction routines. They were manually labeled with NVM software (<http://neuromorphometrics.org:8080/nvm/>), resulting in 84 labeled regions.

CUMC12: 12 subjects were scanned at the Columbia University Medical Center on a 1.5 T GE scanner. Images were resliced coronally to a slice thickness of 3 mm, rotated into cardinal orientation, then segmented and manually labeled by one technician trained according to the Cardviews

Dataset	Subjects	Ages μ =mean	Volume (mm)	Voxel (mm)	TR (ms)	TE (ms)	flip $\angle$
LPBA40	40 (20 σ , 20 φ)	19–40 μ =29.20	256×256×124	38=0.86×0.86×1.5 2=0.78×0.78×1.5	10– 12.5	4.2– 4.5	20°
IBSR18	18 (14 σ , 4 φ)	7–71 μ =38.4 +4 “juve- niles”	256×256×128	8=0.94×0.94×1.5 6=0.84×0.84×1.5 4=1×1×1.5			
CUMC12	12 (6 σ , 6 φ) right-handed	26–41 μ =32.7	256×256×124	0.86×0.86×1.5	34	5	45°
MGH10	10 (4 σ , 6 φ)	22–29 μ =25.3	256×256×128	1×1×1.33	6.6	2.9	8°

Table 2

MRI acquisition parameters. Dataset, number and ages of subjects, volume and voxel dimensions in native space, TR, TE, and flip angle. The images were registered to either the nonlinear MNI152 or MNI305 atlas (see text) in a 181×217×181 volume of 1mm³ voxels.

labeling scheme (Caviness et al., 1996) created at the CMA, and implemented in Cardviews software (<http://www.cma.mgh.harvard.edu/manuals/parcellation/>). The images have 128 labeled regions.

MGH10: 10 subjects were scanned at the MGH/MIT/HMS Athinoula A. Martinos Center for Biomedical Imaging using a 3 T Siemens scanner and standard head coil. The data were inhomogeneity-corrected, affine-registered to the MNI152 template (Evans et al., 1992), and segmented using SPM2 software (Friston et al., 1995). The images were manually labeled by Tourville of Boston University using Ghosh’s ASAP software (Nieto-Castanon et al., 2003); the labeling protocol (Tourville and Guenther, 2003) is similar to Cardviews, and in the version used for this study produces 74 labeled regions.

2.1.2 Brain extraction

To register the brains with each other, we extracted each brain from its whole-head image by constructing a mask from the corresponding manually labeled image (see Fig. 1). However, since white matter and cerebrospinal fluid were not fully labeled in all of the images, they had to be filled to create solid masks. For this, the non-background image in each sagittal slice was dilated by one pixel, any holes were filled, and then the image was eroded by one pixel. This procedure was repeated sequentially on the resulting volume for the coronal, horizontal, and again for the sagittal slices, and resulted in a volume containing the filled brain mask. This manual label-based skull-stripping procedure was performed on each MRI volume in the IBSR18, CUMC12, and MGH10 sets, but not for those in the LPBA40 set; the LPBA40 images had already been similarly prepared, but dilated and eroded with a larger and spherical structural element (neighborhood) (Shattuck et al., 2008). All four SPM algorithms were also run on whole-head images.

2.1.3 File preparation

All image and label volumes were in right-handed orientation and were converted to Analyze 7.5 (.img, .hdr) format (except for MINC format used by ANIMAL) because it was the most common image format accepted by the different software programs, and the only format presently compatible with AIR, ART, JRD-fluid, and SICLE (see Appendix B). This itself was a cause of difficulties, because the different software packages deal with Analyze header information differently, in particular with respect to left-right flipping and origin location. Because of this and because of discrepancies between brain and atlas origins for some of the data sets, origin and orientation information was removed from each of the image and label volumes using FSL’s “fslorient -deleteorient” and “fslchfiletype” commands. The NiFTI data format, accepted by most of the f/MRI software packages, obviates these concerns and is recommended over the Analyze format (<http://nifti.nimh.nih.gov/>). Exceptions to the above steps were made for SPM5’s template-based algorithms (Normalization, Unified Segmentation, and DARTEL Toolbox, but not “SPM2-type” Normalization): Analyze images were flipped right-to-left to left-handed orientation, and header orientation discrepancies were corrected using `spm_get_space.m` (other algorithms were unaffected after the `fslorient` command above).

Some extra preparation had to be done to accommodate the recommendations for running the individual software packages (see Appendix B), which included writing parameter files, intensity correction, padding, smoothing, and reorientation (in the case of SPM). For example, parameter files were required for ROMEO, IRTK, and for each registration pair when using SICLE, and command-line parameters had to be reset to make some of the programs run in less than an hour or so per registration. SICLE required considerable preparation: we wrote a Python script to generate the input parameter files and create output directories, normalized intensities in Matlab, and padded versions of all of the image volumes so that their dimensions were divisible by 16 (e.g., $181 \times 217 \times 181$ files were padded to $224 \times 224 \times 192$).

2.2 Linear registration as initialization

We linearly registered 40 of the brain images to a template using FMRIB Software Library’s (FSL) FLIRT (with the following settings: 9-parameter, correlation ratio, trilinear interpolation; see Fig. 1). The template was the “nonlinear MNI152,” the nonlinear average template in MNI space used by FSL (MNI152_T1_1mm_brain: $181 \times 217 \times 181$ voxels, $1 \times 1 \times 1$ mm/voxel). The remaining 40 images were from the LPBA40 set and had already been registered to the MNI305 atlas.

We then rigidly registered each of the 80 brains in MNI space, I_s , to each of the other brains in its group, I_t , again using FLIRT (6-parameter, correlation ratio, trilinear interpolation). This resulted in 2,168 linear transforms $X_{s \rightarrow t}$ and transformed images in MNI space $I_{s \rightarrow t}$ (a straight arrow denotes linear registration), with 2,088 of them representing non-identical source-target pairs ($40^2 + 18^2 + 12^2 + 10^2 - 80$). These linearly transformed source images, or “linear source images,” serve as the input to each of the algorithms under comparison.

We applied the above linear and rigid-body transforms (with nearest-neighbor interpolation) to the corresponding manually labeled volumes L_s , resulting in the “linear source labels” $L_{s \rightarrow t}$ below (and in Figs. 2 and 3).

2.3 Nonlinear registration

Each of the nonlinear registration algorithms in the study then registered each of the 2,168 linear source images $I_{s \rightarrow t}$ to its corresponding target image I_t . We applied the resulting nonlinear transformation $X_{[s \rightarrow t] \rightsquigarrow t}$ (with nearest-neighbor interpolation) to the corresponding linear source labels $L_{s \rightarrow t}$, producing warped source labels $L_{[s \rightarrow t] \rightsquigarrow t}$ (a curved arrow denotes nonlinear registration). These labels are compared against the manual labels of the target, L_t , for evaluating registration performance. See Figs. 2 and 3 for the context and Supplementary section 7 for the software commands used for each algorithm. Note that some structures were removed during preprocessing prior to computing the transforms, such as the cerebellum in the LPBA40 set, but were included when applying the transforms to the source labels.

2.4 Evaluation measures

We used volume and surface overlap, volume similarity, and distance measures to evaluate how well individual anatomical regions as well as total brain volumes register to one another. For this section and for Fig. 4, source S refers to a registered image to be compared with its registration target T (in our case, the warped source labels $L_{[s \rightarrow t] \rightsquigarrow t}$ and the target labels L_t). These evaluation measures assume the manual label sets are correct, or “silver standards.”

2.4.1 Volume overlap

We used three overlap agreement measures and two overlap error measures, each quantifying some fraction of source S and target T volumes where their labels agree or disagree. For information on overlap measures, including cases for multiple and fractional labels, see (Crum et al., 2005). The first overlap

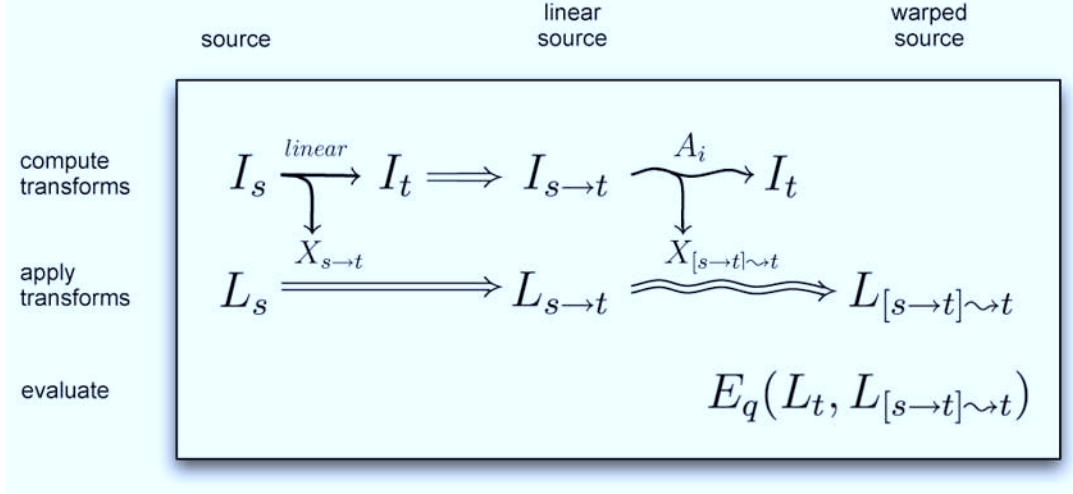


Fig. 2. Registration equations. The three stages of the study were to compute, apply, and evaluate registration transforms. To compute the transforms, we linearly registered each source image I_s to a target image I_t (both already in MNI space), resulting in a “linear source image” $I_{s \rightarrow t}$ as well as a linear transform $X_{s \rightarrow t}$ (a straight arrow denotes linear registration). Each nonlinear algorithm A_i then registered (warped) the linear source image to the same target image, generating a second, nonlinear transform $X_{[s \rightarrow t] \rightsquigarrow t}$ (a curved arrow denotes nonlinear registration). We applied the linear transform to the source labels L_s to give the corresponding “linear source labels” $L_{s \rightarrow t}$, and applied the nonlinear transform to $L_{s \rightarrow t}$ to produce the final warped source labels $L_{[s \rightarrow t] \rightsquigarrow t}$. Finally, we compared these labels to the manual labels for the target, L_t , using a set of evaluation measures E_q .

agreement measure is the “target overlap,” TO , the intersection between two similarly labeled regions r in S and T divided by the volume of the region in T , where $||$ indicates volume computed as the number of voxels:

$$TO_r = \frac{|S_r \cap T_r|}{|T_r|} \quad (1)$$

Target overlap is a measure of sensitivity. When summed over a set of multiple labeled regions, we have the total overlap agreement measure for a given registration:

$$TO = \frac{\sum_r |S_r \cap T_r|}{\sum_r |T_r|} \quad (2)$$

Our second overlap agreement measure is the “mean overlap,” MO , a special case of the Kappa coefficient (Zijdenbos et al., 1994) sometimes called the Dice coefficient; it is the intersection divided by the mean volume of the two regions, which may again be summed over multiple regions:

$$MO = 2 \frac{\sum_r |S_r \cap T_r|}{\sum_r (|S_r| + |T_r|)} \quad (3)$$

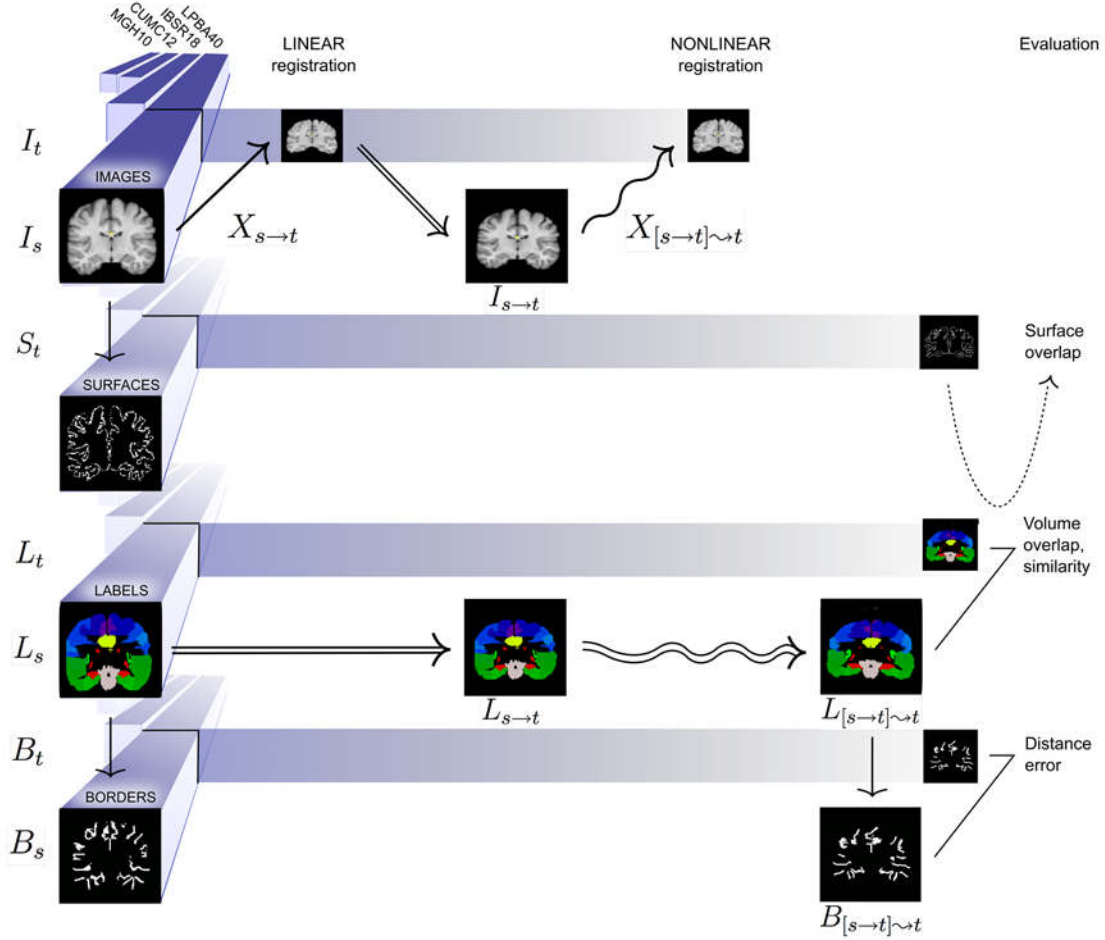


Fig. 3. Overview. This diagram provides an overview of the study for a single nonlinear registration algorithm, placing example preprocessed data from Figure 1 into the equations of Figure 2. The three stages include linear registration, nonlinear registration, and evaluation (left to right). The four different datasets (LBPA40, IBSR18, CUMC12, and MGH10) are aligned along the left in four different versions: images, surfaces derived from the images, labels, and borders derived from the labels. A source and target are drawn from each version (image volumes are shown as coronal slices for clarity). A source image I_s is linearly then nonlinearly registered to a target image I_t . The linear and nonlinear transforms ($X_{s \rightarrow t}$ and $X_{[s \rightarrow t] \rightsquigarrow t}$) are applied to the corresponding source labels L_s . The resulting nonlinearly transformed labels $L_{[s \rightarrow t] \rightsquigarrow t}$ are compared against the target labels L_t . This comparison is used to calculate volume overlap and volume similarity per region. The target surface S_t is intersected with the target labels L_t and warped source labels $L_{[s \rightarrow t] \rightsquigarrow t}$ to calculate surface overlap. Borders between each labeled region and all adjacent labeled regions are constructed from L_t and $L_{[s \rightarrow t] \rightsquigarrow t}$, and average distances between the resulting borders B_t and $B_{[s \rightarrow t] \rightsquigarrow t}$ are calculated per region.

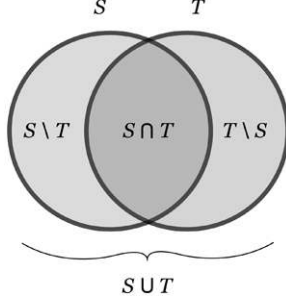


Fig. 4. Overlap. This study uses volume and surface overlap, volume similarity, and distance measures to evaluate the accuracy of registrations. The equations for the three overlap measures: target overlap, mean overlap, and union overlap use the terms in this schematic Venn diagram of two partially overlapping objects, a source S and a target T . Their intersection is denoted by $S \cap T$ and their union by $S \cup T$. $S \setminus T$ indicates the set (theoretic complement) of elements in S but not in T .

Our third overlap agreement measure is the “union overlap,” UO , or Jaccard coefficient (Gee et al., 1993; Jaccard, 1912), the intersection over the union:

$$UO = \frac{\sum_r |S_r \cap T_r|}{\sum_r |S_r \cup T_r|} \quad (4)$$

UO can be converted to MO by the following (Heckemann et al., 2006):

$$MO = \frac{2 \times UO}{1 + UO} \quad (5)$$

To complement the above agreement measures, we also computed false negative (FN) and false positive (FP) errors. For these errors we characterize the source as a tentative set of labels for the target, and again assume that the target’s manual labels are correct. These error measures can range from zero to one; a value of zero is achieved for perfect overlap.

A false negative error for a given region is the measure of how much of that region is incorrectly labeled. It is computed as the volume of a target region outside the corresponding source region divided by the volume of the target region. As before, it is computed in voxels and summed over a set of multiple labeled regions each with index r :

$$FN = \frac{\sum_r |T_r \setminus S_r|}{\sum_r |T_r|} \quad (6)$$

where $T_r \setminus S_r$ indicates the set (theoretic complement) of elements in T_r but not in S_r .

