HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2021 February 15.

Published in final edited form as:

Neuroimage. 2021 February 15; 227: 117657. doi:10.1016/j.neuroimage.2020.117657.

Registration quality filtering improves robustness of voxel-wise analyses to the choice of brain template

Nelson Gil[a,b], **Michael L. Lipton**[c,d,e,f], **Roman Fleysher**[c,d,*]

[a]Department of Systems and Computational Biology, Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, NY, USA

[b]Department of Biochemistry, Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, NY, USA

[c]Department of Radiology, Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, NY, USA

[d]Gruss Magnetic Resonance Research Center, Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, NY, USA

[e]Department of Psychiatry and Behavioral Sciences, Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, NY, USA

[f]Dominick P. Purpura Department of Neuroscience, Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, NY, USA

Abstract

Motivation: Many clinical and scientific conclusions that rely on voxel-wise analyses of neuroimaging depend on the accurate comparison of corresponding anatomical regions. Such comparisons are made possible by registration of the images of subjects of interest onto a common brain template, such as the Johns Hopkins University (JHU) template. However, current image registration algorithms are prone to errors that are distributed in a template-dependent manner. Therefore, the results of voxel-wise analyses can be sensitive to template choice. Despite this problem, the issue of appropriate template choice for voxel-wise analyses is not generally addressed in contemporary neuroimaging studies, which may lead to the reporting of spurious results.

Results: We present a novel approach to determine the suitability of a brain template for voxel-wise analysis. The approach is based on computing a "distance" between automatically-generated atlases of the subjects of interest and templates that is indicative of the extent of subject-to-template registration errors. This allows for the filtering of subjects and candidate templates based

*Corresponding author. roman.fleysher@einsteinmed.org (R. Fleysher).

Credit authorship contribution statement

Nelson Gil: Conceptualization, Methodology, Investigation, Software, Visualization, Writing - original draft. **Michael L. Lipton:** Conceptualization, Supervision, Writing - review & editing. **Roman Fleysher:** Conceptualization, Methodology, Investigation, Supervision, Visualization, Writing - review & editing.

Declaration of Competing Interest

None.

on a quantitative measure of registration quality. We benchmark our approach by evaluating alternative templates for a voxel-wise analysis that reproduces the well-known decline in fractional anisotropy (FA) with age. Our results show that filtering registrations minimizes errors and decreases the sensitivity of voxel-wise analysis to template choice. In addition to carrying important implications for future neuroimaging studies, the developed framework of template induction can be used to evaluate robustness of data analysis methods to template choice.

Keywords

Image registration; Voxel-wise analysis; Brain template; Subject-specific analysis

1. Introduction

Image registration is the mapping of images onto a common coordinate space with the goal of aligning their homologous regions (Oliveira and Tavares, 2014; Sotiras et al., 2013). Accurate image registration is particularly important in voxel-wise analyses of brain magnetic resonance images (MRIs), where the current standard of practice is to have the images of the subjects of interest mapped onto a "template" image such as the Johns Hopkins University (JHU) brain template (Hua et al., 2008a; Mori et al., 2009). Correct and precise mappings are necessary in order to make valid statistical inferences about spatial differences in MRI-derived parameters (e.g. fractional anisotropy (FA)). This represents a central task in medical image analysis: the imaging of a patient can be compared to the imaging of a healthy control group to identify the patient's structural or functional differences that may be indicative of pathology (Despotovic et al., 2015; Han et al., 2017; Kinnunen et al., 2011).

Nonlinear mathematical functions that carry out the mapping from one brain to another during registration are called morphisms. The neuroanatomical variability among individuals and signal noise in the image acquisition process virtually ensure that the anatomy of one brain can never be exactly mapped onto the anatomy of another in a voxel-by-voxel fashion (Klein et al., 2009; Grachev et al., 1999; Ardekani et al., 2005; Suri et al., 2015); morphism/registration misalignments leading to errors will always be present. In addition, due to neuroanatomical variability, the exact set of registration errors differs depending on the brains being aligned (Despotovic et al., 2015; Suri et al., 2015; Crum et al., 2004), implying that the results of template-based voxel-wise analyses are sensitive to the choice of the template image. Thus, voxel-wise analysis may be more accurate for sets of brains that have minimal errors in the morphisms that relate them to one another.

A long-standing approach to mitigating voxel-wise registration errors has been to report MRI measurements in terms of their values at clusters of adjacent voxels rather than individual voxels (Friston et al., 1994). Nevertheless, it was recognized over fifteen years ago that the presumed correspondence of neuroanatomical regions between subjects and a chosen template can harm the validity of brain MRI voxel-wise analyses because of the unknown distributions of registration errors (Crum et al., 2003), which are specific to subject-template pairs. Dependence of the registration errors on the subjects and template highlights the importance of subject-specific analysis and judicious template selection in

voxel-wise imaging studies (Suri et al., 2015; Mayer et al., 2018; Viviani et al., 2007; Douaud et al., 2011; Keihaninejad et al., 2012). As an example, a previous study by our group examined the effects of template choice on the voxel-wise analysis of a set of patients with mild traumatic brain injury, which are expected to have clusters of low FA arising from residual white matter injury (Suri et al., 2015). This study showed that voxel-wise FA cluster analyses over the JHU and Montreal Neurological Institute (MNI) (Aubert-Broche et al., 2006) templates found many more locations of low FA clusters than the analysis using subject-based templates (Fig. 1). Furthermore, the locations of most low FA clusters over the JHU and MNI templates disagreed and were found to correlate with the locations of misregistered voxels. On the other hand, the FA clusters that agreed between the JHU and MNI templates also agreed with the locations found by the subject-based template. A general approach to study-specific template selection is to find or create a template to which registration errors from the subjects can be identified as being below some quantitative threshold.

Methods of registration error quantification are fundamentally based on measuring the displacement between the positions of voxels comprising specific expert-defined anatomical landmarks, and include Bayesian (Risholm et al., 2013), machine-learning-based (Muenzing et al., 2012; Kearney et al., 2018), and analytical (Datteri et al., 2015) approaches. Error estimation methods have been commonly used in a variety of medical applications requiring repeated or real-time image guidance, such as surgery and cancer radiotherapy (Hoffmann et al., 2014; Mascott et al., 2006; Elhawary et al., 2010). Importantly, these applications depend on the quantification of registration errors between different images of the same patient taken at different times, with potential alteration of brain shape due to the procedure. To our knowledge, only a few previous studies have explicitly focused on the effects of template-dependent errors on group-based voxel-wise analysis (Keihaninejad et al., 2012; Acheson et al., 2017). For example, one study of patients exhibiting neurodegeneration due to Alzheimer’s disease found that creating a template based on a morphometric average of the study group led to fewer subject-to-template registration errors on voxel-wise analysis (Keihaninejad et al., 2012). Another study of FA reproducibility in the setting of tract-based spatial statistics assessed registration quality between their subjects and template by computing the average projection distance for a group-wide mean FA skeleton: having too great of a distance to the FA skeleton suggests poor registration quality (Acheson et al., 2017). The protocol in this study used a form of the minimal deformation target (MDT) (Hua et al., 2008 b; Kochunov et al., 2002): the creation of a template that minimizes the average morphometric displacement to a specific set of subjects. While these methods may be effective for groups of subjects that are highly similar to each other, the ultimate output of morphometric averaging and MDT is dependent on the specific morphism algorithm used and is not directly connected to the suitability of the template. To illustrate, consider a mock morphism algorithm that does not transform the study group subjects at all (an identity function): the morphometric average would correspond to directly averaging non-registered study group images, yielding a non-sensical template image to which further registrations would also be non-sensical. MDT would find this template acceptable because deformations are identically zero. The FA skeleton projection step would detect large misalignments. Even though these methods have been found to be helpful for template construction, the

overarching issue of specific template choice and quality of registrations to it is not commonly considered in the current literature, which is concerning given the high potential for the reporting of false positive and negative findings.

Two main difficulties arise in studying and addressing the influence of template choice on the results of voxel-wise analyses based on deformation of images to the template: the absence of “ground truth” answers and the voxel-wise comparison of results obtained over different individual candidate templates. To be compared, all results must first be morphed to a common “master template”, something which cannot be accomplished without morph errors. In the presence of morph errors, observed discrepancies cannot be unambiguously attributed to the poor choice of an individual candidate template because they can be potentially explained by poor registration of that template to the master template. Similarly, results of a voxel-wise analysis over the master template cannot be considered “ground truth” because of morph errors during registration of the subject data to the template. If morph errors could be “turned off” during generation of the ground truth and during registration of the results over individual candidate templates to the master template, then the morph errors to the candidate templates could be isolated for analysis (Fig. 2).

In this ideal scenario, spatial clusters obtained in a voxel-wise analysis of the subject data morphed to the master template represent the “ground truth”. The same analysis is performed over the various candidate templates and clusters are carried over to the master template for comparison without misalignment. Since the subjects used in the analysis are always the same, and because there is no randomness in the analysis, the degree to which these clusters match the ground truth can be directly attributed to the suitability of the candidate templates. Indeed, if the clusters match the ground truth, the candidate template may be termed as “good”. Conversely, if the clusters do not match the ground truth, the candidate template may be termed as “bad”. In turn, the suitability of the candidate is determined by the individual morphisms between subject brains and the template. This allows detailed analysis of the quality of the individual morphisms and development of a criterion with which to screen and filter out aberrant transformations thus converting a “bad” template into a “good” template, albeit for a subset of subjects.

