Fast Predictive Simple Geodesic Regression

Zhipeng Ding[a], Greg Fleishman[c,d], Xiao Yang[a], Paul Thompson[c], Roland Kwitt[e], Marc Niethammer[a,b], The Alzheimer's Disease Neuroimaging Initiative*

[a] *Department of Computer Science, University of North Carolina at Chapel Hill, USA*
[b] *Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, USA*
[c] *Imaging Genetics Center, University of Southern California, USA*
[d] *Department of Radiology, University of Pennsylvania, USA*
[e] *Department of Computer Science, University of Salzburg, Austria*

Abstract

Deformable image registration and regression are important tasks in medical image analysis. However, they are computationally expensive, especially when analyzing large-scale datasets that contain thousands of images. Hence, cluster computing is typically used, making the approaches dependent on such computational infrastructure. Even larger computational resources are required as study sizes increase. This limits the use of deformable image registration and regression for clinical applications and as component algorithms for other image analysis approaches. We therefore propose using a fast predictive approach to perform image registrations. In particular, we employ these fast registration predictions to *approximate* a simplified geodesic regression model to capture longitudinal brain changes. The resulting method is orders of magnitude faster than the standard optimization-based regression model and hence facilitates large-scale analysis on a single graphics processing unit (GPU). We evaluate our results on 3D brain magnetic resonance images (MRI) from the ADNI datasets.

Keywords: Fast prediction, image regression, ADNI dataset, longitudinal data

1. Introduction

Longitudinal image data provides us with a wealth of information to study aging processes, brain development and disease progression. Such studies, for example ADNI [1] and the Rotterdam study [2], involve analyzing thousands of images. In fact, even larger studies will be available in the near future. For example, the UK Biobank [3] targets on the order of 100,000 images once completed. With the number of images increasing, large-scale image analysis typically resorts to using compute clusters for parallel processing. While this is, in principle, a viable solution, increasingly larger compute clusters will become necessary for such studies. Alternatively, more efficient algorithms can reduce computational requirements, which then facilitates computations on individual computers or much smaller compute clusters, interactive (e.g., clinical) applications, efficient algorithm development, and use of these efficient algorithms as components in more sophisticated analysis approaches (which may use them as part of iterative processes).

Image registration is a key task in medical image analysis to study deformations between images. Building on image registration approaches, image regression models [4, 5, 6, 7, 8, 9, 10, 11, 12, 13] have been developed to analyze deformation trends in longitudinal imaging studies. One such approach is geodesic regression (GR) [4, 7, 8] which (for images) build on the large displacement diffeomorphic metric mapping model (LDDMM) [14]. In general, GR generalizes linear regression to Riemannian manifolds. When applied to longitudinal image data, it can compactly express spatial image transformations over time. However, the solution to the underlying optimization problem is computationally expensive. Hence, a simplified, approximate, GR approach has been proposed [15] (SGR) to decouple the computation of the regression geodesic into pairwise image registrations. However, even such a simplified GR approach would require months of computation time on a single graphics processing unit (GPU) to process thousands of 3D image registrations for large-scale imaging studies such as ADNI [1]. The primary reason computational bottleneck for SGR are the optimization required to compute pair-wise registrations.

Recently, efficient approaches have been proposed for deformable image registration [16, 17, 18, 19, 20, 21]. In particular, for LDDMM, which is the basis of GR approaches for images, registrations can be dramatically sped up, by either working with finite-dimensional Lie algebras [22] and frequency diffeomorphisms [21], or by fast

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

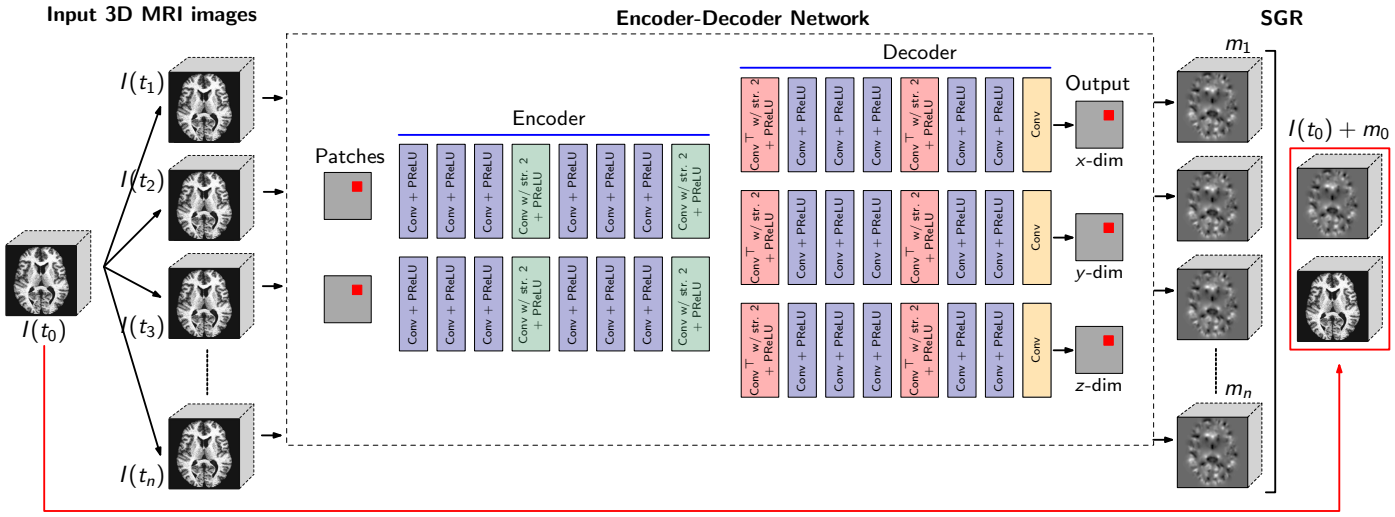

Figure 1: Principle of fast predictive simple geodesic regression (FPSGR). In the encoder-decoder network (middle), the inputs are patches from the moving image and the target image at the *same* spatial location; the outputs are the predicted initial momenta (i.e., $m_1, \ldots, m_n$) of the corresponding patches. `Conv`: Convolutional layer; $\texttt{Conv}^T$: transpose of convolutional layer. In the simple geodesic regression (SGR) part, all the pairwise initial momenta are *averaged* according to Eq. (9) to produce the initial momentum of the regression geodesic (marked red).

predictive image registration (FPIR) [19, 20]. FPIR predicts the initial conditions (specifically, the initial momentum) of LDDMM, which fully characterize the geodesic and the spatial transformation using a *learned* a patch-based deep regression model. Because numerical optimization of standard LDDMM registration is replaced by a *single* prediction step, followed by optional correction steps [20], FPIR is dramatically faster than optimization-based LDDMM without compromising registration accuracy, as measured on several registration benchmarks [23].

Besides FPIR, other predictive image registration approaches have been proposed. Dosovitskiy et al. [24] use a convolutional neural network (CNN) to directly predict optical flow. Liu et al. [25] use an encoder-decoder network to synthesize video frames. Schuster et al. [26] investigate strategies to improve optical flow prediction via a CNN. Cao et al. [16] use a sampling strategy and CNN regression to directly learn the mapping from moving and target image pairs to the final deformation field. Miao et al. [17] use CNN regression for 2D/3D rigid registration. Sokooti et al.[18] use CNNs to directly predict a 3D displacement vector field from input image pairs. An unsupervised approach for image registration was proposed by de Vos et al. [27]; here, the loss function is the image similarity measure between images themselves and a deformation is parameterized via a spatial transformer (which essentially amounts to a parameterized model of deformation in image registration) which generates the sought-for displacement vector field. In [28], Hong et al. employ a low-dimensional band-limited representation of velocity fields in Fourier space [22] to speed up SGR [15] for population-based image analysis.

In this work, we will build on FPIR, as it is a desirable approach for brain image registration for the following reasons: *First*, FPIR predicts the initial momentum of LDDMM and therefore inherits the theoretical properties of LDDMM. Consequently, FPIR results in diffeomorphic transformations, even though predictions are computed in a patch-by-patch manner; this can not be guaranteed by most other prediction methods. *Second*, patch-wise prediction allows for training of the prediction models based on a very small number of images, containing a large number of patches. *Third*, by using a patch-wise approach, even high-resolution image volumes can be processed without running into memory issues on a GPU. *Fourth*, none of the existing predictive methods address longitudinal data. However, as both FPIR and SGR are based on LDDMM, they naturally integrate and hence result in our proposed *fast predictive simple geodesic regression (FPSGR)* approach.

Our *contributions* can be summarized as follows:

Predictive geodesic regression We use a fast predictive registration approach for image geodesic regression. Different to [20], we specifically validate that our approach can indeed capture the frequently subtle deformation trends of *longitudinal* image data.

Large-scale dataset capability Our predictive regression approach facilitates large-scale image regression within a short amount of time on a single GPU, instead of requiring months of computation time for standard optimization-based methods on a single computer, or on a compute cluster.

Accuracy We assess the accuracy of FPSGR by (1) studying linear models of atrophy scores (which are derived from the nonlinear SGR model) over time, as

well as (2) correlations between atrophy scores and various diagnostic groups.

Validation We demonstrate the performance of FPSGR by analyzing > 6000 images of the `ADNI-1` / `ADNI-2` datasets. For comparison, we also perform SGR using numerical optimization for the registrations, again on the complete `ADNI-1` / `ADNI-2` datasets.

This work is an extension of a recent conference paper [29]. In particular, all our experiments are now in 3D. We also added significantly more results to further explore the behavior of FPSGR in comparison to optimization-based SGR. In particular, we added (a) an example to visualize the performance of regression models and associated quantitative comparisons (Sec. 4.1); (b) an analysis of local atrophy score correlated with clinical variables (Sec. 4.3); (c) correlations within diagnostic groups (Sec. 4.3); (d) a comparison with pairwise registration (Sec. 4.4); (e) and experiments on extrapolation on unseen data (Sec. 4.5, Sec. 4.6).

