Journal of Machine Learning for Biomedical Imaging 2024:013 vol. 2, pp. 916–954
Submitted 09/2023
Published 06/2024

Impact of Initialization on Intra-Subject Pediatric Brain MR Image Registration: A Comparative Analysis between SyN ANTs and Deep Learning-Based Approaches

Andjela Dimitrijevic https://orcid.org/0000-0002-8799-6300 andjela.dimitrijevic@polymtl.ca
NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montréal, Montréal, QC, Canada
Research Center, Ste-Justine Hospital University Centre, Montréal, QC, Canada

Vincent Noblet https://orcid.org/0000-0002-3655-3163 vincent.noblet@unistra.fr
ICube-UMR 7357, Université de Strasbourg, CNRS, Strasbourg, France

Benjamin De Leener https://orcid.org/0000-0002-1378-2756 benjamin.de-leener@polymtl.ca
NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montréal, Montréal, QC, Canada
Research Center, Ste-Justine Hospital University Centre, Montréal, QC, Canada
Computer Engineering and Software Engineering, Polytechnique Montréal, Montréal, QC, Canada

Abstract

This study evaluates the performance of conventional SyN ANTs and learning-based registration methods in the context of pediatric neuroimaging, specifically focusing on intra-subject deformable registration. The comparison involves three approaches—without (NR), with rigid (RR), and with rigid and affine (RAR) initializations. In addition to initialization, performances are evaluated in terms of accuracy, speed, and the impact of age intervals and sex per pair. Data consists of the publicly available MRI scans from the Calgary Preschool dataset, which includes 63 children aged 2-7 years, allowing for 431 registration pairs. We implemented the unsupervised deep learning (DL) framework with a U-Net architecture using DeepReg and it was 5-fold cross-validated. The evaluation includes Dice scores for tissue segmentation from 18 smaller regions obtained by SynthSeg, analysis of log Jacobian determinants, and registration pro-rated training and inference times. Learning-based approaches, with or without linear initializations, exhibit slight superiority over SyN ANTs in terms of Dice scores. Specifically, DL-based implementations with RR and RAR initializations significantly outperform SyN ANTs. The lower Dice scores of SyN ANTs are likely due to its lack of population-based optimization, unlike the DL methods which learn optimal parameters through training. Both SyN ANTs and DL-based registration involve parameter optimization, but the choice between these methods depends on the scale of registration—network-based for broader coverage or SyN ANTs for specific structures. Learning-based registration offers fast inference times but needs training, whereas SyN ANTs requires manual fine-tuning, with less clear guidelines, particularly for younger cohorts. Both methods face challenges with larger age intervals due to greater growth changes. Future work will extend the framework to younger populations and explore models that better separate different levels of transformations for improved local brain region registration. The main takeaway is that while DL-based methods show promise with faster and more accurate registrations, SyN ANTs remains robust and generalizable without the need for extensive training, highlighting the importance of method selection based on specific registration needs in the pediatric context. Our code is available at `https://github.com/neuropoly/pediatric-DL-registration`.

Keywords: Deep Learning, MRI, Pediatric, Image Registration, Learning-based Registration

https://doi.org/10.59275/j.melba.2024-151b

1. Introduction

Deformable image registration involves the alignment of a pair of images to establish a shared coordinate reference framework. It is used for both intra and inter-subject analyses within the medical domain, playing a vital role in achieving normalized visualizations across brain scans (Uchida, 2013). This research focuses on the fact that so far, deformable image registration is less adapted to pediatric data, this can arise from bigger volume differences when analyzing longitudinal data of a subject's brain at two different time-points, but also the lack of pediatric data availability (Barkovich et al., 2019). However, improving registration performed on neuroimaging data of pediatric populations remains essential for template creation as well as different diagnostic pipelines. Currently, conventional deformable registration methods such as ANTs (Avants et al., 2011), NiftyReg (Modat et al., 2010) or Elastix (Klein et al., 2010) are functional. Nonetheless, when dealing with extensive datasets, the iterative optimization-based estimation of deformation fields makes the process time-intensive. The emerging deep learning (DL)-based techniques, incorporating convolutional neural networks (CNN), can allow faster registrations by applying a learning-based approach instead. In essence, these recently devised techniques enable the direct estimation of deformation fields from input 3D volume pairs. This study aims to evaluate registration implementations in the pediatric context, comparing the conventional SyN ANTs method with DL-based approaches, with a focus on their performance in terms of accuracy, speed, initialization, and the impact of age intervals per pair and separated by sex within intra-subject pediatric data.

1.1 Current DL-based Registration Frameworks for Deformable Registration and Its Evaluation in the Pediatric Context

A popular DL-based registration approach is VoxelMorph (Balakrishnan et al., 2019), which is designed for brain MRI applications. It uses a U-Net-like architecture (encoder-decoder with skip-connections)(Ronneberger et al., 2015) and employs the scaling and squaring integration method on computed velocity fields to obtain diffeomorphic deformation fields. Kuang et al.(Kuang and Schmah, 2019) developed the fast image registration (FAIM) algorithm, which showed superior results to VoxelMorph. FAIM, composed of a spatial deformation module largely inspired by the spatial transformer networks (STN) (Jaderberg et al., 2015), stacks moving and fixed images as input for the network and uses a training loss composed of a cross-correlation metric and regularization to ensure smooth, non-negative Jacobian determinants. Similarly, Zhang (2018) introduced the inverse-consistent deep network (ICNet) with inverse-consistent and anti-folding constraints added to a mean squared distance similarity metric. Also using a U-Net like architecture and STN, their method outperformed Demons-based registration (Thirion, 1998) as well as symmetric normalization (SyN) based registration (Avants et al., 2008).

All these DL-based registration frameworks often use U-Net-like architectures which frequently outperform conventional registration methods and are trained on adult MRI brain data. Now, these tools are even incorporated into popular neuroimaging analysis tools such as FreeSurfer. EasyReg (Iglesias, 2023) implements SynthSeg's (Billot et al., 2023a) model to predict forward and backward nonlinear fields, achieving symmetric and diffeomorphic

transformations. This approach is the first learning-based registration algorithm publicly available via FreeSurfer (Fischl, 2012).

However, few of these registration implementations have been evaluated in the pediatric longitudinal context (Ghosh et al., 2010). Efficient and accurate registration algorithms could facilitate the comprehensive analysis of longitudinal changes in deformation fields across a larger number of subjects, particularly when handling extensive neuroimaging datasets. This gap presents an opportunity to analyze neurodevelopment in young children, where these advanced DL-based registration methods could be particularly impactful. Recently, Hoffmann et al. (2023) use SynthMorph to evaluate affine and joint registration (affine+deformable) across six different datasets, including only two that encompass pediatric subjects aged 5-21 years from the Lifespan Human Connectome Project Development (HCPD) and MASiVar (MASi) datasets (total of 100 pediatric images). Notably, these pediatric datasets were not used for training, which was conducted solely with adult populations. Moreover, the pediatric datasets did not undergo any preprocessing steps such as skull-stripping or N4 correction, which could impact the registration results. Despite these limitations, SynthMorph demonstrated successful generalization, achieving Dice scores between 0.85 and 0.90 on 23 bilateral brain regions within these pediatric datasets. Nevertheless, further evaluation of learning-based methods contrasted to conventional ones is needed, as significant neurodevelopmental changes can be revealed in longitudinal toddler imaging (Barkovich et al., 2019).

1.2 Importance of Longitudinal Changes

Characterizing longitudinal changes is more present in other applications such as disease detection using normative modeling (Bethlehem et al., 2022; Rutherford et al., 2022; Chen et al., 2021) in adult populations. For example, Alzheimer's disease has been one of the most studied neurodegenerative diseases to characterize those morphological transformations happening through time (Gafuroglu et al., 2018; Ouyang et al., 2021, 2022). Usually, from the presence of atrophy measured by hippocampal volume or cortical thickness measures extracted from structural MRI images, normative measures can be distinguished from pathological cases (Jang et al., 2022).

The deformation field or hidden transformation between two time points contains regions of contraction and expansion as well as potentially being a growth indicator. Hence, neurodevelopmental evolution trajectories could be extracted in the pediatric context. Indeed, changes are exponentially variable from 0 to 6 years of age, where a 6-year-old brain resembles at 95% to an adult brain (Phan et al., 2018). Being dependent on the age interval, an analysis on deformable intra-subject registration performance can be of value to dissect developmental patterns.

The primary aim of this study is to conduct a detailed comparative analysis between SyN ANTs, a conventional registration algorithm and an unsupervised neural network. This involves evaluating the potential of a deep learning framework for intra-subject registration on pediatric longitudinal brain data when separating global and local transformations. Central to our objectives are: 1) examining the performance of unsupervised neural networks, with and without initialization registration tasks, and 2) analyzing how intra-subject age intervals and sex impact the performance metrics used to assess the DL-based registration.

This comparative analysis will inspect the advantages and disadvantages of each method in terms of performance, training, age intervals, and inference times, thereby placing a significant focus on the distinctions between SyN ANTs and a Voxelmorph-like DL-based approach. Hence, one of the contributions of this study is to compare SyN ANTs (Avants et al., 2008) and DL-based methods in achieving a complete deformable registration task on pediatric longitudinal brain magnetic resonance (MR) images when using three pre-alignment approaches. This work builds upon a previous short paper that was presented at the WBIR Workshop (Dimitrijevic et al., 2022). While the primary goal remains consistent, our approach differs in terms of the methodology, and we have introduced a comparative analysis with SyN ANTs, a conventional registration algorithm. The prior paper enabled a comparison between 1.5 mm and 2.0 mm isotropic images, revealing that the 1.5mm version yielded superior registration accuracy. Furthermore, we have expanded our investigation to incorporate age and sex-related analyses, capitalizing on the available longitudinal data to assess performance metrics with respect to age intervals in-between pairs. Additionally, this work includes an extended review of relevant literature.

2. Methods

As illustrated in Figure 1, our investigation explores three initialization approaches: NoReg (NR), which involves no pre-alignment; RigidReg (RR), consisting of a rigid pre-alignment; and RigidAffineReg (RAR), incorporating both rigid and affine pre-alignments. These serve as inputs to either a U-Net for learning-based deformable registration (DL Reg) or SyN ANTs, serving as the conventional state-of-the-art comparison. SyN ANTs was selected due to its widespread use in the literature for various registration tasks in medical imaging (Tustison et al., 2021) and its benchmark performance in numerous competitions (Menze et al., 2015; Murphy et al., 2011). On the other hand, the learning-based architecture and loss function was inspired by Voxelmorph (Balakrishnan et al., 2019), but specifically trained on pediatric data. It is crucial to highlight that the U-Net architectures maintain consistent layer structures (as detailed in section 2.1), ensuring consistent comparison across all three initialization techniques. The exact parameters used for rigid as well as rigid and affine initializations are available in section A.1 in the supplementary material. The reproducible pipeline is also available on the open-source Github repository at `https://github.com/neuropoly/pediatric-DL-registration`.

2.1 Chosen Architecture

This subsection focuses on the unsupervised chosen DL framework to accomplish the full non-rigid registration task. Taking as input a moving (M) and fixed (F) image pair, the network computes a dense displacement field (DDF) which allows the creation of a moved image also called warped moving image or predicted fixed image. This moved image is obtained from aligning the moving image to the fixed image. All calculations are done with 3D volumes as input. The selected CNN is a U-Net for which training optimizes a set of learnable parameters denoted as θ, corresponding to the kernel weights of the network. More specifically, the U-Net architecture which generates the deformation fields is a 3-layer encoder and decoder with 8, 16 and 32 channels each with skip connections. The specific parameters for the architecture are available in the config files on Github which are

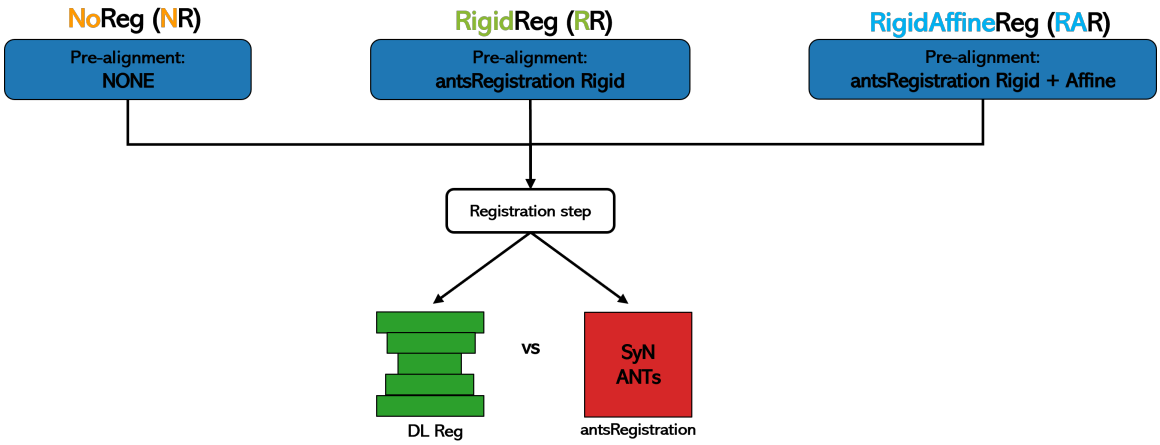

Figure 1: Illustration of the three initialization strategies, NoReg (NR), RigidReg (RR) as well as RigidAffineReg (RAR), (blue) used for comparing deep learning (green) and conventional SyN ANTs (red) registration approaches.

reproducible when using DeepReg. A stochastic gradient descent method is used to find the optimal parameters of the network. In our case, an ADAM optimizer with a learning rate set to 1.0e-4 is used. Each training split was trained for 250 epochs and a batch size of 2 pairs of moving/fixed images. Figure 2 illustrates the loss function of the unsupervised network. This loss function comprises two terms: the first term, the local normalized cross-correlation (LNCC) similarity measure, selected for its robustness to local intensity variations. The second term represents L2-norm gradient regularization, with its weighting factor set to 1 based on an optimal value from the literature (Balakrishnan et al., 2019). While exploring different L2-norm gradient regularization factors for realistic deformation fields would be insightful, this study's primary focus remains on assessing various initialization approaches. The training procedure and architecture initialization are facilitated by the version named develop 0.0.0 from DeepReg (Fu et al., 2020), a DL-based registration framework. A 5-fold cross-validation scheme is used for train and test splits. Finally, training is executed on a system equipped with Ubuntu 18.04.5 (amd64), featuring an octa-core Intel i7-9700KF CPU, 62.7 GB of RAM, and a GeForce RTX 2080 Ti GPU.