A related important and well-known issue that arises in template selection is the potential for selection bias (Thompson and Toga, 2002; Thompson et al., 2000b): a given template may have intrinsic, but not necessarily observable, properties that cause it to be more favorable for comparison with certain subgroups of subjects within a dataset. Conversely, a template may generate inaccurate morphisms and thus be a “bad” template for an “outgroup” within a dataset. For example, a template based on a young brain may introduce selection bias favoring other young brains due to their greater baseline anatomical similarity. Another example would be when different MRI datasets are pooled together for a common analysis: the differences in data collection and post-processing between datasets could cause a template constructed based on a brain from one dataset to be biased against those of other datasets. While one possible approach to mitigate selection bias would be the creation of averaged templates for a dataset (e.g. by implementing MDT as discussed above), this may still not result in adequate morphisms because the averaged template may be too different from any individual subject; in other words, the template may be too “fuzzy” (Wu et al.,

2016). While several approaches have been developed to attempt to minimize the bias introduced by averaging (Wu et al., 2016; Joshi et al., 2004; Lyu et al., 2015), no way of confirming the actual resulting bias in a test dataset currently exists. Therefore, selection bias in the results of these approaches may go unnoticed throughout a data analysis pipeline and lead to false positive or false negative results. A morphism quality measure would allow for the early detection of these errors and adequately inform further data analyses. Potentially, it would also allow investigation of the underlying reasons of the morph errors.

Overall, the present study is organized into three main components: 1) We define a "ground truth" for benchmarking, 2) We propose a morphism quality measure to make voxel-wise analyses more template-independent, and 3) We evaluate effect of morphism quality filtering algorithm against the ground truth. Specifically, for component 1) we developed a template and subject induction process to manufacture the exact, error-free morphisms needed for ground truth generation and results comparison across alternative candidate template choices (Fig. 2). For benchmarking, we elected to examine how template selection affects the results of a voxel-wise analysis based on deformation of images to a common template that seeks to reproduce the well-known decline in white matter integrity with age (Pfefferbaum et al., 2000; Pfefferbaum and Sullivan, 2003; Fleysher et al., 2018; Kochunov et al., 2012). We use FA maps derived from diffusion tensor imaging as a marker of white matter integrity in a set of young to middle-aged subjects that have no brain pathology. For component 2), we investigate the sensitivity of locations of voxel clusters where FA is significantly associated with age (FA clusters) to the choice of template. Poorly-registered images are filtered out using a new morphism quality measure that utilizes an average of Haussdorf-like distances (Garlapati et al., 2015) to compare FreeSurfer-generated atlases of candidate templates and subjects morphed onto them (Fleysher et al., 2017). We hypothesize that this average inter-atlas distance can characterize the extent of subject-to-template registration errors and determine the suitability of a candidate template for voxel-wise analysis of a set of specific subjects. Finally, we evaluate this hypothesis in component 3) by computing the Dice coefficient overlap (Dice, 1945) between the FA clusters of the ground truth and the FA clusters obtained over candidate templates of varying quality. Our results demonstrate that excluding poorly-registered images dramatically increases the robustness of the voxel-wise analysis to the choice of template.

2. Materials and methods

Data and Code Availability Statement:

Data in this study was obtained from the previously conducted Einstein Lifespan Study (ELS). All third-party code used in this work is available online and cited within this work.

Ethics Statement:

The ELS was approved by the institutional review board of Albert Einstein College of Medicine. All participants provided informed consent in writing.

2.1. Imaging protocol

We used 96 whole brain datasets from healthy 18–55-year-old participants (46% female) of the Einstein Lifespan Study, without known brain pathology or history of neurological or psychiatric disorders. All images were reviewed by an experienced neuroradiologist and determined to be free of clinically significant structural abnormalities, including gross changes due to trauma, infection, or neoplasm. Imaging was performed using a 3.0 T Philips Achieva TX scanner (Philips Medical Systems, Best, The Netherlands) and its 32-channel head coil. The imaging protocol included: T1W: TR/TE/TI = 9.9/4.6/900 msec, flip angle 8 deg, 1 mm isotropic resolution, 128 × 116 × 220 matrix; DTI: TR/TE = 10,000 / 65 msec, 32 diffusion directions, b-value = 800 s/mm, 2 mm isotropic resolution, 240 × 188 × 70 matrix; and field map to correct EPI-related distortions in DTI and small distortions in T1W: TR/TE = 20/2.4 msec, delta TE = 2.3 msec, flip angle 20 deg, 4 mm isotropic resolution, 64 × 64 × 50 matrix. DTI data were corrected for eddy current- and EPI-related distortions, followed by registration to the individual's T1W using FSL tools (Jenkinson et al., 2012) as described in (Fleysher et al., 2018). All original images and intermediate results including brain extraction and intra-subject registration were visually inspected by trained raters using a standardized procedure: raw images were inspected for signs of motion; brain extractions were examined in the axial slice traversing them from the superior to the inferior aspect of the brain; inspection of rigid body registrations began with large structures (ventricles and cerebellum) down to thin cortical sulci. All further analyses in the present work considered only the T1W images and FA maps registered to them.

2.2. Registration framework

The overall structure of the present study has been designed to isolate the effects of subject-to-template registration errors on voxel-wise analyses of FA versus age (Fig. 2). This was done by creating "induced" subject and template images which have exactly-known transformations to a "master template" across which all FA analyses are compared. We recognize that morph errors between any two given brains, A and B, can never be turned off to implement the ideal set-up of Fig. 2. However, given an image A, a morphism can be applied to transform it to another image B'. Image B' does not match image B exactly, but the transformation between images A and B' is exact by construction (Fig. 3A). When we take A to be the master template, we refer to image B' as the "induced template" with image B being the "inductor". Specifically, we selected the JHU brain as the master template and morphed it onto the 96 T1W images of the ELS as inductors to produce 96 induced templates, which we subsequently employed in this study as candidate templates (Fig. 3B). This implements one arm of the exact morphisms in Fig. 2.

A similar induction process is used for the second arm of the exact morphisms in Fig. 2 to "turn off" morph errors between subjects and the master template. This is accomplished by "inducing" the subjects and consists of the following 3 steps (Fig. 3C):

1. Morphisms from each of the 96 ELS subjects to the master template are computed and applied to the respective FA maps. Voxel-wise statistical analysis over the master template will be performed using these FA maps to obtain "ground truth" clusters where FA is associated with age. Because registrations

are not perfect, the FA clusters so obtained are influenced by morph errors. However, we are not focused on studying FA dependence on age; we are using this known association to ensure some clusters (real or artifactual) will be identified. Thus, we treat these morphisms as exact.

2. The morphisms from step 1 are inverted; the inverted morphism is applied to the master template to produce an induced T1W image for each subject.

3. Each induced T1W image is paired with its corresponding original FA map to produce induced subject data. This completes subject induction with an exact morphism to the master template.

The procedures up to this point result in two parallel sets of exactly known transformations: *(i)* between the master template and the 96 induced templates and *(ii)* between the master template and 96 induced subjects as required in the ideal set-up (Fig. 2). Finally, the main morphisms of interest for the present study were generated by registering each of the 96 induced subjects to each of the 96 induced templates. This resulted in a total of 9216 induced-subject-to-induced-template morphisms whose errors' effects on voxel-wise FA versus age analysis could be investigated (Fig. 4).

2.3. Registration algorithm and statistical analysis

All registrations were non-linear and were performed using the 3DWarper module from the Automatic Registration Toolbox (ART) package (Ardekani et al., 2005). Initial inverse morphisms were computed using routines from ART, improved upon using an iterative algorithm as follows: given a morphism from brain A to brain B, and its inverse from brain B to brain A', iterations were continued until the displacement error (combined morphism between brains A and A') was less than 0.01 mm in 99.9% of voxels. Clusters of voxels where FA was significantly correlated with age were identified by performing a voxel-wise t -test with gender as a covariate at a significance level of 0.005 and retaining clusters of 100 or more contiguous voxels within the white matter of the master template (Suri et al., 2015; Hoptman et al., 2008).

2.4. Atlas-distance-based morphism quality measure

The ideal measure of quality for a morphism between two brains would be obtained by tracking how far each voxel on one brain maps from its homologous voxel on the other brain. If this were possible, the voxel-wise displacement error obtained in the process would become a correction to the morphism, making it perfect. Therefore, to characterize the extent of induced-subject-to-induced-template morphism errors, we calculated the average "distance" between homologous anatomical landmarks of the induced templates and the induced subjects morphed onto them (Fig. 5A). For this purpose, each of the 96 induced templates and each of the 9216 induced subjects morphed onto them was segmented using the ASEG module of FreeSurfer version 5.3 (Fischl, 2012). The atlases of the induced templates were defined as "reference atlases", while those of the induced subjects morphed onto the induced templates were defined as "query atlases". For each reference atlas corresponding to a specific induced template, there were 96 query atlases coming from the induced subjects targeting that specific induced template. We computed the average distance

between the reference atlas and each of the query atlases to be used as metric of morphism quality between each pair. Mathematically, the computed distances are elements of the Hauss-dorf distance (Fig. 5B): for each voxel assigned to a specific FreeSurfer region in the reference atlas, we computed the distance, d_{min}, from that voxel to the closest edge of its assigned homologous region in the query atlas. This calculation was implemented using the fast Euclidean distance algorithm (Mishchenko, 2015). The voxel-by-voxel distances d_{min} are averaged over the template brain to produce the final morphism quality measure. A smaller distance indicates better morphism quality.