Organization. The remainder of this article is organized as follows: Sec. 2 describes FPSGR, Sec. 3 discusses the experimental setup and the training of the prediction models. In Sec. 4, we present experimental results for 3D MR brain images. The paper concludes with a summary and an outlook on future work.

2. Fast predictive simple geodesic regression

Our fast predictive simple geodesic regression approach is a combination of two methods: *First*, fast predictive image registration (FPIR) and, *second*, integration of FPIR with simple geodesic regression (SGR). Both FPIR and SGR are based on the shooting formulation of LDDMM [7]; Fig. 1 illustrates our overall approach. The individual components are described in the following.

2.1. LDDMM

Shooting-based LDDMM and geodesic regression minimize

$$E(I_0, m_0) = \frac{1}{2}\langle m_0, Km_0\rangle + \frac{1}{\sigma^2}\sum_i d^2(I(t_i), Y^i), \quad (1)$$

$$s.t. \quad m_t + \mathrm{ad}_v^* m = 0, I_t + \nabla I^T v = 0, m - Lv = 0,$$

where I_0 is the initial image (known for image-to-image registration and to be determined for geodesic regression), m_0 is the initial momentum, K is a smoothing operator that connects velocity v and momentum m as $v = Km$ and $m = Lv$ with $K = L^{-1}$, $\sigma > 0$ is a weight, Y^i is the measured image at time t_i (there will be only one such image for image-to-image registration at $t = 1$), and $d^2(I_1, I_2)$ denotes the image similarity measure between I_1 and I_2 (for example L_2 or geodesic distance); ad^* is the dual of the negative Jacobi-Lie bracket of vector fields: $\mathrm{ad}_v w = -[v, w] = Dvw - Dwv$ and D denotes the Jacobian. The deformation of the source image $I_0 \circ \Phi^{-1}$ can be computed by solving $\Phi_t^{-1} + D\Phi^{-1}v = 0,\ \Phi^{-1}(0) = \mathrm{id}$, where id denotes the identity map.

2.2. FPIR

Fast predictive image registration [19, 20] aims at predicting the initial momentum, m_0, between a source and a target image patch-by-patch. Specifically, we use a deep encoder-decoder network to predict the patch-wise momentum. As shown in Fig. 1, in 3D the inputs are two layers of $15 \times 15 \times 15$ image patches (15×15 in 2D), where the two layers are from the source and target images respectively. Two patches are taken at the same position by two parallel encoders, which learn features independently. The output is the predicted initial momentum in the x, y and z directions (obtained by numerical optimization on the training samples). Basically, the network is split into an encoder and a decoder part. An *encoder* consists of 2 blocks of three $3 \times 3 \times 3$ convolutional layers with PReLU activations, followed by another $2 \times 2 \times 2$ convolution+PReLU with a stride of two, serving as a "pooling" operation. The number of features in the first convolutional layer is 64 and increases to 128 in the second. In the *decoder*, three parallel decoders share the same input generated from the encoder. Each decoder is the inverse of the encoder except for using 3D transposed convolution layers with a stride of two to perform "unpooling", and no non-linearity at the end. To speed up computations, we use patch pruning (i.e., for brain imaging, e.g., patches outside the brain are not predicted as the momentum is expected to be zero there) and a large pixel stride (e.g., 14 for $15 \times 15 \times 15$ patches) for the sliding window of the predicted patches.

2.3. Correction network

We follow [20] and use a two-step approach to improve overall prediction accuracy. An additional correction step, i.e., a *correction network*, corrects the prediction of the initial prediction network. Fig. 2 illustrates this two-step approach graphically. The correction network has the same structure as the prediction network. Only the inputs and outputs differ. For the prediction network, the inputs are the original moving image and the original target image; output is the predicted initial momentum. For the correction network, the inputs are the original moving image and the warped target image; the output is the momentum difference.

2.4. SGR

Determining the initial image, I_0, and the initial momentum, m_0, of Eq. (1) is computationally costly. However, in simple geodesic regression, the initial image is fixed to the *first* image of a subject's longitudinal image set (leftmost part of Fig. 1). Furthermore, the similarity measure $d(\cdot, \cdot)$ is chosen as the geodesic distance between images and *approximated* so that the geodesic regression problem can be solved by computing pair-wise image registrations

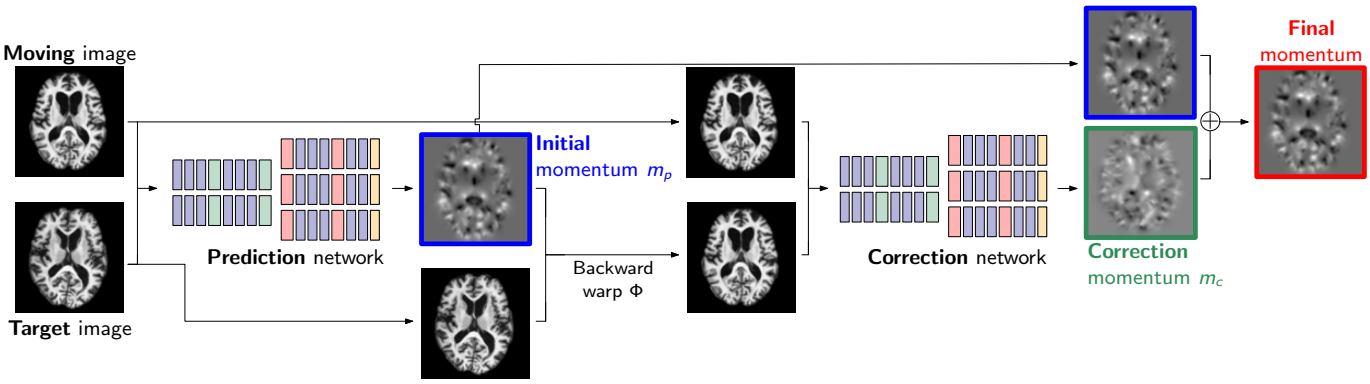

Figure 2: Architecture of the prediction + correction network. Here, we use 2D images and the momentum in the x-direction for illustration. All images are 3D in our experiments. (1) Predict the initial momentum m_p and the corresponding backward deformation, Φ; (2) Predict a correction of the initial momentum, m_c, based on the difference between the moving image and the warped-back target image. The final momentum is $m = m_p + m_c$. The correction network is trained based on the moving images and the warped-back target images of the training dataset.

3D Longitudinal Test Case Deformation Error [pixel]							
Data Percentile	0.3%	5%	25%	50%	75%	95%	99.7%
Longitudinal Training	**0.0156**	**0.0407**	**0.0761**	**0.1098**	**0.1559**	**0.2681**	**0.3238**
Cross-sectional Training	0.0544	0.1424	0.2641	0.3723	0.5067	0.7502	0.8425
3D Cross-sectional Test Case Deformation Error [pixel]							
Data Percentile	0.3%	5%	25%	50%	75%	95%	99.7%
Longitudinal Training	0.1694	0.4802	1.0765	1.7649	2.7630	4.8060	5.6826
Cross-sectional Training	**0.1123**	**0.3024**	**0.5863**	**0.8737**	**1.2743**	**2.2659**	**2.7836**

Table 1: Deformation error of longitudinal and cross-sectional models tested on longitudinal and cross-sectional data. 2-norm deformation errors in pixels w.r.t. the ground truth deformation obtained by numerical optimization for LDDMM. A prediction model trained with longitudinal registration performs better for longitudinal registrations. Conversely, a model trained based on cross-sectional registration is preferred for cross-sectional registrations.

with respect to the first image. Specifically, we define the quadratic distance d^2 between two images A and B as

$$d^2(A,B) = \frac{1}{2}\int_0^1 \|v^*\|_L^2 dt, \tag{2}$$
$$\text{where } v^* = \arg\min_v \frac{1}{2}\int_0^1 \|v\|_L^2 dt + \frac{1}{\sigma^2}\|Q(1) - B\|_2^2,$$
$$\text{s.t.} \quad Q_t + \nabla Q^T v = 0, \text{and } Q(0) = A \ .$$

Assume we have an image $I(t_0)$ at time t_0 as well as two images $A(t_i)$ and $B(t_i)$. Further, assume that the spatial transformation Φ_A maps $A(t_i)$ to $I(t_0)$ and Φ_B maps $B(t_i)$ to I_0. Then $A(t_i) = I(t_0) \circ \Phi_A^{-1}$ and $B(t_i) = I(t_0) \circ \Phi_B^{-1}$. Furthermore, assume that Φ maps $A(t_i)$ to $B(t_i)$, i.e., $B(t_i) = A(t_i) \circ \Phi^{-1}$. Then $\Phi = \Phi_B \circ \Phi_A^{-1}$. Assuming that the geodesic between $I(t_0)$ and $A(t_i)$ is parameterized by the initial velocity v^A and between $I(t_0)$ and $B(t_i)$ by the initial velocity v^B and that we travel between $I(t_0)$ and $A(t_i)$ in time $t_i - t_0$ (and similarly for $B(t_i)$) we can rewrite the map between $A(t_i)$ and $B(t_i)$ based on the exponential map as

$$\Phi = \text{Exp}_{\text{Id}}((t_i - t_0)v^B) \circ \text{Exp}_{\text{Id}}(-(t_i - t_0)v^A), \tag{3}$$

which can be approximated to first order as

$$\Phi \approx \text{Exp}_{\text{Id}}((t_i - t_0)(v^B - v^A)). \tag{4}$$

Hence, the squared geodesic distance between the two images can be approximated as

$$d^2(A(t_i), B(t_i)) \approx \frac{1}{2}(t_i - t_0)^2 \langle K(m^B - m^A), m^B - m^A \rangle, \tag{5}$$

where $v^A = Km^A$ and $v^B = Km^B$. Hence, Eq. (1) becomes

$$E(\bar{I}, \overline{m}) = \frac{1}{2}\langle \overline{m}, K\overline{m} \rangle + \frac{1}{2\sigma^2}\sum_i (t_i - t_0)^2 \langle K(\overline{m} - m_i), \overline{m} - m_i \rangle, \tag{6}$$

where $\overline{m}$ is the sought-for initial momentum of the regression geodesic and m_i are the initial momenta corresponding to the geodesic connecting $\bar{I}$ (the starting image of the geodesic) and the measurements Y_i in time $t_i - t_0$. Differentiating Eq. (6) w.r.t. $\overline{m}$ results in

$$\nabla_{\overline{m}} E = K[\overline{m} + \frac{1}{\sigma^2}\sum_i (t_i - t_0)^2(\overline{m} - m_i)] \stackrel{!}{=} 0. \tag{7}$$