A false positive error for a given region is the measure of how much of the volume outside that region is incorrectly assigned that region’s label. It is

computed as the volume of a source region outside the corresponding target region divided by the volume of the source region:

$$FP = \frac{\sum_r |S_r \setminus T_r|}{\sum_r |S_r|} \quad (7)$$

2.4.2 Surface overlap

We anticipated that imaging artifacts affecting cortical thickness could bias our overlap measures, because (for the same cortical area) thicker regions will have relatively higher volume overlap agreements than thinner regions due to lower surface-to-volume ratios. We tried to reduce this bias by computing overlap agreement only on the target surfaces of the brain images, not throughout the entire target volumes. Computing overlap agreement on the surfaces should also decrease the impact of segmentation biases, when manual labels extend into white matter, especially for the LPBA40 set, where white matter between sulcal structures were also assigned the structures’ labels.

We used Freesurfer software (<http://surfer.nmr.mgh.harvard.edu/>, version 1.41) to construct cerebral cortical surfaces (Dale et al., 1999) for each of the original 80 full-head images, and converted the Freesurfer-generated surfaces to each brain’s native space with Freesurfer’s “mri_surf2vol” command. We then linearly registered each surface to MNI space using the initial affine transform from the original brain image to the MNI template (“Linear registration as initialization” section). Each resulting target surface was intersected with its corresponding target label volume L_t and warped source label volume $L_{[s \rightarrow t] \sim t}$. We compared these target surface labels with the warped source surface labels using the same overlap agreement and error measures used for the volumes.

2.4.3 Volume similarity

The volume similarity coefficient, VS , is a measure of the similarity between source and target volumes. Although this measure does not reflect registration accuracy (source and target regions can be disjoint and still have equal volumes), it is a conventional measure included for retrospective evaluation of prior studies. It is equal to the differences between two volumes divided by their mean volume, here again summed over multiple regions:

$$VS = 2 \frac{\sum_r (|S_r| - |T_r|)}{\sum_r (|S_r| + |T_r|)} \quad (8)$$

2.4.4 Distance error

The above overlap and volume similarity measures do not explicitly account for boundary discrepancies between corresponding source and target regions. So we chose our final evaluation measure, DE is equal to the minimum distance, $mindist$, from each source region boundary point, $S_r B_p$, to the entire set of points making up the target region boundary, $T_r B$, averaged across P points:

$$DE_r = \frac{1}{P} \sum_{p=1}^P mindist(S_r B_p, T_r B) \quad (9)$$

We extracted an approximation of the boundary points for each region of each of the 40 LPBA40 brains by applying a cityblock distance transform² in Matlab and retaining only those voxels of neighboring regions that were within two voxels from the region. This resulted not in a complete shell about a region, but only the portion of the shell abutting other labeled regions. We repeated this procedure for each region of each of the warped LPBA40 source labels generated by each registration algorithm. We chose to construct borders from the warped labels rather than warp borders constructed from the original labels because we were concerned about interpolation artifacts.

We applied the same distance function used to construct the borders to also compute DE between source and target borders. We computed DE for each region as well as for the entire set of label boundaries as a whole.

2.5 Analysis

Testing for significant differences in the performance of the registration methods is not trivial because of non-independency of samples. For example, for the LPBA40 dataset, each of the 40 brain images was registered to the 39 others, resulting in 1,560 pairwise registrations. Each of the brains is represented 39 times as the registration source and 39 times as the target. Because each brain is reused multiple times, independence of observations cannot be assumed. We determined that for most of the registration methods, there is a high correlation between overlap results obtained for pairs that share one or more brains (see Supplementary section 6).

To get around this issue of non-independency of samples, we conducted two separate statistical tests, a permutation test and a one-way ANOVA test, on a small independent sample, and repeated these tests on multiple such samples. We also conducted an indifference-zone ranking on the entire set of results,

² `bwdist.m` in the Image Processing toolbox uses the two-pass, sequential scanning algorithm (Rosenfeld and Pfaltz, 1966; Paglieroni, 1992)

testing for practical rather than statistical significance (see below). For each test, the underlying measure is target overlap averaged across all regions.

2.5.1 Permutation tests

We performed permutation tests to determine if the means of a small set of independent overlap values obtained by each of the registration methods are the same, after (Menke and Martinez, 2004) and according to the following permutation algorithm:

1. Select a subset of P independent brain pairs
2. Select a pair of methods (two vectors of P total overlap values)
3. Subtract the two vectors and compute the mean difference D
4. Select a subset of the elements from one of the vectors
5. Swap this subset across the two vectors
6. Subtract the resulting vectors; compute the mean difference D_p
7. Repeat steps #4-6 N times
8. Count the number of times n where³ $\text{abs}(D_p) \geq \text{abs}(D)$
9. Compute the exact p-value: $p = \frac{n}{N}$
10. Repeat steps #1-9; compute the fraction of times where $p \leq 0.05$

The subset of brain pairs was selected so that each brain was used only once, corresponding to the “no dependence” condition in Supplementary section 6. There were 20, 9, 6, and 5 independent brain pairs for the LPBA40, IBSR18, CUMC12, and MGH10 datasets, respectively, as well as 20, 9, 6, and 5 corresponding average target overlap values obtained by each method.

The number of permutations N for each subset of brain pairs was either the exhaustive set of all possible permutations ($2^{12}=4,096$ for CUMC12 and $2^{10}=1,024$ for MGH10) or 1,000 permutations (LPBA40 and IBSR18) to keep the duration of the tests under 24 hours. The number of p-values calculated was either 100,000 (CUMC12 and MGH10) or 10,000 (LPBA40 and IBSR18).

2.5.2 One-way ANOVA

We also performed a standard one-way ANOVA to test if the means of similar subsets of independent average target overlap values obtained by each of the registration methods are the same. We then subjected these results to a multiple comparison test using Bonferroni correction to determine which pairs of means are significantly different (disjoint 95% confidence intervals about the means, based on critical values from the t distribution). We repeated these

³ $\text{abs}()$ =absolute value

ANOVA and multiple comparison tests 20 times, each time randomly selecting independent samples from each of the datasets. These tests are not expected to be as accurate as the permutation tests because some of the overlap values have skew distributions and because the p-values are not exact.

2.5.3 Indifference-zone ranking

Our third evaluation between methods tested practical significance rather than statistical significance. For example, if a region is registered to another region of equal volume and results in an offset of a single voxel, this is not considered a significant misregistration, but offsets greater than this are considered significant. An evaluation measure of registration accuracy for a given region within a given brain pair is calculated for two different registration methods. If these two values are within delta of one another (referred to as an “indifference zone” when ranking (Bechhofer, 1954)), they are considered equal. The delta must correspond to a practical difference in registration. If we model a region as a cube, then a single-voxel offset along the normal to one of its faces would mean the voxels on that face of the cube reside outside of its target $\tilde{N}$ this is equal to one-sixth of its surface. We therefore set delta to one-sixth of a target region’s surface. For the IBSR18, CUMC12, and MGH10 datasets, we assumed the surface to be that of a cube ($6 \times edge^2 - 12 \times edge$, where $edge$ = the edge length of a cube with the volume of the target region, in voxels). For the LPBA40 dataset, we set the surface to the number of voxels bordering adjacent regions, extracted as in the “Distance error” section.

Our implementation of indifference-zone ranking compared the 15 different registration methods to each other in the following manner. For each region in a given label set and for each pair of registered brains we constructed a 15×15 matrix, where each row and each column corresponded to a registration method. Each element of the matrix was assigned the value -1 , 0 , or 1 , for the cases when the evaluation measure for the method corresponding to its row was at least delta less than, within delta of, or at least delta greater than that of the method corresponding to its column. Then we calculated the mean of these $\{-1, 0, 1\}$ values across all registration pairs for each region to construct Figs. 7, 8, 9, and 10 (the latter three in Supplementary section 3).

3 Results

Results for the initial run are in Supplementary section 1, for the trivial case, where each brain was registered to itself, are in Supplementary section 2, volume similarity results are in Supplementary section 4, and distance error results are in Supplementary section 5.

3.1 *Overlap results*

3.1.1 *Whole-brain averages*

After the initial run and changes described in Supplementary section 1, out of 2,168 registrations per algorithm, the only cases where target overlap values were less than 25% were SPM’s DARTEL (79 cases; the majority were from one brain) Normalize (15 cases), ANIMAL (2 cases), and ROMEO (1 case) for the LPBA40 set and Diffeomorphic Demons (1 case) for the IBSR18 set.

The target, union, and mean overlap values for volumes as well as surfaces (and the inverse of their false positive and false negative values), averaged over all regions, gave almost identical results when corrected for baseline discrepancies. Distributions of target overlap values are shown in Fig. 5. What is remarkable is that the relative performances of these methods appear to be robust not just to type of overlap measure, but also to subject population and labeling protocol, as evidenced by the similar pattern of performances of the methods across the label sets. This is particularly the case across IBSR18, CUMC12, and MGH10 sets. The pattern is more subtle in LPBA40 because that label set has fewer labeled regions that are larger and extend into white matter, and therefore results in higher and more similar absolute overlap values.

We ran all 2,168 registrations again on whole-head images (before skull-stripping) using SPM’s Normalize, Unified Segmentation, and DARTEL, and the results were comparable or better with the skull-stripped images. The relative overlap performance of the SPM programs agrees with Yassa and Stark (Yassa and Stark, 2009): DARTEL performs better than Unified Segmentation which performs better than Normalize. Because the SPM DARTEL results were very similar for its original and pairwise implementations, we have included only the pairwise results; this is a fair comparison because the other methods do not include optimal average template construction.

3.1.2 *Region-based results*

The pattern of region-based overlap values is almost indistinguishable across the methods, discounting baseline differences (data not shown). In Fig. 6 we present volume and surface target overlap data for individual regions in their anatomical context (LPBA40 set). For the most part this figure suggests that the overlap values are approximately the same for volume and surface measures, corroborating whole-brain averages, but also exposes discrepancies at the level of regions (FLIRT and SICL)⁴.

⁴ The worse surface overlaps of the cerebellum (for all the methods except ROMEO) are probably due to the fact that the cerebellum was removed from the LPBA40 set

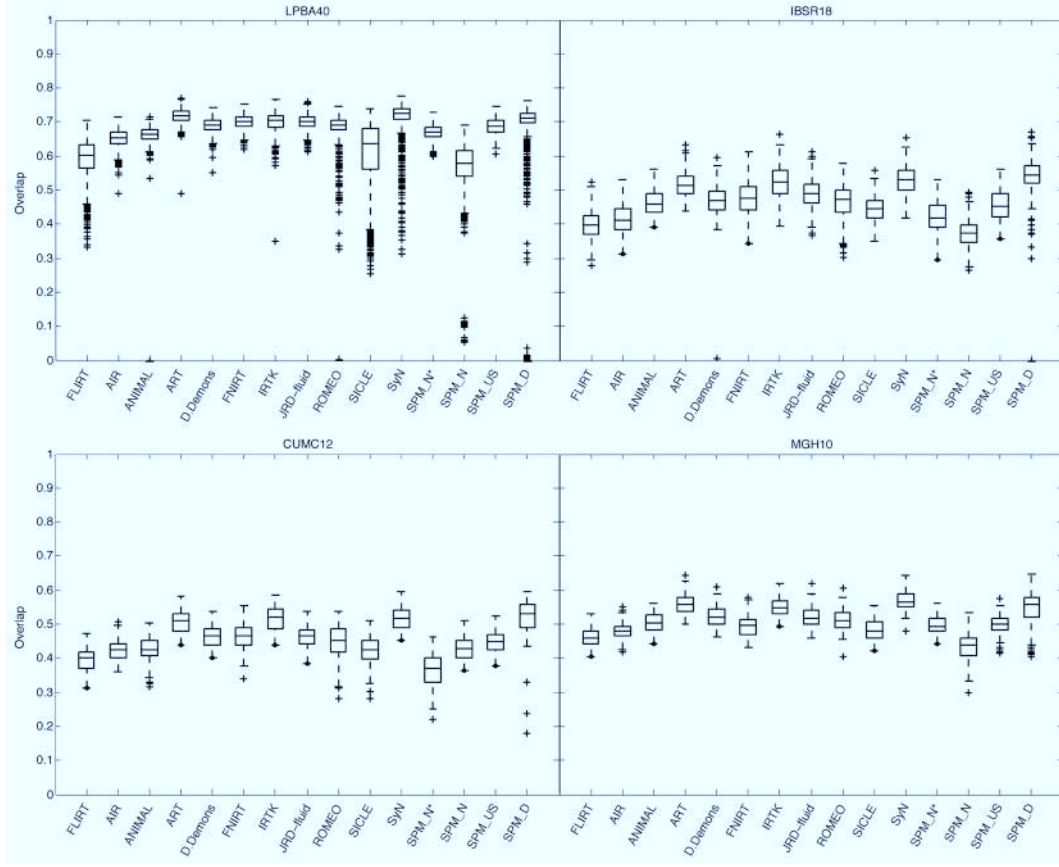


Fig. 5. Overlap by registration method. These box and whisker plots show the target overlap measures between deformed source and target label volumes averaged first across all of the regions in each label set (LPBA40, IBSR18, CUMC12, and MGH10) then across brain pairs. Each box represents values obtained by a registration method and has lines at the lower quartile, median, and upper quartile values; whiskers extend from each end of the box to the most extreme values within 1.5 times the interquartile range from the box. Outliers (+) have values beyond the ends of the whiskers. Target, union and mean overlap measures for volumes and surfaces (and the inverse of their false positive and false negative values) all produced results that are almost identical if corrected for baseline discrepancies. Similarities between relative performances of the different registration methods can even be seen here across the label sets. (SPM_N*="SPM2-type" normalization, SPM_N=SPM's Normalize, SPM_US=Unified Segmentation, SPM_D=DARTEL pairwise)

Most of the regions in the brain volume plots are hidden from view, so for a complete picture at the scale of individual regions, Figs. 7, 8, 9, and 10 present relative performances of the different methods for each region as color-coded tables for each of the four label sets (their construction is described in the "Indifference-zone ranking" section under "Materials and methods"; Figs. 8, 9, and 10 are in Supplementary section 3). If all of the methods had performed

prior to computing the registration transforms, but the transforms were applied to the full label set (including the cerebellum).

equally well, the color tables would be a uniform color. However, some of the methods performed better than average, particularly against simple linear registration (FLIRT). By visual inspection, we can see that ART, IRTK, SyN, and SPM’s DARTEL have consistently high accuracy for the IBSR18, CUMC12, and MGH10 label sets relative to the other methods, and that in addition to ART, IRTK, and SyN, FNIRT and JRD-fluid also appear to have high relative accuracy for the LPBA40 set. As expected, we observed for all of the methods higher overlap values for larger sized regions, because of smaller surface-to-volume ratios (not shown).

3.2 *Rankings*

We ranked the registration methods in three independent ways: permutation tests, confidence intervals obtained from one-way ANOVA tests with Bonferroni correction, and indifference-zone ranking.

3.2.1 *Permutation, ANOVA, and indifference-zone rankings*

Table 3 presents the top three ranks of registration methods according to the percentage of permutation tests whose p-values were less than or equal to 0.05, and Table 4 according to relative target overlap scores. For both tables, members within ranks 1, 2, and 3 have means lying within one, two, and three standard deviations of the highest mean, respectively. Only ART and SyN are in the top rank for all four label sets and for all tests.

For the one-way ANOVA tests, rank 1 methods have means lying within the 95% confidence interval of the best method and rank 2 methods have confidence intervals that overlap the confidence interval of the best method. These rankings were in almost complete agreement among the target, union, and mean overlap values (and distance errors for the LPBA40 set). Because these results were very similar to the permutation test ranks, and because these tests are expected to be less accurate than the permutation tests, they are not included.