3. Data

The chosen publicly available Calgary Preschool dataset (Reynolds et al., 2020) was obtained using a General Electric 3T MR750w system and 32-channel head coil (GE, Waukesha, WI) at the Alberta Children's Hospital in Calgary, Canada. This acquisition process received approval from the University of Calgary Conjoint Health Research Ethics Board. The dataset comprises T1-weighted MRI brain scans from 96 children aged 2 to 8 years. These scans were obtained using an FSPGR BRAVO sequence with the following parameters: TR = 8.23 ms, TE = 3.76 ms, TI = 540 ms, flip angle = 12 degrees, voxel size =

0.4492x0.4492x0.9 mm^3, 210 slices, matrix size = 512x512, and a field of view = 23.0 cm. It includes multiple scans at different time points for 96 subjects, an essential element for single-modality intra-subject DL registration. Given the inherent challenges in obtaining pediatric datasets, the dataset size is particularly noteworthy, especially as it offers longitudinal samples (Lebel and Deoni, 2018). It also includes age, biological sex, handedness, and other parameters which can be further analyzed. From the 96 subjects, it was necessary to choose children with two or more time-point scans from the acquired data. Hence, 64 subjects respected those conditions which brings the used data to a total of 247 T1-weighted images allowing 434 combinations of moving/fixed registration pairs.

Table 1 displays the dataset characteristics and relevant parameters. A graphical representation of the longitudinal age and sex parameters is also available in Figure 7. Additionally, the remaining 64 subjects were inspected for image quality. Each image in the dataset was verified to possess the stated matrix dimension of 512x512x210. For the data preprocessing details, refer to section A.1 in the supplementary material. A full pipeline including steps from data preprocessing to training is also available in Figure 8 in the supplementary material.

4. Experiments

4.1 Segmentations Retrieval

For segmentations, $SynthSeg^+$ (Billot et al., 2023a), also referred as SynthSeg in the following analyses, was used to obtain 32 labels representing various brain structures, resulting in a total of 18 regions for segmentation. Subsequent analyses focus on both the 18 regions individually and globally on three primary tissues: white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF). For global tissues, WM encompasses cerebral white matter, brain stem, cerebellum white matter, pallidum, and ventral DC; GM includes cerebral cortex, crebellum cortex, accumbens area, caudate, thalamus, putamen, hippocampus, and amygdala; and CSF comprises the lateral ventricle, inferior lateral ventricle, 4th ventricle, 3rd ventricle and CSF regions. For QC, initially, three pairs out of the 64 available pairs were excluded using SynthSeg's QC scores, as one subject within each pair had values lower than 0.6 (Billot et al., 2023b). Subsequently, a visual QC check was performed, and no images were excluded.

4.2 Validation Process

Validating registered images is challenging due to the absence of common and normalized decision metrics (Christensen et al., 2006). In this work, we inspect two types of Dice scores and the influence of age interval per pair and sex-specific differences on performances, pro-rated training and inference times, percentage of negative Jacobian determinant (JD) and sum of absolute log JD values. A detailed overview is present for each metric below. Additionally, the equations used for Dice scores and JD related metrics are detailed in section A.2 in the supplementary material.

Unweighted and Weighted Dice Scores: As depicted in Figure 2, the validation process in terms of Dice scores is represented in black dashed lines. Indeed, moving segmentations are warped using the network's output DDFs and then compared to the fixed

Table 1: Characteristics of the chosen subset from the Calgary Preschool dataset. SD is for standard deviation.

	Calgary Preschool subset
No. of subjects	64
No. of images	247
No. of scans/subject	
Average (SD)	3.86 (1.59)
Min	2
Max	10
Total no. of possible combination pairs	434
Age of all time point scans (years)	
Average (SD)	4.49 (1.00)
Min	1.97
Max	6.90
Sex no. of (%)	
Female	107 (∼43)
Male	140 (∼57)
Handedness no. of (%)	
Right	214 (∼87)
Left	25 (∼10)
Both	8 (∼3)

segmentations with the Dice score as the performance metric. Maximal region overlap is indicated by a Dice score of 1 and no overlap by 0. It's crucial to highlight that when averaging Dice scores across regions of different sizes, larger regions like CSF, cerebral white matter, and cerebral cortex may dominate the results. Indeed, overlap scores of small and localized anatomical regions are to be prioritized as reliable indicators that can differentiate between plausible and inaccurate registrations (Rohlfing, 2012). To reduce the influence of larger regions, a weighted Dice score is calculated with the weight consisting of the inverse number of voxels per region. This weighted Dice score is used to assess performances for each one of the 18 available regions. SyN ANTs and DL-based approaches were also compared in terms of the commonly used Dice score in the literature, referred as unweighted Dice score, on each of the 18 regions and on WM, GM and CSF by averaging sub-regions within these tissues from the total 18 regions available from $SynthSeg^{+}$ (Billot et al., 2023a). Unweighted Dice scores for each of the 18 regions and for the three global tissues are plotted

against age intervals in years per pair to measure the influence of age. Additionally, the same age-related analyses are performed separately for each sex.

Pro-rated Training and Inference Times: In our evaluation, we employ the concept of pro-rated training time as a key metric for assessing the computational efficiency of model training relative to dataset size. Pro-rated training time is calculated by dividing the total training time by the number of pairs present in the dataset. Regarding inference times, they denote the duration required to generate a warp field, either through prediction using the trained U-Net model or via running the SyN ANTs algorithm for each pair. They were performed on each pair five times and averaged across the five folds.

Negative JD and Sum of Absolute Log JD values: Calculated deformation fields from both DL-based and SyN ANTs registration methods were evaluated for foldings using the percentage of negative JD values and the average sum of absolute log JD for all initialization approaches. These metrics assess invertibility and local topology preservation of deformation fields, which are crucial for adequate registration quality. To verify if those properties are ensured, the percentage of negative JD and the sum of absolute log JD values are computed on each DDF generated per registration pair. A negative JD indicates non-invertibility properties due to the presence of unwanted local folding. As for log normalized JD values, they show volumetric changes post-registration: negative values indicate local volume contraction, while positive values indicate local volume expansion. For the average sum of absolute log JD over all pairs, it is calculated within local sub-regions extracted from SynthSeg and encompassing global WM, GM, and CSF regions, excluding background regions. Finally, the distribution of the sum of absolute log JD values within the whole brain region is inspected for both DL Reg and SyN ANTs to compare their regularization strengths. These metrics offer insights into the deformation characteristics within specific brain regions and potential distortions, helping assess deformation smoothness or regularization strengths between methods.

5. Results

Figure 3 illustrates Dice scores averaged across all 18 segmented regions for the three initialization methods: NR, RR, and RAR for DL-based registration (in green), contrasted with the results obtained using SyN ANTs (in red). Initial Dice scores prior to both SyN ANTs and U-Net processing are shown in blue. Dice scores weighted based on the inverse number of voxels per region and normalized by the sum of weights for all 18 regions from SynthSeg are available on supplementary Figure 9 which exhibit similar trends as in Figure 3. For comprehensive Dice score breakdowns across the segmented regions and comparisons to SyN ANTs, refer to Table 2 and Figure 10 for results across WM, GM and CSF obtained by averaging SynthSeg sub-regions within these tissues from the total 18 regions available. Figure 11 in the supplementary material also shows Dice scores for all 18 segmented regions separately averaged over all subjects for all three initialization scenarios. Additionally, Table 2 presents a comparison of pro-rated training time and inference durations between DL-based and ANTs registration pipelines. Notably, for all three initialization approaches, the DL-based method demonstrates markedly faster inference times per pair when the model is trained, between 22 to 74 times faster, compared to the required SyN ANTs registration time. However, when comparing pro-rated training times to the same

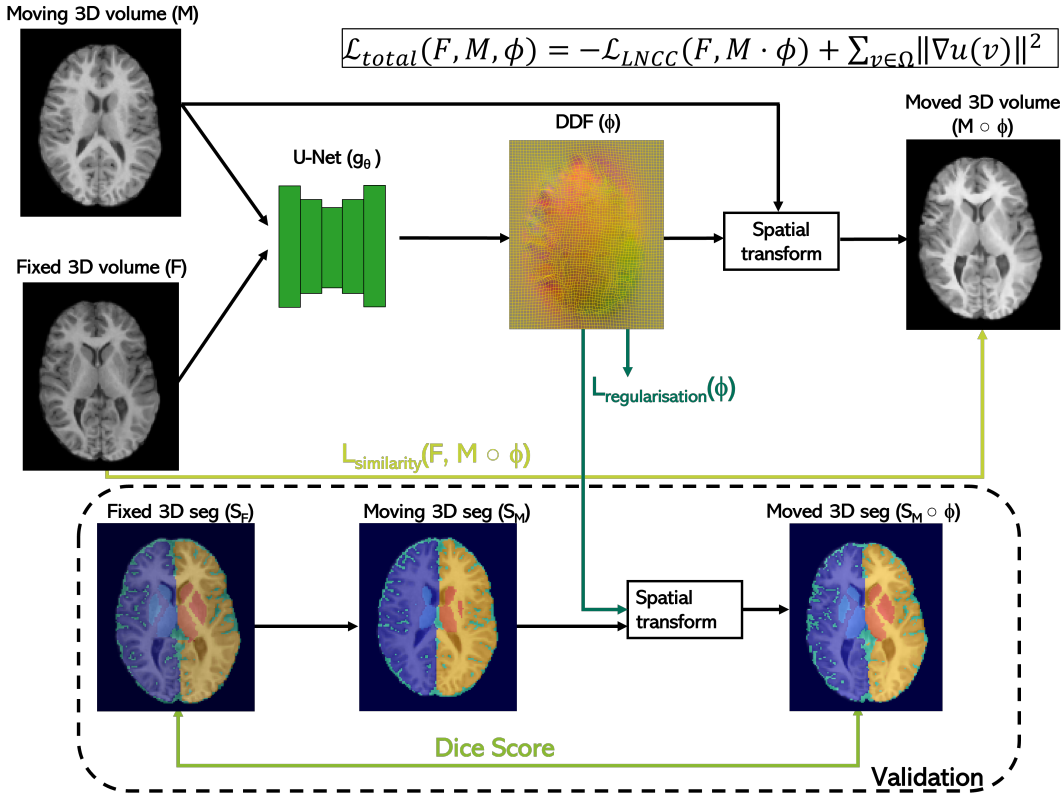

Figure 2: Schema of the training procedure to obtain a deformation field (ϕ) with given moving (M) and fixed (F) 3D pair of images. The validation technique using WM, GM and CSF segmentations (WM depicted in this figure) to calculate Dice scores is also shown in the black dashed region as well as the loss function in the upper right corner where v indicates voxels for the L2 norm of the displacement gradient, ∇u, which encourages a smooth deformation field. ϕ is calculated by adding the identity transform to the displacement field ($\phi = Id + u$). Image inspired by (Balakrishnan et al., 2019).

SyN ANTs registration times per pair, SyN ANTs is around 1.4 times faster for both RR and RAR initializations.

Furthermore, due to the non-parametric nature of the data, two-sided Wilcoxon tests have been conducted for all three initialization approaches between DL Reg and SyN ANTs. For NoReg, the median results from DL-based approaches are not statistically different from those of SyN ANTs across all three segmented regions (p-values > 0.18, $p > 0.05$). In contrast, for RigidReg and RigidAffineReg, the results from DL-based approaches show a statistically significant difference from those of the corresponding ANTs pipelines in all three regions (p-values $<$ 1.27e-17, $p < 0.05$). As shown in Table 2, for both RR and RAR initializations, DL Reg outperforms SyN ANTs for all three segmented regions. However, SyN ANTs has decreased performances compared to the initial alignment. As for NR, only

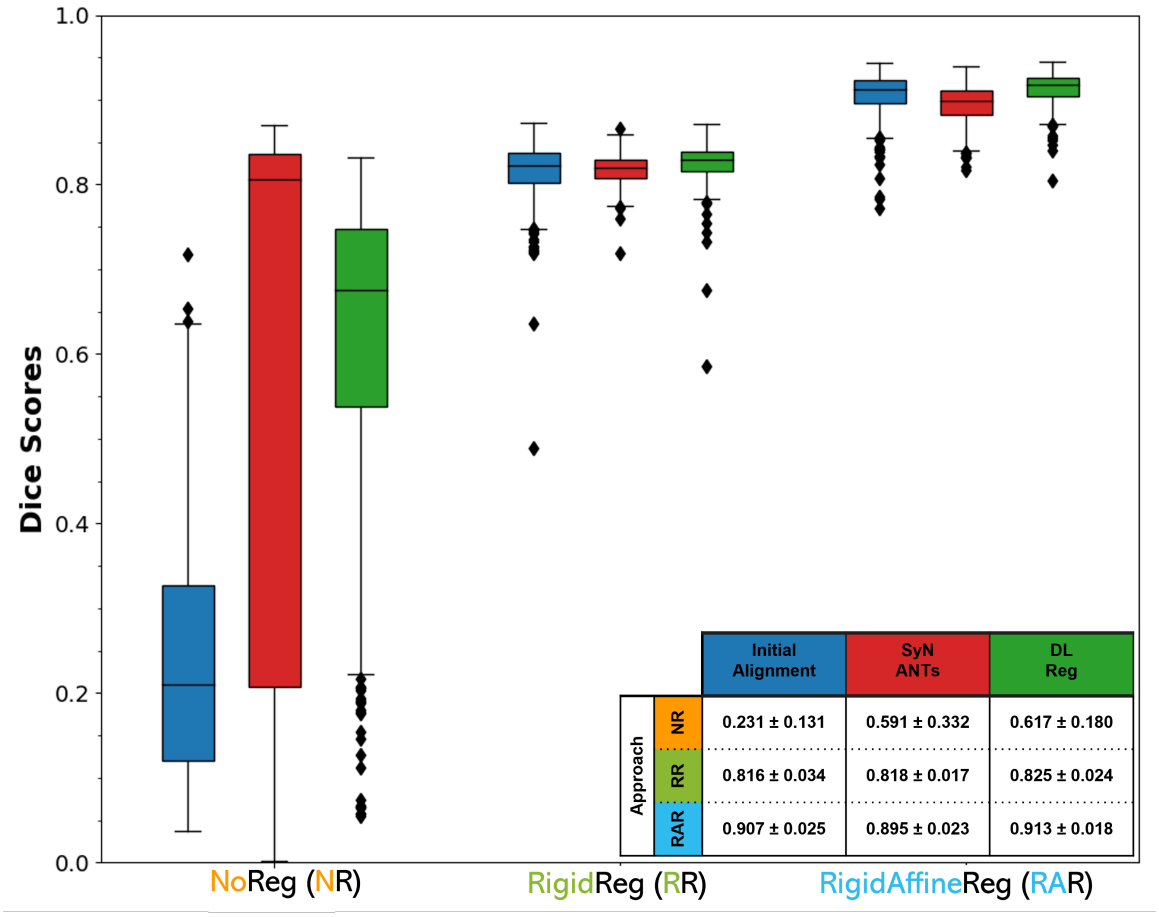

Figure 3: Dice score results on the test sets represented as boxplots for each initialization approach (NoReg, RigidReg and RigidAffineReg) for DL-based methods compared to the initial Dice scores pre-conducting the registration steps. Each method is also compared to the SyN ANTs registration. The Dice scores are averaged over all 18 segmented regions. The table in the lower right corner shows the mean±SD Dice scores for all scenarios.

the average Dice score calculated on CSF tissues for DL Reg are equal in comparison to SyN ANTs however with a smaller standard deviation. When examining each of the 18 segmented regions individually, p-values and statistical significance ($p<0.05$) were determined from Wilcoxon tests, which compare SyN ANTs Dice scores to initial Dice scores before any initialization (Figure 13), DL Reg Dice scores to initial Dice scores before any initialization (Figure 14) and DL Reg Dice scores to SyN ANTs Dice scores (Figure 15), all of which are presented in the supplementary material. Learning-based results show a statistically significant difference across all 18 regions when compared to the initial alignment for NR. However, there are some exceptions for the two other initialization approaches:

for RR initializations in regions such as the caudate, 3rd ventricle, and amygdala, and for RAR initializations in the putamen, pallidum, 3rd ventricle, hippocampus, amygdala, and accumbens area. Also, the percentages of instances where SyN ANTs outperforms DL Reg for NR (64.1%), RR (28.6%), and RAR (10.8%) initializations were calculated across all 18 regions, considering a total of 431 pairs.