Thus,

$$\overline{m} = \frac{\sum_i (t_i - t_0)^2 m_i}{\sigma^2 + \sum_i (t_i - t_0)^2}. \tag{8}$$

In practice, σ^2 is very small and can thus be omitted. Furthermore, m_i is obtained by either registering $\overline{I}$ to Y^i in unit time or, as in our FPSGR approach, by predicting the momenta m_i via FPIR, denoted as $\widetilde{m}_i$. As Equation 8 was derived assuming that images are transformed into each other in time $t_i - t_0$ instead of unit time, the obtained unit-time predicted momenta $\widetilde{m}_i$ correspond in fact to the approximation $\widetilde{m}_i \approx (t_i - t_0)m_i$. Finally, we obtain the approximated optimal $\overline{m}$ of the energy functional in Eq. (1), for a fixed $\overline{I} = I_0$ as

$$\overline{m} \approx \frac{\sum_i (t_i - t_0)\widetilde{m}_i}{\sum_i (t_i - t_0)^2}. \tag{9}$$

3. Setup / Training

All our experiments use 3D images from the `ADNI` dataset[1] which consists of 6471 3D MR brain images of size $220 \times 220 \times 220$ voxels. In particular, `ADNI-1` contains 3479 images from 833 subjects and `ADNI-2` contains 2992 images from 823 subjects. Images belong to various types of diagnostic categories which we will discuss later.

We perform the following two types of studies:

Registration We assess our hypothesis that training FPIR on longitudinal data for longitudinal registrations is preferred over training using cross-sectional data. Vice versa, training FPIR on cross-sectional data for cross-sectional registrations is preferred over training using longitudinal data. Comparisons are with respect to registration results obtained by numerical optimization (i.e., LDDMM).

Regression As for regression, we compare linear models fitted to atrophy scores over time, where scores are either obtained from FPSGR or optimization-based SGR. Additionally, we study correlations between atrophy scores and diagnostic groups. Our hypothesis is that FPSGR is accurate enough to achieve comparable performance to optimization-based SGR, at much lower computational cost, in both situations.

3.1. Training of the prediction models

We use a randomly selected set of 120 patients' MRI images from `ADNI` for training the prediction models and to test the performance of FPIR. We use all of the `ADNI` data for our regression experiments.

Training for registration. We randomly selected 120 subjects from `ADNI-1` and registered their baseline images to their 24 month follow-up images. We used the first 100 subjects for training and the remaining 20 subjects for testing. For *longitudinal training*, we registered the baseline image of a subject to the subject's 24-month image. For *cross-sectional training*, we registered a subject's baseline image to another subject's 24-month image. To assess the performance of prediction models trained on these two types of paired data, we (1) perform the same type of registrations on the held-out 20 subjects and (2) compare the 2-norm of the deformation error computed from the output of the prediction models with respect to the result obtained by numerical optimization of LDDMM[2] (which serves as the "ground-truth"). Table 1 shows the results which confirm our hypothesis that training the prediction model with longitudinal registration cases is preferred for longitudinal registration over training with cross-sectional data. The deformation error is very small for longitudinal training / testing which provides strong evidence that the predictive method exhibits performance comparable to the (costly) optimization-based LDDMM. Another interpretation of these results is, that it is beneficial to train a prediction model with deformations that are to be *expected*, i.e., relatively small deformations for longitudinal registrations and larger deformations for cross-sectional registrations. As we are interested in longitudinal registrations for the `ADNI` data, we only train our 3D models using longitudinal registrations in the following.

Training for regression. The `ADNI-1` dataset contains 228 normal controls, 257 subjects with mild cognitive impairment (MCI), 149 with late mild cognitive impairment (LMCI), as well as 199 subjects suffering from Alzheimer's disease (AD). We randomly picked roughly 1/6 of patients from each diagnostic category to form a set of 139 subjects for training in `ADNI-1`, i.e., 38 normal controls, 43 MCI, 25 LMCI, as well as 33 AD subjects; this results in 139 subjects overall. The baseline images of each subject were registered to *all* the later time-points within the same subject. To maintain the diagnostic ratio, we picked (out of all registrations) 45 registrations from the normal group, 50 registrations from the MCI group, 30 registrations from the LMCI group, and 40 registrations from the AD group, resulting in 165 longitudinal registration cases for training.

The same strategy was applied to `ADNI-2`. In detail, `ADNI-2` contains 200 normal controls, 111 subjects with significant memory complaint (SMC), 182 subjects with early mild cognitive impairment (EMCI), 175 with late mild cognitive impairment (LMCI), and 155 subjects with Alzheimer's disease (AD). We picked 150 subjects and 140 longitudinal registrations, consisting of 35 registrations from the control group, 20 registrations from the SMC group, 30 registrations from the EMCI group, 30

[1] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). 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).

[2] LDDMM results are generated using a vector momentum formulation: https://bitbucket.org/scicompanat/vectormomentum

registrations from the LMCI group, and 25 registrations from the AD group. Note that there are fewer registrations than subjects (140 *vs.* 150) in this setup, as our priority is to maintain the overall diagnostic ratio.

For both, `ADNI-1` and `ADNI-2`, the remaining 5/6 of the data is used for testing. We trained four prediction models and their four corresponding correction models, leading to eight prediction models in total, listed in Table 2. We also note that the training sets within `ADNI-1` and `ADNI-2`, resp., were not overlapping.

`ADNI-1 Pred-1`	Model v1 (no corr.)
`ADNI-1 Pred+Corr-1`	Model v1 +1x corr. step
`ADNI-1 Pred-2`	Model v2 (no corr.)
`ADNI-1 Pred+Corr-2`	Model v2 +1x corr. step
`ADNI-2 Pred-1`	Model v1 (no corr.)
`ADNI-2 Pred+Corr-1`	Model v1 +1x corr. step
`ADNI-2 Pred-2`	Model v2 (no corr.)
`ADNI-2 Pred+Corr-2`	Model v2 +1x corr. step

Table 2: Overview of the trained prediction models.

3.2. Parameter selection

We use the regularization kernel

$$K = L^{-1} = (-a\nabla^2 - b\nabla(\nabla\cdot) + c)^{-2}$$

with $[a, b, c]$ set to $[1, 0, 0.1]$. The parameter σ, from equation (1), is set to 0.1. We train our network (using `ADAM` [30]) over 10 epochs with a learning rate of 0.0001.

3.3. Efficiency

Once trained, the prediction models allow fast computations of registrations. We use a TITAN X (Pascal) GPU and `PyTorch`[3] for our implementation of FPIR. For the 3D `ADNI-1` dataset ($220 \times 220 \times 220$ MR images), FPSGR took about one day to predict 2646 pairwise registrations (i.e., 25 [s]/prediction) and to compute the regression result. Optimization-based LDDMM[4] would require $\approx$ 40 days of runtime. Runtime for FPIR on `ADNI-2` is identical to `ADNI-1` as the images have the same spatial dimension.

Compared to the-state-of-art fast geodesic regression model [28], FPSGR is also at least twice as fast. The model in [28] achieves $\approx$ 16 times speed-up compared with SGR [15] for the same setting (parallel computing with the same number of cores). In our case, we achieve a more than 40 times speed-up compared with SGR for the same setting (a single GPU).

[3] http://pytorch.org

[4] Here, we used 300 fixed iterations for each registration. 300 iterations can guarantee almost all the results converge. Note that the optimization-based LDDMM also uses a GPU implementation.

Distribution of prediction cases in `ADNI-1`						
`Pred-1`	6mo	12mo	18mo	24mo	36mo	48mo
NC	182	172	8	151	128	38
MCI*	274	221	165	122	80	11
AD	153	173	66	163	69	20
Total	609	566	239	436	277	69
`Pred-2`	6mo	12mo	18mo	24mo	36mo	48mo
NC	182	168	9	144	119	33
MCI*	272	224	169	124	70	10
AD	152	168	64	160	67	22
Total	606	560	242	428	256	65

Table 3: Distribution of `Pred/Corr-1` and `Pred/Corr-2` cases in `ADNI-1`. MCI* is the combination of the MCI and LMCI diagnostic groups. 18 month only has one diagnostic group.

Distribution of prediction cases in `ADNI-2`					
`Pred-1`	3mo	6mo	12mo	24mo	36mo
NC*	173	141	153	119	3
MCI*	256	232	207	142	4
AD	93	95	105	66	1
Total	522	468	465	327	8
`Pred-2`	3mo	6mo	12mo	24mo	36mo
NC*	172	142	159	122	3
MCI*	257	230	202	149	4
AD	94	98	101	52	1
Total	523	470	462	323	8

Table 4: Distribution of `Pred/Corr-1` and `Pred/Corr-2` cases in `ADNI-2`. Normal* denotes the combination of the Normal and SMC diagnostic groups; MCI* denotes the combination of the EMCI and LMCI diagnostic groups. Only a small number of images is available for the 36 months time point.