In a typical application of FreeSurfer, one is interested in accurate segmentation of brain regions so that each voxel is assigned a proper, anatomically meaningful label. For the purposes of filtering, the accuracy and meaning of the label are irrelevant. Instead, a much simpler requirement is in place: reliable delineation of homologous regions on similar brains according to some specific criteria without necessary correspondence to a specific anatomical region. Once identified, boundaries between regions are used to assess morphism quality using the average atlas distance.

We computed the average atlas distance corresponding to all 9216 induced-subject-to-induced-template morphisms generated in this study and found that the distance follows a bimodal distribution (Fig. 6), with a larger peak at about 0.11 mm and a smaller peak at approximately 0.18 mm, separated by a trough at around 0.15 mm. Based on the trough in this distribution, we classified morphisms as “superior” or “inferior”: “superior” morphisms have average atlas distance less than 0.14 mm and “inferior” morphisms have average atlas distance greater than 0.15 mm. Those with distances between 0.14 and 0.15 mm were not studied further. We use “superior” and “inferior” to refer to the quality of individual morphisms and reserve the adjectives “good” and “bad” to describe templates based on how well cluster analyses over them match the ground truth.

2.5. Subject and template subselection for cluster analysis

Overall, the distribution of morphism quality is highly subject-template-pair dependent: some induced subjects have superior morphisms to most induced templates, while for others inferior morphisms are predominant (Fig. 7A). To achieve the goal of the study and to demonstrate reduction of sensitivity to the choice of template when only superior morphisms are retained, we algorithmically searched atlas distance results for a subset of induced subjects with an equal number of superior and inferior morphisms to a subset of induced templates. We found 30 induced subjects with superior morphisms to a subset of 25 induced templates and inferior morphisms to another 25 induced templates (Fig. 7B). Consequently, we refer to the first group of 25 induced templates as superior templates and the second 25 as inferior templates. We then show that superior templates are in fact “good” templates and inferior templates are in fact “bad” templates based on the match of the FA cluster analyses of the 30 subjects over them to the ground truth. The degree to which FA clusters match the ground truth can be directly attributed to the quality of the morphisms and to the templates themselves since the 30 induced subjects are held constant and because there is no randomness in the analysis. For this demonstration, all other induced subjects and templates were discarded.

2.6. Comparison to the ground truth

The overall goal of the present study was to examine the dependence of the locations of voxel clusters where FA was statistically significantly associated with age (FA clusters) on the choice of template for voxel-wise analysis. To that end, we defined ground-truth, "gold standard" FA clusters for the purposes of comparison. These were obtained by applying the exactly-known induced-subject-to-master-template morphism (Fig. 4) to the 30 induced subjects' FA maps and performing voxel-wise FA versus age analysis as described above. We denote the set of "gold standard" clusters as X_G.

To evaluate the effect of registration errors introduced by specific template choices, we applied the error-containing induced-subject-to-induced-template morphisms to the 30 induced subjects' FA maps; we subsequently applied the exactly-known morphism from the corresponding induced template to the master template (Fig. 8A). For each induced template, this procedure results in FA clusters, denoted X_C, that have been mapped onto the master template while being influenced by the error-containing induced-subject-to-induced-template morphisms. To evaluate how well X_C matches the "gold standard" clusters X_G, we computed Dice coefficient D of their overlap (Dice, 1945):

$$D = \frac{2|X_G \cap X_C|}{|X_G| + |X_C|}$$

The Dice coefficient ranges between 0 and 1, where 0 corresponds to disjoint sets and 1 corresponds to identical sets. Interpretation of values in between is context dependent; even values that appear low may still indicate substantial overlap.

3. Results

The proposition examined in this work was that inter-atlas distance characterization of subject-to-template registration errors predicts suitability of a candidate template for voxel-wise FA analysis. Furthermore, usage of a suitable template makes the outcomes of voxel-wise analysis more robust to template choice. To support this hypothesis, we compared the FA versus age clusters produced by exact morphisms over the master template, which we denoted as "gold standard", to those first morphed to the "superior" and "inferior" induced templates and subsequently mapped exactly to the master template (Fig. 8A). Our results show that using the "superior" templates results in an average Dice coefficient of approximately 0.56 (range 0.50 – 0.61), while using "inferior" templates results in an average Dice coefficient of approximately 0.42 (range 0.37 – 0.48) (Fig. 8B).

To get a sense of the physical meaning of these values of the Dice coefficient, which may appear "low" even for "superior" templates, we performed a total of twelve single-voxel (1 mm) shifts (along each positive and negative x, y, and z direction, and along each positive and negative xz, yz, and xy plane diagonals) on the gold standard FA clusters and calculated the Dice coefficient representing the overlap of these shifted clusters with the unshifted gold standard clusters (Fig. 8 B). These values range from approximately 0.55 to 0.73, suggesting that the use of "superior" morphisms results in good FA clusters that very closely match the gold standard, up to a margin of error equivalent to a single diagonal 1 mm voxel shift. On

the other hand, the usage of “inferior” morphisms corresponds to a greater degree of error. It is also important to note that their Dice coefficient range is similarly narrow to that of the “good” clusters. This indicates that the atlas-based morphism quality filter lessens the influence of template choice on voxel-wise analysis: if morphisms are filtered to be “superior” for a particular set of subjects, then one can expect similar results no matter what specific template is chosen. The converse also holds: morphisms that fail the quality filter will lead to “robustly bad” results.

4. Discussion

The premise for the present investigation was that registration/morph errors affect the conclusions of voxel-wise template-based analyses in a template-specific manner (Fig. 1). In order to systematically examine how morphism errors vary with template choice, we performed voxel-wise FA analysis on a group of subjects in two ways: by morphing to superior and inferior templates (Fig. 7). FA clusters identified in the analyses over the superior templates matched the ground truth (Fig. 8). Clusters over the inferior templates did not, illustrating and confirming the sensitivity of voxel-wise analysis to the choice of template despite the same subject data set being used throughout. To overcome both main difficulties on the way to studying and addressing the influence of template choice on the results of voxel-wise analyses (the absence of “ground truth” answers and the comparison of results obtained over different templates), we developed and followed an induction process to generate induced subjects and induced templates (Fig. 3, Fig. 4). To identify superior and inferior templates, we developed a morphism quality filter based on inter-atlas distance (Fig. 5).

4.1. Advantages and interpretation of the inter-atlas distance

We implemented an inter-atlas “distance” averaged over brain regions to quantify the degree of registration error between induced templates and induced subjects morphed onto them (Fig. 5); each unique morphism was therefore associated with a specific average atlas distance. We used distance rather than the Dice or Jaccard indexes frequently used to compare atlases (Klein et al., 2009; Avants et al., 2011; Sabuncu et al., 2009) because length is a natural metric for morphism error quantification and has been found to be more sensitive at detecting differences in performance (Avants et al., 2011).

The bimodal distribution of the atlas distances we observed (Fig. 6) provides a convenient approach to dichotomizing quality of morphisms. It implies that pairs of brains can be characterized as either morphable to each other or not, and that as a result, a yes/no answer can be given to questions of template suitability for the comparison of a particular set of subjects. In other words, the average atlas distance can be used to classify morphisms between a subject and a template - and thus the template itself - as “superior” or “inferior”. Although the shape of the distribution suggests that most morphisms in this study are “superior”, this may be due to a limitation of the way our data set was constructed: since the induced subjects and induced templates are derived from the same original template (the JHU brain), they may have more anatomical similarities than in an actual study. The underlying reasons as to why a specific template turned out to be “superior” or “inferior” are

beyond the scope of the present work. Morphism quality will vary due to any combination of factors related to the specific subject-template pair being compared as well as the specific morphism algorithm. Nonetheless, the fact that there are still morph errors between induced subjects and induced templates is what allows our study design to examine morphism quality. In the absence of morph errors, induced templates would exactly match inductors and induced subjects would exactly match the master template and there would be nothing to study. Thus, we both exploit the presence of morph errors and work around them in a controlled fashion.

Additional considerations in our study design relate to interpreting the numerical value of the inter-atlas distance. Dichotomization of any distribution is always possible, even if it is not bimodal. Therefore, the “superior”/”inferior” atlas distance cutoff point may need to be optimized on the basis of individual templates. In addition, the use of alternative registration algorithms may change the morphism error distribution: a subject-template pair with a “superior” morphism constructed by one registration algorithm may be “inferior” when constructed by another and vice versa. This may alter the optimal cutoff for an overall dataset as well. Finally, the exact segmentation procedure used to generate the atlases may affect the optimal cutoff. To illustrate, we computed the ASEG atlas for the JHU brain using FreeSurfer version 6.0 and calculated its distance to the atlas computed with version 5.3, and found a distance of 0.10 mm from one to the other. The amount of empty space cropped around brain images also affects segmentation with FreeSurfer. Thus, the specific value of 0.14 mm as a cutoff point is consistent with the level of segmentation variability of FreeSurfer and might have to be adjusted to the specific segmentation tool employed.

Our inter-atlas distance metric of morphism quality relies on robust brain segmentation. The simplest and most crude automatic segmentation into gray matter, white matter and CSF is the most robust but of little value because misalignments within their boundaries remain un-detected. At the other extreme, fine-grained segmentation, for example by the WMPARC module of FreeSurfer or some other tool, may not be sufficiently reliable and require manual interventions, which would be impractical in our and many other large studies. We, therefore, chose the ASEG module of FreeSurfer as one in between: not too crude and reliable for brains in the age range of our study. It is perhaps possible to join some small regions of a fine segmentation such as WMPARC into larger ones to create reliable segmentation finer than ASEG allowing more sensitive morphism quality metric than we described. Undoubtedly, an optimal algorithm for assessing morphism quality exists; a search for it is left for future work.