Table 2: Dice scores and average sum of absolute log JD values per segmented regions, white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) by averaging SynthSeg sub-regions within these tissues from the total 18 regions available for all three proposed approaches and their comparison to SyN ANTs as well as the initial alignment. This average is calculated by dividing by the number of voxels per region. Pro-rated training (divided by the number of pairs) and inference registration time per pair are shown. Values are presented as mean ± SD.

		Dice score/ Average sum of absolute log JD over all pairs			Time (per pair) (seconds)	
Approach		WM	GM	CSF	Pro-rated training	Inference
NR	DL Reg	0.682 ± 0.174/ 0.368 ± 0.108	0.659 ± 0.191/ 0.366 ± 0.122	0.531 ± 0.183/ 0.387 ± 0.120	128.02 ± 2.90	4.06 ± 0.07
	SyN ANTs	0.621 ± 0.337/ 0.232 ± 0.148	0.622 ± 0.349/ 0.235 ± 0.134	0.531 ± 0.309/ 0.223 ± 0.108	N/A	297.84 ± 164.19
	Initial	0.300 ± 0.150/ N/A	0.267 ± 0.154/ N/A	0.105 ± 0.088/ N/A	N/A	0
RR	DL Reg	0.865 ± 0.021/ 0.103 ± 0.028	0.865 ± 0.022/ 0.101 ± 0.021	0.721 ± 0.037/ 0.130 ± 0.024	125.81 ± 2.90	4.19 ± 0.24
	SyN ANTs	0.856 ± 0.016/ 0.116 ± 0.025	0.858 ± 0.016/ 0.107 ± 0.021	0.716 ± 0.031/ 0.124 ± 0.026	N/A	89.93 ± 17.33
	Initial	0.857 ± 0.030/ N/A	0.859 ± 0.029/ N/A	0.706 ± 0.052/ N/A	N/A	168.6
RAR	DL Reg	0.928 ± 0.016/ 0.092 ± 0.016	0.934 ± 0.014/ 0.098 ± 0.017	0.866 ± 0.030/ 0.155 ± 0.018	128.71 ± 0.28	4.08 ± 0.13
	SyN ANTs	0.907 ± 0.020/ 0.114 ± 0.025	0.917 ± 0.020/ 0.105 ± 0.017	0.848 ± 0.035/ 0.131 ± 0.023	N/A	91.15 ± 14.13
	Initial	0.923 ± 0.024/ N/A	0.931 ± 0.018/ N/A	0.851 ± 0.043/ N/A	N/A	365.8

Also, Table 2 highlights the average sum of absolute log JD values for all three segmentation regions for both DL-based and SyN ANTs approaches. Also, it is to be noted that the percentage of negative JD values is 0% for all initialization approaches. As visible on Table 2, NR contains the larger sum of absolute log JD values for all segmented regions, but remains lower than 0.387. DL Reg seems to lead to more unrealistic deformations compared to SyN ANTs when no prior intialization is done. This is the opposite when looking at RR and RAR initializations, where DL Reg has on average smaller values than

SyN ANTs except in the CSF region. It suggests that the deformation fields are relatively smoother for DL Reg whilst still achieving higher Dice scores. As expected, RR and RAR seem to be more robust to local foldings as their average sum of absolute log JD over all pairs, no matter the segmented region, are always below 0.155 for both DL Reg and SyN ANTs. For further regularization analysis, boxplots in Figure 6 depict the distribution of the sum of absolute log JD values within the whole brain region. This graphical representation highlights that, with the exception of NR, both DL-based registration and SyN ANTs exhibit comparable average absolute sums of log JD for all initialization methods. In the case of NR, DL Reg shows larger values, just as the trends observed in segmented tissues. Lastly, refer to supplementary Figure 12 for an illustration of warped segmentations and their corresponding warping fields for each initialization method and DL-based registration, allowing for a visual examination of their smoothness.

Figure 4 displays an example of registration results on two subjects for all three proposed approaches as well as SyN ANTs counterparts. Moved images are obtained by warping the moving image with the obtained DDFs for both DL-based and SyN ANTs approaches using antsApplyTransforms function from ANTs. Red arrows in the images indicate regions where the warping did not go as expected. On the other hand, green arrows indicate relatively good reproductions of the desired fixed image when inspecting the moved image. Yellow arrows identify areas in proximity to the structures from the fixed image. However, these areas may display blurriness or slight deviations from the expected alignment, indicating regions that are not perfectly matched. For both age intervals, in the RR and RAR scenarios, DL Reg seems to attain similar results as SyN ANTs visually.

To better consider the influence of age intervals on Dice score outcomes, refer to Figure 5, which depicts Dice scores for individual segmented regions against age intervals measured in years per pair. The green data points and corresponding trendlines illustrate outcomes from the proposed initialization followed by DL-based registration. These are contrasted with SyN ANTs Dice scores, represented by red data points and trendlines for all three initialization approaches considered: NR, RR and RAR. The trendlines, visualized through linear regressions, were obtained using ordinary least squares estimation, with the R-squared value reflecting squared linear correlation. In the learning-based context, the average coefficient of determination stands at 0.707. This signifies that, on average, more than 70.7% of the explained Dice score variance is attributed to age interval fluctuations, irrespective of the initialization method employed. Comparatively, the state-of-the-art SyN ANTs pipeline exhibits a slightly lower percentage, with an average R-squared value of 0.653. This discrepancy is primarily due to diminished coefficients when no initialization is applied. In terms of performance, the green trendlines for RR and RAR approaches consistently surpass the red counterparts associated with the conventional ANTs pipeline, across all three segmented regions, except for the CSF and GM in the NR approach. For smaller age intervals, SyN ANTs demonstrates enhanced performance in this particular scenario, specifically for an age interval less than 0.6 for GM and less than 1.6 for CSF. A Plotly HTML displaying trendlines of Dice scores versus age intervals for each one of the 18 segmented regions can be accessed on the GitHub repository.

The same analyses were performed by separating sex information and seeing if it has an impact per initialization approach. These are available in supplementary figures 16 through 18.

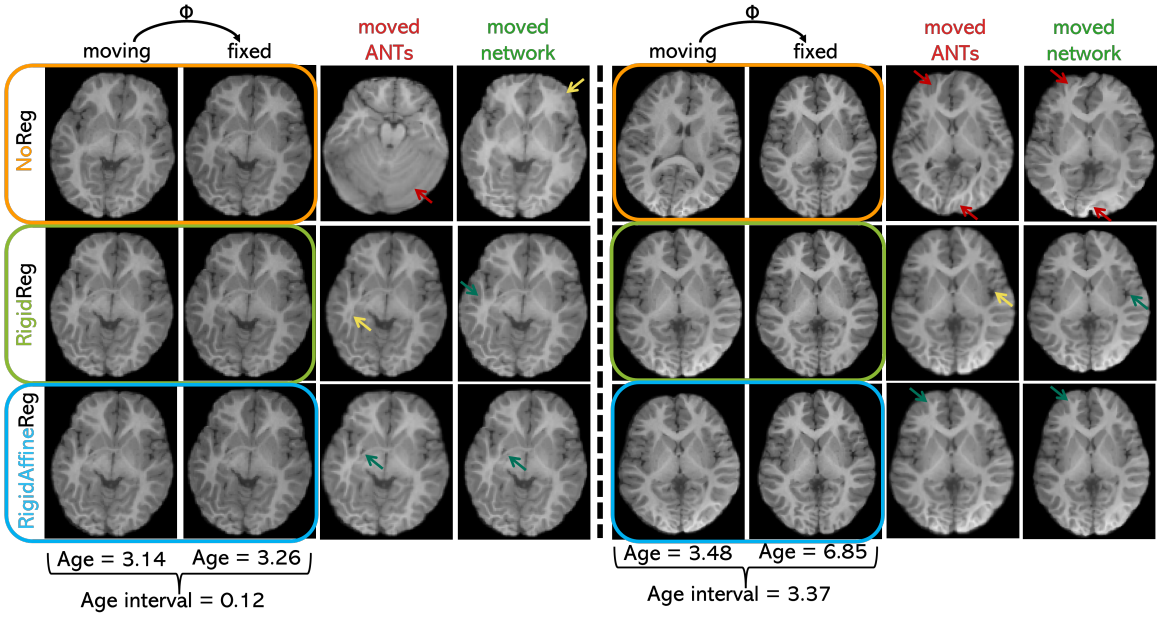

Figure 4: Visual representation of the results obtained with all three DL-based approaches (moved network) compared to SyN ANTs (moved ANTs) results with an age interval between moving and fixed images of 0.116 years on the left and 3.37 years on the right. Red arrows highlight instances of misalignment, yellow arrows indicate blurriness or minor deviations from the fixed image, and green arrows denote successfully aligned areas.

6. Discussion

In this discussion, we examine the registration results obtained using the conventional SyN ANTs alignment method and the U-Net learning-based method. Initially, we analyze Dice scores across all 18 segmented regions, and afterward, we focus on sub-regions within these 18 regions that are included in WM, GM and CSF global tissues. Then, the average sum of absolute JD values are inspected for the whole brain, but also per segmented tissues. Subsequently, we highlight the disparities in both pro-rated training and inference times required for registering intra-subject pairs. Moving forward, we inspect age-related analyses, considering age intervals per pair and any sex-specific differences. We finish by touching upon the generalization capacities of each registration approach and some recommendations for applying intra-subject registration in the pediatric context.

Dice Scores Across All 18 Regions: When applying a pre-alignment strategy before undergoing an unsupervised U-Net, both RR and RAR DL-based approaches exhibit similar performances to the conventional ANTs algorithm in terms of Dice scores. However, when the network is trained, they are faster in terms of registration times compared to the conventional approach. These two DL-based approaches can register two intra-subject images with great quality proven by the high average Dice score results obtained (0.825 ± 0.024 and

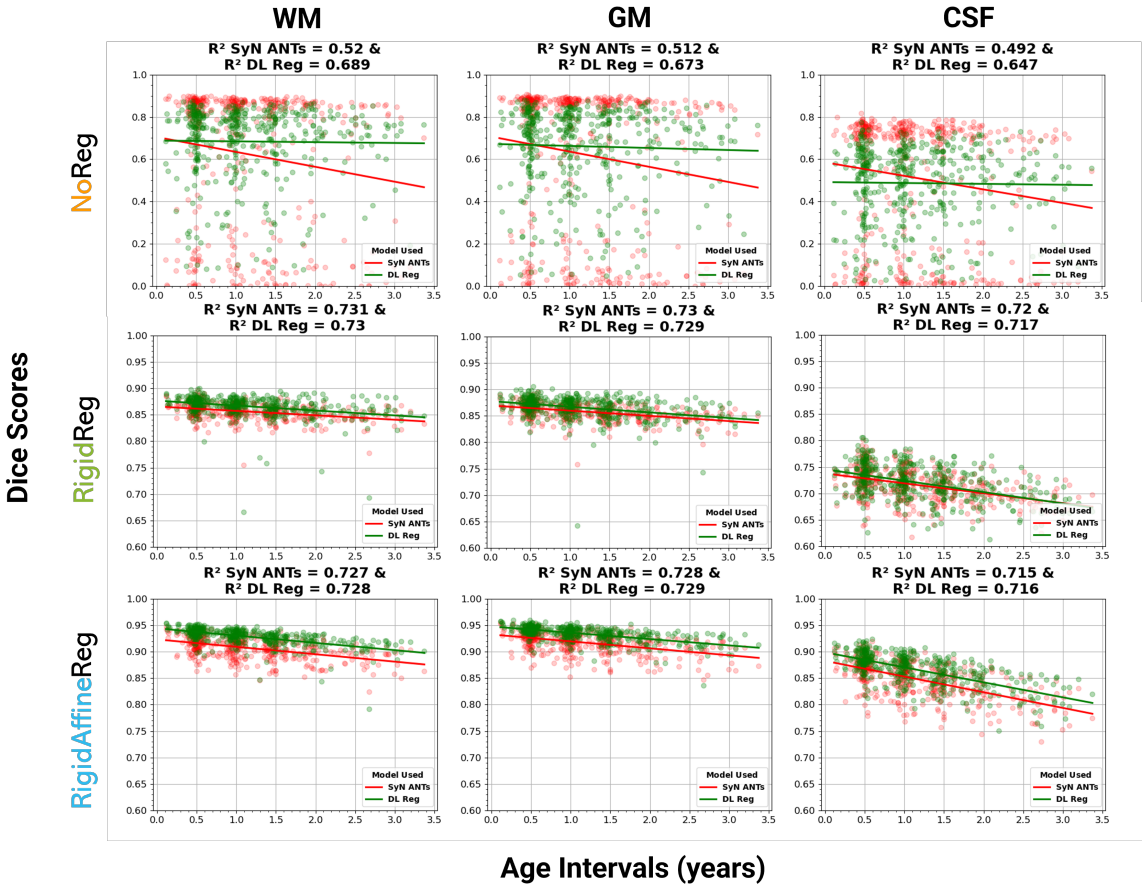

Figure 5: Dice scores against age intervals for all initialization methods compared to SyN ANTs. The three rows correspond to NoReg, RigidReg, and RigidAffineReg, and the three columns represent WM, GM, and CSF segmentations. SyN ANTs Dice scores are shown in red, while the results of DL-based approaches are in green, with corresponding trendlines in the same color. Dice scores in global regions are calculated by averaging SynthSeg sub-regions within these tissues from the total 18 regions available. Figure titles include coefficients of determination (R-squared) for reference. Note: the y-axis scale for NoReg differs due to a distinct range of Dice scores.