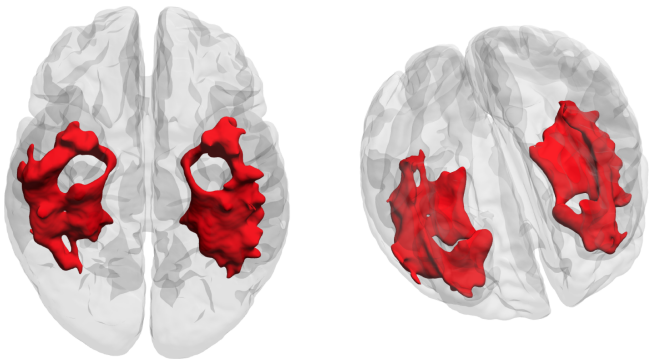

Figure 3: Region of Interest (ROI) significantly associated with atrophy in AD used to compute atrophy scores.

4. Experimental results for 3D ADNI data

For our experiments, we created 10 different (dataset, registration approach) combinations, each combination specifically designed to assess certain properties of our proposed strategy. These combinations are described next.

1) All subjects from the `ADNI-1` dataset in combination

with optimization-based LDDMM.

2) Two subgroups of `ADNI-1` (i.e., different training data portions) in combination with FPSGR *without* a correction network.

3) The same two subgroups as in 2), but in combination with FPSGR *with* a correction network.

4) The same five groups of 1-3, but for `ADNI-2`.

Our general hypothesis is that the prediction models (for `ADNI-1/2`) show similar performance to optimization-based LDDMM and that using the correction network for the predictions improves results. To assess differences, we compare differences in deformations. Specifically, for every deformation produced by the different approaches, we compute its Jacobian determinant (JD). The JDs are then warped to a common coordinate system for the entire `ADNI` dataset using existing non-linear deformations from [31, 32]. Each such spatially normalized JD is then averaged within a region where the rate of atrophy is significantly associated with Alzheimer's disease (AD), i.e., within a *statistical region of interest (stat-ROI)* (see Fig. 3). Specifically, we quantify atrophy as

$$\left(1 - \frac{1}{|\omega|}\int_{\omega} \det(D\phi(x))\ dx\right) \times 100 \tag{10}$$

where $\det(\cdot)$ denotes the determinant and $|\cdot|$ the cardinality/size of a set; ω is an area in the temporal lobes which was determined in prior studies [31, 32] to be significantly associated with accelerated atrophy in Alzheimer's disease. The resulting scalar value is an estimate of the relative volume change experienced by that region between the baseline and a follow-up image. Hence, its sign is positive when the region has lost volume over time and is negative if the region has gained volume over time.

We limited our experiments to the applications in [33, 34], wherein nonlinear registration/regression is used to quantify atrophy within regions known to be associated to varying degrees with AD (2), mild cognitive impairment (MCI) (1) (including LMCI[5]), and normal ageing (NC: normal control) (0) in an elderly population. These are the diagnostic groups for `ADNI-1`. For `ADNI-2`, there are also 3 diagnostic categories[6]: normal ageing (0) (including SMC), mild cognitive impairment (including EMCI and LMCI) (1), and AD (2).

Specifically, we investigate the following *six* questions:

Q1) Can the prediction models for regression qualitatively capture similar trends to the regression model obtained by numerical optimization? (Sec. 4.1)

Q2) Are atrophy measurements derived from FPSGR biased to overestimate or underestimate volume changes? (Sec. 4.2)

Q3) Are FPSGR atrophy measurements consistent with those derived from deformations via numerical optimization (LDDMM) which produced the training dataset? (Sec. 4.3)

Q4) Are regression results more stable and hence capture trends better than pairwise registrations? (Sec. 4.4)

Q5) Is the predictive power of the regression models strong enough to forecast deformations for unseen future timepoints (Sec. 4.5)

Q6) Do the prediction results capture expected trends in deformation? (Sec. 4.6)

If these experiments resolve favorably, then the substantially improved computational efficiency of FPSGR justifies its use for large-scale imaging studies. Tables 3 and 4 show the distributions of the prediction cases per time-point and the diagnostic groups in `ADNI-1` and `ADNI-2`, respectively.

4.1. Regression results

Table 1 indicates that FPIR can predict deformation fields similar to the ones obtained using optimization-based LDDMM, even for the subtle changes seen in longitudinal imaging data. However, it remains to be seen how a predictive model performs for image regression. Fig. 4 shows an exemplary regression result. In this specific case, large changes can be observed around the ventricles. To illustrate differences between the methods, Fig. 4 shows regression results based on optimization-based LDDMM, for FPSGR *without* a correction network, and for FPSGR *with* a correction network. All three methods successfully capture the expanding ventricles and generally capture the image changes. Both FPSGR methods show results that are highly similar to SGR using optimization-based LDDMM. Hence, FPSGR is useful for longitudinal image regression. To further quantify the regression accuracy, we compute the overlay error between measured images and the images on the geodesic as

$$E_{overlay}(I_0 \circ \Phi_{t_i}^{-1}, Y_i) = \frac{1}{|\Omega|}\|I_0 \circ \Phi_{t_i}^{-1} - Y_i\|_{L_1} \tag{11}$$

where Ω is the brain area, $I_0 \circ \Phi_{t_i}^{-1}$ is the regressed image at time t_i and Y_i is the measured image at time t_i. Table 5 shows the overlay error for the population of 100 subjects which includes all diagnostic groups in `ADNI-1`. Both

[5] We combine MCI and LMCI mainly because (a) the diagnostic changes available on the IDA website (https://ida.loni.usc.edu/login.jsp) only provide these three diagnostic groups; (b) to be consistent with the experiments conducted by Hua et al. [33], where only Normal, MCI and AD were used as labels to classify `ADNI-1`. Hereafter, in all discussions of `ADNI-1`, MCI is a combination of MCI and LMCI of `ADNI-1`

[6] Similar to `ADNI-1`, a detailed diagnosis for `ADNI-2` is only available for the baseline images; MR images at later time points are only labeled as NC, MCI, and AD. Thus, we combine SMC and NC, as well as EMCI and LMCI to be consistent with the diagnostic changes in the *ADNI Diagnosis Summary* available on the IDA website. Hereafter, in all discussions of `ADNI-2`, NC includes NC and SMC and MCI includes EMCI and LMCI.

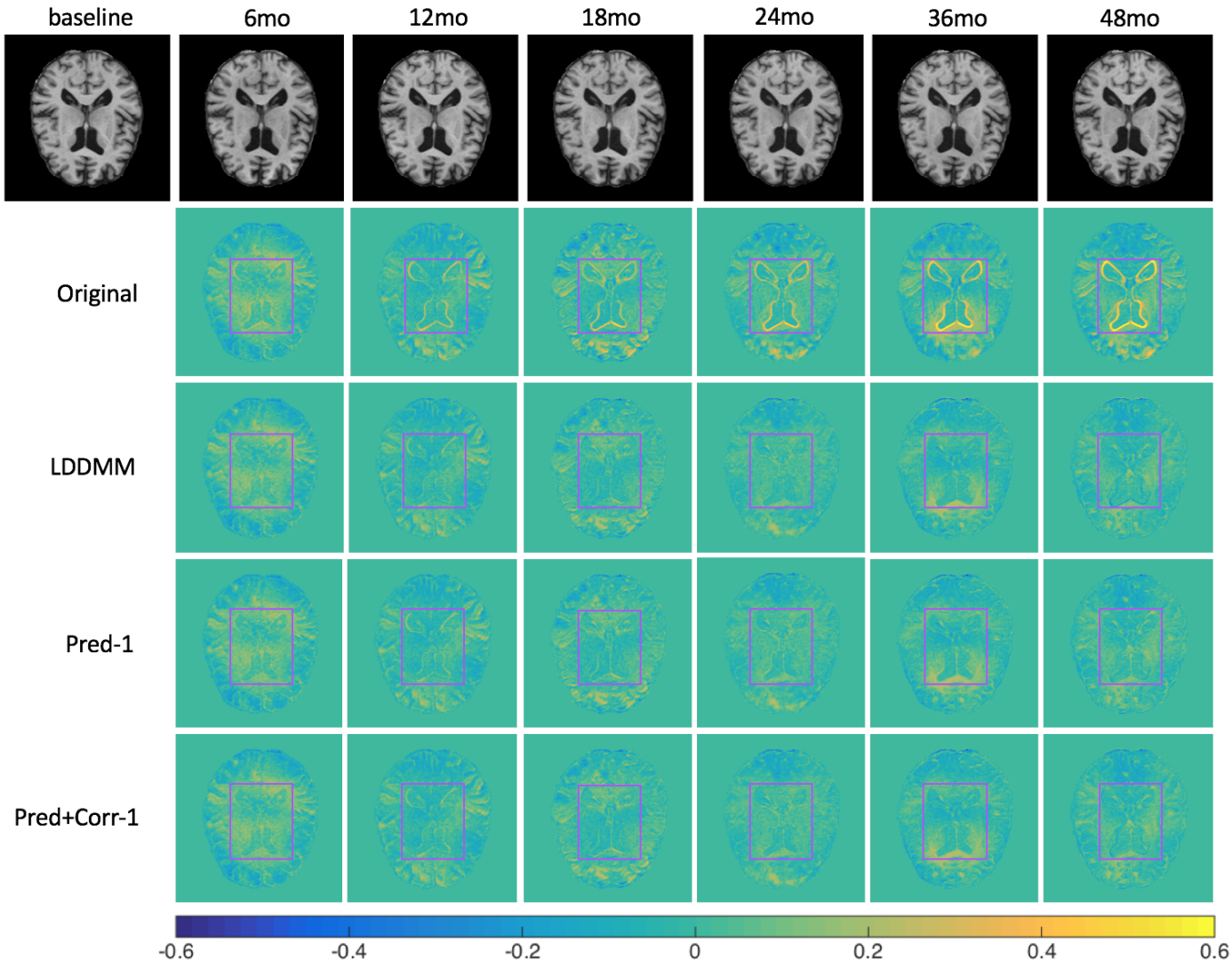

Figure 4: Exemplary regression result: one subject with 6 follow-up images from the `ADNI-1` dataset. Image intensity range is [0, 2.49]. **Top row:** Axial slices extracted from the 3D MR images at the same axial location for different months. **Original:** intensity differences between the baseline image and its 6-month, 12-month, etc. follow-up image. **LDDMM:** intensity differences between the acquired images in the top row and optimization-based regression results at each follow-up month(s). **Pred-1:** intensity differences between the acquired images in the top row and the Pred-1 regression results at each follow-up month(s). **Pred+Corr-1:** Same as for Pred-1, but using the Pred+Corr-1 regression model. Rectangles mark areas of major structural changes. Intensity differences are dramatically reduced, e.g., around the ventricles, demonstrating that these structural changes are captured by all three methods. The *prediction models* (Pred-1, Pred+Corr-1) give very similar results to the regression results obtained by numerical optimization (LDDMM).