Even though we chose the JHU brain as the master template, there is nothing intrinsically special about this choice. Had we selected some other brain, the gold standard clusters would be different because they are influenced by morphism errors. Nevertheless, the main points of our study would remain: FA clusters are sensitive to the choice of template and that this sensitivity is reduced with the help of morph quality filtering.

4.2. Limitations and potential biases

Practical limitations of our atlas-distance-based morphism quality filter include high computational cost and the decrease in sample size of quality-filtered datasets. In practice,

our approach would require the calculation of atlases for the subjects of interest and the candidate templates the subjects will be registered to. Calculations of atlases over entire datasets are computationally very expensive, although this issue may be mitigated by the relatively limited amount of candidate templates generally considered in neuroimaging. In addition, the filtering process can substantially decrease sample size and make statistical analyses of voxel-wise differences more difficult to justify. This is a necessary trade-off to be able to report reliable, robust results: for example, even if the “best” template is chosen in a traditional (subjective) manner for a particular set of subjects, there may still be unacceptably large subject-to-template registration errors present, meaning that any reported results would not be scientifically valid. Our approach would allow datasets to be “cleaned up” such that any voxel-wise analysis would be valid and reproducible: the specific choice of subjects and templates would not matter if the registrations between them pass the quality filter.

Filtering data based on morphism quality as proposed here can potentially lead to selection bias just as filtering based on any other criterion might. However, if selection bias is created by our approach, it only reveals specific features of the dataset and/or analysis itself: any presence of selection bias suggests that morphism quality depends on some metric in the study. This dependence diminishes the reliability of the results in the absence of morphism quality filtering more severely than in its presence. For example, in general older brains do not morph well onto young brains and vice versa (Fleysher et al., 2017). Therefore, results of a voxel-wise analysis of a set of older brains that employs a young brain template without morphism quality filtering would be largely inaccurate because of the influence of morph errors. Application of the proposed morphism quality filtering could reveal inappropriateness of the choice of the template.

Inadvertent selection bias might be caused by the filter itself if segmentation is tuned to specific features of the image causing false rejection of good morphisms or false acceptance of bad ones. Using the same example, if FreeSurfer provides more reliable segmentation of young brains compared to old brains, then a good morphism between young and old brains might be rejected due to erroneous segmentation of the old brains. It is therefore important to verify that a segmentation tool is appropriate to the images at hand. In our study, all brain images are from healthy young to middle-aged subjects that have no brain pathology or structural abnormalities. At the same time, filtering depends on reliable delineation of homologous regions on similar brains according to some specific criteria and does not depend on the accuracy of their assignment to anatomical labels. Therefore, mischaracterization of morphism quality is unlikely. Even if mischaracterization was present, it was not strong enough to erase the beneficial effects of filtering as demonstrated by the marked improvement in the Dice coefficient of the match to the ground truth. In addition, filtering based on a more robust segmentation would lead to even better than reported match between analyses over superior templates and the ground truth. Similarly, the match to the ground truth would become worse than reported for analyses over the inferior templates.

4.3. Future studies and conclusions

An expected benefit of voxel-wise analyses on larger datasets is the “averaging out” of morph errors across subjects that can increase the power of voxel-wise statistical tests. In our study, we see that for errors corresponding to an average atlas distance exceeding 0.15 mm and a sample size of 30, this averaging out does not yet occur, as the FA clusters for the “inferior” induced templates are significantly displaced from the gold standard clusters (Fig. 8B). The stochastic reduction of errors will scale with the inverse of the square root of the sample size: for example, to halve the magnitude of observed errors one would need to quadruple the sample size. Therefore, the total number of scans required to achieve such an improvement would also quadruple, which may be inconvenient, expensive, or impossible for some studies. In an actual study, a “supreme” template could be identified among all candidate templates as the one with the largest number of subjects passing a morphism quality threshold. In addition, instead of focusing on sample size, future studies could incorporate the morphism quality measure into their statistical workflow in place of voxel-wise statistical tests. The subject and template induction process described herein could be used to verify that new statistical approaches are indeed robust to the choice of template. Similarly, existing statistical approaches can be evaluated for their robustness (or lack thereof) to template choice.

The morphism quality measure presented in this work could be used to indicate the suitability of templates for any voxel-wise analysis focused on a specific set of subjects. Although this work was carried out in the context of FA cluster analysis based on deformation of images to a common template, the quality measure could be useful for other types of voxel-wise studies, such as functional MRI or voxel- and tensor-based morphometry (Thompson et al., 2000a; Ashburner and Friston, 2000). Overall, this work addresses an important gap in knowledge, since template choice is an unresolved problem that is seldom addressed in contemporary voxel-wise studies. Indeed, prior work has suggested that conclusions of voxel-wise analyses may need to be revisited to ensure appropriateness of their choice of template (Suri et al., 2015; Crum et al., 2003; Keihaninejad et al., 2012). The present work suggests a method capable of exploring this problem and ensuring that reported results are robust to template choice and scientifically valid.

Acknowledgments

This work was supported by the National Institute of Aging grant P01AG003949 and National Institute of Neurologic Disorders and Stroke grant R01NS082432. NG was supported by the National Research Service Award (NRSA) individual fellowship F31GM116570 and the Medical Scientist Training Program (MSTP) grant T32GM007288.

References

Acheson A, Wijtenburg SA, Rowland LM, Winkler A, Mathias CW, Hong LE, et al., 2017 Reproducibility of tract-based white matter microstructural measures using the ENIGMA-DTI protocol. Brain Behav. 7, e00615. [PubMed: 28239525]

Ardekani BA, Guckemus S, Bachman A, Hoptman MJ, Wojtaszek M, Nierenberg J, 2005 Quantitative comparison of algorithms for inter-subject registration of 3D volumetric brain MRI scans. J. Neurosci. Methods 142, 67–76. [PubMed: 15652618]

Ashburner J, Friston KJ, 2000 Voxel-based morphometry–the methods. Neuroimage 11, 805–821. [PubMed: 10860804]

Aubert-Broche B, Evans AC, Collins L, 2006 A new improved version of the realistic digital brain phantom. Neuroimage 32, 138–145. [PubMed: 16750398]

Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC, 2011 A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54, 2033–2044. [PubMed: 20851191]

Crum WR, Griffin LD, Hill DL, Hawkes DJ, 2003 Zen and the art of medical image registration: correspondence, homology, and quality. Neuroimage 20, 1425–1437. [PubMed: 14642457]

Crum WR, Hartkens T, Hill DL, 2004 Non-rigid image registration: theory and practice. Br. J. Radiol 77 (Spec No 2), S140–S153. [PubMed: 15677356]

Datteri RD, Liu Y, D'Haese PF, Dawant BM, 2015 Validation of a nonrigid registration error detection algorithm using clinical MRI brain data. IEEE Trans. Med. Imaging 34, 86–96. [PubMed: 25095252]

Despotovic I, Goossens B, Philips W, 2015 MRI segmentation of the human brain: challenges, methods, and applications. Comput. Math. Methods Med 2015, 450341. [PubMed: 25945121]

Dice LR, 1945 Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Douaud G, Jbabdi S, Behrens TE, Menke RA, Gass A, Monsch AU, et al., 2011 DTI measures in crossing-fibre areas: increased diffusion anisotropy reveals early white matter alteration in MCI and mild Alzheimer's disease. Neuroimage 55, 880–890. [PubMed: 21182970]

Elhawary H, Oguro S, Tuncali K, Morrison PR, Tatli S, Shyn PB, et al., 2010 Multimodality non-rigid image registration for planning, targeting and monitoring during CT-guided percutaneous liver tumor cryoablation. Acad. Radiol 17, 1334–1344. [PubMed: 20817574]

Fischl B, 2012 FreeSurfer. Neuroimage 62, 774–781. [PubMed: 22248573]

Fleysher R, Kim N, Suri A, Lipton M, Branch C, 2017 Characterization of registration errors to screen aberrant subject results prior to voxel-wise whole brain analysis. In: Proceedings of the 25th ISMRM, p. 4684.

