Surfing the Hippocampus Wave

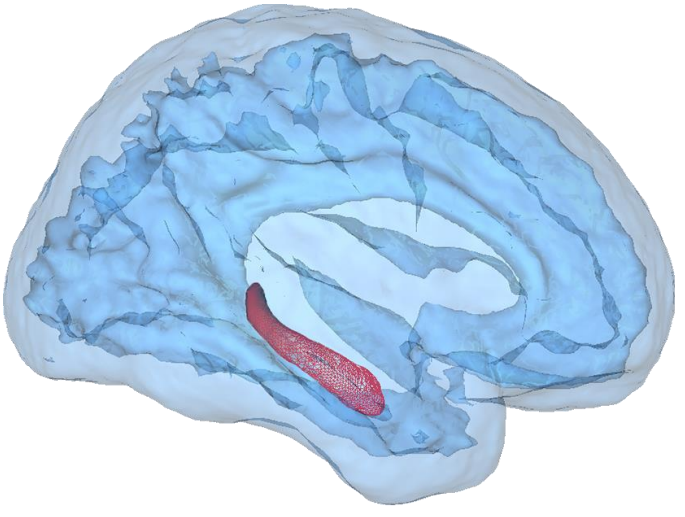
Bartel, F.

2019

document version
Publisher's PDF, also known as Version of record

Chapter 6

Non-linear registration improves statistical power to detect hippocampal atrophy in aging and dementia



Fabian Bartel
Martin Visser
Michiel B. de Ruiter
Jose Belderbos
Frederik Barkhof
Hugo Vrenken
Jan C. de Munck
Marcel B. van Herk

Submitted to NeuroImage: Clinical July 2018

Abstract

Objective: To compare the performance of different methods for determining hippocampal atrophy rates using longitudinal MRI scans in aging and Alzheimer's disease (AD).

Background: Quantifying hippocampal atrophy caused by neurodegenerative diseases is important to follow the course of the disease. In dementia, the efficacy of new therapies can be partially assessed by measuring their effect on hippocampal atrophy. In radiotherapy, the quantification of radiation-induced hippocampal volume loss is of interest to estimate the hippocampus's radiation damage. We evaluated plausibility, reproducibility and sensitivity of eight commonly used methods to determine hippocampal atrophy rates using test-retest scans.

Materials and Methods: Manual, FSL-FIRST, FreeSurfer, multi-atlas segmentation and non-linear registration methods (Elastix, NiftyReg, ANTs and MIRTK) were used to determine hippocampal atrophy rates on longitudinal T1-weighted MRI form the ADNI database. Appropriate parameters for the non-linear registration methods were determined using a small training dataset (N=16) in which two-year hippocampal atrophy was measured using test-retest scans of 8 subjects with low and 8 subjects with high atrophy rates. On a larger dataset of 20 controls, 40 mild cognitive impairment (MCI) and 20 AD patients, one-year hippocampal atrophy rates were measured. Reproducibility of hippocampal atrophy rates was determined using within-session rescans and expressed as an average distance measure D_{Ave}, which expresses the difference in atrophy rate, averaged over all subjects. The same D_{Ave} was used to determine the agreement between different methods. Finally, for each method we calculated effect sizes and the required sample sizes to detect one-year volume change between controls and MCI (N_{CTRL_MCI}) and between controls and AD (N_{CTRL_AD}).

Results: Manually measured hippocampal atrophy rates were poorly reproducible and required the largest sample sizes (N_{CTRL_MCI}=452, N_{CTRL_AD}=87 with D_{Ave}=12.39%). Non-linear registration methods were most consistent in determining hippocampal atrophy on the within-session rescans and atrophy rates of these methods also agreed best with each other. FreeSurfer and MIRTK generally required lowest sample sizes (FreeSurfer: N_{CTRL_MCI}=115, N_{CTRL_AD}=17 with D_{Ave}=3.26%; MIRTK: N_{CTRL_MCI}=97, N_{CTRL_AD}=11 with D_{Ave}=3.76%), while ANTs was most reproducible (N_{CTRL_MCI}=162, N_{CTRL_AD}=37 with D_{Ave}=1.06%), followed by Elastix (N_{CTRL_MCI}=226, N_{CTRL_AD}=15 with D_{Ave}=1.78%) and NiftyReg (N_{CTRL_MCI}=193, N_{CTRL_AD}=14 with D_{Ave}=2.11%).

Discussion and conclusion: Because of their better reproducibility, non-linear registration methods are preferred for determining hippocampal atrophy rates on longitudinal MRI. Since performances of non-linear registration methods are well comparable, the preferred method would depend on computational efficiency.

6.1. Introduction

The hippocampus is a small cortical structure that plays an important role in memory formation. In many neurodegenerative diseases it is impaired due to progressive degeneration and/or death of nerve cells which is reflected in a decrease in hippocampal volume. Hippocampal atrophy has been extensively studied in Alzheimer's disease (AD), where neurodegeneration leads to structural brain changes visible on MRI. Hippocampal atrophy is typically determined by delineating (manually or automatically) the hippocampus on longitudinal MRI scans.

Hippocampal volume and volume change have been studied extensively. For instance, hippocampal atrophy has been reported to be larger in AD compared to mild cognitive impairment (MCI) or healthy controls [1–6]. In clinical trials, in which disease-modifying therapies are studied, hippocampal atrophy is an important biomarker to provide biological evidence of treatment effect and to better understand the underlying mechanism. Completed clinical trials in which hippocampal atrophy is an outcome measure are reviewed in [7].

Recently, there has also been a high interest in investigating hippocampal damage after radiotherapy. Animal studies have shown that the neural stem cell (NSC) compartment in the dentate gyrus of the hippocampus is vulnerable to radiation toxicity and already small doses can damage the NSC [8–12]. Often, in patients with brain tumors or brain metastases brain radiation therapy is the core treatment [13]. Furthermore, in patients with small cell lung cancer (SCLC) prophylactic cranial irradiation (PCI) is used to treat microscopic brain metastases that are statistically likely present, and to reduce the risk of developing larger metastases [14,15]. The magnitude of hippocampal volume loss due to PCI or other brain radiation treatment techniques is currently unknown. For this reason, it is important to assess and compare the sensitivity of currently available processing techniques to detect volume differences in hippocampal volume from longitudinal MRI.

The hippocampus has limited contrast on MR images because adjacent structures have similar intensities [16] and therefore manual delineation is labor intensive and difficult even for experienced observers. Furthermore, this lack of contrast also is an important source of intra- and inter-observer variability. Measuring hippocampal atrophy rates on longitudinal MRI scans is even more challenging, because in segmentations performed on multiple time-points the segmentation errors add up,

whereas the volume change is small. This was shown by [17], in which manually measured atrophy rates were not well reproduced using longitudinal data with within-session rescans. To avoid the burden of manual labor and to reduce observer variability, automatic segmentation methods have been proposed, most of them reviewed in [18] and [19]. For instance, FSL-FIRST [20] and FreeSurfer [16,21] are automatic segmentation methods which are used extensively in the academic community. Both methods were investigated in a longitudinal setting in [17] and showed similarly poor atrophy reproducibility rates. More recently, multi-atlas registration methods were introduced and showed high overlap with manual segmentations [22,23] and outperformed FSL-FIRST or FreeSurfer [24], but these can require long computation times.

An alternative for dedicated longitudinal segmentation methods and multi-atlas registration techniques is to use general purpose non-linear registration algorithms that map a (manual or automatic) segmentation of a baseline (BL) image to a follow up (FU) image. In the presence of MCI and AD, hippocampal atrophy rates have been measured previously using non-linear registration [2,25,26]. In these studies, atrophy rates measured on the basis of non-linear registration showed improved reliability compared to manually measured atrophy rates and significantly different hippocampal atrophy rates between healthy controls, MCI and AD were found. Hippocampal atrophy rates have been measured using symmetric non-linear registration in [27] and [28], which yielded higher sensitivity than a semi-automatic segmentation method [27]. Symmetric non-linear registration methods are not susceptible to directional registration bias and are therefore the preferred registration scheme for robust and sensitive longitudinal analysis. A symmetric registration procedure is for instance also used in FreeSurfer's longitudinal pipeline [21,29,30].

In this study, we compared hippocampal atrophy measurements using eight different methods: manual segmentation, automatic segmentation (FreeSurfer v6.0, FSL-FIRST v5.0.10, multi-atlas segmentation with joint label fusion (MALF, [23])), and four symmetric non-linear registration methods (Elastix [31,32], NiftyReg [33,34], Medical Image Registration ToolKit (MIRTK) [35] and the diffeomorphic registration method from the Advanced Normalization Tools (ANTs, http://stnava.github.io/ANTs/) referred to as Symmetric Normalization (SyN) [36]). We chose these registration methods, because they are publicly available and have been frequently used in the academic community. Furthermore, ANTs and MIRTK scored high in a study in which 14 non-linear registration methods were compared [37].