FPSGR methods obtain results comparable with optimization-based LDDMM. This justifies the use of the proposed methods. The correction network generally increases the prediction accuracy over using the prediction network only.

4.2. Bias

Estimates of atrophy are susceptible to bias [35]. To quantitatively assess this potential bias, we separately considered different diagnostic groups. Specifically, we considered six diagnostic change groups in our experiments: (1) NC for all time points (NC-NC), (2) starting with NC and changing to MCI or AD at a later time point (NC-MCI), (3) MCI for all time points (MCI-MCI), (4) starting with MCI and reversing to NC at later time points (MCI-NC), (5) starting with MCI and changing to AD at later time points (MCI-AD), and (6) AD for all the time points (AD-AD)[7]. In particular, we follow [33] and fit a straight line (i.e., linear regression) through all atrophy measurements over time, conditioned on each diagnostic change category. The intercept term is an estimate of the atrophy one would measure when registering two scans acquired on the same day; hence it should be near zero and its 95% confidence interval should contain zero. Quantitatively, Table 6 lists the slopes, intercepts, and 95% confidence intervals for all

[7] In `ADNI-1/ADNI-2`, there are two patients who show a reversion from AD to MCI. We omitted these cases in our experiment because the number of such cases is too small.

	$E_{overlay}(I_0 \circ \Phi_{t_i}^{-1}, Y_i)$					
Measured Images	I_{6mo}	I_{12mo}	I_{18mo}	I_{24mo}	I_{36mo}	I_{48mo}
Original	0.0770 ± 0.0212	0.0764 ± 0.0207	0.0890 ± 0.0220	0.0810 ± 0.0223	0.0899 ± 0.0341	0.0940 ± 0.0415
LDDMM	0.0750 ± 0.0194	0.0686 ± 0.0176	0.0734 ± 0.0190	0.0609 ± 0.0168	0.0628 ± 0.0177	0.0663 ± 0.0221
Pred-1	0.0754 ± 0.0213	0.0694 ± 0.0182	0.0742 ± 0.0195	0.0621 ± 0.0188	0.0654 ± 0.0184	0.0698 ± 0.0238
Pred+Corr-1	0.0754 ± 0.0211	0.0691 ± 0.0182	0.0734 ± 0.0192	0.0615 ± 0.0166	0.0642 ± 0.0188	0.0688 ± 0.0235

Table 5: Mean+standard deviation of the overlay errors, see Eq. (11), over 100 patients in ADNI-1 dataset. Both prediction models exhibit performance comparable to optimization-based regression results (LDDMM). Including a correction network generally improves the performance of the prediction network.

ten groups of ADNI-1 and ADNI-2, respectively. LDDMM-1 and LDDMM-2 denote the optimization-based results split into the same testing groups used for Pred-1 and Pred-2 to allow for a direct comparison. All of the results show intercepts that are near zero relative to the range of changes observed and all prediction intercept confidence intervals contain zero. For all diagnostic change groups the prediction and prediction+correction models exhibit more stable results than the optimization-based LDDMM method as indicated by the tighter confidence intervals. Furthermore, all slopes are positive, indicating average volume loss over time. This is consistent with expectations for an aging and neuro-degenerative population. The slopes capture increasing atrophy with disease severity. In ADNI-1/ADNI-2, we expect $\text{Slope}_{\text{NC-NC}} < \text{Slope}_{\text{MCI-NC}} < \text{Slope}_{\text{NC-MCI}} < \text{Slope}_{\text{MCI-AD}} < \text{Slope}_{\text{AD-AD}}$ and all six experimental groups (i.e. LDDMM-1, Pred-1, Pred+Corr-1, LDDMM-2, Pred-2, and Pred+Corr-2) are generally consistent with this expectation. Exceptions happen in ADNI-2 for the NC-MCI and MCI-NC cases. As the number of subjects involved is relatively small, i.e., fewer than 20, compared with the other cases (roughly 100), one may speculate that this observation is caused by the limited number of data points for NC-MCI and MCI-NC as shown in the #data column of Table 6. However, the behavior within each starting diagnostic category, is consistent, i.e., for NC $\text{Slope}_{\text{NC-NC}} < \text{Slope}_{\text{NC-MCI}}$ and for MCI $\text{Slope}_{\text{MCI-NC}} < \text{Slope}_{\text{MCI-MCI}} < \text{Slope}_{\text{MCI-AD}}$. Hence, all six groups' slope results in ADNI-1/ADNI-2 are generally consistent with our expectation (and also consistent with results in [33]). The slope estimated from the prediction+correction results is larger than the slope estimated from the prediction model results and closer to the slope obtained from the optimization-based LDDMM results. This indicates that the correction network can improve prediction accuracy. Fig. 5 shows linear regression results for the estimated atrophy scores in ADNI-1/2 for the Pred+Corr-1 model. Both the data points themselves (i.e., the atrophy scores), as well as kernel density estimates for the linear trends for each subject are shown. These results are consistent with the results of Table 6 discussed above. We conclude that (1) neither LDDMM optimization nor FPSGR produced deformations with significant bias to overestimate or underestimate volume change; (2) a linear model of atrophy scores generated by FPSGR can capture intrinsic volume change (i.e., slope) among different diagnostic change groups. Note that our LDDMM optimization results and the prediction results show the same trends. Further, they are directly comparable as the results are based on the same test images (also for the atrophy measurements).

4.3. Atrophy

Atrophy estimates have also been shown to correlate with clinical variables [31]. To quantify this effect, we computed the Spearman rank-order correlation[8] between our atrophy estimates and the diagnostic groups (NC = 0, MCI = 1, AD = 2), and also between our atrophy estimates and the scores of the mini-mental state exam (MMSE). We applied the Benjamini-Hochberg procedure [36] for all the correlation results in this paper to reduce the false discovery rate for multiple comparisons. The overall false discovery rate was set to be 0.01, which resulted in an effective significance level of $\alpha \approx 0.0093$. Detailed results can be found in Table 7 and Fig. 6, respectively. In detail, for ADNI-1/2, we randomly selected 200[9] cases from each diagnostic category at each month and calculated the Spearman rank-order correlation. Fig. 6 shows the results for 50 repetitions. We observe median correlations for all four prediction models in the range of -0.36 to -0.75 for MMSE and 0.36 to 0.65 for diagnostic category. The correlations for all four prediction+correction models were in the range of -0.40 to -0.75 for MMSE and 0.36 to 0.65 for diagnostic category. Previous studies reported Pearson correlations between comparable atrophy estimates and clinical variables as high as -0.7 for MMSE and 0.5 for diagnostic category for 100 subjects[31, 32].Our two optimization-based LDDMM results achieve median correlations ranging from -0.40 to -0.76 for MMSE and 0.40 to 0.66 for diagnostic category, which is very similar to the predction+correlation models. In general, the correction+prediction FPSGR models outperform the models using only the prediction network. Further, using the correction network, FPSGR achieved comparable and sometimes even slightly better performance compared to the optimization-based LDDMM SGR method, see Table 7 for additional quantitative results. Specifically, FPSGR using the prediction+correction network performs best in 8 out

[8] We used Spearman rank-order correlation instead of Pearson correlation, because the diagnostic groups imply an ordering only.

[9] In ADNI-1 48 month, the number was 60 because there was not enough data; ADNI-2 36 month was omitted due to lack of data.