0.913±0.018 for RR and RAR respectively) as seen in Figure 3 whilst operating 22 times faster than the SyN ANTs algorithm. Nevertheless, when comparing pro-rated training times to the corresponding SyN ANTs registration times per pair, SyN ANTs demonstrates approximately 1.4 times greater speed for both RR and RAR initializations. For the NR approach, SyN ANTs achieved an average Dice score of 0.591±0.332, and DL Reg achieved 0.617±0.180. Both are higher than the initial Dice score of 0.231±0.131. Interestingly, although SyN ANTs exhibited a lower mean Dice score, its higher median Dice score (0.806)

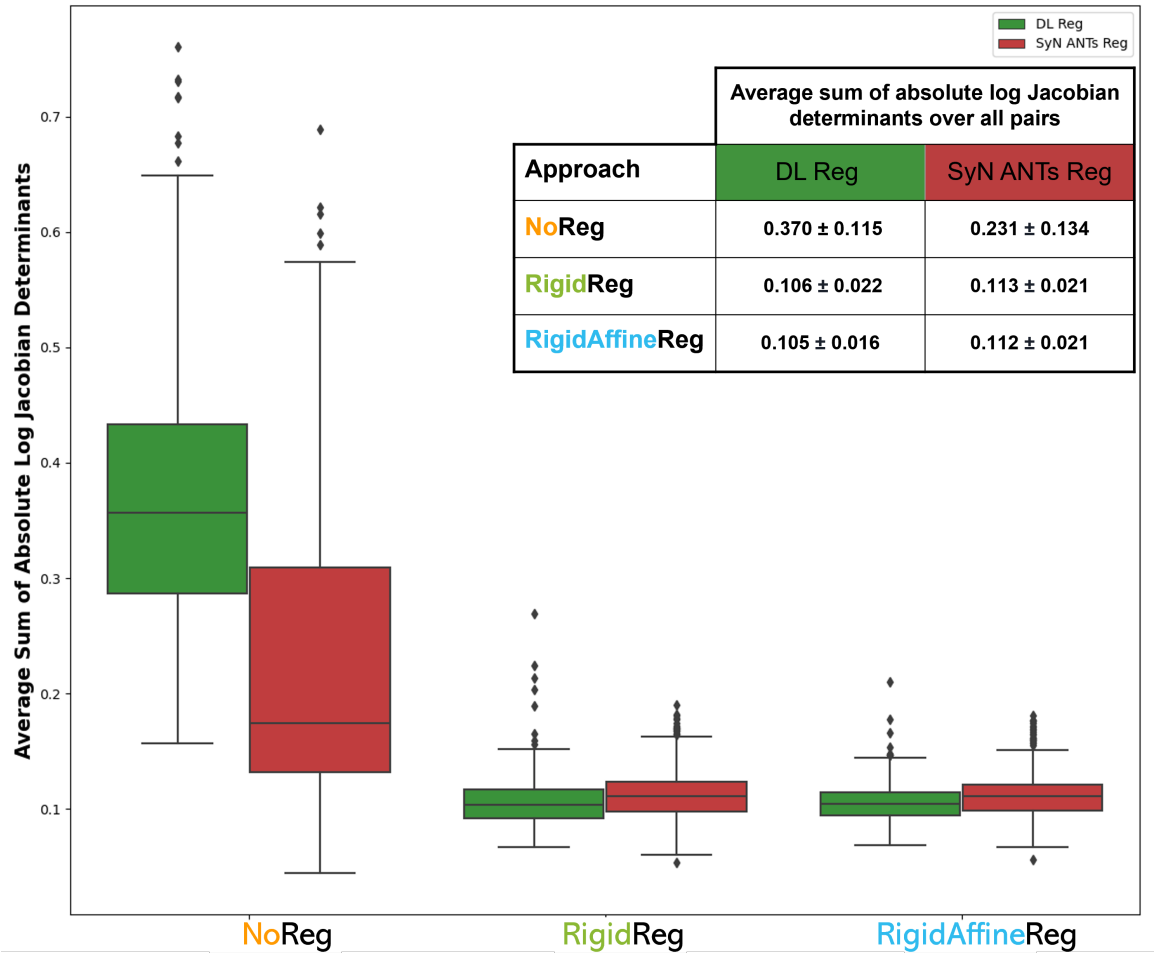

Figure 6: Average sum of absolute Log Jacobian determinants for all three initialization approaches with DL-based in green and SyN ANTs in red. This average is calculated by dividing the sum by the total number of voxels in the brain region per volume. In the upper right corner, a table of the average over all pairs is presented.

compared to DL Reg (0.676) suggests a more robust performance with occasional complete registration failures, highlighting the trade-off between consistency (higher mean) and the potential for superior performance (higher median). It's also crucial to note that for NR, there is no statistical difference between SyN ANTs and DL Reg median Dice scores. This observation may explain why, in Figure 9, the only initialization for which the trend differs from Figure 3 is NR. In this case, when more weight is given to smaller regions based on their number of voxels, SyN ANTs has a mean Dice score of 0.532±0.331, while DL Reg has a mean Dice score of 0.488±0.205, even though SyN ANTs maintains a higher median weighted Dice Score (0.73). For RR, SyN ANTs has the same mean Dice score of 0.735 as

after initial alignment with a smaller standard deviation. Whilst for the unweighted Dice score, SyN ANTs achieved a Dice score of 0.818±0.017, and DL Reg achieved 0.825±0.024 for the RR approach. Again, both are higher than the initial Dice score of 0.816±0.034. Finally, for the RAR approach, SyN ANTs achieved a Dice score of 0.895±0.023, and DL Reg achieved 0.913±0.018. This time, only the DL Reg Dice score is higher than the initial Dice score of 0.907±0.025, not SyN ANTs. It is noteworthy that while SyN ANTs has lower Dice scores for RR and RAR initializations, it demonstrates fewer outliers compared to DL Reg and the initial alignment. Examining more localized regions as illustrated in Figure 11, DL Reg exhibits slightly superior performance compared to SyN ANTs across all 18 regions. Notably, CSF, the 3rd ventricle, and the 4th ventricle show the lowest Dice Scores, while the hippocampus, caudate, and putamen exhibit the highest scores. Frequently, in the context of RR and RAR initializations, SyN ANTs fails to achieve higher Dice scores than those obtained during the initial alignment. Particularly in the case of RAR initialization, where images share substantial similarity in intra-subject registration, SyN ANTs struggles to detect local changes associated with neurodevelopment. This suggests that SyN ANTs may benefit from more nuanced fine-tuning tailored to specific age ranges or individual pairs of images, rather than employing a global approach across a target population. Indeed, to fine-tune ANTs more appropriately for specific age ranges, the primary approach involves using distinct registration parameters tailored to the age group and individual in question. This, however, entails a lengthy and manual process. In contrast, DL-based registration streamlines this by enabling the network to learn and adapt through the training process, eliminating the need for manual adjustments specific to age ranges, although it is trained only for a specific population. Considering the instances where SyN ANTs outperforms DL Reg with NR (64.1%), RR (28.6%), and RAR (10.8%) initializations across all 18 regions and 431 pairs, a hybrid approach could be beneficial. Initially, employing the DL-based approach for all pairs, followed by using SyN ANTs for cases where the registration is unsuccessful, could mitigate processing time while ensuring satisfactory registration results. It is also crucial to note that while all segmentations generated using the DL-based SynthSeg segmentation method underwent quality control evaluation, they still do not constitute ground truth labels. Nevertheless, the method's robustness is evident as it was trained and validated on diverse clinical datasets (Billot et al., 2023b).

Dice Scores in Global Tissues: When examining each segmented tissue separately within the smaller set of 18 regions derived from SynthSeg, where WM and CSF each constitute 27.8% of the total regions, and GM makes up the remaining 44.4%, DL Reg using the RAR initialization outperforms RR for WM (0.928±0.016 vs. 0.865±0.021), GM (0.934±0.014 vs. 0.865±0.022) and CSF (0.721±0.037 vs. 0.866±0.030) segmentations. However, CSF Dice scores consistently tend to be lower compared to WM and GM for both DL Reg and SyN ANTs across all three initialization approaches. Our interpretation is that CSF is more susceptible to the effects of minor misregistration, given its predominantly thin structure around the brain, encompassing the CSF and ventricles. In contrast, WM and GM comprise larger, internal structures, such as the brain stem and cerebral white matter within WM, and the cerebellum cortex within GM. These superior results from RAR are to be expected as all global transformations (rotation, translation, shear, and scaling) are accounted for by the rigid and affine registration pre-alignment steps used for RAR. Therefore, as visualized in Figure 4, the RAR method can uncover local modifications on the

moved image after passing through the network. Contrary to RAR, NR has difficulties doing both local and global transformations as no initialization was applied. Indeed, as presented in Figure 4, parts of the brain microstructures are an unrealistic warping combining both elements from the moving and fixed images for the example pair with an age interval of 3.37 years. The local transformations are not reproduced as well as when using the two other initialization approaches (RR and RAR). Indeed, as visible on both examples in Figure 4, using only a SyN transformation does also not reproduce smaller microstructures correctly, sometimes completely warping incorrectly (for the lower age interval), as no global transformation is done beforehand.

The lower Dice scores for both RR and RAR initialization for each of the 18 regions and looking at the three global tissues may be attributed to the non-optimization of SyN ANTs in a population-based manner, unlike the network. Indeed, the network-based registration relies on learning optimal parameters through training from the provided input population. In contrast, SyN ANTs necessitates individual tweaking for each pair, demanding more human effort. Therefore, the key takeaway is that both SyN ANTs and DL Reg require parameter optimization, but the chosen method varies based on whether registration needs to be performed at a larger scale (encompassing all local regions and multiple subjects) or more locally for a specific small structure. Network-based registration proves fast at inference time, adapting quickly to a given input population, but requires training, potentially adding processing time while being more accurate for the studied population. On the other hand, SyN ANTs requires manual fine-tuning, and clear guidelines are lacking, particularly for younger cohorts, regarding whether to choose age-specific parameters or normalize across ages.

Negative JD and Average Sum of Absolute JD Values: As for the percentage of negative JD values, evaluating if the deformation fields are realistic, the regularization factor of one seems to be sufficient in keeping this percentage to zero for all three approaches. The findings regarding the average sum of absolute log JD values indicate comparability in regularization strengths between DL Reg and SyN ANTs methods, both within the brain region (Figure 6) and for each segmented region (Table 2). Part of the success in DL-based approaches can be attributed to the fact that unsupervised learning is used instead of supervised learning. Indeed, supervised learning, particularly the attempt to predict ANTs-generated deformation fields, might have imposed restrictions on DL-based algorithms by confining their evaluation solely to a comparison with pre-existing ANTs-generated deformation fields.

Pro-rated Training and Inference Times: For the required inference times per pair presented in Table 2, the DL-based methods notably outpace the conventional ANTs pipelines, achieving a speedup of 22 to 74 times once the models are trained. However, the slightly improved Dice scores for RAR initialization compared to RR, observed in both DL Reg and SyN ANTs, come at the cost of doubling the initialization registration times. Concerning the pro-rated training time for DL Reg, calculated by dividing the total training time by the number of training pairs, it averages around 127.51 ± 2.37 across all initialization approaches. In contrast, SyN ANTs doesn't undergo training but necessitates parameter adjustments when a predefined set fails to achieve optimal performance. This process lacks explicit guidelines, usually initiating with default parameters or values recommended by the

literature for the specific population dataset, and subsequently adjusting parameters per pair through quality control procedures.

Age-related Analyses: In Figure 5, as the age interval in-between moving/fixed pairs increases, the Dice score decreases for all three explored initialization approaches for both DL Reg and SyN ANTs. Indeed, comparing how much of explained Dice score variance is due to the age interval when averaging over segmented regions for DL Reg versus SyN ANTs using NR (67.0% vs 50.8%), RR (72.5% vs 72.7%) and RAR (72.4% vs 72.3%) initializations, it shows similar results regardless of the approach. One could think this decrease in Dice score performance is due to the bigger proportion of examples with lower age intervals which the unsupervised network trains on (average age interval of 1.152 ± 0.684, max age interval of 3.372 and min age interval of 0.114). However, the conventional iterative method which takes independent pairs also seems to be following the same negative trend between Dice scores versus age intervals negating this hypothesis. This potentially shows an intrinsic difficulty in registering brains which are quite different in age because of their topology, size and growth factors especially in the pediatric context.

Sex-Specific Differences in Age-related Analyses: As for the supplementary Figures 16 to 18, showing Dice scores versus age intervals separated by sex (45% female pairs and 54% male pairs of total intra-subject pairs) for each initialization approach, the same trends seem to be followed where learning-based registration outperforms SyN ANTs, but both performances diminishing as the age interval increases. The exceptions are for NR, where for all three regions for males, DL Reg does not seem to vary with the age interval, whilst SyN ANTs is slightly more performant for small age intervals and decreases as the age interval increases. However, it's crucial to acknowledge that these differences are based on trendlines reflecting Dice scores, which exhibit significant variability due to the absence of prior initialization. This trend shift is also visible for RR, but this time only for the CSF region only for males, where SyN ANTs is slightly more performant for age intervals bigger than 1.25 years. The coefficients of correlation lie in the same range of around 72% explained variance for all three initialization approaches versus SyN ANTs averaged over regions no matter the sex. The only exception is for NR, where the linear trend is much more visible for the DL Reg approach for males (average of 71.0% explained variance) than females (61.8%). In contrast, for SyN ANTs, the linear trend explains only 48.8% for males and 53.4% for females when averaged across all three segmented regions.

Recommendations: In light of the recommendations for pediatric longitudinal registration, it is crucial to acknowledge the trade-offs between speed, accuracy, and ease of use when selecting an appropriate registration method. Both the RR + ANTs and RAR + ANTs methods provide rapid registration times, approximately 1.5 minutes per pair. Nevertheless, SyN ANTs frequently demands more individualized fine-tuning per pair compared to DL-based approaches, which, in contrast, leverage learned information from a specific population dataset all at once during the training process. It's important to note that the time spent on tuning SyN ANTs parameters is often not accounted for in provided times. Indeed, SyN ANTs exhibited lower performance as it wasn't precisely tuned for each region. For specific small structures like the caudate, amygdala, or pallidum, where precise registration accuracy is crucial, tweaking ANTs parameters might be relevant. On the other hand, if the objective involves studying multiple regions, both local and global, DL-based registration approaches offer a compelling alternative in terms of accuracy and efficiency

when considering no fine-tuning is required when the network is trained for RR and RAR initializations. Figure 11 demonstrates that DL-based methods consistently slightly outperformed ANTs for almost all regions, as they were tailored for the specific population under consideration. It is also to be noted that there are no specific guidelines on how to pick SyN ANTs parameters which makes this task less reproducible. For younger cohorts, where the brain rapidly develops, a question that often arises is, should the parameters be optimized for each developmental stage or age range or standardized across stages? (Turesky et al., 2021). The answer depends on the dataset's characteristics, including available age ranges and the registration requirements, whether it involves segmenting all structures or only a smaller subset of local regions.

Limitations: Finally, it is worth noting that in comparison to SyN ANTs, learning-based algorithms exhibit a notable reliance on their training set, while SyN ANTs proves effective with any intra-subject image pair. The average pro-rated training time for DL Reg, obtained by dividing the total training time by the number of training pairs, is approximately 127.51 ± 2.37 across all initialization approaches. In contrast, SyN ANTs has a zero-second training time since registration is conducted individually for each pair and no learning process is involved. However, achieving optimal registration alignments in terms of Dice scores requires manual tweaking of the SyN parameters. Despite longer inference times, SyN ANTs demonstrates greater generalizability. It is essential to acknowledge the generalization limitations of learning-based algorithms, as the trained network exhibits faster inference times specifically for the chosen pediatric dataset within the provided age range of 2-7 years old subjects. Assessing these strategies on new pediatric MRI datasets within the same age range would shed light on their adaptability to entirely unseen data. While 5-fold cross-validation offers insights into generalization capabilities, testing on a new dataset would provide a more robust assessment. However, given the challenges of obtaining multiple longitudinal time-points in pediatric datasets (Wang et al., 2023), especially within specific age ranges with limited data availability, testing these learning-based strategies on new data becomes notably challenging.