The aim of this study was to determine the accuracy of methods to measure subtle hippocampal volume change on the basis of plausibility (does the method show atrophy where biologically expected?), reproducibility (does the method provide the same atrophy rates for within-session rescans of the same subject?) and sensitivity (how many subjects are required to detect a significant group difference?). Therefore, we measured atrophy rates in different diagnostic groups (controls, MCI and AD) and performed an atrophy rate reproducibility analysis.

6.2. Materials and Methods

Datasets and image acquisition

In this study we determined hippocampal volume change on two different datasets from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Both datasets are described below. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

Training dataset

Training data consisted of a small subset (N=16) of the ADNI data base in which MCI subjects were recorded at BL and at FU two years later. The same dataset was previously used and described in more detail in [38]. Training data were used to tune parameter settings for each non-linear registration method.

At each time-point two sagittal 3D T1 weighted magnetization prepared rapid acquisition gradient echo sequence (MPRAGE) 1.5T MRI were acquired in a single session with just a few seconds apart from each other. In the remainder of this paper we use the term "A and B scans" for those back-to-back (BTB) scan pairs. As described in [38], an extreme-trait design was used, where participants were selected at the extremes of the 2-year longitudinal change distribution of hippocampal volume (eight participants with fast rates of atrophy and eight with slow rates of atrophy). The MRI acquisition is explained in more detail in [39]. "Gold standard" manual hippocampus segmentations were not available for this dataset and the magnitude of the rate of hippocampal volume change was not reported, but groups (slow and fast) differed in the rate of change ($p<0.001$) determined by the statistical parametric mapping (SPM) open source software package (http://www.fil.ion.ucl.ac.uk/spm/).

Validation dataset

This dataset is the same dataset used by [17] and [40]. Eighty subjects were collected from ADNI including 20 healthy controls (CTRL), 40 MCI and 20 AD subjects. BL and FU scans were obtained one year apart and similarly as for the training dataset BTB scans (A and B scans) were acquired with the same imaging sequence within a single session. The MRI acquisition is explained in [39].

Hippocampus segmentation

Manual hippocampus segmentation (only validation dataset)

For the ADNI dataset hippocampi were segmented at the VU University Medical Center (VUmc) Amsterdam) using the outlining protocol from [41], described in [17,41,42]. All BL MRI scans were reformatted perpendicularly to the long axis of the hippocampus with a slice thickness of 2mm, while the in-plane resolution was kept. The M12 scans were then rigidly registered to the BL scans. For hippocampus segmentation on the M12 scans, BL scans and hippocampus segmentations were shown next to the M12 scans. However, the observer (N=1) obtained A and B scans in a random order and did not know the diagnosis. The observer was a well-trained expert of the VUmc and used in-house developed software (Show_Images 3.7.1.0) for hippocampus segmentation.

FSL v.5.0.10 (both datasets)

FSL-FIRST hippocampus segmentation is described in detail in [43] and [20]. It uses shape and appearance models created from a set of manual hippocampus segmentations from the Center for Morphometric Analysis (CMA), Massachusetts General Hospital (MGH) Boston. The manual segmentations were converted to parameterized surface meshes using intensity values around the tissue border and from these a point distribution model was created. For hippocampus segmentation the MRI is registered to MNI152 standard space using a two-stage affine registration. Then, FSL-FIRST searches through linear combinations of shape variation modes to find the most probable shape by using the intensity values of the MRI. The hippocampal mesh is then converted to a voxel-wise segmentation using FAST [44]. For both datasets, we used the run_first_all command without pre-processing the images.

FreeSurfer v.6.0 (only validation dataset)

FreeSurfer subcortical segmentation is described in detail in [16]. FreeSurfer converts MRI scans to their own conformed $1mm^3$ 256^3 space, performs a bias-field correction, intensity normalization and skull-stripping for an atlas registration. Using

prior intensity and tissue class information, voxels are assigned to subcortical structures. We used FreeSurfer's longitudinal stream to determine hippocampal volumes [21,29,30], which includes an unbiased registration procedure and a label fusion technique. FreeSurfer's longitudinal pipeline was introduced in 2007 (FreeSurfer v4.0.0) and since then FreeSurfer has undergone several improvements. We used FreeSurfer v.6.0 released in 2017. In this study, FreeSurfer's default parameters have not been changed and therefore FreeSurfer was only applied on the validation dataset.

Multi-atlas label fusion segmentation (both datasets)

We used the multi-atlas joint label fusion (MALF) segmentation described in [23], implemented in the ANTs software [45]. Briefly, using non-linear registration a set of segmented atlases is deformed to a target image and all transformed atlases are combined to one label using a joint label fusion technique [23]. To speed up registration time we used a registration scheme provided by the ANTs software (antsRegistrationSyNQuick.sh [46,47]), which uses a mutual information metric. Twenty atlases were used as input for MALF (9CTRL, 8MCI, 3AD). The segmentation files and MRI were obtained from the Harmonized Protocol for Hippocampal Segmentation (HarP) project's website (http://www.hippocampal-protocol.net/). HarP is a standardized hippocampus outlining protocol in which hippocampal boundary definitions from different outlining protocols were merged [48–50].

Non-linear registration methods (both datasets)

Before applying non-linear registration, all FU scans were mapped to the corresponding BL scan using a rigid 6 degree of freedom (DOF) registration. These registrations were all performed using FLIRT from the FSL toolbox [51,52]. The resulting linear registration was used to initialize non-linear registrations. Furthermore, brain masks were used for source and target image. These masks were obtained by applying FSL's brain extraction tool (BET) on all subjects' MRI scans. The following four symmetric non-linear registration methods were used:

1. Elastix v4.801 [31,32]
2. NiftyReg v1.4.0 [33,34]
3. Medical image registration toolkit (MIRTK, compiled from the git development tree https://github.com/schuhschuh/MIRTK/tree/develop rev daf2b89, built on Dec 19 2017) [35]
4. Diffeomorphic registration method from ANTs v2.2.0, referred to as symmetric normalization (SyN) [36].

ANTs-SyN is a symmetric diffeomorphic mapping which guarantees topology preservation. Elastix, NiftyReg and MIRTK are based on free-form deformations (FFD) given by the parameters of a cubic B-spline function [53]. MIRTK's and NiftyReg's symmetric registration schemes are similar in approach using a symmetric energy formulation to ensure that the transformation is a diffeomorphism [34,35]. The symmetric registration approach for Elastix is described in [54]. For this approach FU scans were resampled to BL scans using rigid registration and spline interpolation. Then, images were transformed to an average space and the inverse transformation is approximated as described in [54]. This is different to diffeomorphic mapping, in which the inverse is guaranteed to exist. Commands and parameters for each registration method are presented in the supplementary files.

Surface reconstruction and mesh deformation

For the validation data set, hippocampi were manually segmented in cropped MR images, reformatted along the long hippocampal axis. We performed our analysis using the native MR images and therefore hippocampus segmentation needed to be mapped back from "segmentation space" to native MRI scan space. To avoid interpolation errors in this procedure, we converted manual segmentations to meshes using the marching cube algorithm [55] and applied linear transformation parameters directly on the meshes. FSL-FIRST, FreeSurfer and MALF were performed on the native MRIs, but to be consistent we also converted the hippocampal segmentations obtained from these methods to meshes and computed hippocampal volumes from these meshes.

In the validation set, nonlinearly transformed manual hippocampus meshes were used to determine longitudinal volume changes and thus atrophy rates. For the training dataset manual hippocampus segmentations were not available, therefore FSL-FIRST hippocampus segmentations from the BL scans were used to provide the baseline mesh.

Analysis

Hippocampal volumes were obtained by summing signed tetrahedrons created for each triangle in the mesh as described in [56]. Hippocampal atrophy rates were expressed as longitudinal percentage volume change (PVC) defined by:

$$PVC(V_{BL}, V_{FU}) = \frac{V_{FU} - V_{BL}}{V_{BL}} * 100 \quad (1)$$

with V_{BL} and V_{FU} being the volume of the structure from the BL and FU scans, respectively. We calculated PVC for the left and right hippocampus separately, but

we averaged them to obtain the hippocampal atrophy rates for our analysis. The statistical analysis was performed with IBM SPSS Statistics for Windows v. 22 Armonk, NY: IBM Corp. We analyzed atrophy differences between diagnostic groups and atrophy reproducibility using A and B scans measured atrophy rates. For our statistical analysis we reported and excluded all subjects in which hippocampal PVC was larger than +25% or smaller than -25%, because hippocampal atrophy ±25% is an indicator for poor segmentation or registration and would reduce reliability of the analysis.

Training dataset

For the training dataset "gold standard" manual segmentations were not available. Instead, FSL-FIRST BL segmentations were mapped to FU scans with the registration methods. We report two-year PVC in boxplots to graphically illustrate PVC differences between methods and diagnostic groups. In correspondence with the developers of each registration method we tested different registration parameters until we obtained approximately similar PVC in groups for each method, knowing that the 'slow' group should have less atrophy than the 'fast' group [38]. Using PVC measurements from the A and B scans, with repeated measures ANOVA we determined whether there was a significant difference between the 'slow' and the 'fast' group (significance level α=0.05), which indicated if registration methods detected similar atrophy rates.