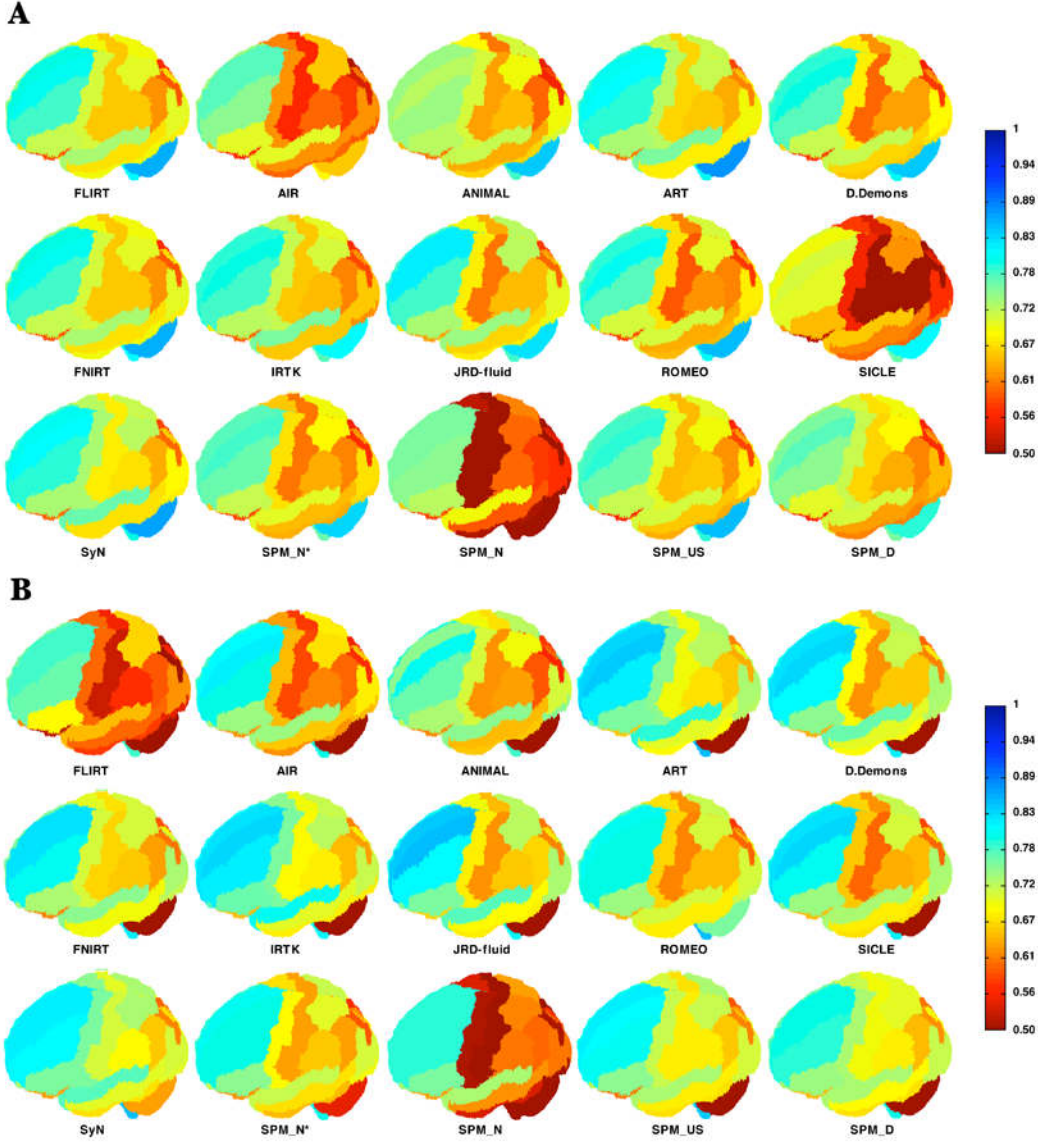


Fig. 6. Volume and surface overlap by registration method: LPBA40 regions. These brain images show the mean target overlap calculated across all 1,560 brain pairs for the (A) volume and (B) surface of each LPBA40 region, and depicts that mean as a color (blue indicates higher accuracy). The values for each registration method are projected on one of the LPBA40 brains, seen from the left, looking down from 30° , with the frontal pole facing left. (SPM_N*="SPM2-type" Normalize, SPM_N=Normalize, SPM_US=Unified Segmentation, SPM_D=DARTEL pairwise)

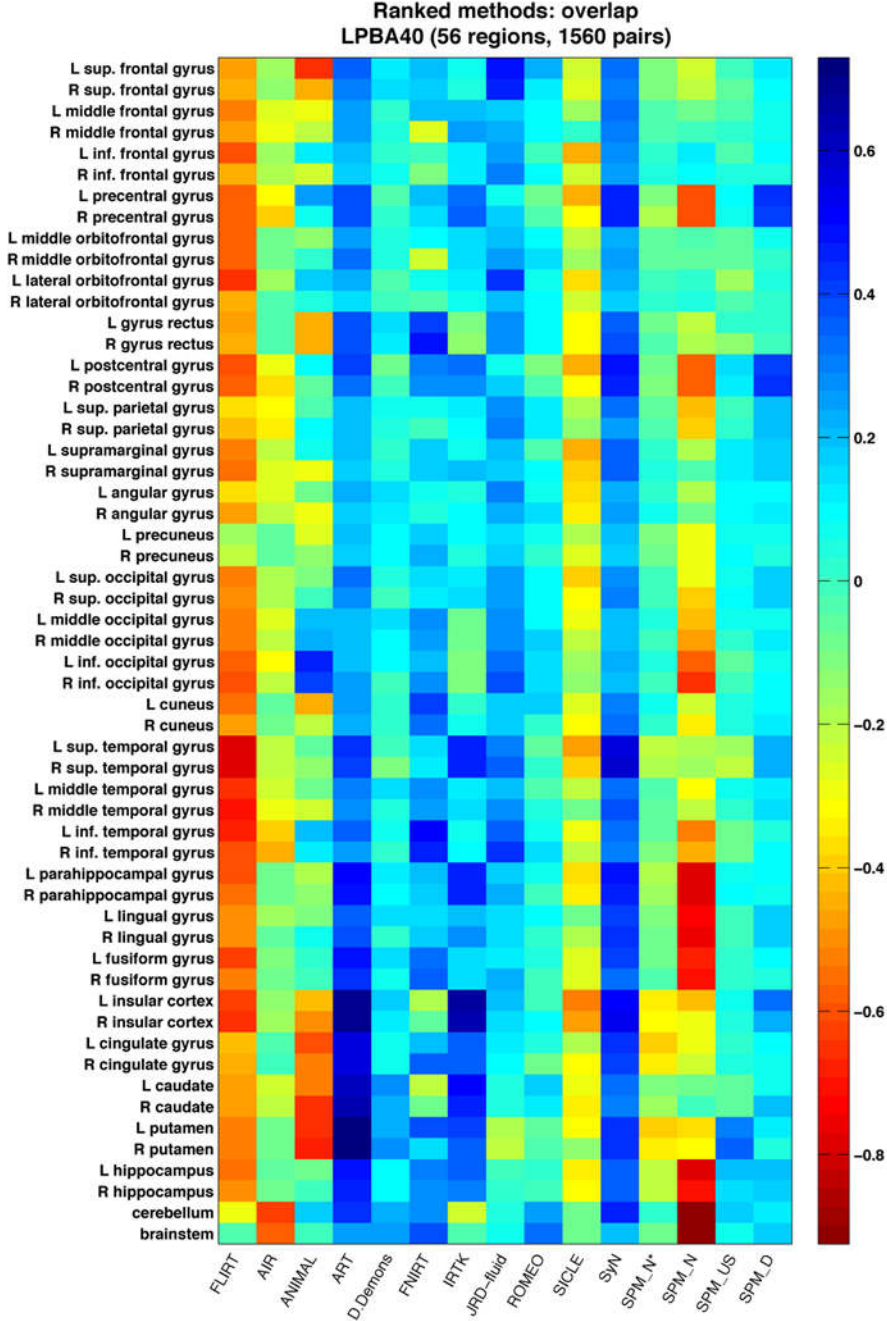


Fig. 7. Indifference-zone ranking of the registration methods: LPBA40 overlaps. This matrix uses a color scale that reflects the relative performance of the registration methods (with blue indicating higher accuracy). Each colored rectangle represents the average score for a given method for a given region, averaged over 1,560 LPBA40 registrations. The scores are $\{-1, 0, 1\}$ values indicating the pairwise performance of the method relative to each of the other methods (see text), according to target volume overlap (union and mean overlap results are almost identical). The colors (and color range) are not comparable to those of the other label sets (Figures 8, 9, and 10 in Supplementary section 3). (SPM_N*="SPM2-type" Normalize, SPM_N=Normalize, SPM_US=Unified Segmentation, SPM_D=DARTEL pairwise)

	LPBA40	μ (SD)	IBSR18	μ (SD)	CUMC12	μ (SD)	MGH10	μ (SD)
rank 1	ART	.82 (.35)	SPM_D	.83 (.27)	SPM_D	.76 (.24)	SyN	.77 (.37)
	SyN	.60 (.38)	SyN	.72 (.51)	SyN	.74 (.51)	ART	.72 (.45)
	FNIRT	.49 (.66)	IRTK	.67 (.53)	IRTK	.74 (.50)	IRTK	.61 (.51)
	JRD-fluid	.49 (.66)	ART	.60 (.70)	ART	.60 (.70)		
2	IRTK	.43 (.63)	JRD-fluid	.30 (.82)			SPM_D	.27 (.23)
	D.Demons	.13 (.82)					D.Demons	.27 (.69)
	SPM_US	.11 (.83)					JRD-fluid	.24 (.66)
							ROMEO	.06 (.63)
3	ROMEO	.08 (.73)	FNIRT	.16 (.82)	D.Demons	.20 (.84)		
	SPM_D	.07 (.29)	D.Demons	.05 (.84)	FNIRT	.18 (.81)		
					JRD-fluid	.17 (.81)		

Table 3

Permutation test ranking of the registration methods by label set. This table lists the methods that attained the top three ranks after conducting permutation tests between mean target overlaps (averaged across regions) for each pair of methods, then calculating the percentage of p-values less than or equal to 0.05 (of 100,000 tests for CUMC12 and MGH10 or of 10,000 tests for LPBA40 and IBSR18; μ =mean, SD=standard deviation). Methods within ranks 1, 2, and 3 have positive mean percentages lying within one, two, and three standard deviations of the highest mean, respectively. Values are not comparable across label sets (columns). (SPM_D=DARTEL pairwise)

	LPBA40	μ (SD)	IBSR18	μ (SD)	CUMC12	μ (SD)	MGH10	μ (SD)
rank 1	ART	.35 (.07)	SPM_D	.50 (.19)	SPM_D	.47 (.17)	SyN	.39 (.06)
	SyN	.34 (.24)	SyN	.40 (.12)	IRTK	.42 (.07)	ART	.36 (.07)
			IRTK	.35 (.15)	SyN	.41 (.06)		
			ART	.33 (.08)	ART	.35 (.05)		
2			JRD-fluid	.18 (.13)				
3	JRD-fluid	.20 (.08)	FNIRT	.06 (.11)	JRD-fluid	.07 (.07)	IRTK	.26 (.07)
	IRTK	.18 (.15)	D.Demons	.01 (.08)	FNIRT	.07 (.09)	SPM_D	.25 (.28)
	FNIRT	.17 (.08)	ROMEO	.01 (.28)	D.Demons	.05 (.05)		
	SPM_D	.14 (.31)						

Table 4

Indifference-zone ranking of the registration methods by label set. This table lists the methods that attained the top three ranks after averaging scores across all brain regions then across all registration pairs (μ =mean, SD=standard deviation). The scores reflect a pairwise comparison between methods, according to target overlap (see text). Methods within ranks 1, 2, and 3 have positive means lying within one, two, and three standard deviations of the highest mean, respectively. Values are not comparable across label sets (columns). (SPM_D=DARTEL pairwise)

4 Discussion

This study evaluates 15 registration algorithms (one linear, 14 nonlinear) based primarily on overlap measures of manually labeled anatomical regions. The scale and thoroughness are unprecedented (over 45,000 registrations, 80 manually labeled brain images representing 4 different labeling protocols, 8 different evaluation measures, and 3 independent analysis methods). We hope that the method of evaluation as well as the results will be useful to the neuroscience community. As they become available, additional materials and updated information will be made publicly available via the website <http://www.mindboggle.info/papers/>.

One of the most significant findings of this study is that the relative performances of the registration methods under comparison appear to be little affected by the choice of subject population, labeling protocol, and type of overlap measure. This is important because it suggests that the findings are generalizable to new healthy subject populations that are labeled or evaluated using different labeling protocols. Furthermore, we ranked the methods according to three completely independent analyses and derived three almost identical top rankings. However, in order to make recommendations, it is important to place these results in the context of the wider range of software packages available and the caveats inherent in registration in general and with respect to this study in particular, as we do below.

Although we were not able to see a pattern in the results that would allow us to rank algorithms by deformation model, similarity measure, or regularization method, there is a modest correlation between the number of degrees of freedom of the deformation and registration accuracy (0.29, or 0.45 if one excludes Diffeomorphic Demons), and between the number of degrees of freedom and year (0.55) (see Table 5). This finding corroborates Hellier’s evaluation: “The global measures used show that the quality of the registration is directly related to the transformation’s degrees of freedom” (Hellier et al., 2003). The four algorithms whose mean rank is less than two (SyN, ART, IRTK, and SPM’s DARTEL Toolbox) all have millions of degrees of freedom and all took at least 15 min per registration, and all but one (IRTK) were created in the last three years. Of the remaining 10 algorithms, seven have fewer than a million degrees of freedom, seven took less than 15 min, and six were created over three years ago.

Algorithm	mean rank	dof	run time: minutes	year
SyN	1.00	28M	77 (15.1)	2008
ART	1.00	7M	20.1 (1.6) [Linux]	2005
IRTK	1.63	1.4M	120.8 (29.3)	1999
SPM5 DARTEL Toolbox	1.88	6.4M	71.8 (6.3)	2007
JRD-fluid	2.50	2M	17.1 (1.0) [Solaris]	2007
Diffeomorphic Demons	3.00	21M	8.7 (1.2)	2007
FNIRT	3.00	30K	29.1 (6.0)	2008
ROMEO	3.50	2M	7.5 (0.5)	2001
ANIMAL		69K	11.2 (0.4)	1994
SICLE		8K	33.5 (6.6)	1999
SPM5 Unified Segmentation		1K	$\simeq 1$	2005
“SPM2-type” Normalize		1K	$\simeq 1$	1999
SPM5 Normalize		1K	$\simeq 1$	1999
AIR		168	6.7 (1.5)	1998

Table 5

Mean rank, degrees of freedom (dof), average run time, and year of publication for each algorithm. The 14 nonlinear deformation algorithms are ordered by mean rank (best at top), which was computed for each algorithm by averaging the target overlap ranks in Tables 3 and 4 (assigned by the permutation tests and indifference-zone rankings). The six algorithms at the bottom are of equal rank (4) since they were not in the top three ranks. For details on architecture and run time, see Appendix. Except for FNIRT and Diffeomorphic Demons, the dof and mean rank sequences roughly match.

4.1 Algorithms excluded from the study

We excluded semi-automated approaches that require even minimal manual intervention to reduce bias. A significant example is the forerunner of modern nonlinear registration methods, the original Talairach coordinate referencing system (Talairach and Szikla, 1967; Talairach and Tournoux, 1988), a piece-wise linear registration method that requires the identification of landmarks in a brain image. Although the Talairach system is well suited to labeling regions proximal to these landmarks (Grachev et al., 1998), it does not deal adequately with nonlinear morphological differences, especially when applied to the highly variable cortex (Grachev et al., 1999; Mandl et al., 2000; Roland et al., 1997; Xiong et al., 2000). Other examples that require landmarks include modern nonlinear algorithms such as Large Deformation Diffeomorphic Metric Mapping (personal communication with Michael Miller)(Beg et al., 2005) and Caret (<http://brainmap.wustl.edu/>, personal communication with David Van Essen and Donna Dierker)(Essen et al., 2001).

We also excluded some of the primary software programs for automatically labeling cortical anatomy: Freesurfer (<http://surfer.nmr.mgh>.

harvard.edu/)(Fischl et al., 2002, 2004), BrainVisa (<http://brainvisa.info>)(Cointepas et al., 2001), HAMMER (<https://www.rad.upenn.edu/sbia/software/index.html>)(Shen and Davatzikos, 2002), and Mindboggle (<http://www.mindboggle.info>)(Klein and Hirsch, 2005; Klein et al., 2005), because their cortical labeling algorithms are tied to their own labeled brain atlas(es). We considered this problematic for three reasons: (1) we wanted to evaluate brain registration algorithms, not brain labeling algorithms or particular atlas-based approaches, (2) their atlas labels are inconsistent with the protocols used to label the brains in this study which would make evaluation difficult, and (3) creating new atlases for each of these requires considerable knowledge of the software. Freesurfer and BrainVisa differ from all of the other methods mentioned in this paper because they register surfaces rather than image volumes. Mindboggle differs from the others because it is based on combinatoric feature-matching and uses multiple independent atlases. And of the four, HAMMER is the only one that can transform an arbitrary set of labels when registering a source brain to a target brain. However, because we were not able to obtain reasonable results, we did not include it in the study. We also tested the PASHA algorithm (Cachier et al., 2003) with and without intensity normalization but because we obtained very inconsistent results across the datasets we decided not to include it in the study either. We also excluded other programs that do not allow one to apply transforms to separate image volumes.

4.2 *Caveats*

4.2.1 *General caveats*

There are numerous caveats that must be taken into account when evaluating registration data. The very question of correspondence between brains that we raised at the beginning of this paper is revisited at every stage: at the level of anatomy, image acquisition, image processing, registration (including similarity measure, transformation model, regularization method, etc.), evaluation measures, and analysis based on these measures. We will focus here on the most fundamental level of correspondence, at the primary level of anatomy, and on the effects of registration on anatomical correspondence.

If we consider the scale of gross anatomy or patterns of functional activity or physiological data, then we may seek correspondences at the level of topographical, functional, or physiological boundaries without assuming one-to-one mapping of the points of the boundaries or the points within these regions of interest. In other words, another way of approaching this “correspondence problem,” and by extension the elusive common reference frame, is as a partial mapping between brains, independent of naming or

spatial conventions. The common reference frame is used simply as a reference of comparison or evaluation, not as a rigid framework for comprehensively annotating brain image data, as is often done.

If we cannot expect every brain to have a one-to-one mapping with every other brain, then if possible we need to compare similar brains. This can easily lead to the confound where image correspondence is mistaken for anatomic correspondence (Crum et al., 2003; Rogelj et al., 2002). Choosing a representative brain with which to establish correspondences with a given brain results in a Catch-22 where determining similarities itself entails determining correspondences between the brains. A few approaches around this dilemma include the use of an established average template or probabilistic atlas as an intermediary registration target (as is standardly done with SPM), construction of such a template from the subject group that includes the brain in question, and decision fusion strategies for combining multiple, tentative brain registrations or labels for a given target brain (Kittler et al., 1998; Rohlfing et al., 2004; Warfield et al., 2004; Klein et al., 2005). With all of these approaches, however, there still remains the distinct possibility that a given brain is not adequately represented by the majority of the set of brains to which it is being compared. Indeed, it is possible that substructures within a brain are most similar to a minority (or even a single, or no instance) of the set of brains, and would be overridden by the majority.