After comparing both DL Reg and SyN ANTs for all three initialization approaches, it is noteworthy that DL-based approaches with RR and RAR initializations, show promising results, delivering Dice scores comparable to SyN ANTs but at significantly faster inference times. However, ANTs' advantage is that it does not require training. RR and RAR excel in registering intra-subject images, particularly due to RAR's robust pre-alignment strategy. On the other hand, NR encounters challenges in capturing both local and global transformations. Age-related analyses reveal a consistent trend of decreasing Dice scores with larger age intervals, a phenomenon observed across all methods. There are some sex-specific shifts in performance noted for NR in all three tissues for males where SyN ANTs is slightly better than DL Reg for smaller age intervals. The exact extent and significance of this influence would require further investigation and analysis. Finally, recommendations highlight the trade-offs between speed, accuracy, and ease of use, providing insights into the suitability of DL-based versus conventional SyN ANTs registration methods for specific applications. The generalization capabilities of learning-based algorithms and the challenges of testing on new pediatric datasets are acknowledged.

7. Conclusion

This study compared the conventional state-of-the-art SyN ANTs registration method with a DL-based approach, evaluating their performance in terms of accuracy, speed, initialization, and their influence on age intervals per pair within the intra-subject pediatric context. Three initialization approaches were explored: NR (without initialization), RR (rigid initialization) and RAR (rigid and affine initialization). Registration quality was evaluated using both unweighted and weighted Dice scores for WM, GM, and CSF segmentations, averaged from corresponding sub-regions of the 18 available regions, while also assessing the individual performance of each region. Additionally, we computed the average sum of absolute log JD values to assess the regularity of the obtained deformation fields for both DL Reg and SyN ANTs methods. These two methods showed comparable regularization strengths for both within the brain region and per segmented region. We demonstrate that learning-based approaches, both with linear pre-alignments (RR and RAR) and without (NR), exhibit slight superiority over the SyN ANTs registration method in terms of Dice scores. Regarding registration quality, the DL-based approaches, specifically with RR and RAR initializations, significantly outperform ($p < 0.05$) SyN ANTs, showing mean Dice scores across all 18 regions of 0.825 ± 0.024 versus 0.818 ± 0.017 and 0.913 ± 0.018 versus 0.895 ± 0.023, respectively. In RR and RAR initializations, SyN ANTs often falls short of surpassing Dice scores achieved during the initial alignment, especially in RAR, indicating challenges in detecting local changes related to neurodevelopment. The main takeaway is that DL-based methods offer faster and more accurate registrations, while SyN ANTs is robust and works well without needing extensive training. Choosing the right method depends on the registration scale needed. DL-based registration adapts quickly but needs training, which can take time, while SyN ANTs requires manual adjustments and lacks clear guidelines, particularly regarding parameter choices for younger cohorts. Both conventional and unsupervised DL-based approaches had their Dice scores decrease as the age interval increased showing intrinsic difficulties to register greater growth changes. Hence, faster registration steps of pairs closer in time can be used to uncover growth characterizations for future pediatric neurodevelopmental pipelines.

Future work will apply this framework on younger populations (0-2 years) from other datasets where the developmental factor is greater. These changes can be inspected in terms of their obtained DDFs to uncover growth patterns in the intra-subject context. It would also be useful to decompose the network into both global and local elements to be able to more accurately register fine-grained and more local brain regions.

Acknowledgments

This study was supported by Polytechnique Montréal, by the Canada First Research Excellence Fund, by CHU Sainte-Justine Research Center and by the TransMedTech Institute.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

References

B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal*, 12(1):26–41, 2008. ISSN 1361-8415 (Print) 1361-8415. .

Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, February 2011.

Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. VoxelMorph: A learning framework for deformable medical image registration. *IEEE Trans. Med. Imaging*, February 2019.

Matthew J Barkovich, Yi Li, Rahul S Desikan, A James Barkovich, and Duan Xu. Challenges in pediatric neuroimaging. *Neuroimage*, 185:793–801, January 2019.

R A I Bethlehem, J Seidlitz, S R White, J W Vogel, K M Anderson, C Adamson, S Adler, G S Alexopoulos, E Anagnostou, A Areces-Gonzalez, D E Astle, B Auyeung, M Ayub, J Bae, G Ball, S Baron-Cohen, R Beare, S A Bedford, V Benegal, F Beyer, J Blangero, M Blesa Cábez, J P Boardman, M Borzage, J F Bosch-Bayard, N Bourke, V D Calhoun, M M Chakravarty, C Chen, C Chertavian, G Chetelat, Y S Chong, J H Cole, A Corvin, M Costantino, E Courchesne, F Crivello, V L Cropley, J Crosbie, N Crossley, M Delarue, R Delorme, S Desrivieres, G A Devenyi, M A Di Biase, R Dolan, K A Donald, G Donohoe, K Dunlop, A D Edwards, J T Elison, C T Ellis, J A Elman, L Eyler, D A Fair, E Feczko, P C Fletcher, P Fonagy, C E Franz, L Galan-Garcia, A Gholipour, J Giedd, J H Gilmore, D C Glahn, I M Goodyer, P E Grant, N A Groenewold, F M Gunning, R E Gur, R C Gur, C F Hammill, O Hansson, T Hedden, A Heinz, R N Henson, K Heuer, J Hoare, B Holla, A J Holmes, R Holt, H Huang, K Im, J Ipser, C R Jack, A P Jackowski, T Jia, K A Johnson, P B Jones, D T Jones, R S Kahn, H Karlsson, L Karlsson, R Kawashima, E A Kelley, S Kern, K W Kim, M G Kitzbichler, W S Kremen, F Lalonde, B Landeau, S Lee, J Lerch, J D Lewis, J Li, W Liao, C Liston, M V Lombardo, J Lv, C Lynch, T T Mallard, M Marcelis, R D Markello, S R Mathias, B Mazoyer, P McGuire, M J Meaney, A Mechelli, N Medic, B Misic, S E Morgan, D Mothersill, J Nigg, M Q W Ong, C Ortinau, R Ossenkoppele, M Ouyang, L Palaniyappan, L Paly, P M Pan, C Pantelis, M M Park, T Paus, Z Pausova, D Paz-Linares, A Pichet Binette, K Pierce, X Qian, J Qiu, A Qiu, A Raznahan, T Rittman, A Rodrigue, C K Rollins, R Romero-Garcia,

L Ronan, M D Rosenberg, D H Rowitch, G A Salum, T D Satterthwaite, H L Schaare, R J Schachar, A P Schultz, G Schumann, M Schöll, D Sharp, R T Shinohara, I Skoog, C D Smyser, R A Sperling, D J Stein, A Stolicyn, J Suckling, G Sullivan, Y Taki, B Thyreau, R Toro, N Traut, K A Tsvetanov, N B Turk-Browne, J J Tuulari, C Tzourio, É Vachon-Presseau, M J Valdes-Sosa, P A Valdes-Sosa, S L Valk, T van Amelsvoort, S N Vandekar, L Vasung, L W Victoria, S Villeneuve, A Villringer, P E Vértes, K Wagstyl, Y S Wang, S K Warfield, V Warrier, E Westman, M L Westwater, H C Whalley, A V Witte, N Yang, B Yeo, H Yun, A Zalesky, H J Zar, A Zettergren, J H Zhou, H Ziauddeen, A Zugman, X N Zuo, E T Bullmore, and A F Alexander-Bloch. Brain charts for the human lifespan. *Nature*, 604(7906):525–533, April 2022.

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, and Juan Eugenio Iglesias. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Med. Image Anal.*, 86:102789, May 2023a.

Benjamin Billot, Colin Magdamo, You Cheng, Steven E Arnold, Sudeshna Das, and Juan Eugenio Iglesias. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. *Proc. Natl. Acad. Sci. U. S. A.*, 120(9): e2216399120, February 2023b.

Li-Zhen Chen, Avram J Holmes, Xi-Nian Zuo, and Qi Dong. Neuroimaging brain growth charts: A road to mental health. *Psychoradiology*, 1(4):272–286, December 2021.

Gary E Christensen, Xiujuan Geng, Jon G Kuhl, Joel Bruss, Thomas J Grabowski, Imran A Pirwani, Michael W Vannier, John S Allen, and Hanna Damasio. Introduction to the non-rigid image registration evaluation project (NIREP). In *Biomedical Image Registration*, pages 128–135. Springer Berlin Heidelberg, 2006.

Andjela Dimitrijevic, Vincent Noblet, and Benjamin De Leener. Deep Learning-Based longitudinal intra-subject registration of pediatric brain MR images. In *Biomedical Image Registration*, pages 206–210. Springer International Publishing, 2022.

Bruce Fischl. FreeSurfer. *Neuroimage*, 62(2):774–781, August 2012.

V Fonov, A C Evans, K Botteron, C R Almli, R C McKinstry, D L Collins, and Group Brain Development Cooperative. Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage*, 54(1):313–327, 2011.

Yunguan Fu, Nina Brown, Shaheer Saeed, Adrià Casamitjana, Zachary Baum, Rémi Delaunay, Qianye Yang, Alexander Grimwood, Zhe Min, Stefano Blumberg, Juan Iglesias, Dean Barratt, Ester Bonmati, Daniel Alexander, Matthew Clarkson, Tom Vercauteren, and Yipeng Hu. Deepreg: a deep learning toolkit for medical image registration. *Journal of Open Source Software*, 5(55), 2020. ISSN 2475-9066. .

Can Gafuroglu, I Rekik, and Alzheimer's Disease Neuroimaging Initiative. Joint prediction and classification of brain image evolution trajectories from baseline brain image with application to early dementia. *MICCAI*, 2018.

Satrajit S Ghosh, Sita Kakunoori, Jean Augustinack, Alfonso Nieto-Castanon, Ioulia Kovelman, Nadine Gaab, Joanna A Christodoulou, Christina Triantafyllou, John D E Gabrieli, and Bruce Fischl. Evaluating the validity of volume-based and surface-based brain image registration for developmental cognitive neuroscience studies in children 4 to 11 years of age. *Neuroimage*, 53(1):85–93, October 2010.

K Gorgolewski et al. NIPYPE: Neuroimaging in python: Pipelines and interfaces. `https://github.com/nipy/nipype`, 2021.

Malte Hoffmann, Andrew Hoopes, Douglas N Greve, Bruce Fischl, and Adrian V Dalca. Anatomy-aware and acquisition-agnostic joint registration with SynthMorph. January 2023.

Juan Eugenio Iglesias. A ready-to-use machine learning tool for symmetric multi-modality registration of brain MRI. *Sci. Rep.*, 13(1):1–15, April 2023.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In C Cortes, N Lawrence, D Lee, M Sugiyama, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Ikbeom Jang, Binyin Li, Joost M Riphagen, Bradford C Dickerson, and David H Salat. Multiscale structural mapping of alzheimer's disease neurodegeneration. *Neuroimage Clin*, 33:102948, January 2022.

Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien P W Pluim. elastix: a toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging*, 29(1):196–205, January 2010.

Dongyang Kuang and Tanya Schmah. FAIM – a ConvNet method for unsupervised 3D medical image registration. In *Machine Learning in Medical Imaging*, Lecture notes in computer science, pages 646–654. Springer International Publishing, Cham, 2019.

Catherine Lebel and Sean Deoni. The development of brain white matter microstructure. *Neuroimage*, 182:207–218, November 2018.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B Avants, Nicholas Ayache, Patricia Buendia, D Louis Collins, Nicolas Cordier, Jason J Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M Iftekharuddin, Raj Jena, Nigel M John, Ender Konukoglu, Danial Lashkari, José Antonió Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J Price, Tammy Riklin Raviv, Syed M S Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A Silva, Nuno Sousa, Nagesh K Subbanna, Gabor Szekely, Thomas J Taylor, Owen M Thomas, Nicholas J Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye

Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging*, 34(10):1993–2024, October 2015.

Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Comput. Methods Programs Biomed.*, 98(3):278–284, June 2010.

Keelin Murphy, Bram van Ginneken, Joseph M Reinhardt, Sven Kabus, Kai Ding, Xiang Deng, Kunlin Cao, Kaifang Du, Gary E Christensen, Vincent Garcia, Tom Vercauteren, Nicholas Ayache, Olivier Commowick, Grégoire Malandain, Ben Glocker, Nikos Paragios, Nassir Navab, Vladlena Gorbunova, Jon Sporring, Marleen de Bruijne, Xiao Han, Mattias P Heinrich, Julia A Schnabel, Mark Jenkinson, Cristian Lorenz, Marc Modat, Jamie R McClelland, Sébastien Ourselin, Sascha E A Muenzing, Max A Viergever, Dante De Nigris, D Louis Collins, Tal Arbel, Marta Peroni, Rui Li, Gregory C Sharp, Alexander Schmidt-Richberg, Jan Ehrhardt, René Werner, Dirk Smeets, Dirk Loeckx, Gang Song, Nicholas Tustison, Brian Avants, James C Gee, Marius Staring, Stefan Klein, Berend C Stoel, Martin Urschler, Manuel Werlberger, Jef Vandemeulebroucke, Simon Rit, David Sarrut, and Josien P W Pluim. Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge. *IEEE Trans. Med. Imaging*, 30(11):1901–1920, November 2011.

Jiahong Ouyang, Qingyu Zhao, Ehsan Adeli, Edith V Sullivan, Adolf Pfefferbaum, Greg Zaharchuk, and Kilian M Pohl. Self-supervised longitudinal neighbourhood embedding. *Med. Image Comput. Comput. Assist. Interv.*, 12902:80–89, September 2021.

Jiahong Ouyang, Qingyu Zhao, Ehsan Adeli, Greg Zaharchuk, and Kilian M Pohl. Disentangling normal aging from severity of disease via weak supervision on longitudinal MRI. *IEEE Trans. Med. Imaging*, 41(10):2558–2569, October 2022.

D. Paniukov, R. M. Lebel, G. Giesbrecht, and C. Lebel. Cerebral blood flow increases across early childhood. *Neuroimage*, 204:116224, 2020. ISSN 1095-9572 (Electronic) 1053-8119 (Linking). . URL `https://www.ncbi.nlm.nih.gov/pubmed/31561017`.

T V Phan, D Smeets, J B Talcott, and M Vandermosten. Processing of structural neuroimaging data in young children: Bridging the gap between current practice and state-of-the-art methods. *Dev. Cogn. Neurosci.*, 33:206–223, 2018.

J. E. Reynolds, X. Long, D. Paniukov, M. Bagshawe, and C. Lebel. Calgary preschool magnetic resonance imaging (mri) dataset. *Data Brief*, 29:105224, 2020. ISSN 2352-3409 (Electronic) 2352-3409 (Linking). . URL `https://www.ncbi.nlm.nih.gov/pubmed/32071993`.