We used an average distance measure to quantify the agreement between different methods to determine atrophy rates:

$$D_{Ave,Method} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(PVC_{Method1,i} - PVC_{Method2,i}\right)^2} \quad (2)$$

in which *n* is the number of measured PVC values in the pooled A and B scans. These "distances", which can be interpreted as the root mean squared difference between two methods, were plotted in a color-coded distance matrix.

To be reproducible, the PVCs measured for the A scans should be approximately the same as for the B scans. To study and quantify reproducibility we plotted PVC values from the A scans against PVC values from the B scans and used again the average distance for quantification:

$$D_{Ave,Scans} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(PVC_{AScan,i} - PVC_{BScan,i}\right)^2} \quad (3)$$

A commonly used method to determine volume change of a non-linear deformation field is integration of the local Jacobian determinants of the field. For comparison, we determined volume changes using Jacobian integration, which should give very similar results to our mesh-based approach. Using FSL-FIRST BL segmentations, this was also tested with the training dataset. Finally, we investigated if PVC values depend on the specific choice of the BL segmentation by replacing the FSL-FIRST by the MALF BL segmentation.

Validation dataset

For the validation dataset we used the registration parameter settings determined with the training dataset and performed a similar analysis as for the training dataset. We report one-year PVC in mean, standard deviations (SD) and boxplots for all methods and diagnostic groups. Differences between CTRL, MCI and AD were assessed with repeated measures ANOVA using Tukey's honestly significant difference (HSD) post hoc test. Furthermore, we performed a power analysis and estimated the sample size needed to detect a reduction in atrophy relative to the CTRL or MCI atrophy using:

$$N = \frac{2(z_{\alpha/2} + z_{1-\beta})^2}{(Effect\ Size)^2} = \frac{2(z_{\alpha/2} + z_{1-\beta})^2}{\left(\frac{\mu_2 - \mu_1}{\sigma_{pooled}}\right)^2} \quad (4)$$

where μ_i are the means of the two groups that are compared (CTRL, MCI or AD). Statistical power (1-β) was set to 80%, significance level α=0.05 using a two-sided alternative hypothesis. Z_t is the t-th quantile of the normal distribution, i.e. $z_{\alpha/2}$=1.96 and $z_{1-\beta}$=0.8416. Finally, we also presented the distance matrix ($D_{Ave,\ Method}$) and atrophy reproducibility measure ($D_{Ave,\ Scans}$) using A and B scans' determined PVC values.

6.3. Results

Training dataset

Parameter settings and descriptions of parameters we optimized for each registration method can be requested from the author of this book. PVC results obtained with Elastix, NiftyReg, ANTs or MIRTK of the training data set are shown in figure 1, where results have been separated in 'slow' and 'fast' groups for ease of comparison. Standard deviations of PVC for pooled A and B measured hippocampal atrophy are presented in table 1. Mean and SD PVC for the A and B scans separately can be found in the supplementary table 1. Despite the low sample size (N=16 for each box), median and interquartile ranges overlap well for all registration methods. The 5% volume loss in two years found by all methods is in agreement with the annual hippocampal atrophy of approximately 2.5% found in a meta-analysis by [57]. All methods showed a similar PVC trend in the 'slow' and 'fast' group as determined by repeated measures ANOVA (table 1). ANTs showed the smallest group differences, while also having the lowest SD.

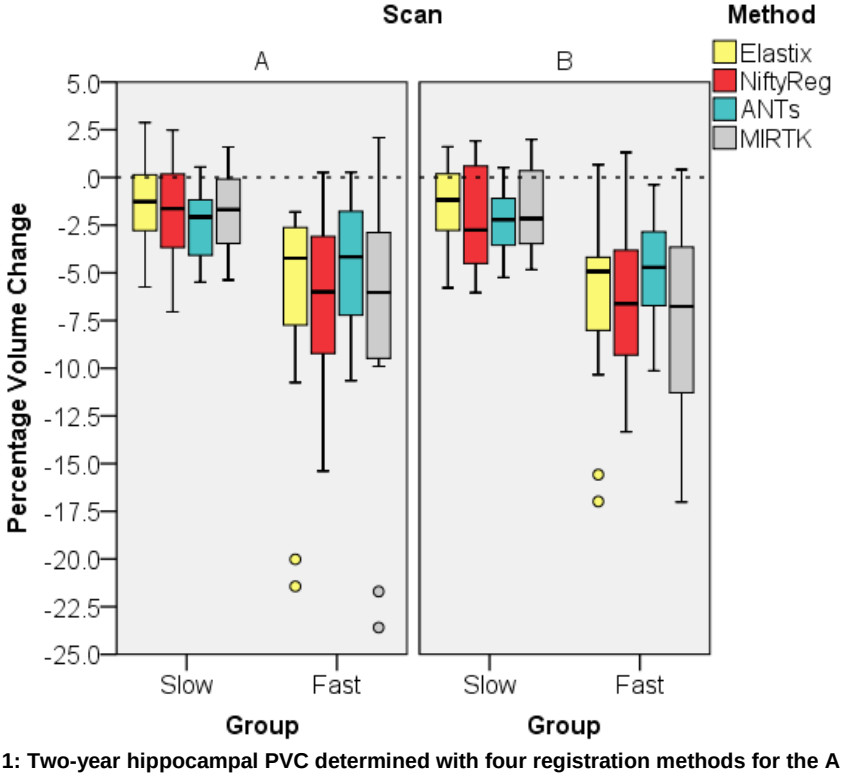


Figure 1: Two-year hippocampal PVC determined with four registration methods for the A and B longitudinal scans. FSL-FIRST BL segmentation was used to determine PVC with the registration methods. Both groups are subjects with MCI, one group with 'slow' and one with 'fast' progressing atrophy. The small circles are outliers defined by the SPSS software.

Table 1: Mean (μ) and standard deviation (σ) of two-year hippocampal PVC determined with four registration methods for pooled A and B longitudinal scans. Both groups are subjects with MCI, one group with 'slow' and one with 'fast' progressing atrophy. Using repeated measures ANOVA significance between the 'slow' and the 'fast' group was determined. The F-statistic is the ratio of the between group variance and the within group variance and the numbers in the brackets are the degrees of freedom.

Method	Slow		Fast			
	μ	σ	μ	σ	F(1,30)	p-value
Elastix	-1.36	2.153	-6.33	5.448	11.82	0.002
NiftyReg	-1.96	2.781	-6.24	4.154	12.91	0.001
ANTs	-2.39	1.721	-4.66	3.050	7.04	0.013
MIRTK	-1.75	2.218	-7.32	6.124	12.26	0.001

In figure 2 we plotted hippocampal PVC measured with the A scans against hippocampal PVC measured with the B scans for the training dataset. The average distance (D_{Ave}) between A and B scans measured atrophy rates is shown in table 2. All non-linear registration methods show a similar reproducibility trend with D_{Ave} ranging from 0.86% to 3.08%. ANTs reproduced data best with D_{Ave}=0.86%.

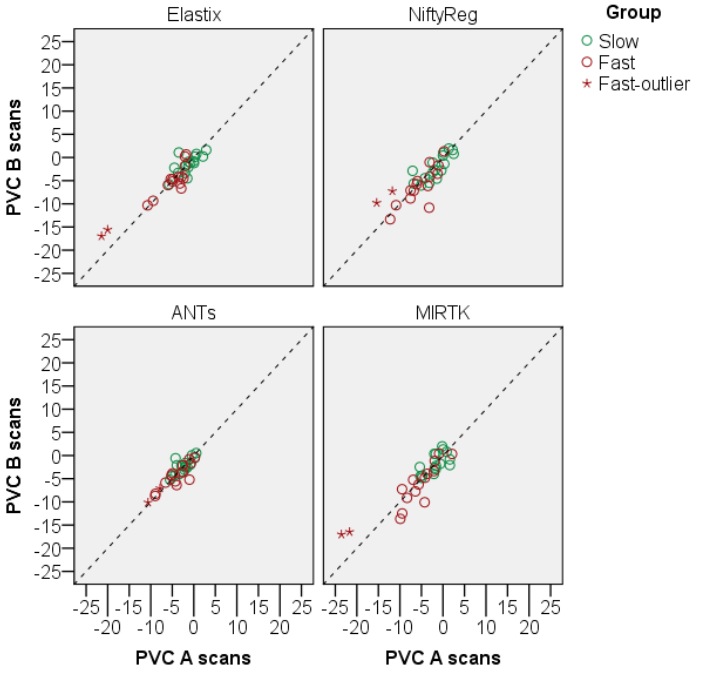


Figure 2: For all four registration methods, hippocampal PVC measured with the A scans is plotted against hippocampal PVC measured with the B scans. The dashed line is the identity line. For Elastix and MIRTK there was one subject in the 'fast' group with larger PVC compared to the other cases (two stars correspond to left and right hippocampal PVC). This subject was specifically highlighted (Fast-outlier) to observe if the other methods measured similar high hippocampal atrophy.

Table 2: Calculated average distance defined in (3) by using mean and standard deviation of pooled ('slow' and 'fast') hippocampal PVC. All units are %.