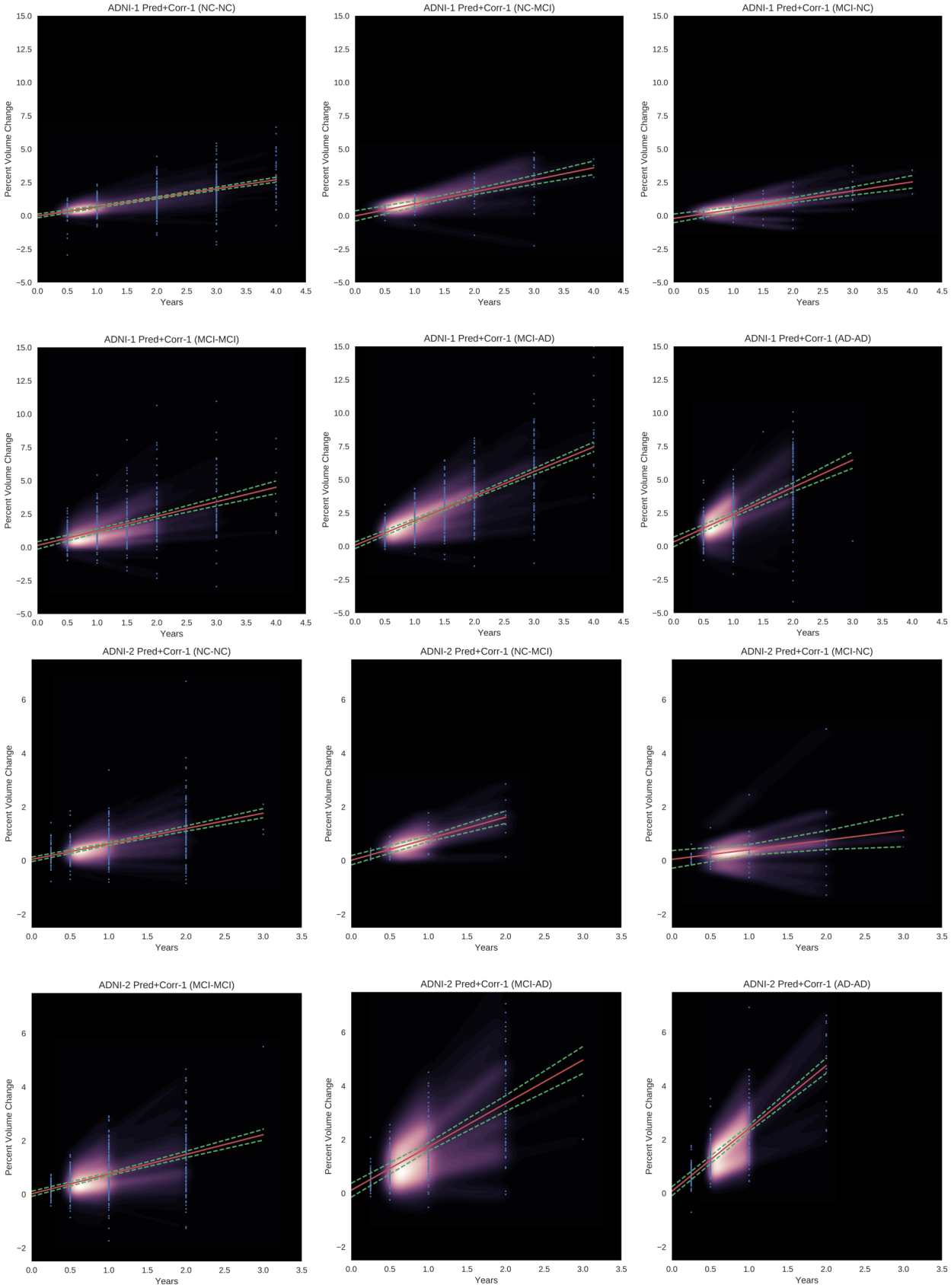

Figure 5: Linear regression of atrophy scores with respect to time for different diagnostic changes of `ADNI-1 Pred+Corr-1` and `ADNI-2 Pred+Corr-1`. Red line is the estimated regression line. green curves are the lower and upper bounds of the 95% confidence interval. Blue dots indicate actual data points. Bright white / purple images indicate kernel density estimations for all real data points illustrating dominant longitudinal trends in the data.

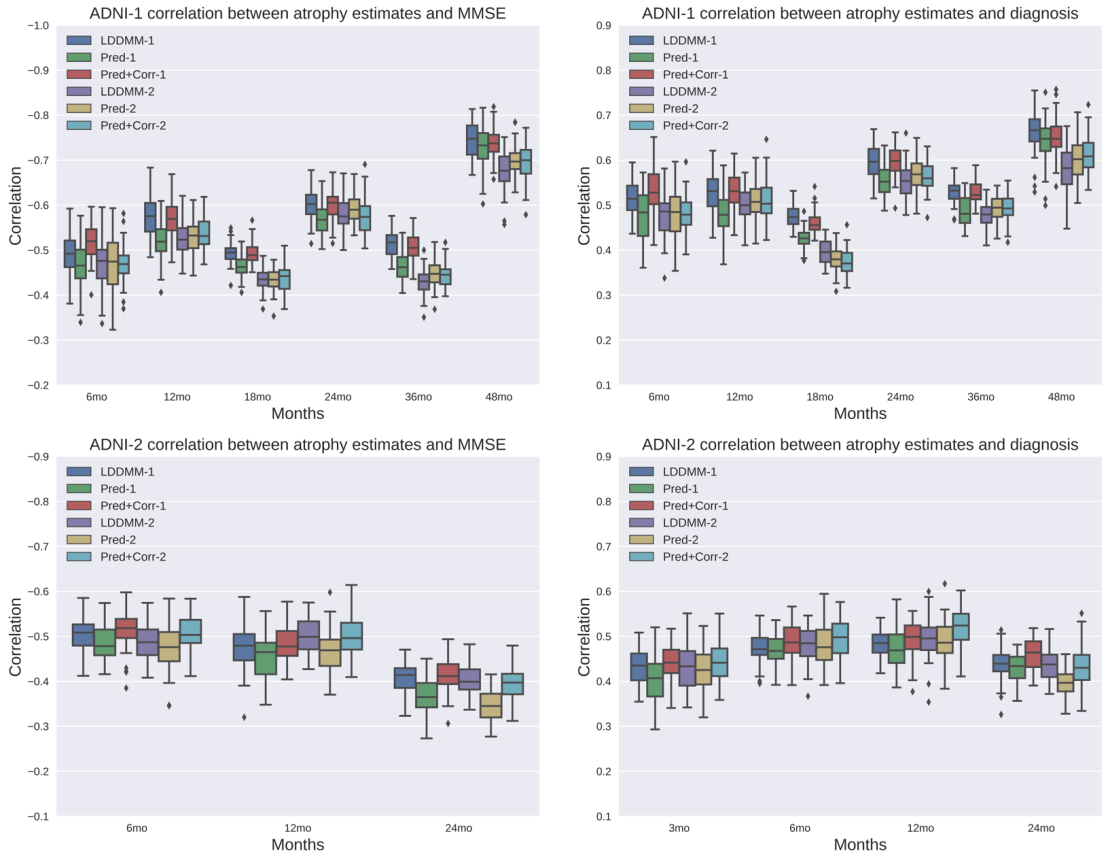

Figure 6: Boxplot of FPSGR-derived correlations with clinical variables in ADNI-1 and ADNI-2. Prediction results are comparable with optimization-based LDDMM. Adding the correction network generally improves prediction results.

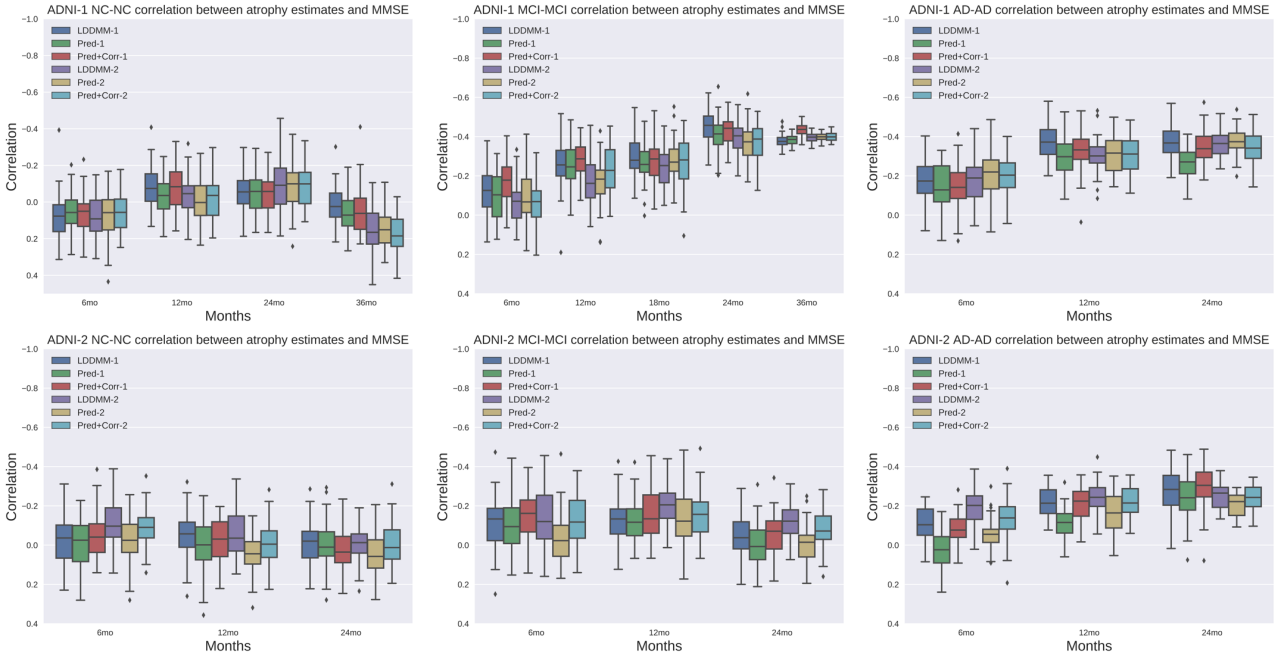

Figure 7: Boxplot of Spearman rank-order correlations between atrophy measures and MMSE with respect to time in ADNI-1 and ADNI-2. **Top row:** ADNI-1 NC-NC group (left), ADNI-1 MCI-MCI group (middle), ADNI-1 AD-AD group (right). **Bottom row:** ADNI-2 NC-NC group (left), ADNI-2 MCI-MCI group (middle), ADNI-2 AD-AD group (right). ADNI-1 MCI-MCI and ADNI-1 AD-AD show stronger correlations with time. In comparison, correlations remain relatively stable over time for the diagnostic groups in ADNI-2.