Fleysher R, Lipton ML, Noskin O, Rundek T, Lipton R, Derby CA, 2018 White matter structural integrity and transcranial Doppler blood flow pulsatility in normal aging. Magn. Reson. Imaging 47, 97–102. [PubMed: 29158187]

Friston KJ, Worsley KJ, Frackowiak RS, Mazziotta JC, Evans AC, 1994 Assessing the significance of focal activations using their spatial extent. Hum. Brain Mapp 1, 210–220. [PubMed: 24578041]

Garlapati RR, Mostayed A, Joldes GR, Wittek A, Doyle B, Miller K, 2015 Towards measuring neuroimage misalignment. Comput. Biol. Med 64, 12–23. [PubMed: 26112607]

Grachev ID, Berdichevsky D, Rauch SL, Heckers S, Kennedy DN, Caviness VS, et al., 1999 A method for assessing the accuracy of intersubject registration of the human brain using anatomic landmarks. Neuroimage 9, 250–268. [PubMed: 9927554]

Han X, Yang X, Aylward S, Kwitt R, Niethammer M, 2017 Efficient registration of pathological images: a joint Pca/image-reconstruction approach. Proc. IEEE Int. Symp. Biomed. Imaging 2017, 10–14. [PubMed: 29887971]

Hoffmann C, Krause S, Stoiber EM, Mohr A, Rieken S, Schramm O, et al., 2014 Accuracy quantification of a deformable image registration tool applied in a clinical setting. J. Appl. Clin. Med. Phys 15, 4564. [PubMed: 24423856]

Hoptman MJ, Nierenberg J, Bertisch HC, Catalano D, Ardekani BA, Branch CA, et al., 2008 A DTI study of white matter microstructure in individuals at high genetic risk for schizophrenia. Schizophr. Res 106, 115–124. [PubMed: 18804959]

Hua K, Zhang J, Wakana S, Jiang H, Li X, Reich DS, et al., 2008a Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. Neuroimage 39, 336–347. [PubMed: 17931890]

Hua X, Leow AD, Parikshak N, Lee S, Chiang MC, Toga AW, et al., 2008b Ten-sor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. Neuroimage 43, 458–469. [PubMed: 18691658]

Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM, 2012 Fsl. Neuroimage 62, 782–790. [PubMed: 21979382]

Joshi S, Davis B, Jomier M, Gerig G, 2004 Unbiased diffeomorphic atlas construction for computational anatomy. Neuroimage 23 (Suppl 1), S151–S160. [PubMed: 15501084]

Kearney V, Haaf S, Sudhyadhom A, Valdes G, Solberg TD, 2018 An unsupervised convolutional neural network-based algorithm for deformable image registration. Phys. Med. Biol 63, 185017. [PubMed: 30109996]

Keihaninejad S, Ryan NS, Malone IB, Modat M, Cash D, Ridgway GR, et al., 2012 The importance of group-wise registration in tract based spatial statistics study of neurodegeneration: a simulation study in Alzheimer's disease. PLoS One 7, e45996. [PubMed: 23139736]

Kinnunen KM, Greenwood R, Powell JH, Leech R, Hawkins PC, Bonnelle V, et al., 2011 White matter damage and cognitive impairment after traumatic brain injury. Brain 134, 449–463. [PubMed: 21193486]

Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang MC, et al., 2009 Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage 46, 786–802. [PubMed: 19195496]

Kochunov P, Lancaster J, Thompson P, Toga AW, Brewer P, Hardies J, et al., 2002 An optimized individual target brain in the Talairach coordinate system. Neuroimage 17, 922–927. [PubMed: 12377166]

Kochunov P, Williamson DE, Lancaster J, Fox P, Cornell J, Blangero J, et al., 2012 Fractional anisotropy of water diffusion in cerebral white matter across the lifespan. Neurobiol. Aging 33, 9–20. [PubMed: 20122755]

Lyu I, Kim SH, Seong JK, Yoo SW, Evans A, Shi Y, et al., 2015 Robust estimation of group-wise cortical correspondence with an application to macaque and human neuroimaging studies. Front. Neurosci 9, 210. [PubMed: 26113807]

Mascott CR, Sol JC, Bousquet P, Lagarrigue J, Lazorthes Y, 2006 Lauwers-Cances V. Quantification of true in vivo (application) accuracy in cranial image-guided surgery: influence of mode of patient registration. Neurosurgery 59, ONS146–ONS156 discussion ONS-56.

Mayer AR, Dodd AB, Ling JM, Wertz CJ, Shaff NA, Bedrick EJ, et al., 2018 An evaluation of Z-transform algorithms for identifying subject-specific abnormalities in neuroimaging data. Brain Imaging Behav. 12, 437–448. [PubMed: 28321608]

Mishchenko Y, 2015 A fast algorithm for computation of discrete Euclidean distance transform in three or more dimensions on vector processing architectures. Signal Image Video Process. 9, 19–27.

Mori S, Oishi K, Faria AV, 2009 White matter atlases based on diffusion tensor imaging. Curr. Opin. Neurol 22, 362–369. [PubMed: 19571751]

Muenzing SE, van Ginneken B, Murphy K, Pluim JP, 2012 Supervised quality assessment of medical image registration: application to intra-patient CT lung registration. Med. Image Anal 16, 1521–1531. [PubMed: 22981428]

Oliveira FP, Tavares JM, 2014 MediCAL IMAGE REGISTRATION: A REVIew. Comput. Methods Biomech. Biomed. Eng 17, 73–93.

Pfefferbaum A, Sullivan EV, Hedehus M, Lim KO, Adalsteinsson E, Moseley M, 2000 Age-related decline in brain white matter anisotropy measured with spatially corrected echo-planar diffusion tensor imaging. Magn. Reson. Med 44, 259–268. [PubMed: 10918325]

Pfefferbaum A, Sullivan EV, 2003 Increased brain white matter diffusivity in normal adult aging: relationship to anisotropy and partial voluming. Magn. Reson. Med 49, 953–961. [PubMed: 12704779]

Risholm P, Janoos F, Norton I, Golby AJ, Wells WM 3rd, 2013 Bayesian characterization of uncertainty in intra-subject non-rigid registration. Med. Image Anal 17, 538–555. [PubMed: 23602919]

Sabuncu MR, Yeo BT, Van Leemput K, Vercauteren T, Golland P, 2009 Asymmetric image-template registration. Med. Image Comput. Comput. Assist. Interv 12, 565–573. [PubMed: 20426033]

Sotiras A, Davatzikos C, Paragios N, 2013 Deformable medical image registration: a survey. IEEE Trans. Med. Imaging 32, 1153–1190. [PubMed: 23739795]

Suri AK, Fleysher R, Lipton ML, 2015 Subject based registration for individualized analysis of diffusion tensor MRI. PLoS One 10, e0142288. [PubMed: 26580077]

Thompson P, Toga AW, 2002 A framework for computational anatomy. Comput. Vis. Sci 5, 13–34.

Thompson PM, Giedd JN, Woods RP, MacDonald D, Evans AC, Toga AW, 2000a Growth patterns in the developing brain detected by using continuum mechanical tensor maps. Nature 404, 190–193. [PubMed: 10724172]

Thompson PM, Woods RP, Mega MS, Toga AW, 2000b Mathematical/computational challenges in creating deformable and probabilistic atlases of the human brain. Hum. Brain Mapp 9, 81–92. [PubMed: 10680765]

Viviani R, Beschoner P, Jaeckle T, Hipp P, Kassubek J, Schmitz B, 2007 The bootstrap and cross-validation in neuroimaging applications: estimation of the distribution of extrema of random fields for single volume tests, with an application to ADC maps. Hum. Brain Mapp 28, 1075–1088. [PubMed: 17266105]

Wu G, Peng X, Ying S, Wang Q, Yap PT, Shen D, et al., 2016 eHUGS: enhanced hierarchical unbiased graph shrinkage for efficient groupwise registration. PLoS One 11, e0146870. [PubMed: 26800361]

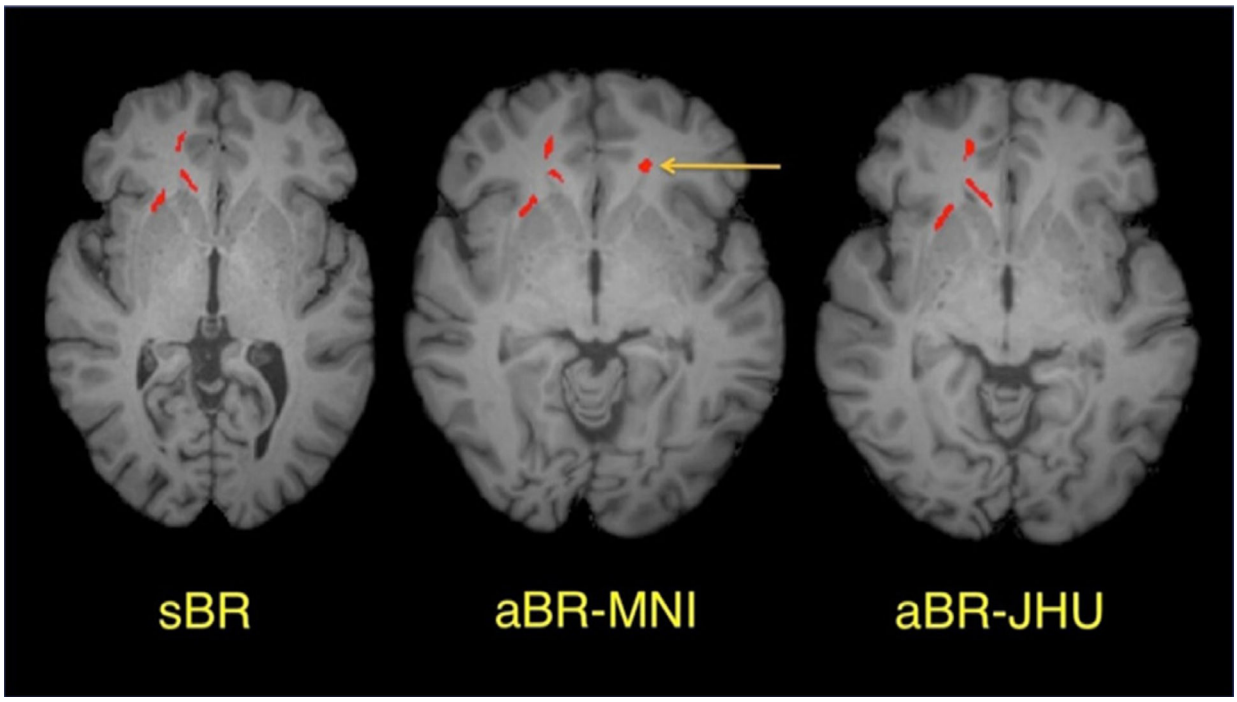

Fig. 1.