Method	A Scans		B Scans		
	μ	σ	μ	σ	D_{Ave}
Elastix	-3.88	5.269	-3.81	4.391	1.93
NiftyReg	-3.98	4.373	-4.22	3.913	2.88
ANTs	-3.48	2.832	-3.57	2.629	0.86
MIRTK	-4.52	5.780	-4.55	5.005	3.08

μ Mean, *σ* standard deviation, D_{Ave} Average distance as defined in (3)

Consistency checks with the training dataset

For all ANTs deformation fields, we computed the Jacobian determinants and determined hippocampal PVC by integrating local Jacobian determinants within the region of FSL-FIRST BL hippocampus segmentations. These PVC values were plotted against the PVC values obtained by deforming FSL-FIRST hippocampus meshes (supplementary figure 1). PVC measured with Jacobian integration was nearly identical to PVC measured by deforming the meshes (linearly fitted line equation: y=-0.03+1.0*x and a D_{Ave} of 0.0027%), illustrating the consistency of the mesh and volume-based methods. BL hippocampus segmentations were also computed with MALF and converted to meshes. Using the ANTs deformation fields, we deformed MALFs' BL hippocampal meshes and plotted these hippocampal PVCs against PVCs obtained by deforming FSL-FIRST BL hippocampal meshes (supplementary figure 2). Both determinations were highly correlated, with a R^2 obtained from a linearly fitted line of 0.915 and a D_{Ave} of 0.31%. To indicate how similar FSL-FIRST and MALF BL segmentations were, we calculated the Jaccard index (Jacc=(A∪B)/(A∩B)) between both BL segmentations. The mean Jaccard indices for the 'slow' and the 'fast' groups were $Jacc_{slow}$= 0.70(±0.056) and $Jacc_{fast}$=0.73(±0.029) respectively.

Validation dataset

For the validation dataset, FSL-FIRST's segmentation failed in 35 subjects out of 80, which was surprising because in the training dataset FSL-FIRST did not fail. To include these cases, we used FSL's brain extraction tool (BET, [58]) to extract the brains of all subjects and ran FSL-FIRST again. This second FSL-FIRST run still failed in six subjects. These cases were different from the first FSL-FIRST run. To address this incongruity, we included the segmentation files of these 6 subjects from the first FSL-FIRST run for our analysis. In nine subjects, FSL-FIRST exceeded our set PVC threshold ±25% in which four subjects were from the A scans (3 CTRL, 1

AD) and five from the B scans (2 CTRL, 1 MCI, 2 AD). For one subjects' A and B scans, MALF and Elastix also exceeded the ±25% threshold.

Results of PVCs computations are presented in figure 4, separated for three diagnostic groups and separated for A and B scans. Corresponding mean and SDs for pooled A and B scans' PVC are given in table 3 and mean PVC for A and B scans separated can be found in the supplementary table 3. Table 3 also includes results of the repeated measures ANOVA and post hoc analysis, in which PVCs are compared between diagnostic groups for each method. Manual, FSL-FIRST and FreeSurfer showed larger PVC variability compared to the other methods, whereas FSL-FIRST had largest standard deviations (table 3). Except for MALF, for all other methods a statistically significant difference was detected between groups using the repeated measures ANOVA. The post hoc analysis for MALF was therefore irrelevant, but for completeness results are presented in table 3. Tukey post hoc analysis revealed that for all methods (MALF excluded) there was no significant difference between the CTRL and the MCI groups. All methods showed statistically significant higher atrophy rates in the AD group compared to the CTRL group. The distance correlation matrix used to visualize overall differences between methods is shown is figure 5 and the corresponding numerical average distances in supplementary table 4. These distances show that manually determined atrophy rates and atrophy rates determined with FSL-FIRST are generally very different from the other methods. Results from our sample size calculation defined in equation (3) are presented in table 4. Because we grouped A and B scans and left and right hippocampi together, means and the pooled standard deviations for the sample size calculation were obtained from 80 CTRL, 160 MCI and 80 AD samples (for FSL-FIRST, Elastix and MALF outliers were discarded). Between the CTRL and MCI group FreeSurfer and MIRTK clearly had lowest estimated sample sizes ($N_{CTRL-MCI}$) with approximately 30% to 40% less samples than the next method with low estimated sample sizes (ANTs, $N_{CTRL-MCI} = 162$). For $N_{CTRL-AD}$ FreeSurfer, Elastix, NiftyReg and MIRTK showed lowest estimated sample sizes with approximately 55% to 70% less samples than ANTs estimated sample sizes ($N_{CTRL-AD} = 37$) and 79% to 86% less samples than used in this study.

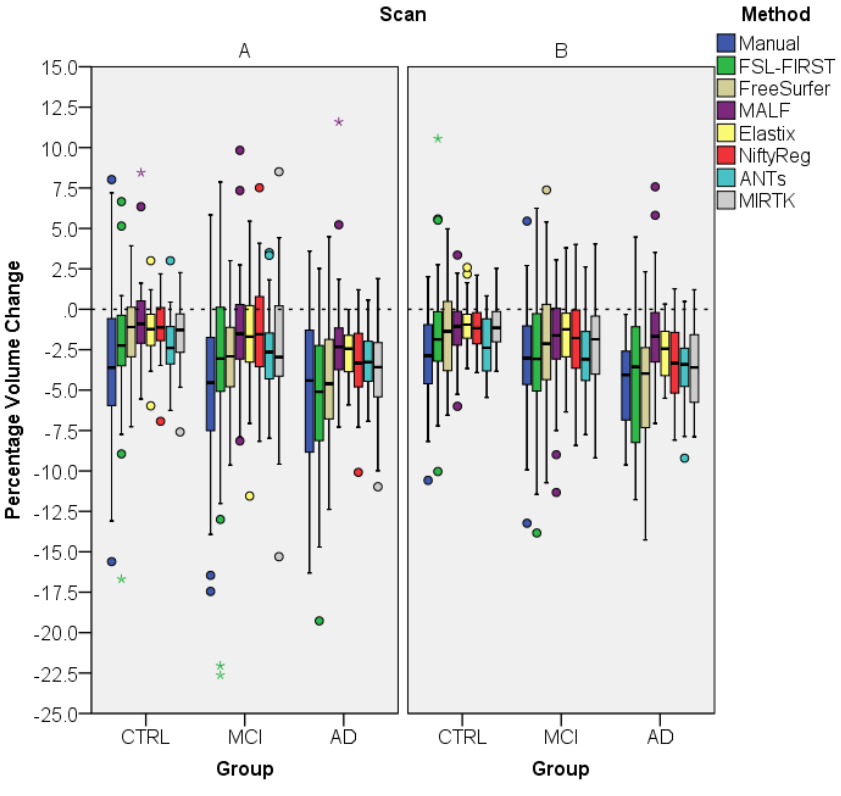


Figure 3: One-year hippocampal PVC measured with eight methods for the A and B longitudinal scans. PVC was determined separately for the CTRL, MCI and AD groups. The small circle and the star sign are outliers defined by the SPSS software, with the star sign being a "far out" outlier.

Table 3: For all methods mean (μ), standard deviations (σ) of one-year PVC measured on pooled A and B scans are shown. Using repeated measures ANOVA overall group significance was determined and with Tukey's post-hoc analysis in between group differences were analysed. The F-statistic is the ratio of the between group variance and the within group variance and the numbers in the brackets are the degrees of freedom.

							Overall group significance		post-hoc analysis	
Meth-od	μ_{CTRL}	σ_{CTRL}	μ_{MCI}	σ_{MCI}	μ_{AD}	σ_{AD}	F[a]	p-value	$p_{CTRL-MCI}$	$p_{CTRL-AD}$
Man.	-3.18	3.838	-3.89	3.813	-4.81	3.833	3.15	0.045	0.416	0.036
FSL	-1.88	3.785	-3.00	4.592	-4.94	4.506	5.59	0.003	0.681	0.005
FS	-1.42	2.650	-2.53	3.335	-4.47	3.598	10.71	**	0.137	**
MALF	-0.86	2.121	-1.60	2.737	-1.88	3.145	2.04	0.134	0.243	0.138
Elastix	-1.08	1.480	-1.60	2.359	-2.70	1.665	7.94	0.001	0.262	*
NR	-1.06	1.516	-1.68	2.719	-3.25	2.474	11.02	**	0.301	**
ANTs	-2.27	1.798	-2.90	2.216	-3.49	1.941	4.01	0.02	0.215	0.014
MIRTK	-1.30	1.752	-2.32	3.133	-3.90	2.599	12.49	***	0.068	***

Man. Manual, *FSL* FSL-FIRST, *FS* FreeSurfer, *NR* NiftyReg

[a]F(2,157), For FSL-FIRST, Elastix and MALF outliers were removed for the analysis resulting in $F_{FIRST}(2,145)$, $F_{Elastix}(2,155)$, $F_{MALF}(2,155)$