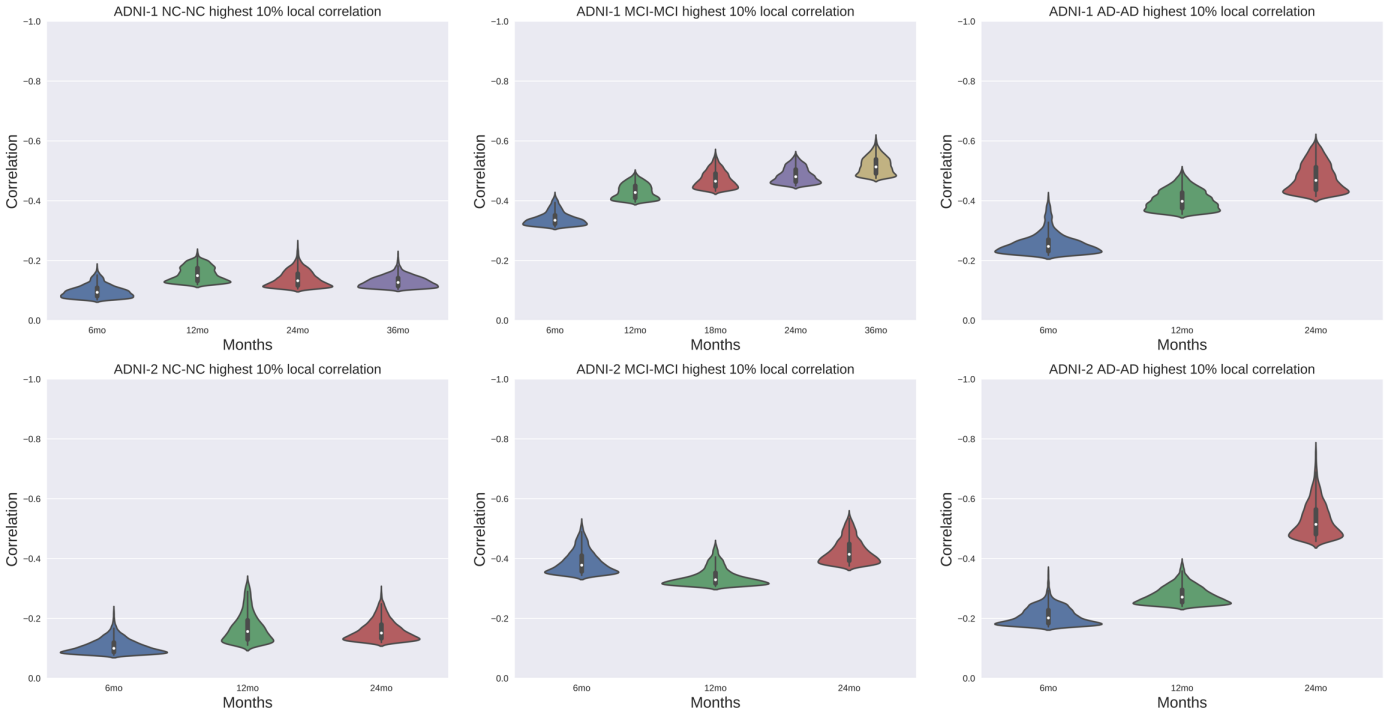

Figure 8: Kernel density estimates of highest 10% local correlations of atrophy with MMSE within the ROI depicted in Fig. 3. **Top row:** results of NC group, MCI group and AD group from ADNI-1. **Bottom row:** results of NC group, MCI group and AD group from ADNI-2. Results show a shifting pattern for the ADNI-1 MCI case, the ADNI-1 AD case and the ADNI-2 AD case.

of 18 comparisons for MMSE and in 12 out of 20 comparisons for diagnostic group. In the cases where FPSGR with prediction+correction network did not perform best its difference to the best method was generally very small. In general FPSGR using the correction network performs better than FPSGR without the correction network. To check for statistical differences in the performance of FPSGR, we use a paired t-test. Table 8 shows the resulting p-values for the three methods: optimization-based SGR (i.e., LDDMM), FPSGR without correction network (i.e., Pred) and FPSGR with correction network (i.e., Pred+Corr). In both correlation with MMSE and DX, FPSGR with correction network shows significantly better performance than LDDMM and FPSGR without correction network, which justifies the use of the FPSGR method. In summary, FPSGR captures correlations between atrophy and clinical measures well.

To further explore the correlations of atrophy with MMSE scores, we visualize them separated by diagnostic group where diagnosis did not change (i.e., NC-NC, MCI-MCI, AD-AD) in Fig. 7. For the `ADNI-1` dataset, we observe (as expected) very low correlations for the normal diagnostic group (with no clear trend), and much stronger correlations for the MCI and AD groups. MCI and AD also exhibit increasingly stronger correlations with time. In case of `ADNI-2`, the MCI group shows modest correlations, which remain consistent across time. Correlations are relatively low for the normal groups. The AD groups show increasingly strong correlations over time. In contrast to `ADNI-1`, `ADNI-2` focuses mainly on earlier stages of the diagnostic groups [34]. Hence, the deformations in `ADNI-2` are generally smaller than in `ADNI-1`. This may explain why the NC and MCI diagnostic groups show consistent correlation values over time (instead of stronger correlations as for AD in `ADNI-2` or the MCI and AD groups in `ADNI-1`).

To address the question how stat-ROI specific measures behave over time, we explore how atrophy *locally* (i.e., voxel-by-voxel) correlates with MMSE. The local atrophy is defined as

$$(1 - \det(D\phi(x))) \times 100 \ .$$

I.e., each voxel in a stat-ROI has an associated atrophy score. Fig. 8 shows kernel density estimates of the highest 10% local correlations in a violin plot. For the `ADNI-1` MCI and AD groups, a clear shift toward stronger correlations can be observed over time, consistent with the boxplots of Fig. 7. This indicates the progression of the disease. Correlations for the normal groups in `ADNI 1/2` are mostly centered around a modest correlation (as expected). In `ADNI-2`, only the AD diagnostic group shows a shift towards stronger correlations over time. All the other diagnostic groups show a relatively consistent distribution over time. This is also consistent with Fig. 7.

4.4. Justification of SGR

For simple geodesic regression to be a useful model it should outperform pairwise image registration. The