Locations of low FA clusters (red) in a patient with mild traumatic brain injury obtained by voxel-wise analysis using subject-based registration (sBR) with the subject's T1W image as template and atlas-based registration with the MNI (aBR-MNI) and JHU (aBR-JHU) templates. The arrow highlights an FA cluster found only when using the MNI template. (*Adapted from* (Suri et al., 2015). *Permission to reuse granted by Creative Commons Attribution License CC BY).*

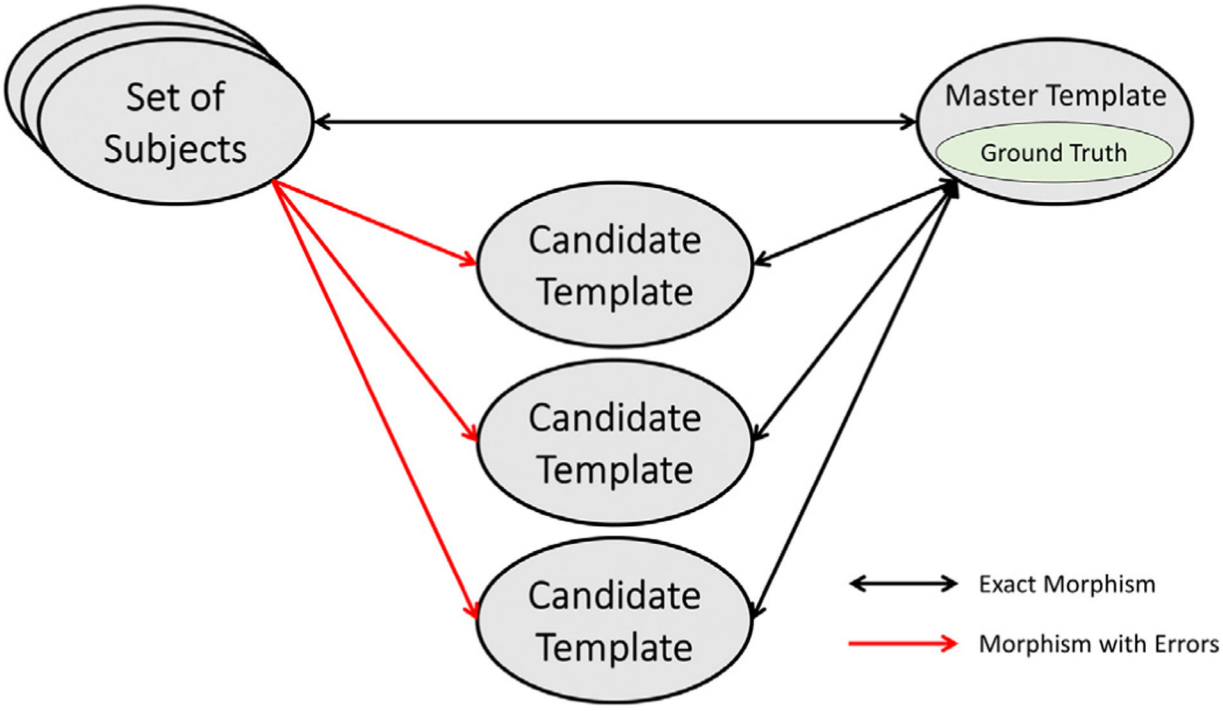

Fig. 2. Ideal study design.

An ideal study of the effects of template choice on voxel-wise analyses would allow the comparison of "ground truth" defined over a master template to results obtained by first morphing data from a set of subjects to different candidate templates. The "ground truth" would be obtained by applying an error-free, reversible morphism (black, double-headed arrows) to the set of subjects of interest. Template choice would then be uniquely indicated by differences in the errors of the morphisms from the subjects to the candidate templates (red, single-headed arrows). Results of analyses over the candidate templates would then be morphed exactly to the master template for comparison with the "ground truth".

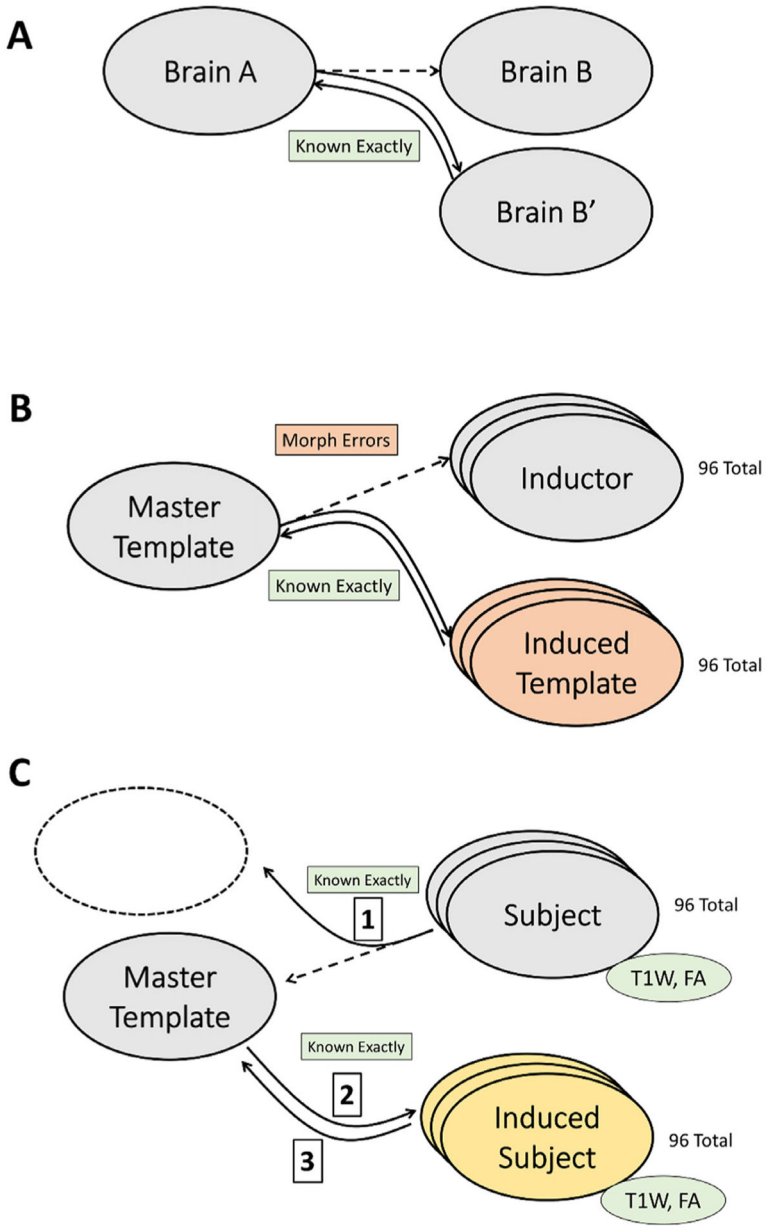

Fig. 3. Construction of induced templates and subjects.

A) Registration algorithms do not allow for exactly-known morphisms between two brain images to be directly computed (dashed arrow). Given two brains A and B, what can in fact be known exactly is a morphism from brain A to brain B' (i.e. "brain B plus error"); because this morphism is known exactly, it can be inverted (solid arrows). We utilize this principle to create exactly-known morphisms between the master template and the induced subjects and induced templates. **B)** Generation of induced templates. An error-free morphism between the "master template" - the JHU brain - and an "inductor" - an Einstein Lifespan Study (ELS) brain image - would transform the master template perfectly to the inductor (dashed arrow). However, as in A), morph errors cause the output of the morphism to differ from the actual inductor brain, producing an "induced template" (curved solid arrows) with exact mapping

on the master template. This process is repeated for all 96 ELS brain images as inductors to produce 96 induced templates. **C)** Generation of induced subjects through a three-step protocol. **(1)** Morphisms are computed intending to bring 96 ELS brain images including T1W anatomy and FA data onto the master template (dashed arrow), but errors do not allow this computation to be direct (curved solid arrow to empty ellipse). **(2)** Instead, inversions of the morphisms produced in (1) are applied to the master template to generate T1W images of induced subjects. An induced T1W image is paired with original subject's FA map to complete an induced subject dataset. **(3)** The inverted morphisms from (2) are paired with those in (1) to create an exactly-known set of reversible transformations between the master template and induced subjects (curved solid arrows).