The evaluation measures and analysis methods used in this paper are predicated on the assumption that, at the macroscopic scale of topographic anatomical regions, there are correspondences across a majority of brains that can effectively guide registrations. It is very important to stress that we cannot make inferences about the accuracy of registrations within these macroscopic regions. Therefore our overlap evaluation measures not only ignore misregistration within a labeled region but are insensitive to folding in the deformations, which would impact studies such as deformation-based morphometry. More generally, our evaluation measures rely on information which is not directly included in the images, which is good for evaluating the registrations, but they do not inform us about the intrinsic properties of the spatial transformations. Example measures of the intrinsic properties of spatial transformations include inverse consistency error, transitivity error, and “mean harmonic energy” (where the Jacobian determinant of the transformation is averaged over the volume).

Another general caveat comes from recent evidence that nonlinear registration to average templates affects different brain regions in different ways that lead to relative distortions in volume that are difficult to predict (Allen et al., 2008). The evidence was based on varying the target template and registration method (AIR and piecewise linear). Although our study was not concerned with absolute volumetry, and nonlinear registrations were conducted from one

brain to another without the use of a template, we share the caution raised by their study.

4.2.2 *Specific caveats*

Caveats that are specific to our study mirror the general caveats raised above: anatomical and labeling variability of the subject brains, quality of their images, the preprocessing steps the images were subjected to, the implementation of the registration algorithms, and our evaluation and analysis methods. With regard to the first three caveats, we made the assumption that each label set consists of a subject group of normal individuals whose brain images were acquired, preprocessed, and labeled in a consistent manner. Some of the co-authors have commented that the quality of the images in this study is worse than the quality of the images that they are used to applying their algorithms to. Some of the reasons for this are that the images for these label sets were acquired years ago, are incomplete (for example, only the CUMC12 set includes the cerebellum in registered images and labels), many are of low contrast, and all of them were linearly transformed to a template space that involved two trilinear interpolation steps (see below). All of the algorithms performed worst on the IBSR18 set, whose images were acquired from various sources and are of varying quality, flouting our assumption above regarding consistency.

Each brain image was labeled only once. Because there are no intra- or inter-labeler data for these images, we cannot know how accurately and consistently they were labeled, let alone have an idea of the degree of confidence for any of the label boundaries. We can only estimate based on labeling tests for two of the labeling protocols (Caviness et al., 1996; Shattuck et al., 2008). We therefore had to treat these label sets as “silver standards” whose hard label boundaries are considered correct.

Regarding pre-processing, the brain images of each label set were consistently preprocessed, and each registration method that performed preprocessing steps did so in a consistent manner across all images. However, these preprocessing steps may be suboptimal for particular registration methods. For example, aside from SPM’s algorithms, we did not test registration accuracy for whole-head images. Although most of the co-authors indicated that they believe their registration methods would perform better on properly skull-stripped images than on whole-head images⁵, we are not aware of any published study that has made this comparison. Likewise, we are aware

⁵ FNIRT is an exception: In the beta version used in this study, zero values are interpreted as missing data; FNIRT will not use the information for the edge of the cortex in the registration with this setting, which may result in misregistration of the surface of the brain.

of no comparisons between the registration of interpolated versus non-interpolated (bias-field corrected and uncorrected, intensity normalized and non-normalized, etc.) images. All of the images in this study were linearly interpolated twice, once to linearly register each brain to a template, and a second time to linearly register each source brain to a target brain in the template space, prior to nonlinear registration. We did this to be consistent, because all of the registration methods we compared do not accept an affine transform to initialize registration. The first author has observed much more accurate nonlinear registrations with ART (on a separate set of brain images) when using nearest-neighbor (or no) interpolation on a preliminary linear registration step, most noticeably in occipital-parietal boundaries. This suggests that, at the very least, ART would perform much better than this study suggests. More work will need to be conducted to see how consistent the improvements are and which algorithms are affected most by interpolation.

Regarding the registration methods themselves, each one has a similarity measure, transformation model, regularization method, and optimization strategy. Unfortunately, we could only evaluate each algorithm in its entirety. A superior transformation model coupled with an unsuitable similarity measure, for example, would most likely lead to suboptimal results. By extension, a poor selection of parameter settings will lead to poor registrations. We could only evaluate each algorithm using the software parameters that were recommended by their authors. Perhaps the most crucial assumption of our study is that these parameter settings for each method were appropriate for all of our brain images. We fully expect that each registration algorithm could perform better given the opportunity to experiment with these settings. This is one aspect of our study that sets it apart from comparison studies such as Hellier’s (Hellier et al., 2001a, 2002, 2003), where the authors of the software packages were allowed to tweak and run their own programs on the full test set of brain images. The commands that were run for this study were recommended by the authors of their respective software programs after having seen only one or two of the 80 images from one of the four datasets (FNIRT, IRTK, SICLE, SyN, and SPM’s DARTEL Toolbox⁶), or no images at all.

When reslicing the source label volumes, we used nearest-neighbor interpolation to preserve label values. An alternative approach is recommended where the source label volume is first split into N binary volumes, with one label per volume (Collins, personal communication). Each volume is then

⁶ Updated versions of these software packages were used after the authors of the packages saw an image or two, or their recommended commands or parameter files were altered to set the number of iterations or control point spacing to reduce computation time, or the authors needed to determine if intensity correction was warranted (see Supplementary section 1).

resampled using the nonlinear transformation with a tricubic or truncated sinc kernel instead of nearest-neighbor interpolation. The resulting N temporary volumes are finally combined into a single volume, where each voxel label is set to the label of the structure that has the highest value. This presumably gives more consistent behavior at structure edges, especially in areas where the deformation changes local volumes or where more than three structures meet. Others have implemented variants of this approach (Crum et al., 2004; Shattuck et al., 2008). We were unable to follow this recommendation due to computational and storage constraints, and were advised that the results would be only marginally different.

4.3 Recommendations

Bearing in mind the caveats mentioned above, particularly those regarding parameter settings, the first author makes the following recommendations based on the results of this study. All of the software packages under comparison are freely available via the Internet or from the authors themselves (except for JRD-fluid, run on LONI’s servers) and all but one (SICLE) are easy to install. They vary in the extent of their documentation, primarily because the pre-release software packages are new and very much under active development.

The highest-ranking registration methods were SyN, ART, IRTK, and SPM’s DARTEL Toolbox (see Tables 3, 4, and 5). SyN and ART gave consistently high-ranking results and were the only methods that attained top rank for all tests and for all label sets. IRTK and SPM’s DARTEL were competitive with these two methods.

All four of these methods are available on Unix-type systems, and all but ART are available for the Windows operating system. Of the four, only SPM requires a commercial software package (Matlab) and has a graphical user interface (which was not used in the study). If flexibility is desired, SyN provides the most options and the closest documentation to a manual for command-line parameters. If resources are an issue, note that SyN requires at least 1 GB RAM and 87 MB storage per x, y, z set of transform files (followed by ART at 67 MB for our data). If time is a constraint, ART is the fastest of the four. If consistency is the top priority, ART had the fewest outliers and among the tightest distributions of the four methods. If interested in particular regions, please refer to Figs. 7, 8, 9, and 10 (the latter three are in Supplementary section 3) to determine which of the 15 methods had the highest relative accuracy for those regions across the label sets.

For time-sensitive scenarios, such as intraoperative imaging, and in high-

throughput environments that need to process dozens to thousands of brain images, Diffeomorphic Demons and ROMEO are reasonable candidates.

With regard to the evaluation protocol, based on the experience of conducting this study the first author recommends caution when choosing an image format and preprocessing steps, particularly when comparing across methods, recommends avoiding interpolation prior to running nonlinear registration, and recommends the model of Pierre Jannin et al. for defining and reporting reference-based validation protocols (Jannin et al., 2006).

With regard to designing and distributing registration algorithms, the first author recommends where possible creating separable components for the similarity measure, transformation model, regularization method, and optimization strategy. This would aid users and evaluators who would want to alter or improve upon these individual components.

5 Acknowledgments

The first author would like to extend his sincere gratitude to the participants in this study for their guidance and support in the use of their software, which in some cases took the form of new pre-release software and reslicing algorithms. He is grateful to his colleagues in the Division of Molecular Imaging and Neuropathology, and thanks Steve Ellis, Todd Ogden, Satrajit Ghosh, and Jack Grinband for their helpful discussions. And of course he thanks his two closest colleagues Deepanjana and Ellora. This work was partially funded by the National Institutes of Health through NIH grant P50-MH062185. The LPBA40 MR and label data were provided by the Laboratory of Neuro Imaging at UCLA and are available at <http://www.loni.ucla.edu/Atlases/LPBA40>. The IBSR18 MR and label data were provided by the Center for Morphometric Analysis at Massachusetts General Hospital and are available at <http://www.cma.mgh.harvard.edu/ibsr/>. The CUMC12 data were provided by Brett Mensh, and the MGH10 data were provided by Satrajit Ghosh and Jason Tourville. The contributions to this paper by Babak A. Ardekani were supported by Grant Number R03EB008201 from the National Institute of Biomedical Imaging And Bioengineering (NIBIB) and the National Institute of Neurological Disorders and Stroke (NINDS). The contributions to this paper by Gary E. Christensen and Joo Hyun Song were supported by NIH grant EB004126. Mark Jenkinson would like to thank the UK BBSRC (David Phillips Fellowship). John Ashburner is funded by the Wellcome Trust.

A Supplementary data

Supplementary data associated with this article can be found below, in the online version, at [doi:10.1016/j.neuroimage.2008.12.037](https://doi.org/10.1016/j.neuroimage.2008.12.037), and on <http://www.mindboggle.info/papers/>.

B Algorithm requirements

Algorithm	Code	Computer	Input	Setup	Run time: minutes
FLIRT (FSL 4.0)	C++	OSX, Linux, Win,...	Analyze, NiFTI		
AIR 5.25	C	OSX, Unix, Win,... ANSI C compiler	Analyze 8-/16-bit	Remove nonbrain structures	6.7 (1.5)
ANIMAL (AutoReg 0.98k)	C, Perl	OSX, Linux, Unix	MINC	Intensity correction (option)	11.2 (0.4)
ART	C++	OSX, Linux	Analyze		20.1 (1.6) [Linux]
Diffeomorphic Demons	C++	Most (ITK compilable)	Analyze, NiFTI, DICOM,... (ITK)		8.7 (1.2)
FNIRT beta	C++	OSX, Linux, Unix	Analyze, NiFTI (writes to Analyze)		29.1 (6.0)
IRTK	C++	OSX, Linux, Win	Analyze, NiFTI, VTK, GIPL	Parameter file	120.8 (29.3)
JRD-fluid	C++	Sun	Analyze		17.1 (1.0) [Solaris]
ROMEO	C++	OSX, Linux, Win 900+MB RAM	Analyze, NiFTI, DICOM,... (ITK)	Parameter file Intensity correction (Hellier, 2003)	7.5 (0.5)
SICLE	C++	OSX, Linux, Solaris, Alpha, Win g77/gfortran lapack, f2c 1+GB RAM	Analyze (7.5) 8-bit	Dimensions divisible by 16 Intensity correction Isotropic Individual parameter files	33.5 (6.6)
SyN beta	C++	Most (ITK compilable) 1+GB RAM	Analyze, NiFTI, DICOM,... (ITK)		77 (15.1)
SPM5:	Matlab	Most (Matlab)			
“SPM2-type” Normalization		Matlab 6.5 onwards	Analyze, NiFTI	Smooth targets (Gaussian 8mm FWHM)	<1
Normalization		Matlab 6.5 onwards	Analyze, NiFTI	Left-handed orientation	<1
Unified Segmentation		Matlab 6.5 onwards	Analyze, NiFTI	Left-handed orientation	$\simeq$ 1
DARTEL Toolbox (pairs)		Matlab 7.0 onwards	Analyze, NiFTI	Left-handed orientation Origin near anterior commissure	71.8 (6.3)*

Table B.1

Algorithm requirements, input, and run time. The run time average (and standard deviation) is estimated from a sample of registrations and includes the time to compute the source-to-target transform but not to apply it to resample the source labels. *SPM’s DARTEL Toolbox requires time to construct a template per subject group. The time listed is for the pairwise implementation; for the normal toolbox implementation, it took 17 minutes per brain, or 17.5 hours to run all 80 brains (LPBA40: 480 min., IBSR18: 220 min., CUMC12: 195 min., MGH10: 158 min.). All programs were run on an OSX system (Mac Pro Quad-Core Intel Xeon, 3GHz, 6GB RAM) with a 10.4 operating system, except for ROMEO (10.5 operating system), ART (the OSX version was made available after the study; Dell PowerEdge 6600 Enterprise server with four 2.8GHz Intel Xeon processors and 28GB of RAM running Redhat linux, approximately 1.25-1.5 times slower than the OSX machine), and JRD-fluid (run on LONI’s servers: SUN Microsystem workstations with a dual 64-bit AMD Opteron 2.4 GHz processor running Solaris).

Supplementary section 1: Initial run

A few algorithms resulted in consistently low accuracies or occasional failures in their registrations. We did our best to rectify these problems. For example, ROMEO’s initial run produced inconsistent results, so Hellier provided intensity correction code which we applied to all of the data prior to registering with ROMEO. Of the 2,168 registrations, there were 19 failures for AIR, accompanied by the message: “Registration terminated due to a Hessian matrix that was not positive definite” (even with the “-q” option). We were able to correct all of these cases by skipping the “alignlinear” step (and relying on the preliminary linear alignment with FLIRT). SyN had seven failures, which we corrected by first running Avants’ TranslateRegistration program (this program had no effect on other registrations).

SPM’s DARTEL Toolbox resulted in highly variable results for the LPBA40 set (and low results obtained with one of the MGH10 images), most likely because of the inconsistent way that an older version of the code dealt with zeros in the images. DARTEL estimates its spatial transformation by registering gray and white matter maps produced by the segmentation step. Regions in these images, which had been set to zero in the skull-stripped data, sometimes contained information from the tissue probability maps used by the segmentation. We downloaded an updated version of DARTEL that corrects for this and ran it again (in both a pairwise as well as average template manner).

In addition to addressing variable results, we had to set reasonable time constraints. SyN’s default parameters were found to be too computationally intensive, so Avants recommended a different number of iterations to keep computation under an hour per registration. Likewise, Rueckert provided a parameter file with reduced control point spacing to reduce IRTK’s computation time. IRTK, SICLE, and SPM’s Unified Segmentation and DARTEL Toolbox were run by their authors on one registration pair to ensure that the first author’s implementation and execution of their software was correct.⁷

⁷ Likewise for PASHA and HAMMER, but these were excluded from the study (see Discussion).

Supplementary section 2: The trivial case: self-registration results

For the trivial case, where each brain was registered to itself, all of the methods performed nearly perfectly as measured by volume similarity, with the exception of SPM’s DARTEL Toolbox for the LPBA40 set. According to target volume overlap, the methods that gave less than perfect results were: ANIMAL and ROMEO (for all four label sets), “SPM2-type” Normalization (IBSR18, CUMC12, and MGH10), and ROMEO, SICLE, and SPM’s DARTEL (LPBA40). The few cases of low values were obtained with SPM’s DARTEL, and only for the LPBA40 set according to overlap and distance measures.

The results were almost identical for the target, union, and mean volume and surface overlap measures. Not surprisingly, the deviations from perfect registration in the trivial case are revisited in the non-trivial, inter-brain registrations results.