*<0.0005, **<0.00005, ***<0.000005

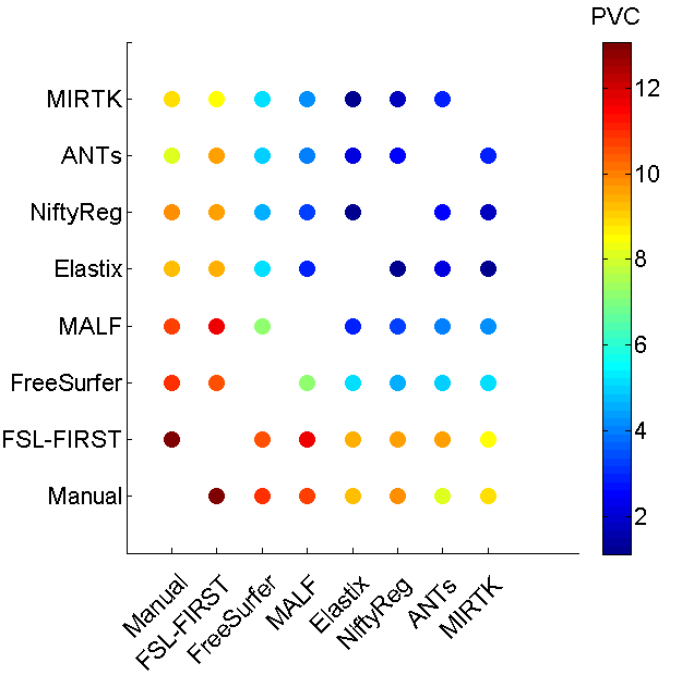


Figure 4: For one-year hippocampal PVC of the validation dataset, average "distance" atrophy rate agreement was calculated using (2) and plotted in a colour coded distance matrix. Registration based method outperform the other methods with lower distance in PVC (in %).

Table 4: Estimated Cohen's d effect size and sample size (N) between groups for all methods defined in equation (3). Lowest sample sizes are highlighted with bold font.

	Cohen's d effect size		Sample size	
Method	$d_{CTRL\text{-}MCI}$	$d_{CTRL\text{-}ADI}$	$N_{CTRL\text{-}MCI}$	$N_{CTRL\text{-}AD}$
Manual	0.19	0.42	452	87
FSL-FIRST	0.27	0.74	222	29
FreeSurfer	0.37	0.96	**115**	**17**
MALF	0.30	0.38	174	108
Elastix	0.26	1.03	226	**15**
NiftyReg	0.29	1.07	193	**14**
ANTs	0.31	0.65	162	37
MIRTK	0.40	1.17	**97**	**11**

In figure 5, PVC determined using the A scans was plotted against those based on the B scans. Table 5 presents corresponding D_{Ave}. ANTs had lowest D_{Ave} (1.06%) followed by Elastix (1.78%) and NiftyReg (2.11%). Manually determined atrophy rates showed poorest reproducibility, D_{Ave}=12.39%. Using the same hippocampus segmentations and MRI scans but a different statistical analysis, Mulder and

colleagues observed similarly poor reproducibility based on manual segmentations [17].

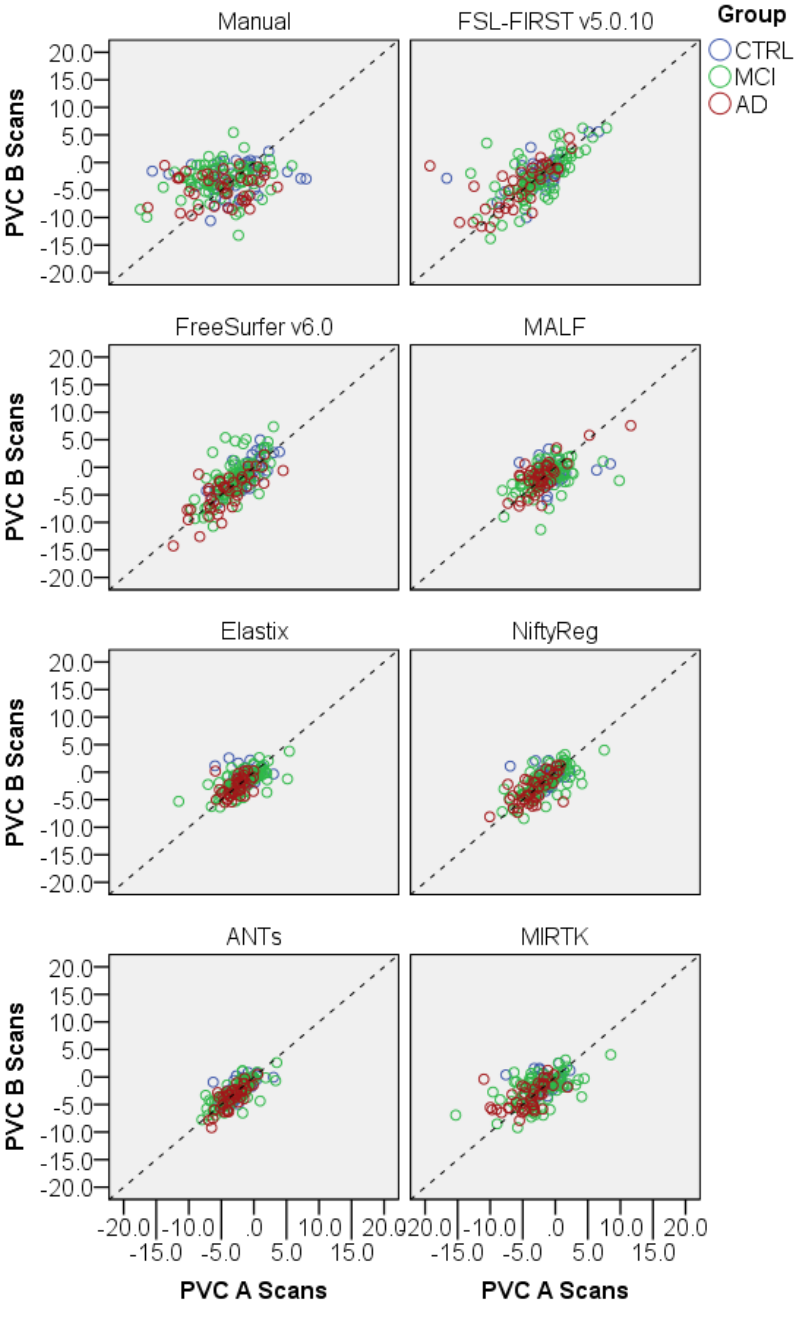


Figure 5: For all eight methods, one-year hippocampal PVC determined with the longitudinal A scans is plotted against hippocampal PVC determined with the longitudinal B scans. The dashed line is the identity line.

Table 5: Calculated average distance (D_{Ave}) defined in (3) by using mean (μ) and standard deviation (σ) of pooled (CTRL, MCI and AD) hippocampal PVC. All units are %.

	A Scans		B Scans		
Method	μ	σ	μ	σ	D_{Ave}
Manual	-4.40	4.574	-3.49	2.914	12.39
FSL-FIRST	-3.33	4.398	-2.93	3.962	6.15
FreeSurfer	-2.92	3.150	-2.56	3.672	3.26
MALF	-1.34	2.908	-1.62	2.524	3.60
Elastix	-1.79	2.243	-1.64	1.898	**1.78**
NiftyReg	-1.89	2.647	-1.95	2.428	**2.11**
ANTs	-2.84	2.097	-2.93	2.088	**1.06**
MIRTK	-2.62	3.163	-2.30	2.525	3.76

Consistency checks with the validation dataset

PVC values for the non-linear registration methods were determined by deforming manual BL segmentations (PVC_{Manual_ANTs}). To assess dependency on BL segmentation we also used BL segmentations from MALF and deformed these with ANTs deformation fields to obtain new PVC values (PVC_{MALF_ANTs}). PVC_{Manual_ANTs} was plotted against PVC_{MALF_ANTs} (supplementary figure 3). The R^2 was 0.837 and the average distance was 0.46%. For the outlier AD cases (red circles in supplementary figure 3), which are further away from the identity line, we inspected the BL segmentations. MALF BL segmentations slightly overestimated the hippocampal region and also outlined a small part of the cerebral fluid next to the hippocampal boundaries. Removing these cases yielded a R^2 of 0.927 and a D_{Ave} of 0.26%, illustrating that BL segmentations are interchangeable to compute atrophy rates.

6.4. Discussion

In this study we compared hippocampal atrophy rates for eight different methods and evaluated their differences based on plausibility over diagnostic groups, their estimated errors and their reproducibility over back to back scans. All methods showed largest PVCs for AD, followed by MCI and controls. The non-linear registration-based FreeSurfer and MIRTK showed highest sensitivity in terms of predicted sample size, and ANTs, Elastix and NiftyReg showed largest reproducibility. The segmentation-based technique FSL-FIRST and manual segmentations scored lowest in these aspects.

Our atrophy rates reproducibility analysis is in agreement with [26] who used directional non-linear registration (biased) to measure hippocampal atrophy rates and showed that hippocampal atrophy measured with non-linear registration is more

reproducible than manually measured atrophy. We investigated this with multiple symmetric non-linear registration (unbiased) methods and automatic segmentation methods and we also conclude that hippocampal atrophy rates should preferably be computed with such methods. Manually measured atrophy rates showed large variation and atrophy rates were poorly reproduced. Therefore, large sample sizes are needed when using manual segmentation to detect differences between diagnostic groups.