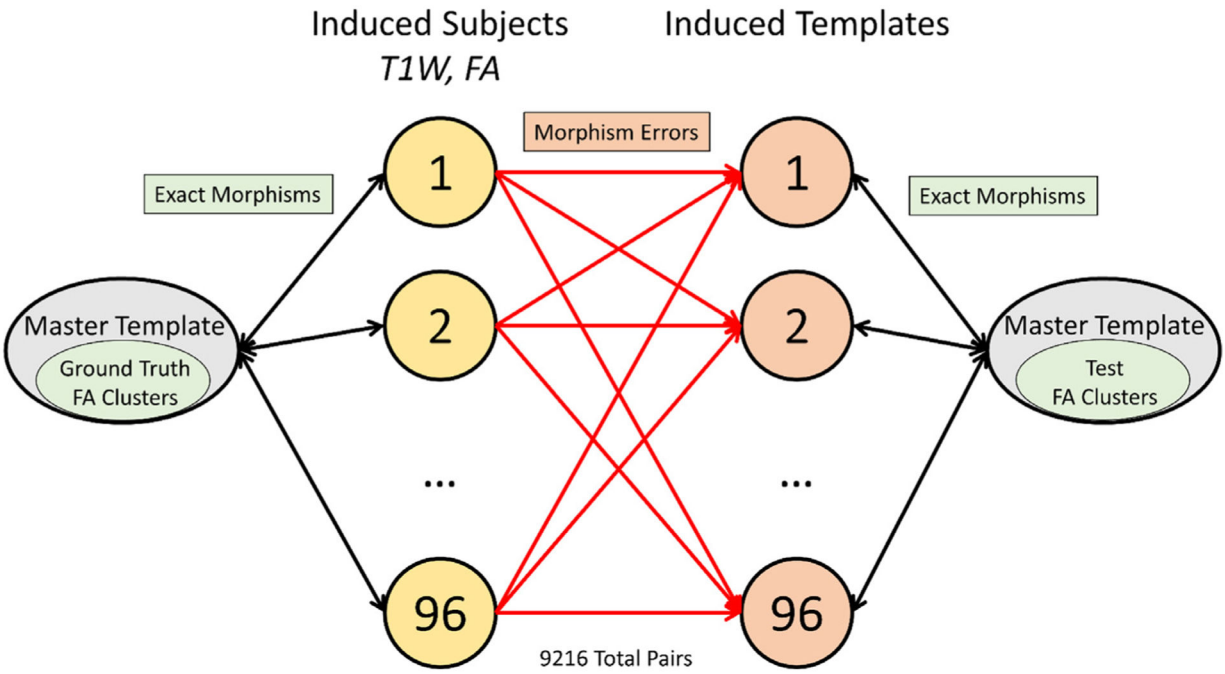

Fig. 4. Overall registration framework.

We constructed two sets of "exact", reversible morphisms between the master template and the induced subjects and induced templates (black double-headed arrows). The morphisms of interest for this study are those from each of the induced subjects to each of the induced templates (red arrows). The outcome this study measures is the concordance of the locations of the induced subjects' FA versus age clusters between when they are directly morphed onto the master template ("ground truth FA clusters") and when they are first morphed onto an induced template and then to the master template ("test FA clusters").

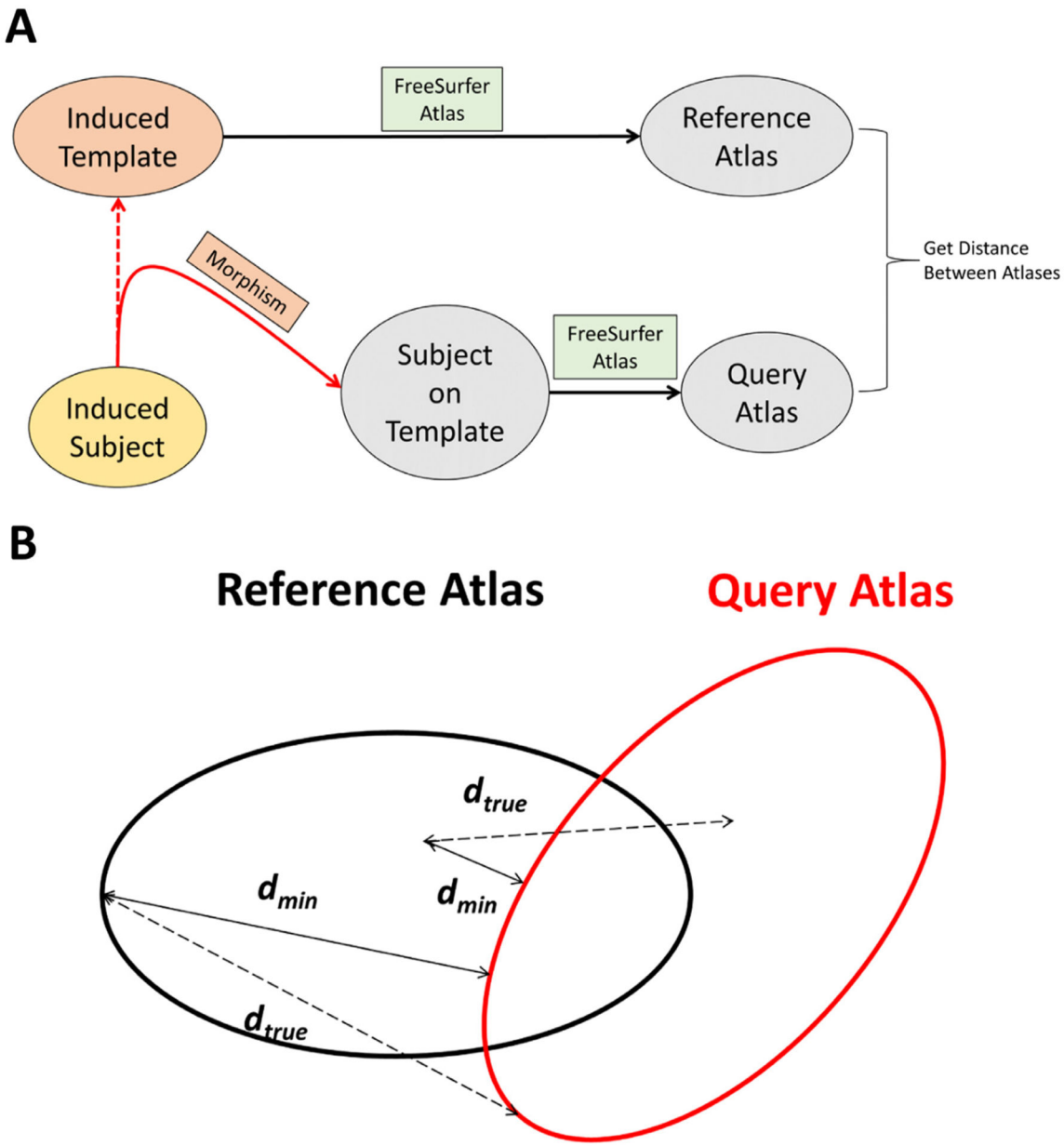

Fig. 5. Characterization of registration error by average atlas distance.

A) Atlas distance computation is shown for a single morphism between an induced subject and an induced template. The attempted registration of the induced subject to the induced template (dashed red arrow) in fact produces an image that resembles the induced template but has some degree of error (solid red arrow). This induced-subject-on-induced-template image and the induced template then have atlases computed by FreeSurfer, respectively generating query and reference atlases. The "distance" between atlases is then computed by a Haussdorf-like method. **B)** Conceptual illustration of Haussdorf-like method for average inter-atlas distance calculation. The reference atlas (black oval) corresponds to the FreeSurfer-generated atlas of the induced template, while the query atlas (red oval) corresponds to the FreeSurfer-generated atlas of the induced-subject-on-induced-template

image; the query atlas has been shifted and rotated during the error-prone morphing process. The closest-edge distance d_{min} (solid double-headed arrows) from each point in the reference atlas to the query atlas is computed for all voxels in the reference atlas. The d_{min} are averaged over all voxels to compute the final quality measure. Using the closest-edge distance underestimates the "true" atlas distance (d_{true}, dashed double-headed arrows) that would be obtained by computing distances between the voxels in the reference atlas to those exactly corresponding to them in the query atlas. Note that the true voxel to corresponding voxel distance cannot be assessed: if it could, it would be included in the morphism in the first place.