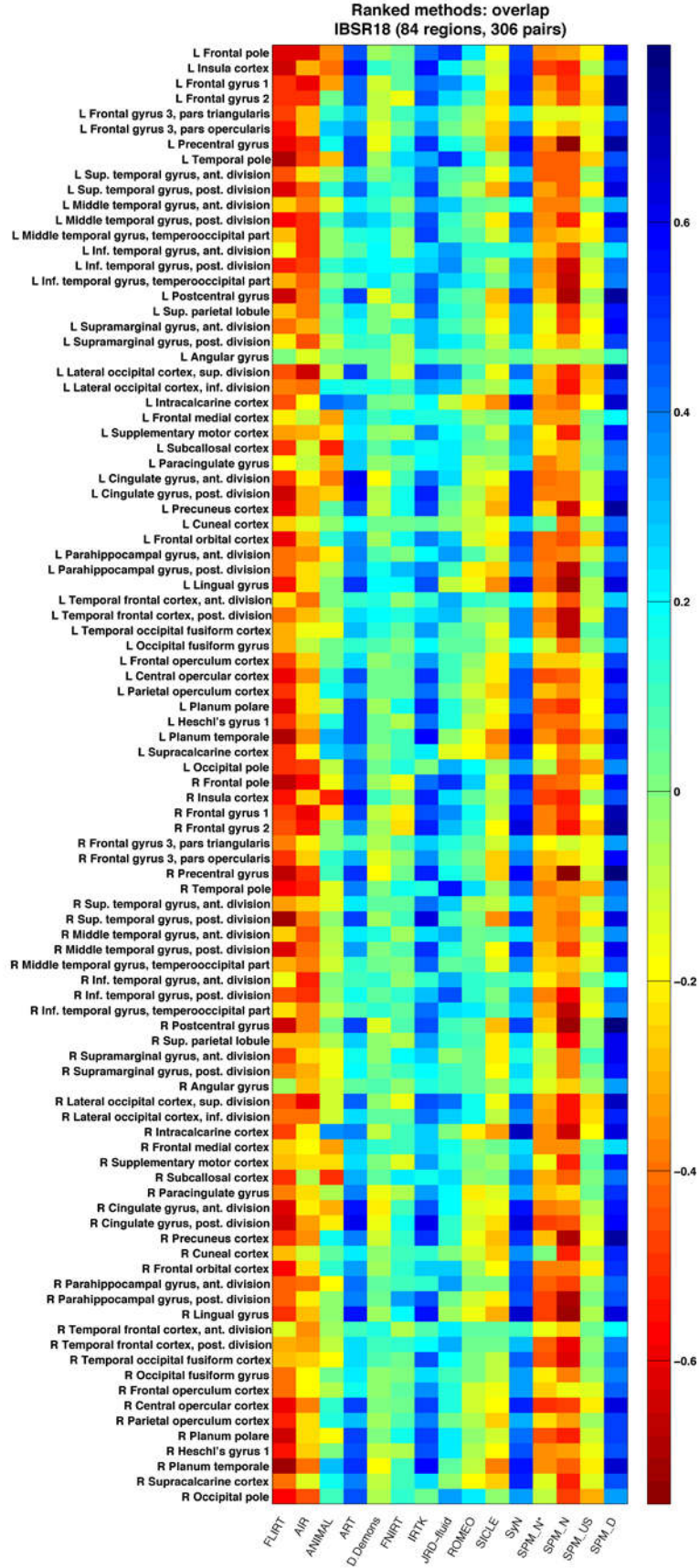
Supplementary section 3: More indifference-zone rankings

See Figures 8, 9, and 10 for indifference-zone ranking of the registration methods for the IBSR18, CUMC12, and MGH10 data, respectively. These rankings are based on target overlap results, and correspond to the indifference-zone ranking for the LPBA40 data in Figure 7.

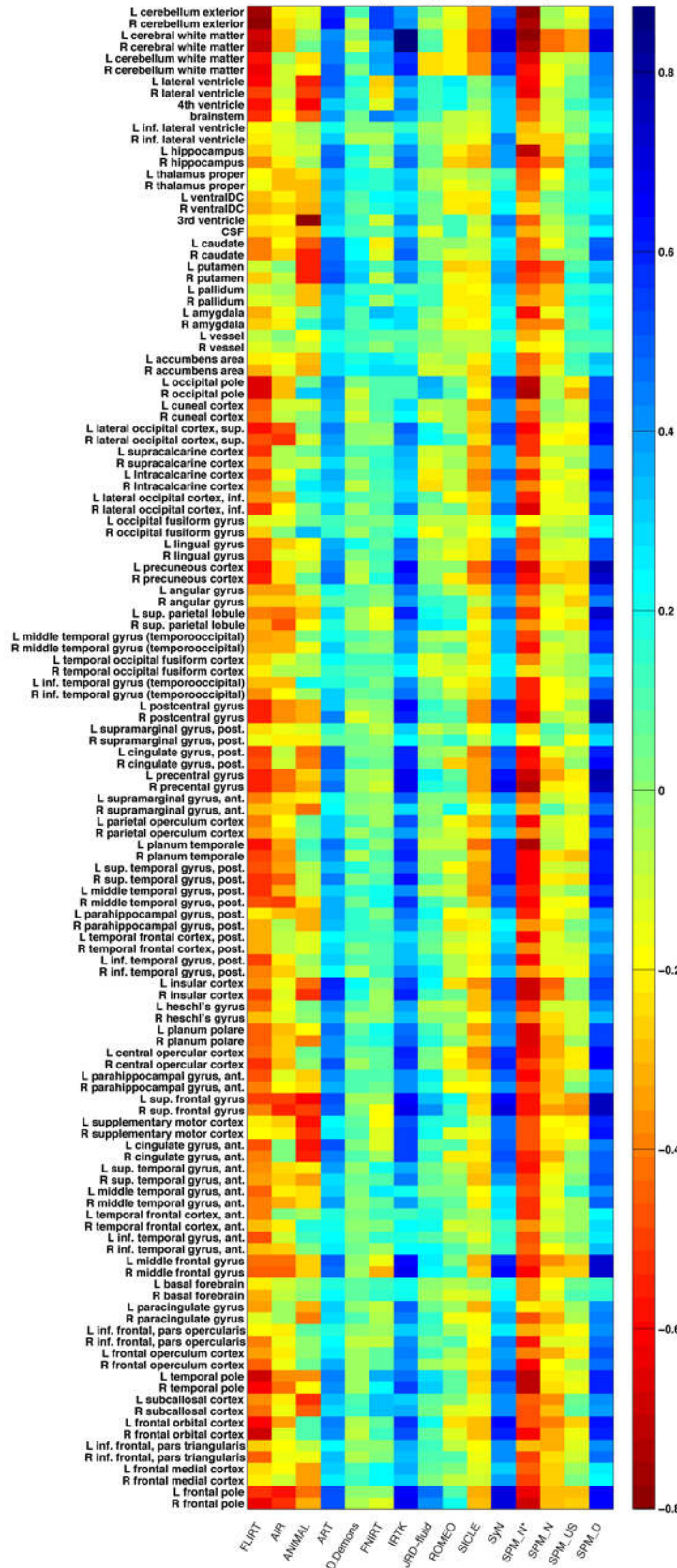
Fig.8 Indifference-zone ranking of the registration methods: IBSR18 overlaps. This matrix was constructed as in Figure 7 for target overlap rankings averaged across 306 registration pairs using the 84 regions of the IBSR18 dataset (union and mean overlap results are almost identical). Blue indicates higher accuracy; the colors (and color range) are not comparable to those of the other label sets (Figures 7, 9, and 10). (SPM_N*="SPM2-type" Normalize, SPM_N=Normalize, SPM_US=Unified Segmentation, SPM_D=DARTEL pairwise)

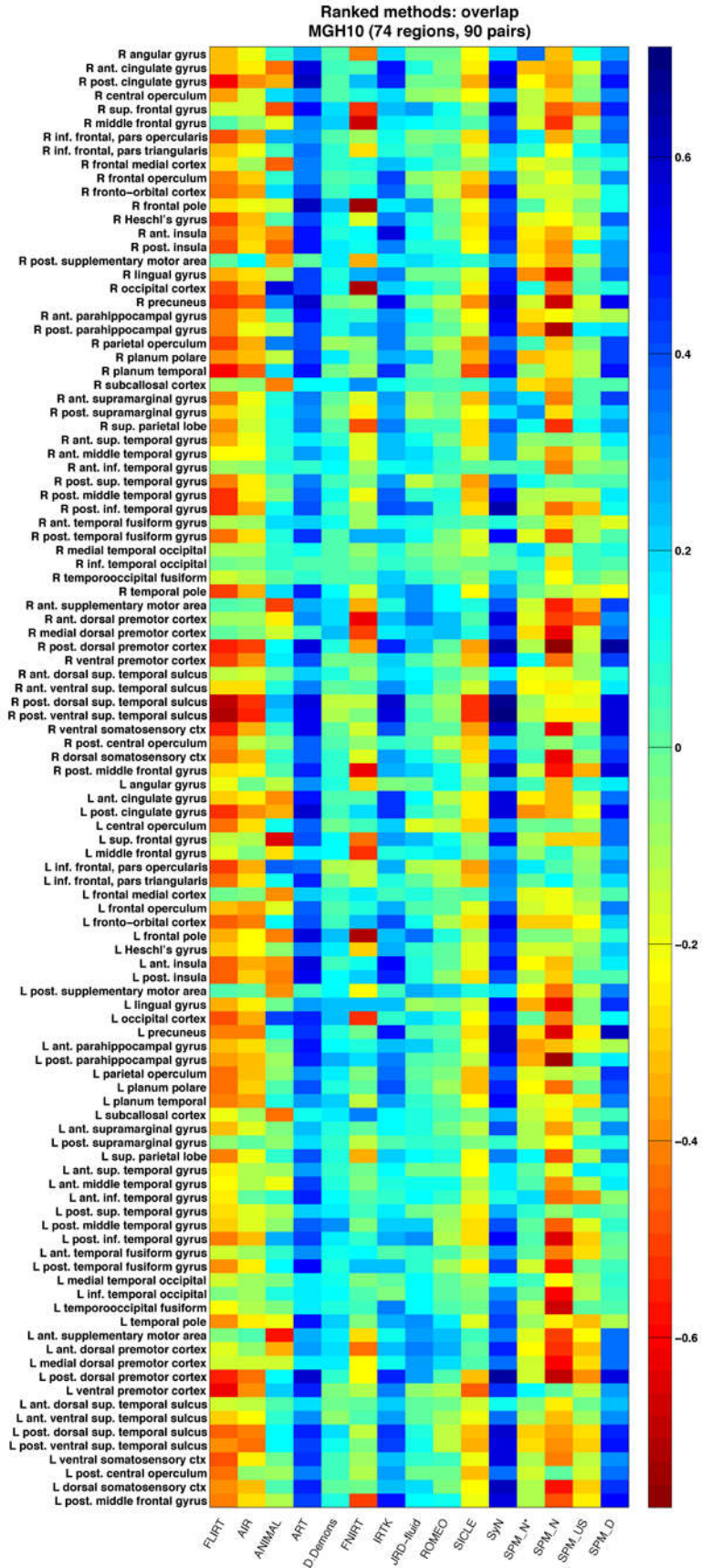
Fig.9 Indifference-zone ranking of the registration methods: CUMC12 overlaps. This matrix was constructed as in Figure 7 for target overlap rankings averaged across 132 registration pairs using the 128 regions of the CUMC12 dataset (union and mean overlap results are almost identical). Blue indicates higher accuracy; the colors (and color range) are not comparable to those of the other label sets (Figures 7, 8, and 10). (SPM_N*="SPM2-type" Normalize, SPM_N=Normalize, SPM_US=Unified Segmentation, SPM_D=DARTEL pairwise)

Fig.10 Indifference-zone ranking of the registration methods: MGH10 overlaps. This matrix was constructed as in Figure 7 for target overlap rankings averaged across 90 registration pairs using the 74 regions of the MGH10 dataset (union and mean overlap results are almost identical). Blue indicates higher accuracy; the colors (and color range) are not comparable to those of the other label sets (Figures 7, 8, and 9). (SPM_N*="SPM2-type" Normalize, SPM_N=Normalize, SPM_US=Unified Segmentation, SPM_D=DARTEL pairwise)



Ranked methods: overlap
CUMC12 (128 regions, 132 pairs)





Supplementary section 4: Volume similarity results

The volume similarity measures, averaged across all of the regions in each label set, are shown in Figure 11. Volume similarity is only an indirect measure of registration accuracy, however it can be used to expose gross discrepancies in volume, if not shape. For each label set, all of the methods resulted in very similar median values. The LPBA40 set and SPM DARTEL exhibited the largest number of outliers.

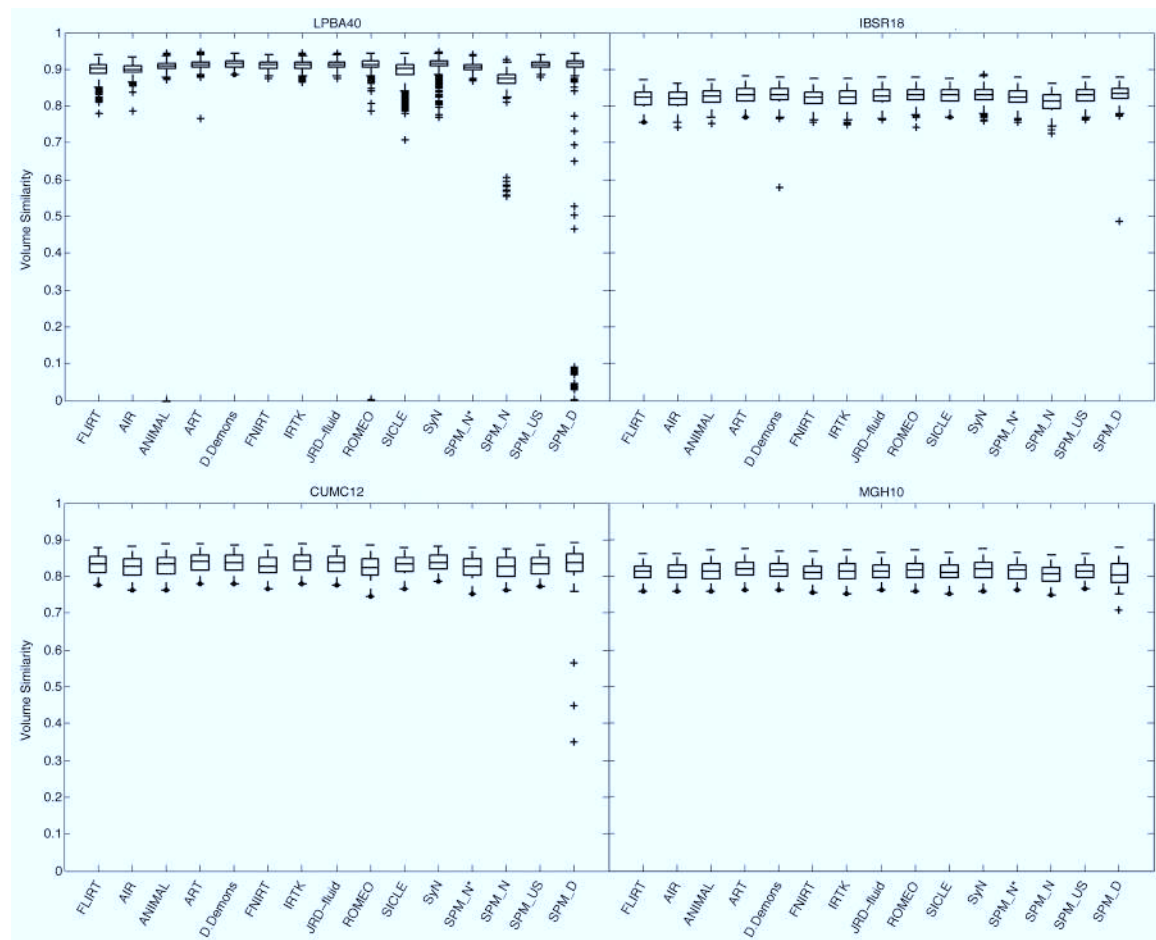


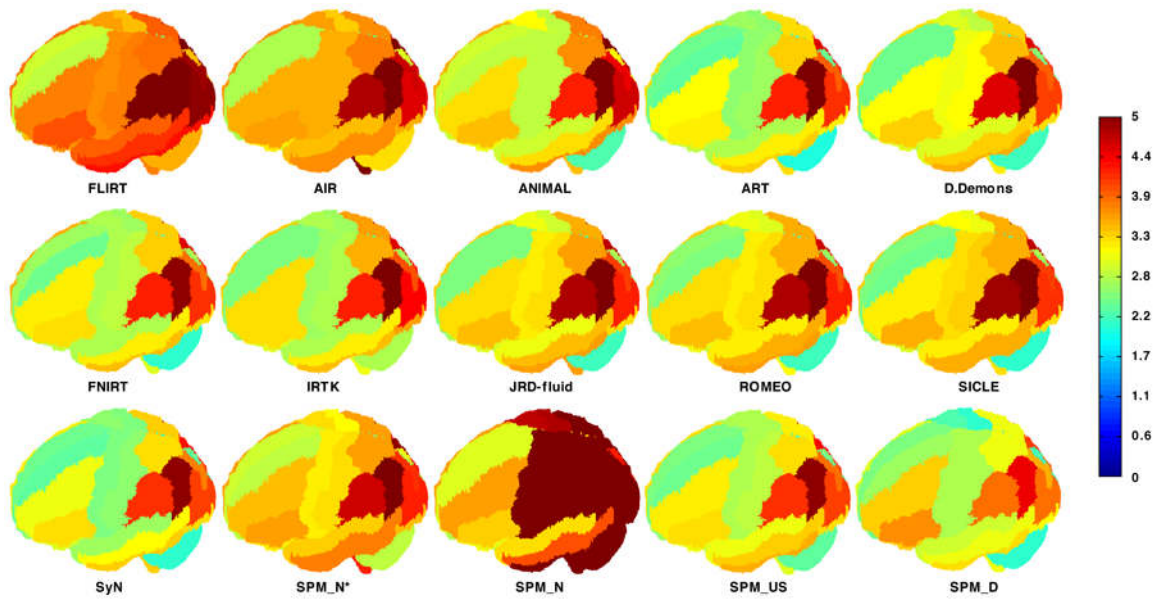
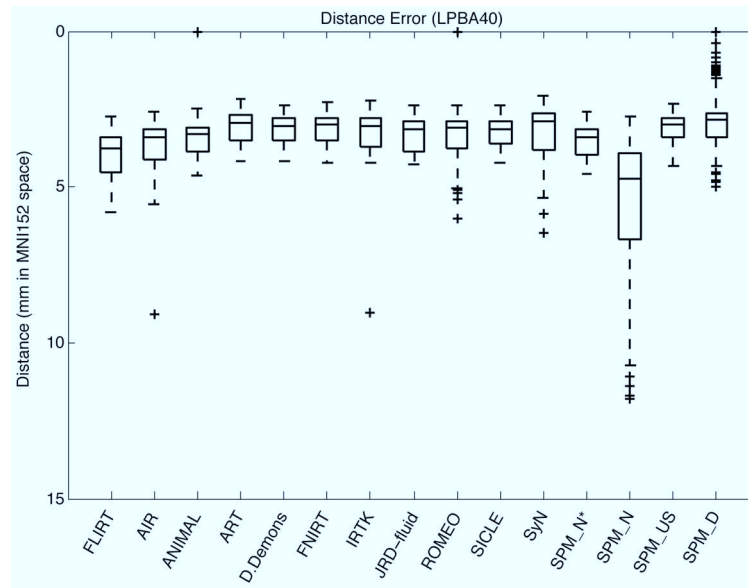
Fig.11 Volume similarity by registration method. These box and whisker plots (constructed as in Figure 5) show the volume similarity measures between deformed source and target labels, averaged first across all of the regions in each label set (LPBA40, IBSR18, CUMC12, and MGH10) then across brain pairs, with highest similarity at the top. (SPM_N*="SPM2-type" Normalize, SPM_N=Normalize, SPM_US=Unified Segmentation, SPM_D=DARTEL pairwise)