Jess E. Reynolds, Melody N. Grohs, Deborah Dewey, and Catherine Lebel. Global and regional white matter development in early childhood. *bioRxiv*, 2019. . URL `https://www.biorxiv.org/content/early/2019/01/18/524785`.

Torsten Rohlfing. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Trans. Med. Imaging*, 31(2):153–163, February 2012.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015.

Saige Rutherford, Seyed Mostafa Kia, Thomas Wolfers, Charlotte Fraza, Mariam Zabihi, Richard Dinga, Pierre Berthet, Amanda Worker, Serena Verdi, Henricus G Ruhe, Christian F Beckmann, and Andre F Marquand. The normative modeling framework for computational psychiatry. *Nat. Protoc.*, June 2022.

J.-P. Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998. ISSN 1361-8415. . URL `https://www.sciencedirect.com/science/article/pii/S1361841598800224`.

T K Turesky, J Vanderauwera, and N Gaab. Imaging the rapidly developing brain: Current challenges for MRI studies in the first five years of life. *Dev. Cogn. Neurosci.*, 2021.

Nicholas J Tustison, Philip A Cook, Andrew J Holbrook, Hans J Johnson, John Muschelli, Gabriel A Devenyi, Jeffrey T Duda, Sandhitsu R Das, Nicholas C Cullen, Daniel L Gillen, Michael A Yassa, James R Stone, James C Gee, and Brian B Avants. The ANTsX ecosystem for quantitative biological and medical imaging. *Sci. Rep.*, 11(1):9068, April 2021.

S. Uchida. Image processing and recognition for biological images. *Dev Growth Differ*, 55(4):523–49, 2013. ISSN 1440-169X (Electronic) 0012-1592 (Linking). . URL `https://www.ncbi.nlm.nih.gov/pubmed/23560739`.

Jian Wang, Jiaji Wang, Shuihua Wang, and Yudong Zhang. Deep learning in pediatric neuroimaging. *Displays*, 80:102583, December 2023.

Jun Zhang. Inverse-Consistent deep networks for unsupervised deformable image registration. *arXiv.org*, 2018.

Appendix A. Supplementary Material

A.1 Data Preprocessing

Prior to using DL models, preprocessing of MR images is crucial. This involves N4 bias correction for intensity uniformity and skull-stripping for isolating the brain region. Skull-stripping was done by first doing rigid, affine and SyN registration steps to the Montreal's Neurological Institute (MNI) 4.5-8.5 template (Fonov et al., 2011) to obtain a brain mask for each subject scan. For rigid and the affine transforms, the parameters used are a gradient of 0.1, mattes similarity metric, 1000x11110x11110 multi-resolution steps, a threshold of 1e-7 for 20 iterations as a convergence criteria, 3x2x1 shrink factors and 4x2x1 voxels as smoothing sigmas. As for SyN, its gradient step, updateFieldVarianceInVoxelSpace and parameters are respectively 0.2, 3 and 0. A cross-correlation similarity metric is used with 100x100x50 multi-resolution steps, a threshold of -0.01 for 5 iterations as a convergence criteria, 4x2x1 shrink factors and 1x0.5x0 voxels as smoothing sigmas. The registrations performed were done at full resolution to ensure better alignment results. However, keeping the images at full resolution also increases registration time. Paniukov et al. (2020)'s techniques, executed through Nipype in Python pipelines (Gorgolewski et al., 2021), were adopted for these tasks, as the analysis was successfully done on the same dataset. Moreover, a manual quality control (QC) step was incorporated to ensure accurate registration-based skull-stripping for each scan, and no pairs were excluded as a result. The final skull-stripped images were used as inputs to both SyN ANTs and the U-Net after rescaling them to 1.5x1.5x1.5 mm isotropic.

As for the parameters used for rigid as well as rigid and affine initializations, they are a gradient of 0.1, mattes similarity metric, 500x250x100 multi-resolution steps, a threshold of 1e-6 for 10 iterations as a convergence criteria, 4x2x1 shrink factors and 2x1x0 voxels as smoothing sigmas. If some registrations failed with specific image pairs with the given parameters, the multi-resolution steps were increased, and the threshold decreased. For SyN ANTs, we employed a gradient step of 0.1, and updateFieldVarianceInVoxelSpace and totalFieldMeshSizeAtBaseLevel parameters were respectively set to 3 and 0.

A.2 Metrics

A.2.1 Unweighted and Weighted Dice Scores

We use the term *unweighted Dice score* to refer to the Dice score equation commonly used in registration studies, which assesses registration performance in terms of volume overlap between the warped image and ground truth. This metric is also interchangeably referred to as simply the Dice score throughout the article. These unweighted Dice scores are also inspected with respect to the age interval per pair and separated by sex. Dice scores are calculated in the white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) by averaging SynthSeg sub-regions within these tissues from the total 18 regions available. Given v, the voxels for the fixed (F) and moved volume (M), the Dice score for a region, r is calculated as follows:

$$Dice(v_{\mathrm{F}}^{\mathrm{r}}, v_{\mathrm{M}}^{\mathrm{r}} \circ \phi) = 2 \cdot \frac{|v_{\mathrm{F}}^{\mathrm{r}} \cap (v_{\mathrm{M}}^{\mathrm{r}} \circ \phi)|}{|v_{\mathrm{F}}^{\mathrm{r}}| + |v_{\mathrm{M}}^{\mathrm{r}} \circ \phi|} \tag{1}$$

An additional analysis was introduced to reduce the influence of larger regions, preventing them from disproportionately elevating the mean Dice score across all 18 regions segmented using SynthSeg. Hence, a weighted Dice score, D_{weighted}, is calculated per region, r, using a weight, w_{r} by incorporating the inverse number of voxels as presented in the equations 2 to 4 below. The average weighted Dice scores are obtained by summing the scores across all 18 segmented regions and subjects.

$$D_{\text{weighted}} = \sum_{\text{r}=1}^{18} w_{\text{r}} \times D_{\text{unweighted}}(\text{r}) \tag{2}$$

w_r is calculated as such per region where the denominator ensures that the weights collectively add up to 1 per subject,

$$w_{\text{r}} = \frac{W_{\text{r}}}{\sum_{\text{r}=1}^{18} W_{\text{r}}}, \tag{3}$$

and the main weighting factor is the inverse of the number of voxels in that specific region as follows,

$$W_{\text{r}} = \frac{1}{V_{\text{r}}} \tag{4}$$

A.2.2 Negative JD and Sum of Absolute Log JD values

We compute two key metrics related to the JD to assess deformation fields within local sub-regions extracted from SynthSeg and encompassing global white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) regions, excluding background regions. The JD quantifies volumetric changes following registration processes. The sum of absolute log JD values allows us to have a measure of the spread of the distribution of the JDs. For a volume of size *mxnxp*, it is calculated as represented in equation 5 below. The average absolute sum of log Jacobian determinants is obtained by dividing the sum by the total number of voxels in the brain region per volume.

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{p} |\log(\text{JD}(i,j,k))| \tag{5}$$

As a negative JD signifies non-invertible properties, indicating the presence of undesired local folding, we also quantify the number of negative JDs using the following expression:

$$\%\text{ of Negative JD values} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{p} 1(\text{JD}(i,j,k) < 0)}{m \cdot n \cdot p} \cdot 100 \tag{6}$$

A.3 Representation of the Longitudinal Data

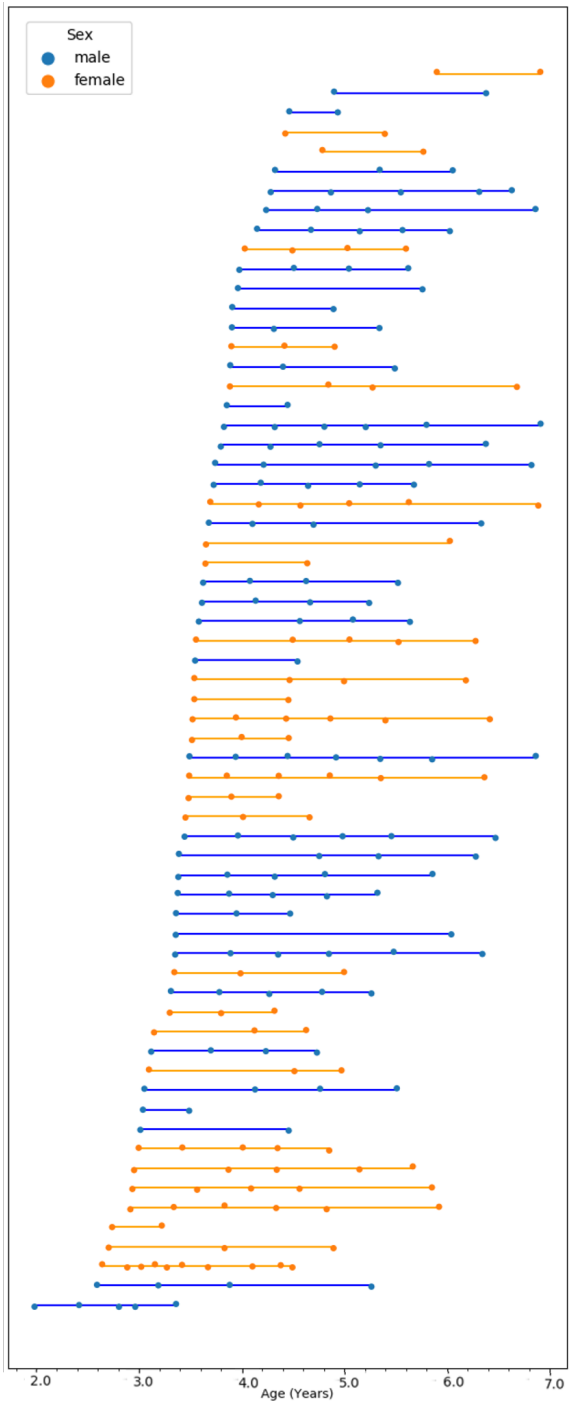

Figure 7: Representation of the 64 subjects used with age at scans and biological sex information, image inspired by (Reynolds et al., 2019).

A.4 Full Pipeline

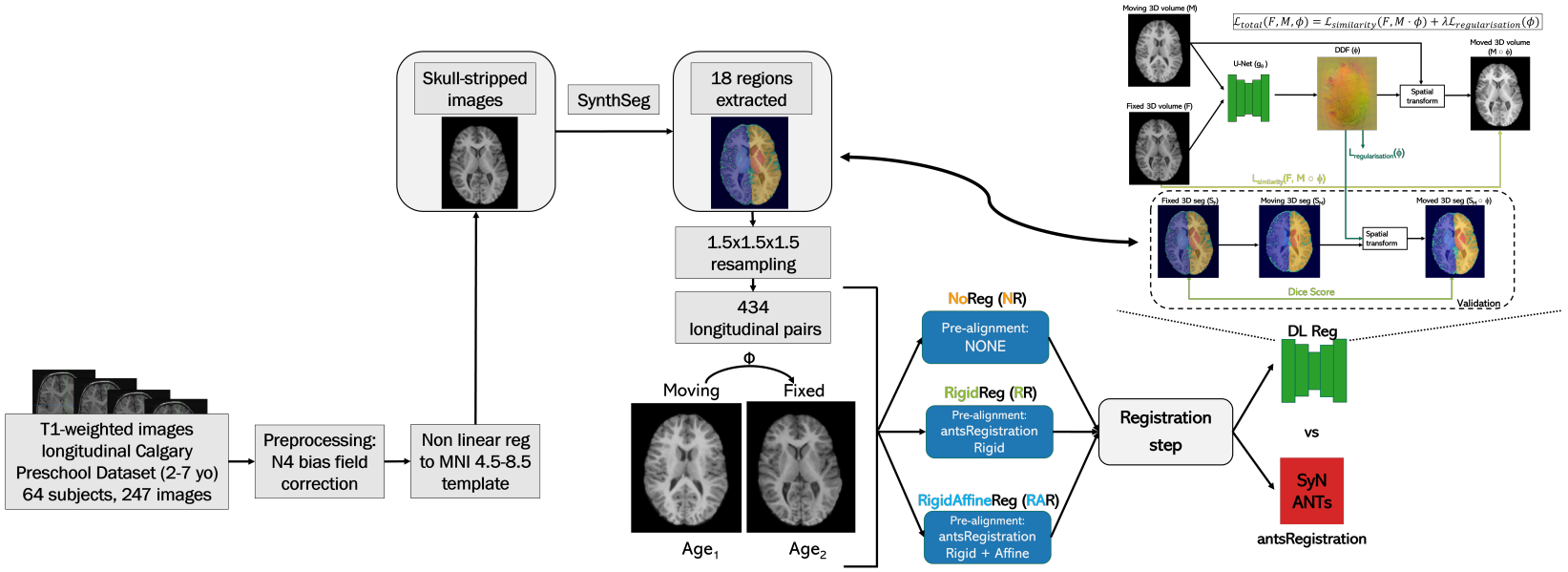

Figure 8: Full pipeline: 247 images from the longitudinal Calgary Preschool dataset are first N4 bias field corrected then non-linearly registered to the MNI 4.5-8.5 template to obtain skull-stripped. These skull-stripped images are segmented by DL-based $SynthSeg^{+}$ (Billot et al., 2023a), a robust segmentation method, then resampled to 1.5 mm isotropic resolution. All longitudinal pairs per subject (average: 3.86±1.59 time points/subject) are pre-aligned considering three initialization approaches (NoReg, RigidReg and RigidAffineReg). Deformations obtained by a unsupervised registration scheme using DeepReg are compared to the conventional SyN ANTs method using Dice scores.