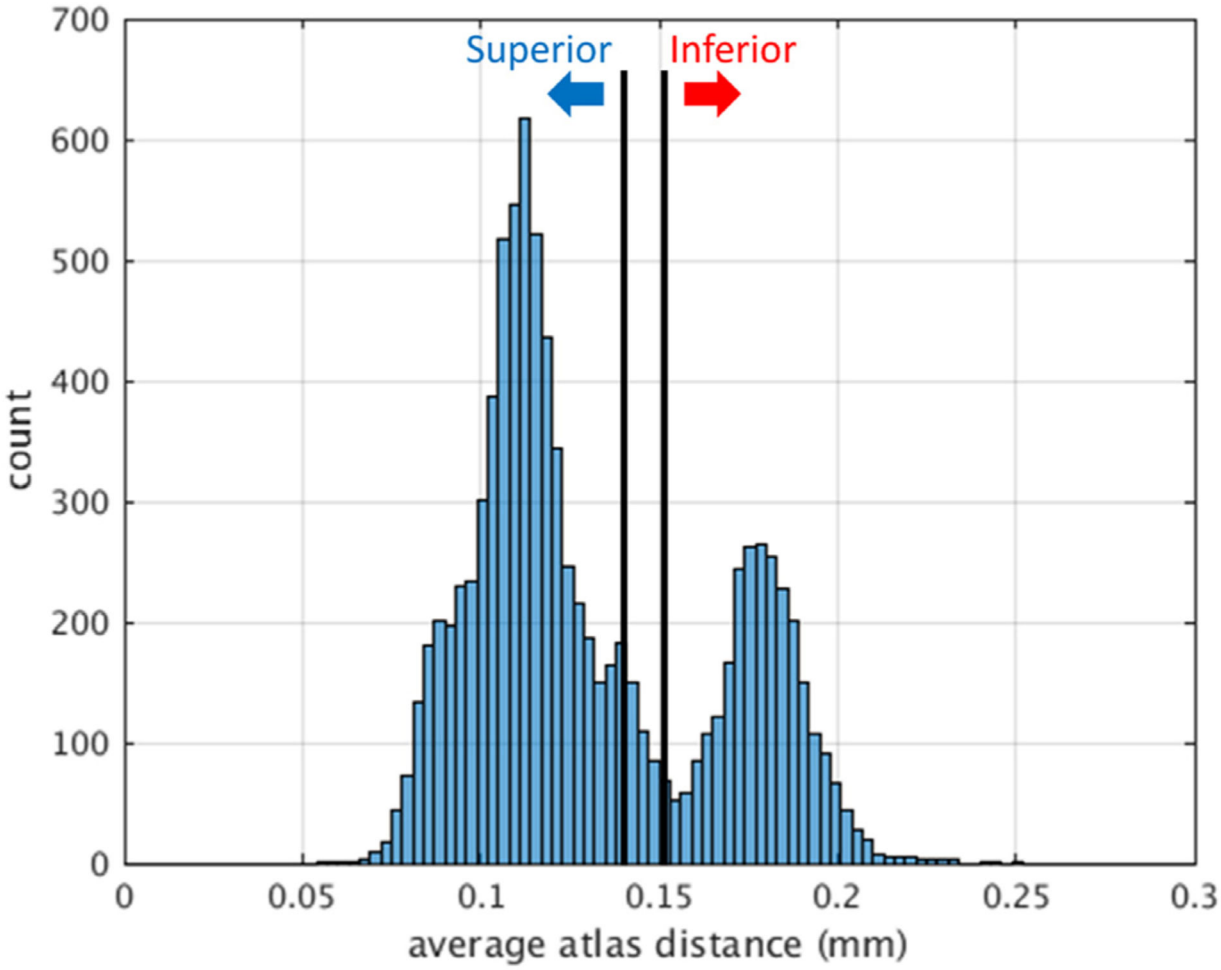

Fig. 6. Morphism quality is bimodally distributed.

Shown is the distribution of average atlas distance for all 9216 morphisms corresponding to the induced subject - induced template pairs generated in the present study. The average atlas distance is computed between atlases of the induced template and the induced-subject-on-induced-template (Fig. 5). Based on this bimodal distribution, we defined morphisms as "superior" if they had an average atlas distance of less than 0.14 mm, and as "inferior" if they had an average atlas distance greater than 0.15 mm. The relatively small set of morphisms with in-between average atlas distances was not evaluated further.

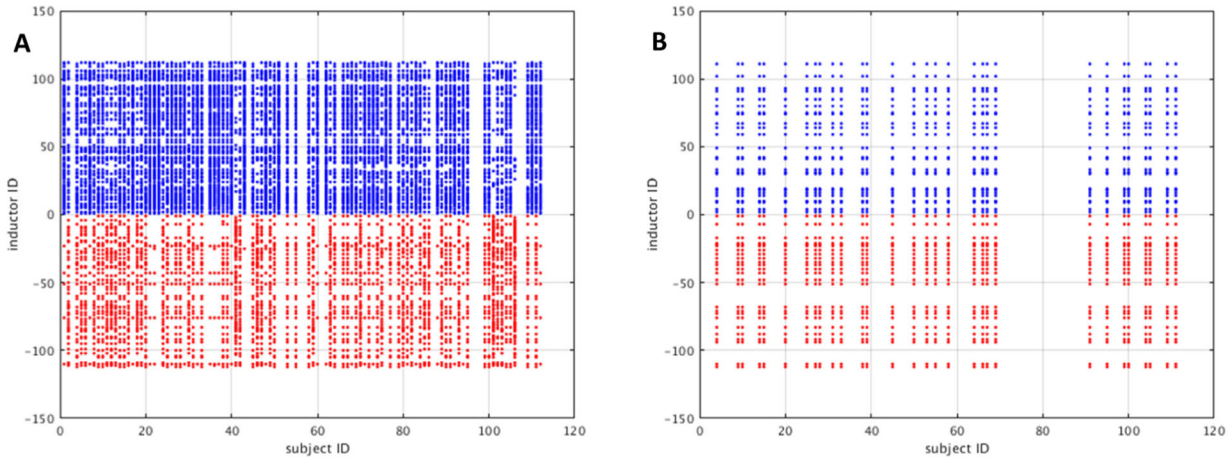

Fig. 7. Morphism quality is subject-template-pair dependent.

A) The subject-template dependence of morphism quality is shown for all induced subjects and templates combinations generated in this study. The horizontal axes on the plots list the IDs of the 96 induced subjects, using original numbering from the ELS. The vertical axis lists the inductor IDs corresponding to the induced templates, with positive (blue points) and negative (red points) numbers respectively indicating “superior” and “inferior” morphisms from the subjects listed on the horizontal axis. This plot demonstrates the subject-template-pair dependence of morphism quality: some subject IDs display mostly blue template IDs, and others display mostly red. **B)** Output of algorithmic selection of 30 specific subjects that have superior morphisms to a first subset of 25 induced templates and inferior morphisms to a second subset of 25 induced templates for FA cluster comparison. Consequent to this selection, the first 25 templates are called “superior templates” and the second 25 are called “inferior templates”. The sets of “superior” and “inferior” templates are identical across all 30 subjects.

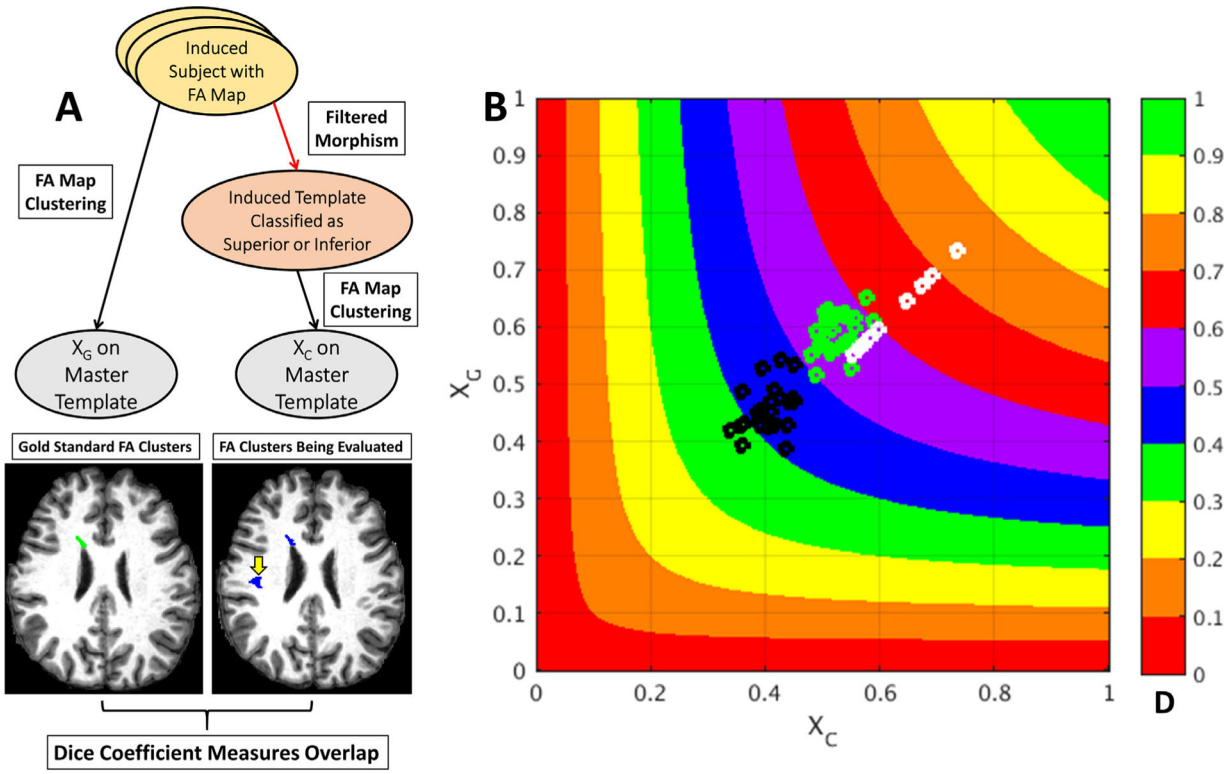

Fig. 8. Evaluation of morphism quality filter.

A) Overall schematic of morphism quality filter evaluation. The chosen 30 induced subjects (Fig. 7B) have their FA maps morphed directly onto the master template; their voxel-wise analysis produces "gold standard" FA clusters X_G that represent a ground truth (left arrow). To evaluate the effect of morphism quality on FA cluster location, the same 30 FA maps were morphed onto the induced templates that were classified (Fig. 7B) as superior or inferior (red arrow); results of voxel-wise analyses over them are subsequently morphed onto the master template, producing test FA clusters X_C (right black arrow). The "gold standard" clusters are compared to the test FA clusters over the 25 superior and 25 inferior templates using the Dice coefficient. Shown at the bottom of the figure are examples of "gold standard" (green) and test (blue) FA clusters projected onto the master template. The specific example shown demonstrates agreement between a test and gold standard FA cluster anterior to the right ventricle, but also the presence of a false positive cluster in the right cerebral hemisphere (yellow arrow). **B)** Dice index D (colored contour map) representing the spatial overlap between "gold standard" FA clusters X_G and three types of test FA clusters of interest X_C : the 25 superior (green points), the 25 inferior (black points), and a set of 12 "control" clusters (white points). The 12 control clusters were obtained by perturbing the "gold standard" clusters by single-voxel shifts in each of the x, y, and z orthogonal directions as well as in the diagonal directions. Calculation of the Dice coefficient in three-dimensional space can lead to values much less than 1 while still indicating close overlap. This is exemplified by the overlap between the green and white points, which suggests that superior morphisms produce good clusters that in fact closely agree with the gold standard ones.