Supplementary section 5: Distance results

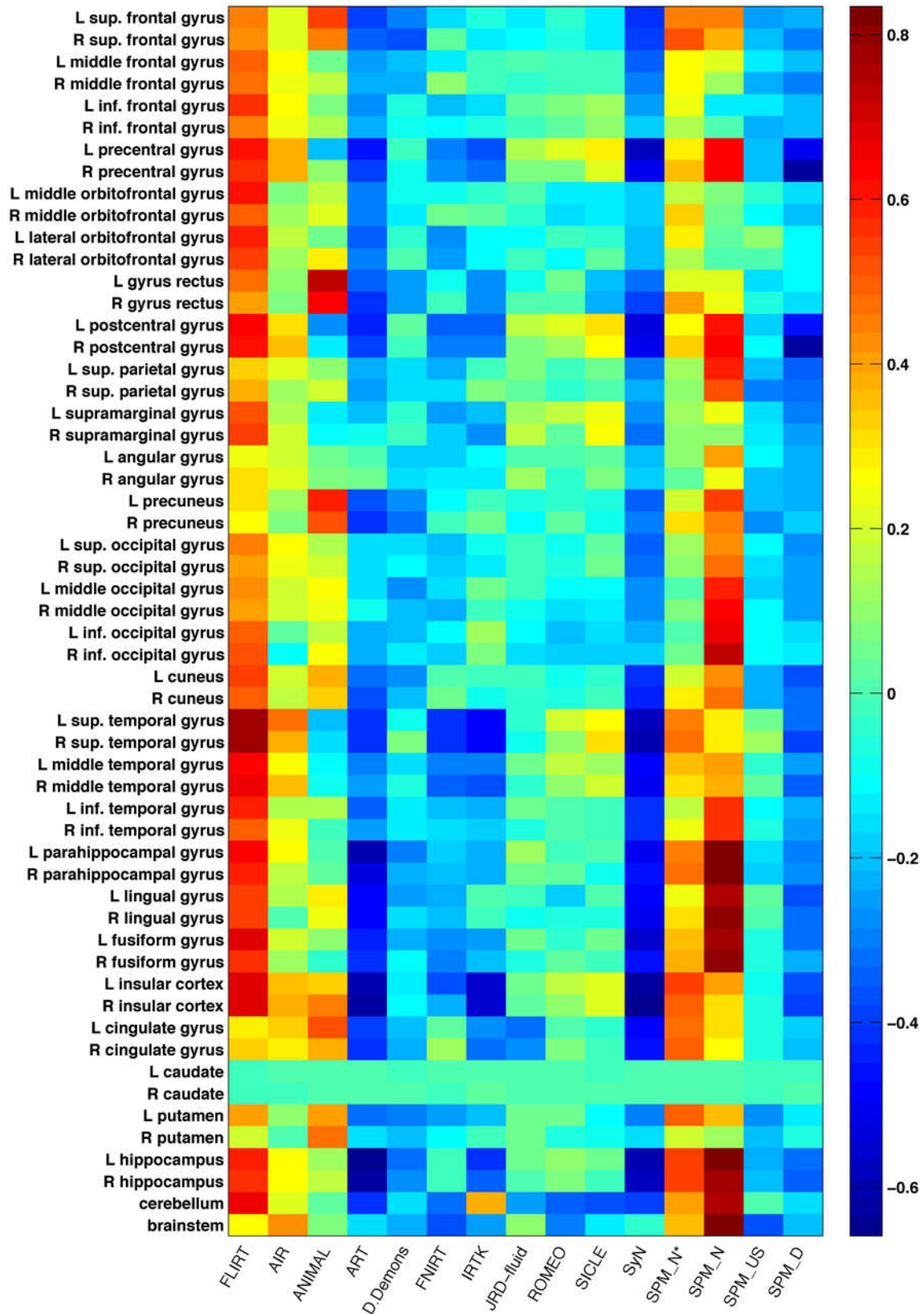
The box and whisker plot in Figure 12 shows the distance errors for the LPBA40 label set. They exhibit a roughly similar pattern across the methods as the overlap values in Figure 5, although regional differences may be seen across Figures 13 and 7.

Fig.12 Distance error by registration method. The box and whisker plot was constructed as in Figure 5 except that the measure is distance error between deformed source and target label boundaries, averaged first across all of the regions in the LPBA40 label set then across brain pairs, with lowest errors toward the top. The brain images (constructed as in Figure 6) show the mean distance error per region as a color (blue indicates higher accuracy). (SPM_N*="SPM2-type" normalization, SPM_N=SPM's Normalize, SPM_US=Unified Segmentation, SPM_D=DARTEL pairwise)

Fig.13 Indifference-zone ranking of the registration methods: LPBA40 distance errors. This matrix was constructed as in Figure 7, except for distance error rankings rather than overlap rankings. Blue indicates lower error and higher accuracy. (SPM_N*="SPM2-type" normalization, SPM_N=SPM's Normalize, SPM_US=Unified Segmentation, SPM_D=DARTEL pairwise)



Ranked methods: distance error
LPBA40 (56 regions, 1560 pairs)



Supplementary section 6: Dependence test

To determine the degree of correlation across source-target registration pairs, we selected and organized pairs from the LPBA40 brains (e.g., $a \rightarrow b$) into two columns ($a \rightarrow b$, $c \rightarrow d$), where rows were independent of one another (no brain appeared in multiple rows), but the two columns had one of four dependency relationships to one another, as follows:

- (1) No dependence; each brain was used only once (10 rows):
 $a \rightarrow b$, $c \rightarrow d$
- (2) source and target dependence; both brains were used twice (20 rows):
 $a \rightarrow b$, $b \rightarrow a$
- (3) source dependence; the source is used twice (13 rows):
 $a \rightarrow b$, $a \rightarrow c$
- (4) target dependence; the target is used twice (13 rows):
 $a \rightarrow b$, $c \rightarrow b$

We then replaced each registration pair with the target overlap calculated for that pair for each registration method, and computed the correlation between the left and right columns for each dependency condition. We repeated this procedure 1,000 times, using a new set of pairs satisfying each condition, and averaged the results. The values for the no dependence condition are close to zero as we would expect for independent pairs. All of the other conditions, however, result in high correlations for most of the methods. One interesting detail is that there is an asymmetry between conditions 3 and 4. For example, for some methods, even when pairs that share a source brain are correlated, pairs that share a target brain are not necessarily correlated.

Supplementary section 7: Software setup and commands

Most of the programs offer a range of flexible options for their similarity metric, regularization method, etc. Therefore the following commands each represent but a single implementation of its underlying algorithm. The nonlinear deformation programs are listed in alphabetical order after the linear registration program FLIRT.

Python and Matlab programs were used to call all of the commands below, and were run on an OSX system (Mac Pro 2-Quad-Core (8-processor) Intel Xeon, 3GHz, 6GB RAM) with a 10.4 operating system, except for ROMEO (10.5 operating system, same hardware), ART (Dell PowerEdge 6600 Enterprise server with four 2.8GHz Intel Xeon processors and 28GB of RAM running Redhat linux), and JRD-fluid (run on LONI's servers: SUN Microsystem workstations with a dual 64-bit AMD Opteron 2.4 GHz processor running Solaris). IRTK was run on some of the LPBA40 brains on servers at Imperial College, and the results were confirmed to be identical to those obtained by the first author.

FLIRT linear and rigid alignment commands were run before any of the other commands, and the resulting linearly transformed brain images were resliced using trilinear interpolation. For all algorithms, manual labels were resliced using nearest-neighbor interpolation.

FLIRT (FSL v4.0)

*Linearly align **to** MNI space (9-parameter):*

```
flirt -in <source> -ref <MNI152> -omat <output transform>.mat -out <output brain> -bins 256 -cost corratio -searchrx -90 90 -searchry -90 90 -searchrz -90 90 -dof 9 -interp trilinear
```

where <MNI152> is the nonlinear MNI152 template used by FSL: MNI152_T1_1mm_brain

*Rigidly align **in** MNI space (6-parameter):*

```
flirt -in <above output brain> -ref <target> -omat <output transform>.mat -out <starting point for all other algorithms> -bins 256 -cost corratio -searchrx -30 30 -searchry -30 30 -searchrz -30 30 -dof 6 -interp trilinear
```

Reslice using 6-parameter output transform:

```
flirt -in <labeled source> -ref <labeled target> -applyxfm -init <transform>.mat -out <output labels> -padding 0.0 -interp nearestneighbour
```

AIR v5.25

Align (12-parameter affine):

```
alignlinear <target>.img <source>.img <output affine transform>.air -m 12 -t1 1 -t2 1 -q
```

Warp (2nd-order, 3rd-order, 4th-order and then 5th-order 168-parameter):
align_warp <target>.img <source>.img <output transform>.warp -m 2 5 -f
<above affine transform> -t1 1 -t2 1 -q

Reslice:

reslice_warp <align_warp transform>.warp <output labels>.img -a <labeled
source>.img -n 0 -o

ANIMAL (MNI AutoReg v0.98k)

A Perl script performs non-linear fitting. The original ANIMAL parameters were optimized Robbins (Robbins et al., 2004). The resulting Perl script was modified again by Janke and Lepage. The script calls minctracc and the parameters and hierarchical steps are as follows:

'-nonlinear', '-corcoeff', '-weight',1, '-stiffness',1, '-similarity',0.3,
'-sub_lattice',6

'step': 32 16 12 8 6 4

'blur_fwhm': 16 8 6 4 3 2

'iterations': 20 20 20 20 20 10

Warp: <Perl script>.pl -clobber -normalize <source>.mnc <target>.mnc <output
transform>.xfm

Reslice: mincresample -transform <transform>.xfm -like <labeled target>.mnc
<source>.mnc <output labels>.mnc -nearest_neighbour -keep_real_range -
short

Convert to Analyze format: mnc2nii -short <above output labels>.mnc <output
labels>.nii

ART

Warp: 3dwarper -R -trg <target>.img -obj <source> -u <output transform>.wrp
-o <output brain>.img -A -sd 8.0

Reslice: applywarp3d -nn -w <transform>.wrp <labeled source>

Diffeomorphic Demons:

Warp: DemonsRegistration -f <target>.hdr -m <source>.hdr -0 <output
transform>.hdr -e -s2 -i30x20x10

Reslice: ResampleImage -i <labeled source>.hdr -f <transform>.hdr -o <output
labels>.img

FNIRT

Warp: fnirt --config=schedule_01.cnf --ref=<target> --in=<source> --cout
<output transform>

Reslice: applywarp -i <labeled source> -r <target> -w <transform> -o <output
labels> --interp=nn

IRTK

In the current study no regularization was used.

The parameter file called for a 2.5 mm minimum control point spacing, and
the non-rigid registration parameters were:

Lambda1, 2, and 3 = 0

Control point spacing in X, Y, and Z = 20

Rigidly align: rreg <target>.hdr <source>.hdr -dofout <output
transform>_rreg.dof

Affinely align: areg <target>.hdr <source>.hdr -dofin <rreg transform>_rreg.dof
-dofout <output transform>_areg.dof

Warp: nreg <target>.hdr <source>.hdr -dofin <areg transform>_areg.dof -dofout
<output transform>_nreg.dof -parin ITK_parameters_2.5mm.txt

Reslice: transformation <labeled source>.hdr <output labels>.hdr -dofin <nreg
transform>_nreg.dof -target <target>.hdr

JRD-Fluid:

Warp: mix_fluidmap_nohassle_column.out <source>.img
<dimensions> <target>.img <dimensions> <output directory> source_le 0.95 0
<parameter>,

where *parameter* is set to 4 for LONI LPBA40 images (181x217x181 voxels)
and set to 20 for all other images registered to the nonlinear MNI152 template.

Reslice: resample_8bit_image_NN.out <labeled source>.img <above output
directory>/trn_DXP_ftle <above output directory>/trn_DYP_ftle <above
output directory>/trn_DZP_ftle <dimensions> <output labels>

ROMEO

Warp and reslice: Romeo.py -t <target> -s <source> -l <labeled source> -o
<output brain> -d <output labels>

SICLE

Preparation: lereg2.exe -gf global.param -p input_files/<source-target pair
parameter file>.in | tee <source-target pair output file>.out

Warp: rpp2.exe <source-target pair parameter file>.in -def -jac -gf global.param

Reslice: deform3d.exe -coeff <source-target pair output stem>_res10000_iter00020.coeffs -t <labeled source>.hdr -o <output labels>.hdr -interp N

SPM5's "SPM2-type" Normalization

Warp (in Matlab): spm_normalise("<target>.img", "<source>.img", "<output transform>.mat");

Reslice (in Matlab): spm_write_sn("<labeled source>.img", "<output labels>.mat", struct("interp",0));

SPM5's Normalize, Unified Segmentation, and DARTEL Toolbox

Matlab scripts were used to compose the transforms for these methods, and may be obtained from <http://www.mindboggle.info/papers/>.

SyN

Linearly align (optional): TranslationRegistration <target>.hdr <source>.hdr <output linear transform>

Warp: perl arnoMVSN.pl <target>.hdr <source>.hdr <above linear transform> "-c 5 -n3 -i30x99x11 -l 0.5 -s 2 -a 0.05 " <SyN directory> 0 0 0 <user name> 1

Reslice: WarpImageBackward <labeled source>.hdr <transform filestem>warp <output labels>.hdr donearestneighbor

We corroborated results obtained with the above commands for the newer, publicly released ANTS software (<http://www.picsl.upenn.edu/ANTS/>) using the following commands:

Warp: ANTS 3 -m PR[<target>.nii, <source>.nii, 1, 2] -o <output transform>.nii -r Gauss[2,0] -t SyN[0.5] -i 30x99x11 -use-Histogram-Matching

Reslice: WarpImageMultiTransform 3 <labeled source>.nii <output labels>.nii -R <target>.nii <transform>Warp.nii <transform>Affine.txt -use-NN

Supplementary section 8: Algorithm descriptions

8.1 FLIRT: FMRIB's Linear Image Registration Tool

FLIRT was developed by Jenkinson (Jenkinson and Smith, 2001) at the FMRIB Centre at the University of Oxford, UK. FLIRT can be downloaded with the FMRIB Software Library (FSL) at <http://www.fmrib.ox.ac.uk/fsl/>.

FLIRT is an automated linear (affine) registration tool based around a multi-start, multi-resolution global optimisation method. It can be used for inter- and intra-modal registration with 2-D or 3-D images. In addition, it can be run with a number of different transformation models (degrees of freedom) and it implements a general cost function weighting scheme for all cost functions.

8.2 AIR: Automated Image Registration

AIR was developed by Woods (Woods et al., 1998) at the David Geffen School of Medicine at UCLA, CA. The C source code may be downloaded at <http://air.bmap.ucla.edu>.

AIR aligns a pair of images by minimizing the mean squared difference between the image designated for resampling and the image to which it is being registered. First and second derivatives of the cost function are computed and used to iteratively adjust the model parameters. In addition to an optional intensity scaling parameter, the algorithm adjusts the elements of polynomials of the desired order. Registration begins with a first order polynomial (i.e., an affine transformation) and the order is incremented sequentially until the order specified by the user is reached. Polynomial orders up to twelfth order are implemented. At each order, registration begins with sparse sampling of the data and proceeds to denser sampling, per user specifications, with subsequent iterations. Criteria for advancing to denser sampling or to higher order polynomials include the magnitude of the predicted improvement in the cost function, the number of iterations without improvement in the cost function, and the total number of iterations. In some instances, full Newton-based minimization cannot proceed because the Hessian matrix of second derivatives is not positive definite. The likelihood of a non-positive definite Hessian matrix can optionally be reduced by omitting certain contributions to the second derivatives using a strategy similar to that employed by the Levenberg-Marquardt algorithm. The optimized polynomial transformation is stored, and separate programs allow either of the two original images to be resampled into the space defined by the other image using a variety of interpolation algorithms (nearest neighbor interpolation, trilinear interpolation, sinc interpolation, and hybrids of these methods). Diagnostics are available to verify that no regions with non-positive Jacobians are present. Polynomial transformations can be mathematically combined with any number of affine linear transformations,

allowing data to be resampled directly from or to any space that has an affine relationship to the original images.

(degrees of freedom (dof) = 168)

8.3 ANIMAL: Automatic Nonlinear Image Matching and Anatomical Labeling

ANIMAL was developed by Collins et al. (Collins et al., 1994, 1995; Collins and Evans, 1997; Robbins et al., 2004) of the Montreal Neurological Institute, Canada. The original ANIMAL is available for download at http://www.bic.mni.mcgill.ca/users/louis/MNI_ANIMAL_home/readme/. Lepage provided a Perl script (see below) for implementing the multi-resolution strategy.