Of all automatic segmentation methods, FreeSurfer performed best. FSL-FIRST had several failed segmentations and MALF was not able to detect hippocampal atrophy rate differences between diagnostic groups. We used FreeSurfer v6.0 and compared to other studies in which v5.3 was used [17,59], FreeSurfer clearly improved and atrophy rates were well reproduced in comparison with the other methods (4^{th} lowest of all methods; D_{Ave}=3.26%). Comparing only the non-linear registration methods, MIRTK had highest effect sizes, but reproducibility rates in the validation dataset were the worst (D_{Ave}=3.76%). Compared to the other registration methods, ANTs determined on average almost 1% more hippocampal atrophy in the CTRL and MCI group (table 3), which explains the higher required sample size of $N_{CTLR-AD}$=37, despite being most reproducible (D_{Ave}=1.06%). Reproducing hippocampal atrophy rates is an important quality measurement and therefore we find that ANTs performed overall as the best among the non-linear registration methods.

Somewhat surprisingly given the previous good performance of Multiple-Atlas Propagation and Segmentation with Hippocampal Boundary Shift Integral (MAPS-HBSI) in [59] is the poor performance of the multi-atlas method MALF. This discrepancy could be due to the fact that the registration used in our implementation of MALF was fast but not maximally optimized, and to the fact that MAPS-HBSI included a boundary shift integral calculation to compute volume change, whereas in this paper only the BL and FU cross-sectional segmentations were used to calculate volume change. Furthermore, different label fusion techniques might improve results for multi-atlas segmentation methods as for example in [22,60].

Because differences between best performing methods are small, our results suggest using either ANTs, Elastix or NiftyReg to determine hippocampal atrophy rates. In the validation dataset we deformed manual BL segmentations with ANTs to determine hippocampal atrophy rates. But we also showed that when changing the BL segmentation from manual to MALF in the validation dataset, or from FSL-FIRST to MALF using the training dataset, atrophy rates almost stayed the same (supplementary figure 2 and 3), showing that the accuracy of the BL segmentations is not that important. Therefore, determining hippocampal atrophy rates with non-

linear registration methods can be completely automated. Such a procedure is similar to FreeSurfer's longitudinal pipeline in which images are first pre-processed (resampling, skull stripping, intensity normalization), then an unbiased atlas registration procedure is used and finally labels from different time-points are fused [21,29,30].

From the color-coded distance matrix, we could see that the registration-based methods, including MALF and FreeSurfer, agreed best with each other. Within those, Elastix, MIRTK and NiftyReg had smallest average distances ranging from 1.1% to 1.7%, possibly reflecting that these are based on the same underlying principle (FFD). Comparing ANTs to these methods yielded average distances ranging from 2.2% to 2.9%.

Manual segmentation was performed on resliced MRI with 2mm slices. Supplementary figure 3 and the corresponding analysis illustrated that this reslicing did not have a large influence in our analysis, because MALF was performed on native MR image resolution. The low reproducibility of manually measured atrophy rates was to some extent surprising, because on the contrary hippocampus segmentations of the A and B scans showed high outline reproducibility (Jaccard~0.8 equivalent to a Dice~0.89) as found in a previous study using the same manual segmentations [40]. Apparently, such small differences in outlines with uncorrelated errors, results in large uncertainties in volume, making manual segmentation a poor choice for measuring atrophy rates.

Even in elderly healthy aging subjects, hippocampal atrophy is to be expected [61]. However, in all groups (CTRL, MCI and AD) and for all methods we also measured unexpected hippocampal volume increase for a few subjects (figure 3). Unexpected hippocampal volume increase for a few subjects can also be observed in other studies in which hippocampal atrophy rates were measured [17,26,27,59]. We did not further investigate these subjects with positive PVCs, but most probably it is due to noise in the images.

Except of brain extraction to reduce image registration time, we did not include any other image pre-processing steps. Intensity normalization might improve results for all non-linear registration methods. In FreeSurfer's pipeline intensity normalization is already included. Further improvements could be achieved by removing directional registration bias during the global/rigid registration step and by applying a symmetric global/rigid registration as discussed and investigated in [27,62].

Clinical applicability

Our findings have potential clinical applicability in e.g. radiotherapy, where hippocampal volume loss has recently become a point of concern when patients are given prophylactic cranial irradiation (PCI). In a recent study, in which 22 patients with SCLC received PCI, short-term longitudinal brain changes were reported using statistical parametric mapping (SPM) software package (open source, http://www.fil.ion.ucl.ac.uk/spm/) [63]. There, hippocampal volume decrease was associated with PCI, but the magnitude of hippocampal volume loss was not reported. In a small case series study (n=9) patients with melanoma brain metastasis showed a mean hippocampal volume loss of 7.81% due to whole brain irradiation therapy after 6 month [64]. Additionally, in patients with primary brain tumors dose was significantly correlated with hippocampal volume loss one year after radiotherapy, while high doses of >40Gy showed a mean hippocampal volume loss of 5.8% and low doses of <10Gy a mean volume loss of 1.2% [65]. These reported hippocampal volume losses exceed one-year hippocampal atrophy in AD of 2.5-3.5% as determined in this study. Hence, we expect that radiation induced annual hippocampal volume loss should be detectable with for instance ANTs, Elastix or NiftyReg even with a fairly small sample size (table 4).

Each registration method might be improved by adjusting registration parameters. We used the training dataset and changed parameters until we measured similar group-based atrophy rates for each method. A larger investigation to obtain optimal registration parameters would be preferable but also challenging because a ground truth does not exist, and manual segmentations are not reliable enough.

6.5. Conclusion

Hippocampal volume loss measured on longitudinal T1-weighted MRI should preferably be computed with symmetric non-linear registration methods such as ANTs, Elastix or NiftyReg, because these were least susceptible to noise.

Supplementary data

Supplementary Table 1: Mean (μ) and standard deviation (σ) of two-year hippocampal PVC determined with four registration methods for the A and B longitudinal scans. Both groups are subjects with MCI, one group with 'slow' and one with 'fast' progressing atrophy.

Scan	Method	Slow		Fast	
		μ	σ	μ	σ
A	Elastix	-1.27	2.311	-6.49	6.125
	NiftyReg	-1.76	2.934	-6.20	4.520
	ANTs	-2.47	1.835	-4.49	3.322
	MIRTK	-1.76	2.324	-7.28	6.886
B	Elastix	-1.46	2.054	-6.17	4.874
	NiftyReg	-2.15	2.702	-6.29	3.902
	ANTs	-2.30	1.654	-4.83	2.851
	MIRTK	-1.74	2.183	-7.37	5.486

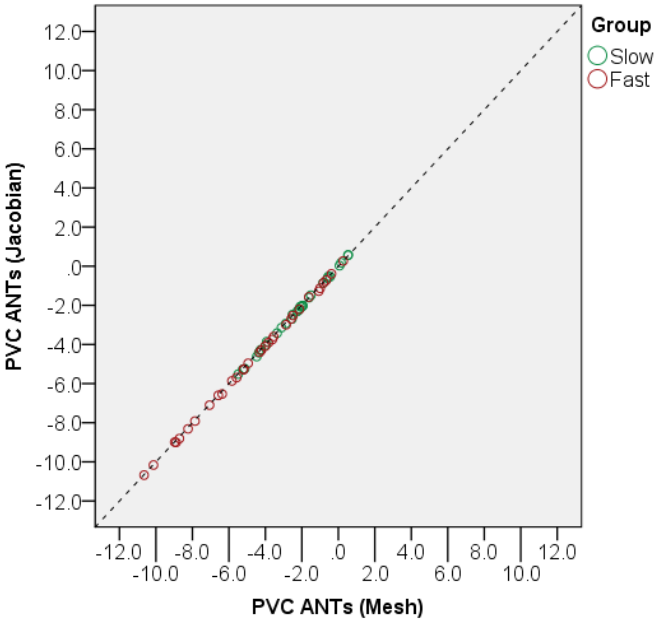
Supplementary Table 2: Mean (μ) and standard deviation (σ) of two-year hippocampal PVC determined with eight methods for the A and B longitudinal scans.