A.5 Weighted Dice Scores Across All 18 Regions

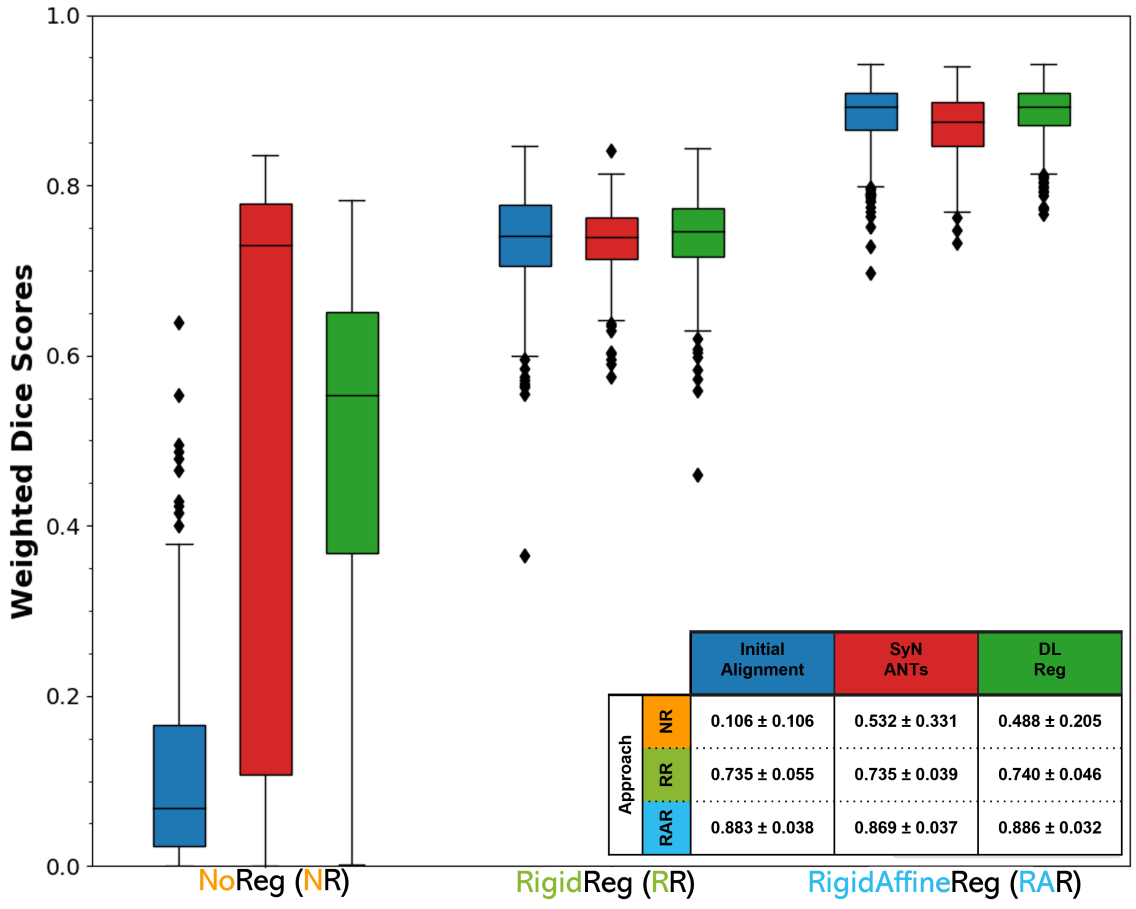

Figure 9: Weighted Dice score results on the test sets represented as boxplots for each initialization approach (NoReg, RigidReg and RigidAffineReg) for DL-based methods compared to the initial Dice scores pre-conducting the registration steps. Each method is also compared to the SyN ANTs registration. The weighting is determined by the inverse number of voxels per region, normalized by the sum of weights for all 18 regions, ensuring they collectively add up to 1 per subject. The average weighted Dice scores are computed by summing across all 18 segmented regions and subjects, considering their respective weights. The table in the lower right corner provides the mean±SD Dice scores for all scenarios.

A.6 Dice Scores in Global Tissues

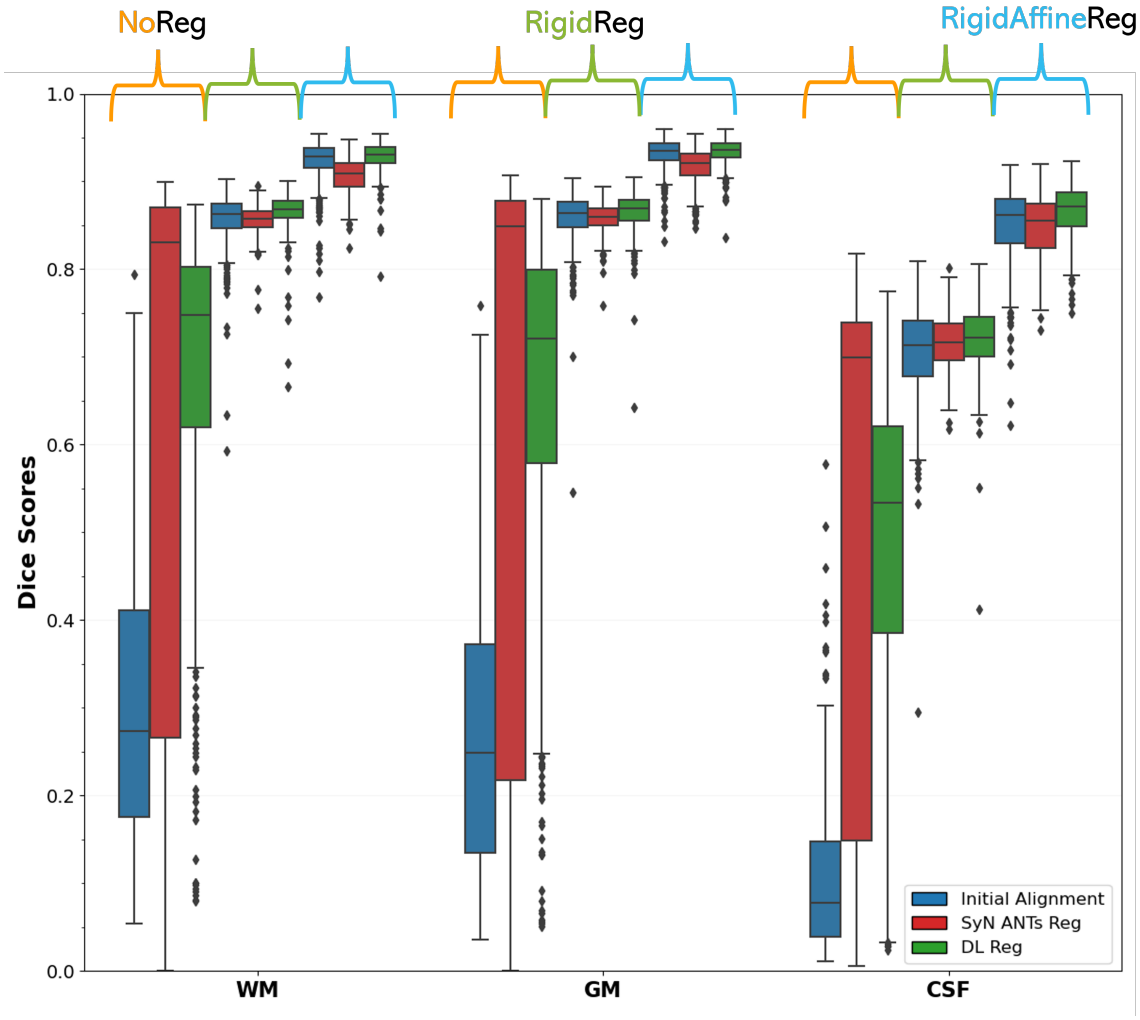

Figure 10: Dice score results on the evaluation set represented as boxplots for each method (NoReg, RigidReg and RigidAffineReg indicated as respectively orange, light green or light blue braces) compared to the initial Dice scores pre-conducting the registration steps. Each method is compared to SyN ANTs registration (refer to Figure 1). Dice scores are computed for the white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) by averaging SynthSeg sub-regions within these tissues from the total 18 regions available.

A.7 Dice Scores per Region

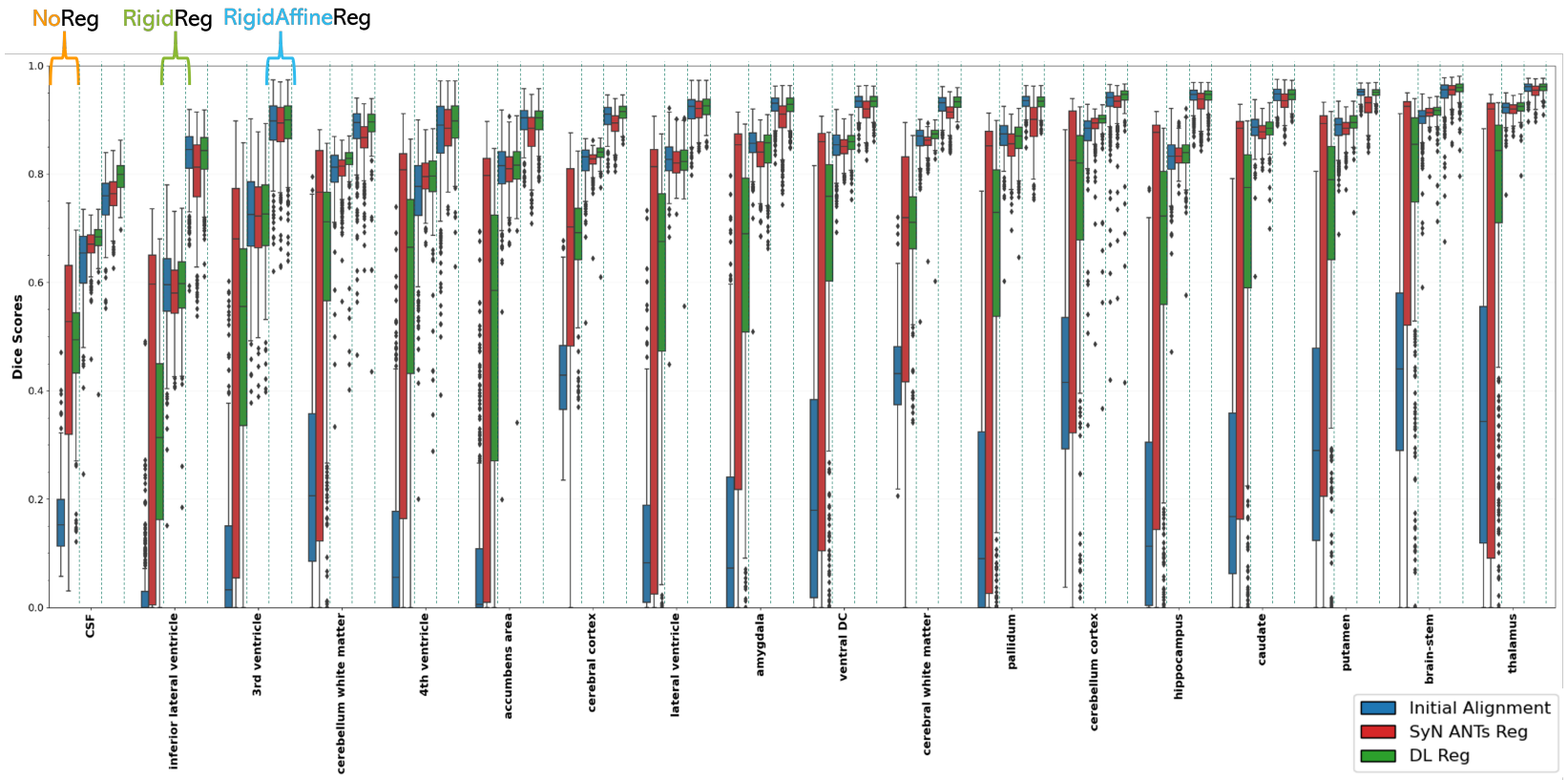

Figure 11: Boxplots depict Dice score results on the evaluation set for each method (NoReg, RigidReg, and RigidAffineReg denoted by orange, light green, or light blue braces, respectively) in comparison to the initial Dice scores before executing the registration steps. Each method is contrasted with SyN ANTs registration (refer to Figure 1). Dice scores are computed for every region amongst the 18 regions extracted from SynthSeg, averaging left and right brain hemisphere labels for all structures except brain stem and CSF. Regions are arranged in ascending order according to the Dice scores obtained with DL-based RigidAffineReg. For reference, the 18 brain regions, ordered from smallest to largest by the number of voxels, are as follows: 3rd ventricle, inferior lateral ventricle, accumbens area, 4th ventricle, amygdala, pallidum, ventral DC, hippocampus, caudate, lateral ventricle, putamen, thalamus, brain-stem, cerebellum white matter, cerebellum cortex, CSF, cerebral white matter, and cerebral cortex.

A.8 Visualisations of Example Pairs

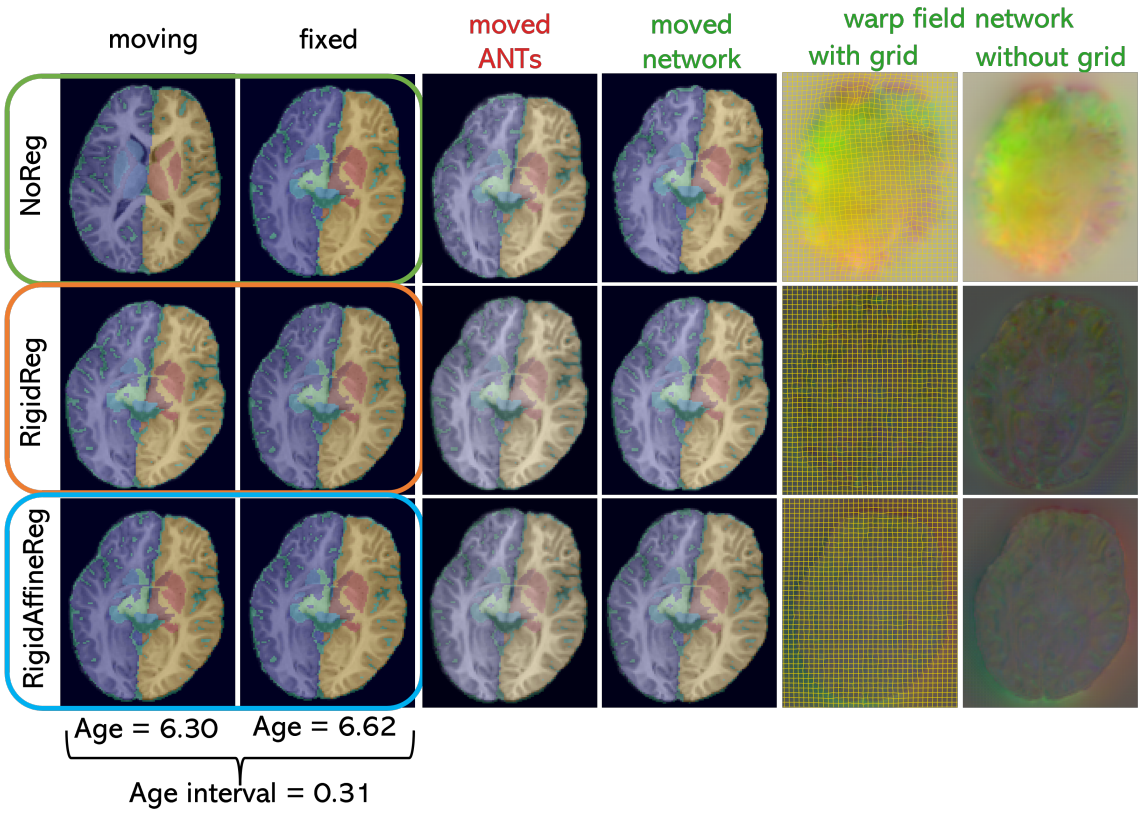

Figure 12: Illustration of overlaid moving, fixed, and transformed segmentations on the intensity volumes for each of the three initialization methods (NoReg, RigidReg and RigidAffineReg) and both ANTs and DL-based pipelines for a subject with an age interval of 0.31. The fourth and fifth columns depict the warping grid overlaid with the RGB image of displacement values in each spatial dimension and the RGB image itself, respectively.