ANIMAL is based on multi-scale, 3-D cross-correlation. Spatial registration is completed automatically as a two step process. The first accounts for the linear part of the transformation by using correlation between Gaussian-blurred features extracted from both volumes. In the second step, ANIMAL estimates the 3-D deformation field required to account for this variability. The deformation field is built by sequentially stepping through the target volume in a 3-D grid pattern. At each grid-node i , the deformation vector required to achieve local registration between the two volumes is found by optimization of three translational parameters (tx_i, ty_i, tz_i) that maximize the objective function evaluated only in the neighborhood region surrounding the node. The algorithm is applied iteratively in a multi-scale hierarchy, so that image blurring and grid size are reduced after each iteration, thus refining the fit. The multi-scale approach also makes the procedure very robust and ensures that the algorithm converges to the global minimum.

A Perl script (nlfit) implements the multi-resolution fitting strategy to map brains into stereotaxic space at the Montreal Neurological Institute. At the heart of this procedure is minctracc, the program that automatically finds the best non-linear transformation to map one volumetric data set (stored in MINC format) to another. The program uses optimization over a user selectable number of parameters to identify the best transformation mapping voxel values of the first data set into the second.

(dof $\leq 3 \times \simeq 23,000$ nodes (1.5M brain voxels / 4mm each direction) = 69,000)

8.4 ART: Automated Registration Tool

ART was developed by Ardekani et al. (Ardekani et al., 2005) at the Nathan Kline Institute, NY. The executables may can be downloaded at <http://www.nitrc.org/projects/art/>. Ardekani revised the registration program to allow one to specify an output file and revised the resampling code to enable nearest-neighbor interpolation.

ART uses local normalized cross-correlation between the source and target images as its similarity measure. It determines a displacement vector field defined for each grid point on the target image using a non-parametric free-form multi-resolution approach. The displacement vector field obtained at each resolution level is regularized by median and low-pass filtering.

Prior to non-linear registration, there are options for the program to determine a linear rigid-body (6-parameter) registration between the target and subject images followed by a linear affine (12-parameter) registration (Ardekani et al., 1995).

(dof $\simeq 3 \times \# \text{brain voxels} \simeq 7 \text{million}$)

8.5 Demons: Diffeomorphic Demons

Diffeomorphic Demons was developed by Vercauteren et al. (Vercauteren et al., 2007) at Mauna Kea Technologies and at INRIA Sophia Antipolis, France; it is implemented as part of the finite difference solver framework within the Insight Toolkit (ITK) and can be downloaded at <http://hdl.handle.net/1926/510>. A graphical user interface is provided as part of MedINRIA (Toussaint et al., 2007). Vercauteren provided a version of the resampling code with an option for nearest-neighbor interpolation that doesn't require origin information.

This non-parametric algorithm generalizes Thirion's Demons algorithm (Thirion, 1998) to produce a diffeomorphic spatial transformation (Vercauteren et al., 2007). This method alternates between the computation of warping forces inspired from optical flow theory and the regularization of these forces by a simple Gaussian smoothing.

The Demons algorithm may be characterized as the optimization of a global energy function (Cachier et al., 2003) where correspondences act as a hidden variable in the registration process. The regularization criterion is then considered as a prior on the smoothness of the spatial transformation s , and point correspondences between image pixels (a vector field c) are allowed to have some error.

Given a *fixed image* $F(\cdot)$ and a *moving image* $M(\cdot)$, the following global energy is optimized:

$$E(c, s) = \frac{1}{\sigma_i^2} \text{Sim}(F, M \circ c) + \frac{1}{\sigma_x^2} \text{dist}(s, c)^2 + \frac{1}{\sigma_T^2} \text{Reg}(s), \quad (\text{B.1})$$

where σ_i accounts for the noise on the image intensity, σ_x accounts for spatial uncertainty on the correspondences and σ_T controls the amount of regularization. Classically, $\text{Sim}(F, M \circ s) = \frac{1}{2} \|F - M \circ s\|^2$, $\text{dist}(s, c) = \|c - s\|$ and $\text{Reg}(s) = \|\nabla s\|^2$ but the regularization can also be modified

to handle fluid-like constraints.

Within this framework, the Demons registration can be explained as an alternate optimization over s and c . The optimization is performed within the complete space of dense non-rigid transformations by taking a series of additive steps, $s \leftarrow s + \mathbf{u}$. In contrast, the diffeomorphic demons algorithm optimizes $E(c, s)$ over a space of diffeomorphisms. This is done in (Vercauteren et al., 2007) by using an intrinsic update step, $s \leftarrow s \circ \exp(\mathbf{u})$, on the group of diffeomorphisms.

(dof $\simeq 3 \times \# \text{voxels} = 21 \text{million}$)

8.6 FNIRT: FMRIB's Nonlinear Image Registration Tool

FNIRT was developed by Andersson et al. (Andersson et al., 2008) of the FMRIB Centre at the University of Oxford, UK. Andersson provided pre-release software for use in this study.

FNIRT is the FMRIB tool for small-displacement non-linear registration. The displacement fields are modelled as linear combinations of a basis set of splines of order two or higher, with a default of three (cubic splines). Regularisation of the field is based on bending energy (default) or membrane energy (the prior default used in this study), with optimization by multi-scale Levenberg-Marquardt minimization. The registration is initialized and run to convergence with sub-sampled images, a field of low resolution and a high regularization weight. The images and the fields from the first step are then up-sampled, the regularization modified and it is again run to convergence. This is repeated until the required warp resolution and level of regularization is achieved. After each resolution step the field is projected back onto the space of fields with Jacobians within a predefined range (Karacali and Davatzikos, 2004).

The important and unique aspects of FNIRT are with respect to its cost function. Normally when using a sum-of-squares (SOS) cost function one minimizes the difference between a warped source image and some target. FNIRT minimizes the SOS between a warped source image and the expectation of an image in target space. This expectation is based on an actual image in target space and an intensity model. The intensity model will have some set of parameters that is determined along with the warp parameters as part of the optimization. This model can be very simple, for example just a linear scaling of the target image, in which case the parameters will be a single scale factor. The model can also be quite complex, if for example one wants to model a spatially varying flip-angle (which is a real problem with high-field scanners) it will consist of some set (e.g. 5) fields. Each of these fields is modeled as a linear combination of some basis set (also splines) and will consist of a few thousand parameters. For each voxel the expected intensity would then be a 5th-order polynomial of the intensity in the target

image, where the coefficients are given by the values of the five fields at that voxel. This way it is possible to accurately model an image where (due to inhomogeneities) one area is strongly T1-weighted and another area has a strong T2-component. There is a set of different intensity models in FNIRT and the decision of which to use is made based on the properties of the two images one attempts to match.

(dof $\simeq$ 30,000)

8.7 IRTK: Image Registration Toolkit

IRTK was developed by Rueckert et al. (Rueckert et al., 1999; Studholme et al., 1999; Rueckert et al., 2006) of Imperial College, UK. The executables are available for download at <http://www.doc.ic.ac.uk/~dr/software/>.

IRTK uses a combined transformation $\mathbf{T}$ which consists of a global transformation and a local transformation:

$$\mathbf{T}(\mathbf{x}) = \mathbf{T}_{global}(\mathbf{x}) + \mathbf{T}_{local}(\mathbf{x}) \quad (\text{B.2})$$

The global transformation describes the overall differences between the two subjects and is represented by an affine transformation. The local transformation describes any local deformation required to match the anatomies of the subjects. IRTK uses a free-form deformation (FFD) model based on B-splines. The basic idea of FFDs is to deform an object by manipulating an underlying mesh of control points. The resulting deformation controls the shape of the 3-D object and can be written as the 3-D tensor product of the familiar 1-D cubic B-splines,

$$\mathbf{T}_{local}(\mathbf{x}) = \sum_{l=0}^3 \sum_{m=0}^3 \sum_{n=0}^3 B_l(u) B_m(v) B_n(w) \mathbf{c}_{i+l, j+m, k+n} \quad (\text{B.3})$$

where $\mathbf{c}$ denotes a $m_x \times m_y \times m_z$ lattice of control points which parameterise the free-form deformation, i, j, k denote the indices of the control points and u, v, w correspond to the relative positions of $\mathbf{x}$ in lattice coordinates. The lattice of control points is defined as a grid with uniform spacing which is placed on the underlying reference image. The optimal transformation is found using a gradient descent minimisation of a cost function associated with the global transformation parameters as well as the local transformation parameters. The cost function comprises two competing goals: The first term represents the cost associated with the voxel-based similarity measure, in this case normalised mutual information (Studholme et al., 1999), while the second term corresponds to a regularization term which constrains the transformation to be smooth.

In the current study no regularization was used. This will most likely not affect results evaluated with overlap measures, but may affect studies interested in folding in the deformation fields. A penalty term for folding in the transformation is described in (Rueckert et al., 2006).

(dof $\simeq 3 \times \# \text{control points} = 1,422,843$)

8.8 JRD-fluid: Jensen-Rényi Divergence fluid

JRD-fluid was developed by Chiang et al. (Chiang et al., 2007) at LONI, UCLA, CA. Chiang provided the command options for the executables and revised the resampling code to include an option for nearest-neighbor interpolation. All registrations were run on LONI’s servers.

JRD-fluid is based on an information-theoretic measure, the Jensen-Rényi divergence. JRD is derived from the joint histogram of two images. Using variational calculus methods, the driving forces are defined throughout the deforming image to maximize the JRD between it and the target image. A viscous fluid regularizer was applied to guarantee diffeomorphic (i.e., smooth, one-to-one) deformation mappings. The resulting partial differential equation was solved iteratively by convolving the applied force field with the Green’s function of the linear differential operator.

(dof $\simeq 128^3 = 2,097,152$)

8.9 ROMEO: Robust multigrid elastic registration based on optical flow

ROMEO was developed by Hellier et al. (Hellier et al., 2001b) at INRIA Rennes, France. Hellier provided the executables for this study.

The ROMEO registration method expresses the registration process as the minimization of a cost function depending on two terms: an optical flow-based similarity measure and a regularization term. The optical flow hypothesis, introduced by Horn and Schunck (Horn and Schunck, 1981), assumes that the luminance of a physical point does not change when the point moves with the flow:

$$f(s + \mathbf{w}_s, t_1) - f(s, t_2) = 0 \quad (\text{B.4})$$

where s is a voxel of the volume, t_1 and t_2 are the indexes of the volumes (temporal indexes for a dynamic acquisition, indexes in a database for multi-subject registration), f is the luminance function and $\mathbf{w}$ the expected 3-D displacement field.

Generally, a linear expansion of this equation is preferred: $\nabla f(s, t) \cdot \mathbf{w}_s + f_t(s, t) = 0$, where $\nabla f(s, t)$ stands for the spatial gradient of luminance and $f_t(s, t)$ is the voxelwise difference between the two volumes. The resulting set of undetermined equations has to be complemented with some prior on the

deformation field. This prior is defined according to the quadratic difference of the deformation field computed between neighbours. Using an energy-based framework the regularization problem may be formulated as the minimization of the following cost function:

$$U(\mathbf{w}; f) = \sum_{s \in S} [\nabla f(s, t) \cdot \mathbf{w}_s + f_t(s, t)]^2 + \alpha \sum_{\langle s, r \rangle \in \mathcal{C}} \|\mathbf{w}_s - \mathbf{w}_r\|^2 \quad (\text{B.5})$$

where S is the voxel lattice, $\mathcal{C}$ is the set of neighboring pairs w.r.t. a given neighborhood system $\mathcal{V}$ on S ($\langle s, r \rangle \in \mathcal{C} \Leftrightarrow s \in \mathcal{V}(r)$), and α controls the balance between the two energy terms. The first term is the linear expansion of the luminance conservation equation and represents the interaction between the field and the data. The second term is the smoothness constraint. In order to cope with large displacements, an incremental multi-resolution procedure is used to construct a pyramid of volumes by successive Gaussian blurring and subsampling.

(dof $\simeq 2$ million for 300,000 rigid and 50,000 affine estimates)

8.10 SICLE: *Small-deformation, Inverse-Consistent, Linear-Elastic image registration*

SICLE was developed by Christensen et al. (Christensen, 1999; Christensen and Johnson, 2001; Johnson and Christensen, 2002) at the University of Iowa. Song provided the executables and helped to install the software.

SICLE is based on the principle of jointly estimating the forward h and reverse g transformations between two images while minimizing the inverse consistency error $\|h(x) - g^{-1}(x)\| + \|g(x) - h^{-1}(x)\|$. Ideally, the forward transformation h from image T to S and the reverse transformation g from S to T should be uniquely determined and should be inverses of one another. However, estimating h and g independently as with most uni-directional image registration algorithms rarely results in an inverse consistent set of transformations due to a large number of local minima. Jointly estimating the forward and reverse transformations provides additional correspondence information helping to minimize correspondence errors.

The SICLE image registration algorithm iteratively minimizes the following cost function

$$\begin{aligned} C = & \sigma \int_{\Omega} |T_i(h_{i,j}(x)) - T_j(x)|^2 + |T_j(h_{j,i}(x)) - T_i(x)|^2 dx \\ & + \rho \int_{\Omega} \|\mathcal{L}u_{i,j}(x)\|^2 + \|\mathcal{L}u_{j,i}(x)\|^2 dx \\ & + \chi \int_{\Omega} \|u_{i,j}(x) - \tilde{u}_{j,i}(x)\|^2 + \|u_{j,i}(x) - \tilde{u}_{i,j}(x)\|^2 dx. \end{aligned} \quad (\text{B.6})$$

where the parameters σ , ρ , and χ are weighting constants used to vary the influence of each term of the cost function and Ω is the image domain. The intensity of the images are normalized between 0 and 1. The first integral of the cost function defines the correspondence (squared intensity difference) between the deformed template and target images and between the deformed target and template images, respectively. The second integral is used to regularize the forward and reverse displacement fields $u_{i,j}$ and $u_{j,i}$ respectively, and is minimized when the forward and reverse displacement fields satisfy the properties of the linear elastic model. In the linear elastic model, the linear differential operator $\mathcal{L}u(x) = \alpha \nabla^2 u(x) + \beta \nabla(\nabla \cdot u(x)) + \gamma u(x)$ penalizes large second and cross derivatives in the x , y , and z directions in the displacement fields. The constant weighting parameters α , β , and γ are used to vary the influence of the terms of $\mathcal{L}u(x)$. The third integral is called the inverse consistency constraint and is minimized when the forward and reverse transformations $h_{i,j}$ and $h_{j,i}$, respectively, are inverses of each other.

The cost function in Equation B.6 is minimized using the gradient descent method described in (Christensen and Johnson, 2001) where each component of the displacement field is parameterized in terms of a 3-D Fourier series. A multi-resolution approach in the frequency and spatial domains is taken to estimate the Fourier Series coefficients by first estimating the low frequency components then increasing the number of harmonics as the estimation progresses.

(dof = 7,986 for 10 harmonics)

8.11 SPM5: Statistical Parametric Mapping

SPM5 registration algorithms were developed by Ashburner et al. (Ashburner and Friston, 1999, 2005; Ashburner, 2007) at the Functional Imaging Laboratory, UK. Normalize, Unified Segmentation, and the DARTEL Toolbox are part of the SPM5 package and may be downloaded from <http://www.fil.ion.ucl.ac.uk/spm/software/spm5/>. Ashburner provided Matlab scripts for composing the Normalization, Unified Segmentation and DARTEL Toolbox transforms.

Five methods were evaluated from the SPM5 package: regular and “SPM2-type” Normalization, Unified Segmentation, and regular and pairwise implementations of the DARTEL Toolbox. It is expected that the SPM user will register original brain images (non-skull-stripped and in their native space) to idealized templates. The template images supplied with SPM5 conform to the space defined by the ICBM, NIH P-20 project, and approximate that of the space described in the atlas of Talairach and Tournoux (Talairach and Tournoux, 1988). In this study, the “SPM2-type” Normalization and the pairwise DARTEL were the only SPM methods that were applied in the same

manner as all of the other non-SPM algorithms, that is, by directly registering one skull-stripped brain image in MNI space to another (rather than to one another via a coregistration template).

SPM5 regular and “SPM2-type” Normalization (Ashburner and Friston, 1999)

The algorithms work by minimizing the sum of squares difference between the image which is to be normalized and a linear combination of one or more template images (in the case of “SPM2-type” Normalization, a single target image). The first step of the normalization is to determine the optimum 12-parameter affine transformation. A Bayesian framework is used, such that the registration searches for the solution that maximizes the *a posteriori* probability of it being correct. That is, it maximizes the product of the likelihood function (derived from the residual squared difference) and the prior function (which is based on the probability of obtaining a particular set of zooms and shears).

The affine registration is followed by estimating nonlinear deformations, whereby the deformations are defined by a linear combination of 3-D discrete cosine transform (DCT) basis functions. The parameters represent coefficients of the deformations in three orthogonal directions. The matching involves simultaneously minimizing the bending energies of the deformation fields and the residual squared difference.

(dof $\simeq$ 1,000)

SPM5 Unified Segmentation (Ashburner and Friston, 2005)

This approach uses a probabilistic generative model that combines image registration, tissue classification, and bias correction. The log-likelihood objective function is based on a mixture of Gaussians, and is extended to incorporate a smooth intensity variation and nonlinear registration with tissue probability maps. A small-deformation registration model is used, which is parameterised by a linear combination of around 1,000 cosine transform basis functions.