Scan	Method	μ_{CTRL}	σ_{CTRL}	μ_{MCI}	σ_{MCI}	μ_{AD}	σ_{AD}
A	Manual	-3.26	4.676	-4.66	4.391	-5.01	4.745
	FSL-FIRST	-2.36	3.891	-3.24	5.068	-5.51	4.781
	FreeSurfer	-1.50	2.529	-2.96	2.877	-4.28	3.647
	MALF	-0.52	2.360	-1.41	2.952	-2.03	3.192
	Elastix	-1.26	1.605	-1.63	2.660	-2.75	1.605
	NiftyReg	-1.09	1.587	-1.59	2.893	-3.30	2.495
	ANTs	-2.30	1.817	-2.87	2.324	-3.31	1.777
	MIRTK	-1.55	1.946	-2.37	3.523	-4.18	2.842
B	Manual	-3.10	2.822	-3.13	2.964	-4.60	2.675
	FSL-FIRST	-1.42	3.677	-2.74	4.064	-4.34	4.179
	FreeSurfer	-1.34	2.796	-2.11	3.708	-4.66	3.584
	MALF	-1.20	1.821	-1.78	2.509	-1.73	3.134
	Elastix	-0.87	1.338	-1.57	2.031	-2.65	1.742
	NiftyReg	-1.02	1.461	-1.78	2.547	-3.21	2.485
	ANTs	-2.24	1.801	-2.92	2.118	-3.66	2.099
	MIRTK	-1.05	1.518	-2.27	2.710	-3.62	2.332

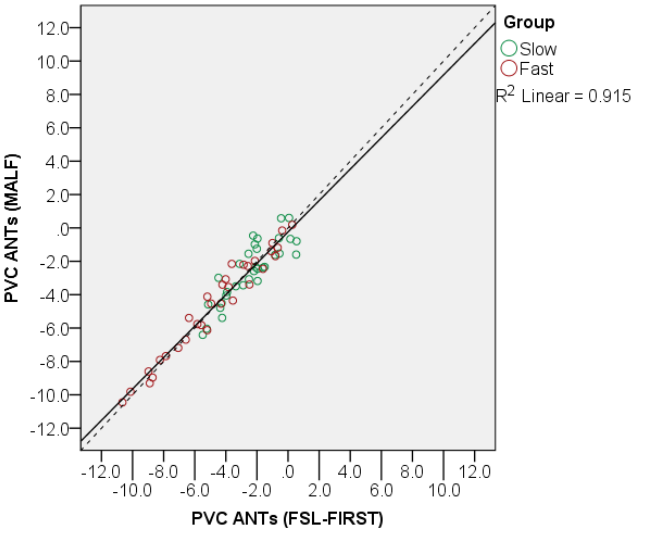
CTRL Healthy controls, *MCI* Mild cognitive impairment, *AD* Alzheimer's disease

Supplementary Table 3: Average "distance" atrophy rates in % between methods using equation defined in (2).

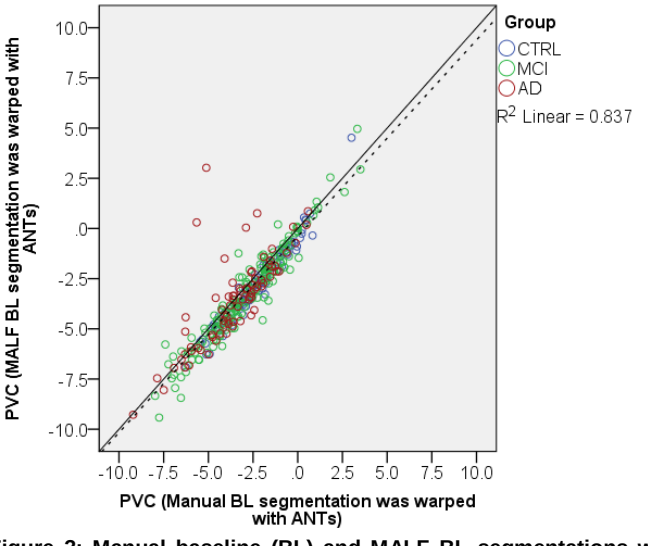
Method	Manual	FSL-FIRST	FS	MALF	Elastix	NiftyReg	ANTs	MIRTK
Manual	0.0	13.1	10.9	10.7	9.2	9.8	8.1	8.8
FSL-FIRST	13.1	0.0	10.5	11.6	9.3	9.6	9.6	8.6
FreeSurfer	10.9	10.5	0.0	7.2	5.1	4.6	5.0	5.1
MALF	10.7	11.6	7.2	0.0	2.9	3.2	4.0	4.2
Elastix	9.2	9.3	5.1	2.9	0.0	1.3	2.2	1.1
NiftyReg	9.8	9.6	4.6	3.2	1.3	0.0	2.6	1.7
ANTs	8.1	9.6	5.0	4.0	2.2	2.6	0.0	2.9
MIRTK	8.8	8.6	5.1	4.2	1.1	1.7	2.9	0.0



Supplementary Figure 1: For the training dataset FSL-FIRST baseline (BL) hippocampus segmentations were used to measure two-year PVC in two ways: 1) The deformation field obtained from ANTs non-linear registration was applied on BL hippocampal meshes and percentage volume change (PVC) was calculated between the BL and the deformed mesh. 2) Local Jacobian determinants of the deformation field was integrated in the BL hippocampus segmentation area to measure PVC. Both PVC measurements were plotted against each other. The dashed line is the identity line.



Supplementary Figure 2: For the training dataset FSL-FIRST and MALF baseline (BL) hippocampus segmentations were available for the same subjects. These segmentations were converted to meshes and mapped to follow up image using deformation fields obtained from the ANTs registrations. Percentage volume change (PVC) was calculated and ANTs PVC using FSL-FIRST BL segmentation was plotted against ANTs PVC using MALF BL segmentation. The dashed line is the identity line and the solid line a linearly fitted line.



Supplementary Figure 3: Manual baseline (BL) and MALF BL segmentations were warped to follow up images and percentage volume change (PVC) was calculated and plotted against each other. In this figure one subject was removed due to a failed segmentation (MALF, described in the result section). The solid line is the identity line and the dashed line is the linearly fitted line. The average distance was 0.46%. For the outlier AD cases (red circles) which are further away from the identity line, we inspected the BL segmentations. MALF BL segmentations slightly overestimated the hippocampal region and also outlined a small part of the cerebral fluid next to the hippocampal boundaries. Removing these yielded a R^2 of 0.927 and a D_{Ave} of 0.26%.

References

[1] Apostolova LG, Thomson paul M. Mapping progressive brain structural changes in early Alzheimer´s Disease and mild cognitive impairment. Neuropsychologia 2009;46:1597–612.

[2] Henneman WJP, Sluimer JD, Barnes J, Van Der Flier WM, Sluimer IC, Fox NC, et al. Hippocampal atrophy rates in Alzheimer disease: Added value over whole brain volume measures. Neurology 2009;72:999–1007.

[3] Likeman M, Anderson VM, Stevens JM, Waldman AD, Godbolt AK, Frost C, et al. Visual assessment of atrophy on magnetic resonance imaging in the diagnosis of pathologically confirmed young-onset dementias. Arch Neurol 2005;62:1410–5.

[4] Mouiha A, Duchesne S. Hippocampal atrophy rates in Alzheimer's disease: automated segmentation variability analysis. Neurosci Lett 2011;495:6–10.

[5] Sabuncu MR. The Dynamics of Cortical and Hippocampal Atrophy in Alzheimer Disease. Arch Neurol 2011;68:1040.

[6] Schuff N, Woerner N, Boreta L, Kornfield T, Shaw LM, Trojanowski JQ, et al. MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. Brain 2009;132:1067–77.

[7] Cash DM, Rohrer JD, Ryan NS, Ourselin S, Fox NC. Imaging endpoints for clinical trials in Alzheimer's disease. Alzheimers Res Ther 2014;6:87.

[8] Ferrer I, Serrano T, Alcantara S, Tortosa A, Graus F. X-ray-induced cell death in the developing hippocampal complex involves neurons and requires protein synthesis. J Neuropathol Exp Neurol 1993;52:370–8.

[9] Raber J, Rola R, LeFevour A, Morhardt D, Curley J, Mizumatsu S, et al. Radiation-induced cognitive impairments are associated with changes in indicators of hippocampal neurogenesis. Radiat Res 2004;162:39–47.

[10] Nagai R, Tsunoda S, Hori Y, Asada H. Selective vulnerability to radiation in the hippocampal dentate granule cells. Surg Neurol 2000;53:503-6; discussion 506-7.

[11] Madsen TM, Kristjansen PEG, Bolwig TG, Wörtwein G. Arrested neuronal proliferation and impaired hippocampal function following fractionated brain irradiation in the adult rat. Neuroscience 2003;119:635–42.

[12] Mizumatsu S, Monje ML, Morhardt DR, Rola R, Palmer TD, Fike JR. Extreme sensitivity of adult neurogenesis to low doses of X-irradiation. Cancer Res 2003;63:4021–7.

[13] Makale MT, McDonald CR, Hattangadi-Gluth JA, Kesari S. Mechanisms of radiotherapy-associated cognitive disability in patients with brain tumours. Nat Rev Neurol 2016;13:52–64.

[14] Gondi V, Tomé WA, Mehta MP. Why avoid the hippocampus? A comprehensive review. Radiother Oncol 2010;97:370–6.

[15] Péchoux C Le, Sun A, Slotman BJ, De Ruysscher D, Belderbos J, Gore EM. Prophylactic cranial irradiation for patients with lung cancer. Lancet Oncol 2016;17:e277–93.

[16] Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, et al. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33:341–55.

[17] Mulder ER, de Jong R a., Knol DL, van Schijndel R a., Cover KS, Visser PJ, et al. Hippocampal volume change measurement: Quantitative assessment of the reproducibility of expert manual

outlining and the automated methods FreeSurfer and FIRST. Neuroimage 2014;92:169–81.