A.9 P-Values and Statistical Significance Analysis

	NR		RR		RAR	
Region	P-Value	Significant	P-Value	Significant	P-Value	Significant
cerebral white matter	3.03636E-28	yes	8.33245E-11	yes	5.20819E-64	yes
cerebral cortex	5.58339E-38	yes	0.666152131	**no**	4.33949E-57	yes
lateral ventricle	4.13988E-51	yes	1.84045E-16	yes	0.204170608	**no**
inferior lateral ventricle	8.23746E-57	yes	0.010856065	yes	2.15578E-54	yes
cerebellum white matter	1.90971E-41	yes	0.597890228	**no**	8.42952E-31	yes
cerebellum cortex	4.10159E-24	yes	1.3322E-24	yes	1.96013E-09	yes
thalamus	3.95628E-31	yes	0.006231265	yes	3.20016E-40	yes
caudate	2.35023E-47	yes	9.46423E-19	yes	9.52174E-48	yes
putamen	1.30286E-37	yes	0.003350506	yes	1.0193E-69	yes
pallidum	1.19705E-48	yes	5.75179E-24	yes	6.56591E-67	yes
3rd ventricle	8.37827E-54	yes	0.632308794	**no**	0.67563223	**no**
4th ventricle	1.05671E-54	yes	1.65677E-14	yes	0.103480785	**no**
brain-stem	1.15691E-34	yes	2.35206E-16	yes	0.026295907	yes
hippocampus	5.42914E-51	yes	0.014733658	yes	2.0636E-54	yes
amygdala	3.15231E-54	yes	1.57074E-26	yes	4.9128E-59	yes
CSF	3.02915E-61	yes	3.91488E-15	yes	0.002384396	yes
accumbens area	2.77112E-55	yes	0.082012561	**no**	8.72515E-45	yes
ventral DC	2.40851E-47	yes	0.550664305	**no**	4.40345E-57	yes

Figure 13: P-values and statistical significance ($p<0.05$) derived from a Wilcoxon test comparing SyN ANTs Dice scores to initial Dice scores before any initialization, across all 18 individual regions. The colors in the table correspond to global regions, with white matter in yellow, gray matter in turquoise, and cerebrospinal fluid in purple, matching the color scheme outlined in Table 2. Each set of two columns represents data from one of the three initialization approaches (NoReg, RigidReg, and RigidAffineReg).

	NR		RR		RAR	
Region	P-Value	Significant	P-Value	Significant	P-Value	Significant
cerebral white matter	7.63541E-73	yes	3.58821E-20	yes	9.06494E-59	yes
cerebral cortex	7.63541E-73	yes	1.56511E-41	yes	5.32474E-70	yes
lateral ventricle	1.64943E-71	yes	0.00020016	yes	1.31415E-33	yes
inferior lateral ventricle	5.29644E-69	yes	9.82491E-05	yes	0.011579884	yes
cerebellum white matter	8.52543E-72	yes	1.28693E-48	yes	2.28947E-23	yes
cerebellum cortex	3.34988E-72	yes	6.85202E-69	yes	1.25917E-69	yes
thalamus	4.289E-71	yes	3.30132E-13	yes	1.17552E-18	yes
caudate	2.77053E-72	yes	0.241134824	**no**	5.62168E-05	yes
putamen	1.37917E-72	yes	2.00567E-16	yes	0.327798676	**no**
pallidum	1.14637E-70	yes	0.001049835	yes	0.219017182	**no**
3rd ventricle	2.12768E-65	yes	0.725454173	**no**	0.582750135	**no**
4th ventricle	6.446E-64	yes	5.96637E-11	yes	1.81568E-05	yes
brain-stem	4.1518E-60	yes	6.30116E-24	yes	1.09261E-10	yes
hippocampus	3.00264E-71	yes	1.12249E-10	yes	0.313852311	**no**
amygdala	1.5967E-68	yes	0.200567365	**no**	0.289404928	**no**
CSF	1.04983E-72	yes	1.19125E-31	yes	6.77239E-45	yes
accumbens area	1.87621E-67	yes	3.48992E-05	yes	0.959299082	**no**
ventral DC	2.58983E-71	yes	8.35306E-17	yes	6.01504E-06	yes

Figure 14: P-values and statistical significance ($p<0.05$) derived from a Wilcoxon test comparing DL Reg Dice scores to initial Dice scores before any initialization, across all 18 individual regions. The colors in the table correspond to global regions, with white matter in yellow, gray matter in turquoise, and cerebrospinal fluid in purple, matching the color scheme outlined in Table 2. Each set of two columns represents data from one of the three initialization approaches (NoReg, RigidReg, and RigidAffineReg).

	NR		RR		RAR	
Region	P-Value	Significant	P-Value	Significant	P-Value	Significant
cerebral white matter	0.001473155	yes	1.18792E-62	yes	1.52127E-71	yes
cerebral cortex	0.080742571	**no**	9.32005E-60	yes	1.52127E-71	yes
lateral ventricle	0.550889111	**no**	2.03043E-16	yes	8.14115E-37	yes
inferior lateral ventricle	1.9576E-12	yes	9.2995E-08	yes	8.72263E-56	yes
cerebellum white matter	0.293411772	**no**	3.05524E-45	yes	6.40907E-65	yes
cerebellum cortex	0.004321782	yes	2.14689E-48	yes	1.00819E-63	yes
thalamus	0.815222593	**no**	4.6973E-18	yes	2.16944E-54	yes
caudate	0.256293989	**no**	3.5308E-17	yes	3.63828E-62	yes
putamen	0.834877304	**no**	3.25964E-27	yes	7.75715E-71	yes
pallidum	0.116954073	**no**	8.17348E-14	yes	1.27265E-68	yes
3rd ventricle	0.001444037	yes	0.002903719	yes	4.96684E-05	yes
4th ventricle	0.000997363	yes	0.378204469	**no**	1.20127E-32	yes
brain-stem	0.125000508	**no**	2.35328E-08	yes	7.44135E-43	yes
hippocampus	0.145777354	**no**	0.001397245	yes	1.96331E-57	yes
amygdala	0.000441306	yes	1.09331E-08	yes	1.25864E-63	yes
CSF	0.65644862	**no**	2.81381E-46	yes	1.52127E-71	yes
accumbens area	4.27622E-07	yes	3.31066E-09	yes	1.97515E-48	yes
ventral DC	0.154029847	**no**	1.42966E-12	yes	1.2134E-62	yes

Figure 15: P-values and statistical significance ($p<0.05$) derived from a Wilcoxon test comparing DL Reg Dice scores to SyN ANTs Dice scores, across all 18 individual regions. The colors in the table correspond to global regions, with white matter in yellow, gray matter in turquoise, and cerebrospinal fluid in purple, matching the color scheme outlined in Table 2. Each set of two columns represents data from one of the three initialization approaches (NoReg, RigidReg, and RigidAffineReg).

A.10 Sex-Specific Age-Related Analyses

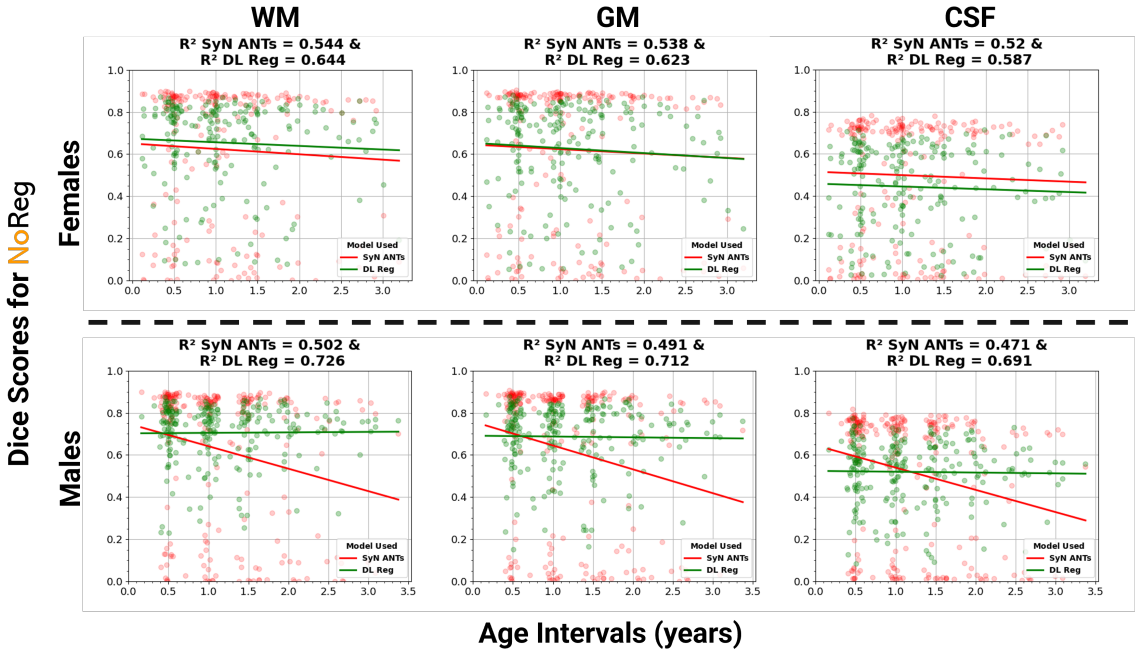

Figure 16: Dice score results compared to the age interval between moving and fixed pairs calculated on the test set separated by sex, females on the first row and males on the second for no prior initialization and DL-based registration (DL Reg) in green compared to SyN ANTs registration in red. The Dice scores are calculated in the white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) by averaging SynthSeg sub-regions within these tissues from the total 18 regions available and are presented on each column. Coefficients of determination (R-squared) are also presented in each figure title.

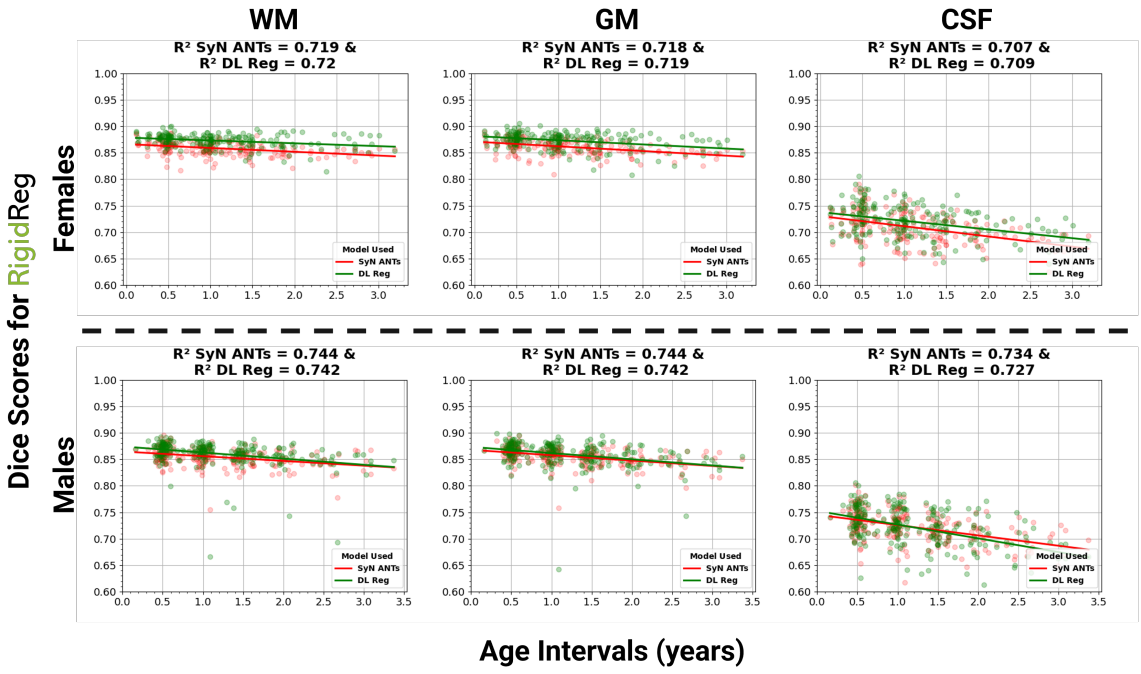

Figure 17: Dice score results compared to the age interval between moving and fixed pairs calculated on the test set separated by sex, females on the first row and males on the second for rigid initialization and DL-based registration (DL Reg) in green compared to SyN ANTs registration in red. The Dice scores are calculated in the white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) by averaging SynthSeg sub-regions within these tissues from the total 18 regions available and are presented on each column. Coefficients of determination (R-squared) are also presented in each figure title.

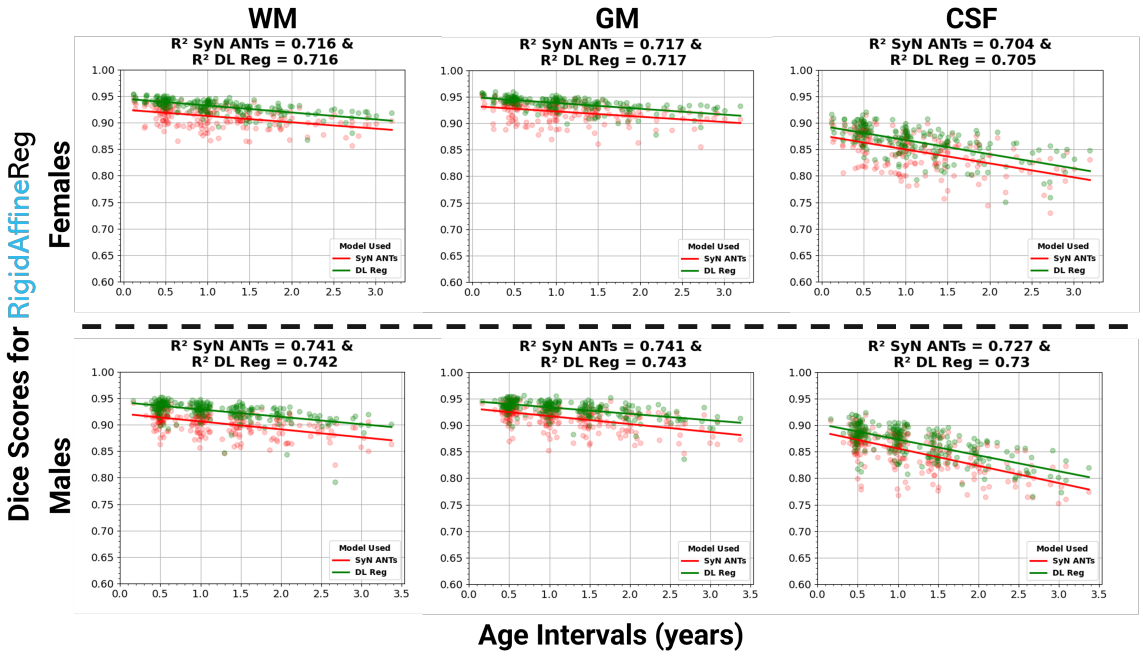

Figure 18: Dice score results compared to the age interval between moving and fixed pairs calculated on the test set separated by sex, females on the first row and males on the second for rigid and affine initialization and DL-based registration (DL Reg) in green compared to SyN ANTs registration in red. The Dice scores are calculated in the white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) by averaging SynthSeg sub-regions within these tissues from the total 18 regions available and are presented on each column. Coefficients of determination (R-squared) are also presented in each figure title.