In addition to simple inter-subject registration, the Unified Segmentation approach also performs classification of brain tissues into gray and white matter, as well as bias correction and rudimentary skull-stripping.

(dof $\simeq$ 1,000)

SPM5 DARTEL Toolbox: Diffeomorphic Anatomical Registration using Exponentiated Lie algebra (Ashburner, 2007)

This approach is an extension of the approach described by Ashburner (Ashburner, 2007), and is intended to register tissue class images from multiple subjects with a common template. The tissue class images are typically gray and white matter, which have been extracted using the SPM5 Unified Segmentation approach. The generative model assumes that the template encodes the mean of a multinomial distribution. Pre-computed templates can be used, but the toolbox also allows them to be iteratively generated from a population of subjects using a “congealing” approach. Nonlinear registration is considered as a local optimization problem, which is done using a Levenberg-Marquardt strategy. The necessary matrix solutions are obtained in reasonable time using a multi-grid method. A constant Eulerian velocity framework is used, which allows a rapid scaling and squaring method to be used in the computations.

$$(\text{dof} = 121 \times 145 \times 121 \times 3 = 6,368,835)$$

8.12 SyN: Symmetric Normalization

SyN was developed by Brian Avants et al. (Avants et al., 2008) at the University of Pennsylvania. Brian Avants provided the executables for a beta version for this study. Several implementation improvements have been made since the algorithm was evaluated in this work. SyN is available in the ANTS toolkit; a new release with tutorial is available at <http://www.picsl.upenn.edu/ANTS/>.

The symmetric normalization (SyN) methodology uses a symmetric parameterization of the shortest path of diffeomorphisms connecting two neuroanatomical configurations. The SyN formulation uses a bidirectional gradient descent optimization which gives results that are unbiased with respect to the input images. SyN also provides forward and inverse continuum mappings that are consistent within the discrete domain and enables both large and subtle deformations to be captured. Specific performance characteristics depend upon the range of similarity metrics chosen for the study and the velocity field regularization. The current study uses Gaussian smoothing of the velocity field and a gradient-based optimization of an approximate cross-correlation (CC) similarity metric with CC evaluated in a window of size 5x5x5 voxels. A variety of other similarity measures are available, including robust optical flow, mutual information and additional correlation measures.

$$(\text{dof} \simeq 4 \times \# \text{voxels} = 28 \text{million})$$

References

- Allen, J. S., Bruss, J., Mehta, S., Grabowski, T., Brown, C. K., Damasio, H., 2008. Effects of spatial transformation on regional brain volume estimates. *NeuroImage* 42, 535–547.
- Andersson, J., Smith, S., Jenkinson, M., 2008. FNIRT - FMRIB's Non-linear Image Registration Tool. *Human Brain Mapping* 2008, Poster #496.
- Ardekani, B., Braun, M., Hutton, B. F., Kanno, I., Iida, H., Aug. 1995. A fully automatic multimodality image registration algorithm. *Journal of Computer Assisted Tomography* 19, 615–623.
- Ardekani, B. A., Guckemus, S., Bachman, A., Hoptman, M. J., Wojtaszek, M., Nierenberg, J., Mar. 2005. Quantitative comparison of algorithms for inter-subject registration of 3D volumetric brain MRI scans. *Journal of Neuroscience Methods* 142, 67–76.
- Ashburner, J., Oct. 2007. A fast diffeomorphic image registration algorithm. *NeuroImage* 38, 95–113.
- Ashburner, J., Friston, K. J., 1999. Nonlinear spatial normalization using basis functions. *Human Brain Mapping* 7, 254–266.
- Ashburner, J., Friston, K. J., Jul. 2005. Unified segmentation. *NeuroImage* 26, 839–851.
- Avants, B. B., Epstein, C. L., Grossman, M., Gee, J. C., 2008. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis* 12, 26–41.
- Bechhofer, R. E., Mar. 1954. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics* 25, 16–39.
- Beg, M. F., Miller, M. I., Trounev, A., Younes, L., Feb. 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision* 61, 139–157.
- Cachier, P., Bardinet, E., Dormont, D., Pennec, X., Ayache, N., 2003. Iconic feature based nonrigid registration: the PASHA algorithm. *Computer Vision and Image Understanding* 89, 272–298.
- Caviness, V., Meyer, J., Makris, N., Kennedy, D., 1996. MRI-based topographic parcellation of human neocortex: an anatomically specified method with estimate of reliability. *Journal of Cognitive Neuroscience* 8, 566–587.
- Chiang, M.-C., Dutton, R. A., Hayashi, K. M., Lopez, O. L., Aizenstein, H. J., Toga, A. W., Becker, J. T., Thompson, P. M., 2007. 3D pattern of brain atrophy in HIV/AIDS visualized using tensor-based morphometry. *NeuroImage* 34, 44–60.
- Christensen, G., Johnson, H., 2001. Consistent image registration. *IEEE Transactions on Medical Imaging* 20, 568–582.
- Christensen, G. E., 1999. Consistent linear-elastic transformations for image matching. *Proceedings of Information Processing in Medical Imaging: IPMI*

- 99 1613, 224–237.
- Clouchoux, C., Coulon, O., Rivière, D., Cachia, A., Mangin, J.-F., Régis, J., 2005. Anatomically constrained surface parameterization for cortical localization. *Medical Image Computing and Computer-Assisted Intervention: MICCAI 2005* 3750, 344–351.
- Cointepas, Y., Mangin, J.-F., Garnero, L., Poline, J.-B., Benali, H., Jun. 2001. BrainVISA: Software platform for visualization and analysis of multi-modality brain data. *NeuroImage* 13, 98.
- Collins, D. L., Evans, A. C., 1997. ANIMAL: validation and applications of non-linear registration-based segmentation. *International Journal of Pattern Recognition and Artificial Intelligence* 11, 1271–1294.
- Collins, D. L., Holmes, C. J., Peters, T. M., Evans, A. C., 1995. Automatic 3-d model-based neuroanatomical segmentation. *Human Brain Mapping* 3, 190–208.
- Collins, D. L., Neelin, P., Peters, T. M., Evans, A. C., Apr. 1994. Automatic 3D intersubject registration of mr volumetric data in standardized talairach space. *Journal of Computer Assisted Tomography* 18, 192–205.
- Crum, W. R., Camara, O., Rueckert, D., Bhatia, K. K., Jenkinson, M., Hill, D. L. G., 2005. Generalised overlap measures for assessment of pairwise and groupwise image registration and segmentation. *Medical Image Computing and Computer-Assisted Intervention: MICCAI 2005* 3749, 99–106.
- Crum, W. R., Griffin, L. D., Hill, D. L. G., Hawkes, D. J., 2003. Zen and the art of medical image registration: correspondence, homology, and quality. *NeuroImage* 20, 1425–1437.
- Crum, W. R., Rueckert, D., Jenkinson, M., Kennedy, D., Smith, S. M., 2004. A framework for detailed objective comparison of non-rigid registration algorithms in neuroimaging. *Medical Image Computing and Computer-Assisted Intervention: MICCAI 2004* 3216, 679–686.
- Dale, A. M., Fischl, B., Sereno, M. I., Feb. 1999. Cortical surface-based analysis I: Segmentation and surface reconstruction. *NeuroImage* 9, 179–194.
- Drury, H. A., Essen, D. C. V., Anderson, C. H., Lee, C. W., Coogan, T. A., Lewis, J. W., 1996. Computerized mappings of the cerebral cortex: A multiresolution flattening method and a surface-based coordinate system. *Journal of Cognitive Neuroscience* 8, 1–28.
- Essen, D. C. V., Drury, H. A., Dickson, J., Harwell, J., Hanlon, D., Anderson, C. H., 2001. An integrated software suite for surface-based analyses of cerebral cortex. *Journal of the American Medical Informatics Association : JAMIA* 8, 443–59.
- Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., Peters, T. M., Oct. 1993. 3D statistical neuroanatomical models from 305 MRI volumes. *Nuclear Science Symposium and Medical Imaging Conference* 3, 1813–1817.
- Evans, A. C., Collins, D. L., Milner, B., 1992. An MRI-based stereotactic brain atlas from 300 young normal subjects. *Proc. of the 22nd Symposium of the Society for Neuroscience, Anaheim* 408.

- Fiez, J. A., Damasio, H., Grabowski, T. J., 2000. Lesion segmentation and manual warping to a reference brain: intra- and interobserver reliability. *Human Brain Mapping* 9, 192–211.
- Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A. M., 2002. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355.
- Fischl, B., Sereno, M. I., Tootell, R. B. H., Dale, A. M., 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping* 8, 272–284.
- Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., Dale, A. M., 2004. Automatically parcellating the human cerebral cortex. *Cerebral Cortex* 14, 11–22.
- Friston, K. J., Ashburner, J., Poline, J. B., Frith, C. D., Heather, J. D., Frackowiak, R. S. J., 1995. Spatial registration and normalization of images. *Human Brain Mapping* 2, 165–189.
- Gee, J. C., Reivich, M., Bajcsy, R., 1993. Elastically deforming 3D atlas to match anatomical brain images. *Journal of Computer Assisted Tomography* 17, 225–236.
- Gerig, G., Jomier, M., Chakos, M., 2001. Valmet: A new validation tool for assessing and improving 3D object segmentation. *Lecture Notes in Computer Science* 2208, 516–524.
- Gholipour, A., Kehtarnavaz, N., Briggs, R., Devous, M., Gopinath, K., 2007. Brain functional localization: A survey of image registration techniques. *IEEE Transactions on Medical Imaging* 26, 427–451.
- Grachev, I. D., Berdichevsky, D., Rauch, S. L., Heckers, S., Alpert, N. M., 1998. Anatomic landmark-based method for assessment of intersubject image registration techniques: Woods vs Talairach [abstract]. 8th Annual Meeting of the Organization for Human Brain Mapping, Brighton, England.
- Grachev, I. D., Berdichevsky, D., Rauch, S. L., Heckers, S., Kennedy, D. N., Caviness, V. S., Alpert, N. M., 1999. A method for assessing the accuracy of intersubject registration of the human brain using anatomic landmarks. *NeuroImage* 9, 250–268.
- Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D., Hammers, A., Oct. 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage* 33, 115–126.
- Hellier, P., 2003. Consistent intensity correction of MR images. *International Conference on Image Processing: ICIP 2003* 1, 1109–1112.
- Hellier, P., Ashburner, J., Corouge, I., Barillot, C., Friston, K., 2002. Inter subject registration of functional and anatomical data using SPM. *Medical Image Computing and Computer-Assisted Intervention: MICCAI 2002* 2489, 590–597.
- Hellier, P., Barillot, C., Corouge, I., Gibaud, B., Goualher, G. L., Collins,

- D. L., Evans, A., Malandain, G., Ayache, N., Christensen, G. E., Johnson, H. J., Sep. 2003. Retrospective evaluation of intersubject brain registration. *IEEE transactions on medical imaging* 22, 1120–30.
- Hellier, P., Barillot, C., Corouge, I., Gibaud, B., Goualher, G. L., Collins, L., Evans, A., Malandain, G., Ayache, N., 2001a. Retrospective evaluation of inter-subject brain registration. *Lecture Notes in Computer Science* 2208, 258–266.
- Hellier, P., Barillot, C., Mémin, E., Pérez, P., 2001b. Hierarchical estimation of a dense deformation field for 3D robust registration. *IEEE Transactions on Medical Imaging* 20, 388–402.
- Horn, K., Schunck, B., 1981. Determining optical flow. *Artificial Intelligence* 17, 185–203.
- Jaccard, P., 1912. The distribution of flora in the alpine zone. *The New Phytologist* 11, 37–50.
- Jannin, Grova, Maurer, 2006. Model for defining and reporting reference-based validation protocols in medical image processing. *International Journal of Computer Assisted Radiology and Surgery* 1, 63–73.
- Jenkinson, M., Smith, S., Jun. 2001. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis* 5, 143–156.
- Johnson, H., Christensen, G., 2002. Consistent landmark and intensity-based image registration. *IEEE Transactions on Medical Imaging* 21, 450–461.
- Karacali, B., Davatzikos, C., Jul. 2004. Estimating topology preserving and smooth displacement fields. *IEEE Transactions on Medical Imaging* 23, 868–880.
- Kittler, J., Hatef, M., Duin, R. P. W., Matas, J., 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 226–239.
- Klein, A., Hirsch, J., 2005. Mindboggle: a scatterbrained approach to automate brain labeling. *NeuroImage* 24(2), 261–280.
- Klein, A., Mensh, B., Ghosh, S., Tourville, J., Hirsch, J., 2005. Mindboggle: Automated brain labeling with multiple atlases. *BMC Medical Imaging* 5:7.
- Mandl, R., Jansma, J., Slagter, H., Collins, D., Ramsey, N., 2000. On the validity of associating stereotactic coordinates with anatomical nomenclature. *Human Brain Mapping: HBM 2000: S539*.
- Menke, J., Martinez, T., 2004. Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons. In: *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*. Vol. 2. pp. 1331–1335 vol.2.
- Nieto-Castanon, A., Ghosh, S. S., Tourville, J. A., Guenther, F. H., 2003. Region of interest based analysis of functional imaging data. *NeuroImage* 19, 1303–1316.
- Paglieroni, D. W., 1992. Distance transforms: properties and machine vision applications. *CVGIP: Graph. Models Image Process.* 54, 56–74.
- Robbins, S., Evans, A. C., Collins, D. L., Whitesides, S., Sep. 2004. Tuning and comparing spatial normalization methods. *Medical Image Analysis* 8,

- 311–323.
- Rogelj, P., Kovacic, S., Gee, J., 2002. Validation of a nonrigid registration algorithm for multimodal data. *Proc. SPIE* 4684, 299–307.
- Rohlfing, T., Russakoff, D. B., Maurer, C. R., 2004. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. *IEEE Transactions on Medical Imaging* 23, 983–994.
- Roland, P., Geyer, S., Amunts, K., Schormann, T., Schleicher, A., Malikovic, A., Zilles, K., 1997. Cytoarchitectural maps of the human brain in standard anatomical space. *Human Brain Mapping* 5, 222–227.
- Rosenfeld, A., Pfaltz, J. L., 1966. Sequential operations in digital picture processing. *J. ACM* 13, 471–494.
- Rueckert, D., Aljabar, P., Heckemann, R. A., Hajnal, J. V., Hammers, A., 2006. Diffeomorphic registration using B-splines. *9th International Conference on Medical Image Computing and Computer-Assisted Intervention: MICCAI 2006* 4191, 702–709.
- Rueckert, D., Sonoda, L., Hayes, C., Hill, D., Leach, M., Hawkes, D., 1999. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging* 18, 712–721.
- Shattuck, D. W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K. L., Poldrack, R. A., Bilder, R. M., Toga, A. W., Feb. 2008. Construction of a 3D probabilistic atlas of human cortical structures. *NeuroImage* 39, 1064–1080.
- Shen, D., Davatzikos, C., Nov. 2002. Hammer: hierarchical attribute matching mechanism for elastic registration. *IEEE transactions on medical imaging* 21, 1421–39.
- Studholme, C., Hill, D. L. G., Hawkes, D. J., 1999. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition* 32, 71–86.
- Talairach, J., Szikla, G., 1967. *Atlas d’anatomie stereotaxique du telencephale: etudes anatomo-radiologiques*. Masson & Cie, Paris.
- Talairach, J., Tournoux, P., 1988. *Co-planar stereotaxic atlas of the human brain*. Thieme Medical Publishers, New York.
- Thirion, J. P., Sep. 1998. Image matching as a diffusion process: an analogy with Maxwell’s demons. *Medical Image Analysis* 2, 243–260.
- Tourville, J., Guenther, F., 2003. A cortical and cerebellar parcellation system for speech studies. *Boston University Technical Reports CAS/CNS-03-022*.
- Toussaint, N., Souplet, J.-C., Fillard, P., 2007. MedINRIA: Medical image navigation and research tool by INRIA. *Proceedings of MICCAI’07 Workshop on Interaction in Medical Image Analysis and Visualization* 4791, 1–8.
- Towle, V. L., Khorasani, L., Uftring, S., Pelizzari, C., Erickson, R. K., Spire, J.-P., Hoffmann, K., Chu, D., Scherg, M., Jul. 2003. Noninvasive identification of human central sulcus: a comparison of gyral morphology, functional mri, dipole localization, and direct cortical mapping. *NeuroImage* 19, 684–97.

- Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2007. Non-parametric diffeomorphic image registration with the demons algorithm. *Medical Image Computing and Computer-Assisted Intervention: MICCAI 2007* 4792, 319–326.
- Warfield, S., Zou, K., Wells, W., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging* 23, 903–921.
- Woods, R. P., Grafton, S. T., Holmes, C. J., Cherry, S. R., Mazziotta, J. C., Feb. 1998. Automated image registration: I. general methods and intrasubject, intramodality validation. *Journal of Computer Assisted Tomography* 22, 139–152.
- Xiong, J., Rao, S., Jerabek, P., Zamarripa, F., Woldorff, M., Lancaster, J., Fox, P. T., Sep. 2000. Intersubject variability in cortical activations during a complex language task. *NeuroImage* 12, 326–339.
- Yassa, M. A., Stark, C. E. L., 2009. A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe. *NeuroImage* 44, 319–327.
- Zijdenbos, A., Dawant, B., Margolin, R., Palmer, A., 1994. Morphometric analysis of white matter lesions in mr images: method and validation. *IEEE Transactions on Medical Imaging* 13, 716–724.