[18] Dill V, Franco AR, Pinho MS. Automated methods for hippocampus segmentation: the evolution and a review of the state of the art. Neuroinformatics 2015;13:133–50.

[19] González-Villà S, Oliver A, Valverde S, Wang L, Zwiggelaar R, Lladó X. A review on brain structures segmentation in magnetic resonance imaging. Artif Intell Med 2016;73:45–69.

[20] Patenaude B, Smith SM, Kennedy DN, Jenkinson M. A Bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 2011;56:907–22.

[21] Reuter M, Schmansky NJ, Rosas HD, Fischl B. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 2012;61:1402–18.

[22] Zhu H, Cheng H, Yang X, Fan Y, Alzheimer's Disease Neuroimaging Initiative. Metric Learning for Multi-atlas based Segmentation of Hippocampus. Neuroinformatics 2017;15:41–50.

[23] Wang H, Suh JW, Das SR, Pluta JB, Craige C, Yushkevich PA. Multi-Atlas Segmentation with Joint Label Fusion. IEEE Trans Pattern Anal Mach Intell 2013;35:611–23.

[24] Pipitone J, Park MTM, Winterburn J, Lett TA, Lerch JP, Pruessner JC, et al. Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. Neuroimage 2014;101:494–512.

[25] Crum WR, Scahill RI, Fox NC. Automated hippocampal segmentation by regional fluid registration of serial MRI: validation and application in Alzheimer's disease. Neuroimage 2001;13:847–55.

[26] van de Pol LA, Barnes J, Scahill RI, Frost C, Lewis EB, Boyes RG, et al. Improved reliability of hippocampal atrophy rate measurement in mild cognitive impairment using fluid registration. Neuroimage 2007;34:1036–41.

[27] Yushkevich P a., Avants BB, Das SR, Pluta J, Altinay M, Craige C. Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: An illustration in ADNI 3 T MRI data. Neuroimage 2010;50:434–45.

[28] Das SR, Avants BB, Pluta J, Wang H, Suh JW, Weiner MW, et al. Measuring longitudinal change in the hippocampal formation from in vivo high-resolution T2-weighted MRI. Neuroimage 2012;60:1266–79.

[29] Reuter M, Fischl B. Avoiding asymmetry-induced bias in longitudinal image processing. Neuroimage 2011;57:19–21.

[30] Reuter M, Rosas HD, Fischl B. Highly accurate inverse consistent registration: A robust approach. Neuroimage 2010;53:1181–96.

[31] Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging 2010;29:196–205.

[32] Shamonin DP, Bron EE, Lelieveldt BPF, Smits M, Klein S, Staring M. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. Front Neuroinform 2013;7:50.

[33] Modat M, Ridgway GR, Taylor ZA, Lehmann M, Barnes J, Hawkes DJ, et al. Fast free-form deformation using graphics processing units. Comput Methods Programs Biomed 2010;98:278–84.

[34] Modat M, Daga P, Cardoso MJ, Ourselin S, Ridgway GR, Ashburner J. Parametric non-rigid registration using a stationary velocity field. Proc. Work. Math. Methods Biomed. Image Anal., IEEE; 2012, p. 145–50.

[35] Schuh A, Murgasova M, Makropoulos A, Ledig C, Counsell SJ, Hajnal J V, et al. Construction of a 4D Brain Atlas and Growth Model Using Diffeomorphic Registration. Third Int. Work. Spat. Image Anal. Longitud. Time-Series Image Data, Boston, MA: 2014, p. 27–37.

[36] Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal 2008;12:26–41.

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

[38] Nho K, Corneveaux JJ, Kim S, Lin H, Risacher SL, Shen L, et al. Whole-exome sequencing and imaging genetics identify functional variants for rate of change in hippocampal volume in mild cognitive impairment. Mol Psychiatry 2013;18:781–7.

[39] Jack CR, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging 2008;27:685–91.

[40] Bartel F, Vrenken H, Bijma F, Barkhof F, Van Herk M, De Munck JC. Regional analysis of volumes and reproducibilities of automatic and manual hippocampal segmentations. PLoS One 2017;12:e0166785.

[41] Jack CR. MRI-Based Hippocampal Volume Measurements in Epilepsy. Epilepsia 1994;35:S21–9.

[42] van de Pol LA, van der Flier WM, Korf ESC, Fox NC, Barkhof F, Scheltens P. Baseline predictors of rates of hippocampal atrophy in mild cognitive impairment. Neurology 2007;69:1491–7.

[43] Patenaude B. Bayesian Statistical Models of Shape and Appearance for Subcortical Brain Segmentation. Dep Clin Neurol 2007;Doctor of:247.

[44] Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 2001;20:45–57.

[45] Avants BB, Tustison NJ, Stauffer M, Song G, Wu B, Gee JC. The Insight ToolKit image registration framework. Front Neuroinform 2014;8:44.

[46] Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee C. A Reproducible Evaluation of ANTs Similarity Metric Performance in Brain Image Registration 2012;54:2033–44.

[47] Tustison NJ, Avants BB. Explicit B-spline regularization in diffeomorphic image registration. Front Neuroinform 2013;7:39.

[48] Boccardi M, Ganzola R, Bocchetta M, Pievani M, Redolfi A, Bartzokis G, et al. Survey of protocols for the manual segmentation of the hippocampus: Preparatory steps towards a joint EADC-ADNI harmonized protocol. Adv Alzheimer's Dis 2011;2:111–25.

[49] Boccardi M, Bocchetta M, Apostolova LG, Barnes J, Bartzokis G, Corbetta G, et al. Delphi definition of the EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance. Alzheimer's Dement 2015;11:126–38.

[50] Frisoni GB, Jack CR, Bocchetta M, Bauer C, Frederiksen KS, Liu Y, et al. The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: evidence of validity. Alzheimers Dement 2015;11:111–25.

[51] Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate

linear registration and motion correction of brain images. Neuroimage 2002;17:825–41.

[52] Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Med Image Anal 2001;5:143–56.

[53] Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imaging 1999;18:712–21.

[54] Metz CT, Klein S, Schaap M, Van Walsum T, Niessen WJ. Nonrigid registration of dynamic medical imaging data using nD + t B-splines and a groupwise optimization approach. Med Image Anal 2011;15:238–49.

[55] Lorensen WE, Cline HE. Marching cubes: A high resolution 3D surface construction algorithm. Proc. 14th Annu. Conf. Comput. Graph. Interact. Tech. - SIGGRAPH '87, New York, New York, USA: ACM Press; 1987, p. 163–9.

[56] Cha Zhang, Tsuhan Chen. Efficient feature extraction for 2D/3D objects in mesh representation. Proc. 2001 Int. Conf. Image Process. (Cat. No.01CH37205), vol. 2, IEEE; 2001, p. 935–8.

[57] Tabatabaei-Jafari H, Shaw ME, Cherbuin N. Cerebral atrophy in mild cognitive impairment: A systematic review with meta-analysis. Alzheimer's Dement Diagnosis, Assess Dis Monit 2015;1:487–504.

[58] Popescu V, Battaglini M, Hoogstrate WS, Verfaillie SCJ, Sluimer IC, van Schijndel R a., et al. Optimizing parameter choice for FSL-Brain Extraction Tool (BET) on 3D T1 images in multiple sclerosis. Neuroimage 2012;61:1484–94.

[59] Cover KS, van Schijndel RA, Versteeg A, Leung KK, Mulder ER, Jong RA, et al. Reproducibility of hippocampal atrophy rates measured with manual, FreeSurfer, AdaBoost, FSL/FIRST and the MAPS-HBSI methods in Alzheimer's disease. Psychiatry Res Neuroimaging 2016;252:26–35.

[60] Song Y, Wu G, Bahrami K, Sun Q, Shen D. Progressive multi-atlas label fusion by dictionary evolution. Med Image Anal 2017;36:162–71.

[61] Fraser MA, Shaw ME, Cherbuin N. A systematic review and meta-analysis of longitudinal hippocampal atrophy in healthy human ageing. Neuroimage 2015;112:364–74.

[62] Modat M, Cash DM, Daga P, Winston GP, Duncan JS, Ourselin S. Global image registration using a symmetric block-matching approach. J Med Imaging 2014;1:024003.

[63] Simó M, Vaquero L, Ripollés P, Gurtubay-Antolin A, Jové J, Navarro A, et al. Longitudinal Brain Changes Associated with Prophylactic Cranial Irradiation in Lung Cancer. J Thorac Oncol 2016;11:475–86.

[64] Hong AM, Hallock H, Valenzuela M, Lo S, Paton E, Ng D, et al. Hippocampal Avoidance Whole Brain Radiation Therapy is Associated with Preservation of Hippocampal Volume at Six Months: A Case Series. Neuro-Oncology Open Access 2017;2.

[65] Seibert TM, Karunamuni R, Bartsch H, Kaifi S, Krishnan AP, Dalia Y, et al. Radiation Dose-Dependent Hippocampal Atrophy Detected With Longitudinal Volumetric Magnetic Resonance Imaging. Int J Radiat Oncol Biol Phys 2017;97:263–9.