ADNI-1		Slope	Intercept	#data
NC-NC	LDDMM-1	[0.62, **0.70**, 0.78]	[-0.25,**-0.08**, 0.09]	154
	Pred-1	[0.37, **0.44**, 0.50]	[-0.21, **-0.08**, 0.05]	
	Pred+Corr-1	[0.61, **0.68**, 0.75]	[-0.15, **-0.01**, 0.13]	
	LDDMM-2	[0.57, **0.66**, 0.75]	[-0.21, **-0.04**, 0.14]	156
	Pred-2	[0.43, **0.50**, 0.57]	[-0.16, **-0.02**, 0.11]	
	Pred+Corr-2	[0.51, **0.58**, 0.65]	[-0.12, **0.01**, 0.15]	
NC-MCI	LDDMM-1	[0.72, **0.94**, 1.16]	[-0.45, **-0.03**, 0.39]	24
	Pred-1	[0.39, **0.58**, 0.78]	[-0.43, **-0.05**, 0.33]	
	Pred+Corr-1	[0.71, **0.90**, 1.10]	[-0.40, **-0.01**, 0.37]	
	LDDMM-2	[0.88, **1.19**, 1.50]	[-0.65, **-0.05**, 0.55]	22
	Pred-2	[0.72, **0.99**, 1.26]	[-0.68, **-0.16**, 0.36]	
	Pred+Corr-2	[0.80, **1.07**, 1.34]	[-0.66, **-0.14**, 0.38]	
MCI-MCI	LDDMM-1	[0.97, **1.17**, 1.38]	[-0.28, **0.05**, 0.39]	146
	Pred-1	[0.65, **0.80**, 0.96]	[-0.29, **-0.03**, 0.22]	
	Pred+Corr-1	[0.92, **1.09**, 1.26]	[-0.14, **0.14**, 0.42]	
	LDDMM-2	[0.83, **1.00**, 1.17]	[-0.21, **0.06**, 0.33]	148
	Pred-2	[0.69, **0.82**, 0.96]	[-0.20, **0.02**, 0.24]	
	Pred+Corr-2	[0.77, **0.90**, 1.04]	[-0.15, **0.07**, 0.29]	
MCI-NC	LDDMM-1	[0.48, **0.72**, 0.96]	[-0.85, **-0.42**, 0.01]	16
	Pred-1	[0.26, **0.44**, 0.62]	[-0.61, **-0.29**, 0.03]	
	Pred+Corr-1	[0.51, **0.68**, 0.86]	[-0.52, **-0.20**, 0.13]	
	LDDMM-2	[0.54, **0.79**, 1.03]	[-0.79, **-0.36**, 0.07]	17
	Pred-2	[0.40, **0.61**, 0.83]	[-0.62, **-0.24**, 0.14]	
	Pred+Corr-2	[0.49, **0.70**, 0.91]	[-0.59, **-0.21**, 0.17]	
MCI-AD	LDDMM-1	[1.94, **2.10**, 2.27]	[-0.28, **0.02**, 0.31]	148
	Pred-1	[1.28, **1.40**, 1.53]	[-0.24, **-0.02**, 0.20]	
	Pred+Corr-1	[1.70, **1.84**, 1.98]	[-0.17, **0.08**, 0.33]	
	LDDMM-2	[1.75, **1.92**, 2.09]	[-0.16, **0.14**, 0.44]	147
	Pred-2	[1.42, **1.56**, 1.70]	[-0.11, **0.14**, 0.39]	
	Pred+Corr-2	[1.49, **1.64**, 1.78]	[-0.08, **0.17**, 0.43]	
AD-AD	LDDMM-1	[1.97, **2.33**, 2.69]	[-0.17, **0.27**, 0.70]	143
	Pred-1	[1.23, **1.50**, 1.77]	[-0.13, **0.21**, 0.54]	
	Pred+Corr-1	[1.74, **2.05**, 2.35]	[-0.04, **0.33**, 0.70]	
	LDDMM-2	[1.92, **2.28**, 2.65]	[-0.20, **0.24**, 0.68]	140
	Pred-2	[1.56, **1.85**, 2.15]	[-0.13, **0.22**, 0.57]	
	Pred+Corr-2	[1.65, **1.95**, 2.24]	[-0.10, **0.25**, 0.60]	
ADNI-2		Slope	Intercept	
NC-NC	LDDMM-1	[0.55, **0.65**, 0.75]	[-0.08, **0.03**, 0.13]	170
	Pred-1	[0.41, **0.48**, 0.55]	[-0.03, **0.04**, 0.12]	
	Pred+Corr-1	[0.50, **0.57**, 0.65]	[-0.04, **0.05**, 0.13]	
	LDDMM-2	[0.51, **0.62**, 0.72]	[-0.10, **0.01**, 0.12]	175
	Pred-2	[0.47, **0.55**, 0.62]	[-0.03, **0.05**, 0.13]	
	Pred+Corr-2	[0.35, **0.44**, 0.52]	[-0.09, **-0.00**, 0.08]	
NC-MCI	LDDMM-1	[0.56, **0.79**, 1.02]	[-0.22, **0.01**, 0.25]	16
	Pred-1	[0.53, **0.68**, 0.82]	[-0.14, **0.01**, 0.16]	
	Pred+Corr-1	[0.63, **0.80**, 0.97]	[-0.16, **0.02**, 0.19]	
	LDDMM-2	[0.62, **0.90**, 1.18]	[-0.32, **-0.02**, 0.28]	17
	Pred-2	[0.58, **0.77**, 0.97]	[-0.19, **0.01**, 0.22]	
	Pred+Corr-2	[0.46, **0.68**, 0.91]	[-0.25, **-0.02**, 0.22]	
MCI-MCI	LDDMM-1	[0.71, **0.83**, 0.94]	[-0.13, **-0.00**, 0.12]	184
	Pred-1	[0.53, **0.61**, 0.68]	[-0.06, **0.02**, 0.10]	
	Pred+Corr-1	[0.64, **0.73**, 0.82]	[-0.08, **0.02**, 0.11]	
	LDDMM-2	[0.71, **0.82**, 0.92]	[-0.14, **-0.02**, 0.09]	183
	Pred-2	[0.58, **0.66**, 0.73]	[-0.05, **0.03**, 0.12]	
	Pred+Corr-2	[0.50, **0.59**, 0.67]	[-0.12, **-0.02**, 0.07]	
MCI-NC	LDDMM-1	[0.03, **0.39**, 0.74]	[-0.38, **0.05**, 0.47]	16
	Pred-1	[0.05, **0.29**, 0.52]	[-0.24, **0.05**, 0.33]	
	Pred+Corr-1	[0.08, **0.36**, 0.64]	[-0.28, **0.05**, 0.38]	
	LDDMM-2	[0.14, **0.40**, 0.67]	[-0.28, **0.04**, 0.35]	21
	Pred-2	[0.24, **0.42**, 0.61]	[-0.17, **0.05**, 0.28]	
	Pred+Corr-2	[0.05, **0.26**, 0.48]	[-0.22, **0.03**, 0.29]	
MCI-AD	LDDMM-1	[1.65, **1.95**, 2.25]	[-0.21, **0.13**, 0.47]	70
	Pred-1	[1.09, **1.27**, 1.46]	[-0.12, **0.09**, 0.30]	
	Pred+Corr-1	[1.39, **1.62**, 1.85]	[-0.15, **0.11**, 0.37]	
	LDDMM-2	[1.59, **1.91**, 2.23]	[-0.16, **0.19**, 0.53]	65
	Pred-2	[1.15, **1.35**, 1.56]	[-0.09, **0.14**, 0.36]	
	Pred+Corr-2	[1.20, **1.45**, 1.69]	[-0.13, **0.14**, 0.41]	
AD-AD	LDDMM-1	[2.49, **2.76**, 3.04]	[-0.15, **0.07**, 0.30]	101
	Pred-1	[1.74, **1.90**, 2.07]	[-0.09, **0.04**, 0.18]	
	Pred+Corr-1	[2.14, **2.34**, 2.54]	[-0.09, **0.08**, 0.24]	
	LDDMM-2	[2.72, **2.99**, 3.27]	[-0.15, **0.07**, 0.29]	103
	Pred-2	[1.97, **2.14**, 2.31]	[-0.07, **0.07**, 0.21]	
	Pred+Corr-2	[2.16, **2.36**, 2.56]	[-0.15, **0.02**, 0.18]	

Table 6: Slope and intercept values for simple linear regression of volume change over time. Our notation for *slope* and *intercept* indicate [lower bound of 95% C.I., **point estimate**, upper bound of 95% C.I.]. The interval of intercept estimates all contain zero. The slope changes between the different diagnostic groups. The #data column lists the number of data points analyzed.

main conceptual difference is that the regression model will recover an *average trend* based on multiple image time-points, i.e., the resulting regression geodesic will be a compromise between all the measurements. In contrast, for pairwise image registration (which can be seen as a trivial case of geodesic regression with two images only)

ADNI-1		MMSE	p-value	DX	p-value	#data
6mo	LDDMM-1	-0.4957	5.17e-39	0.5140	2.66e-42	608
	Pred-1	-0.4642	8.09e-34	0.4754	1.30e-35	
	Pred+Corr-1	-0.5104	1.22e-41	0.5259	1.53e-44	
	LDDMM-2	-0.4667	4.17e-34	0.4814	1.75e-36	606
	Pred-2	-0.4711	8.48e-35	0.4849	4.58e-37	
	Pred+Corr-2	-0.4734	3.54e-35	0.4890	9.67e-38	
12mo	LDDMM-1	-0.5749	5.23e-51	0.5313	1.81e-42	565
	Pred-1	-0.5328	9.46e-43	0.4898	1.97e-35	
	Pred+Corr-1	-0.5799	4.39e-52	0.5406	3.44e-44	
	LDDMM-2	-0.5301	6.81e-42	0.5055	1.17e-37	560
	Pred-2	-0.5351	9.79e-43	0.5120	1.11e-38	
	Pred+Corr-2	-0.5374	3.73e-43	0.5155	2.89e-39	
18mo	LDDMM-1	-0.4939	4.86e-16	0.4776	5.76e-15	238
	Pred-1	-0.4659	3.18e-14	0.4313	3.37e-12	
	Pred+Corr-1	-0.4924	6.16e-16	0.4643	3.98e-14	
	LDDMM-2	-0.4385	9.50e-13	0.4000	1.12e-10	241
	Pred-2	-0.4389	9.06e-13	0.3818	8.80e-10	
	Pred+Corr-2	-0.4384	9.75e-13	0.3790	1.19e-9	
24mo	LDDMM-1	-0.6064	5.01e-45	0.5978	1.69e-43	435
	Pred-1	-0.5664	2.83e-38	0.5607	2.18e-37	
	Pred+Corr-1	-0.6001	6.55e-44	0.5943	6.82e-43	
	LDDMM-2	-0.5822	4.11e-40	0.5534	1.24e-35	427
	Pred-2	-0.5911	1.41e-41	0.5714	2.26e-38	
	Pred+Corr-2	-0.5898	2.28e-41	0.5709	2.65e-38	
36mo	LDDMM-1	-0.5142	4.29e-20	0.5300	1.81e-21	277
	Pred-1	-0.4731	7.38e-17	0.4926	2.42e-18	
	Pred+Corr-1	-0.5069	1.71e-19	0.5296	1.99e-21	
	LDDMM-2	-0.4334	3.79e-13	0.4815	2.93e-16	256
	Pred-2	-0.4425	1.07e-13	0.4894	7.99e-17	
	Pred+Corr-2	-0.4393	1.67e-13	0.4863	1.34e-16	
48mo	LDDMM-1	-0.7456	2.01e-13	0.6635	5.20e-10	69
	Pred-1	-0.7294	1.18e-12	0.6458	2.08e-9	
	Pred+Corr-1	-0.7443	2.30e-13	0.6575	8.43e-10	
	LDDMM-2	-0.6889	2.25e-10	0.5927	1.98e-7	65
	Pred-2	-0.6995	9.08e-11	0.6048	9.49e-8	
	Pred+Corr-2	-0.7005	8.31e-11	0.6067	8.49e-8	
ADNI-2		MMSE	p-value	DX	p-value	#data
3mo	LDDMM-1	N/A	N/A	0.4254	2.34e-24	522
	Pred-1	N/A	N/A	0.4142	4.72e-23	
	Pred+Corr-1	N/A	N/A	0.4353	1.52e-25	
	LDDMM-2	N/A	N/A	0.4409	2.77e-26	523
	Pred-2	N/A	N/A	0.4280	1.05e-24	
	Pred+Corr-2	N/A	N/A	0.4445	9.64e-27	
6mo	LDDMM-1	-0.4989	8.01e-31	0.4688	6.09e-27	468
	Pred-1	-0.4768	6.22e-28	0.4625	3.47e-26	
	Pred+Corr-1	-0.5128	9.64e-33	0.4846	6.19e-29	
	LDDMM-2	-0.5072	4.29e-32	0.4883	1.58e-29	470
	Pred-2	-0.4718	2.02e-27	0.4742	9.96e-28	
	Pred+Corr-2	-0.5066	5.25e-32	0.4913	6.33e-30	
12mo	LDDMM-1	-0.4756	1.43e-27	0.4859	7.22e-29	464
	Pred-1	-0.4530	7.32e-25	0.4771	9.39e-28	
	Pred+Corr-1	-0.4908	1.67e-29	0.5064	1.37e-31	
	LDDMM-2	-0.4937	1.07e-29	0.5026	7.05e-31	461
	Pred-2	-0.4626	7.94e-26	0.4913	2.21e-29	
	Pred+Corr-2	-0.4987	2.35e-30	0.5149	1.44e-32	
24mo	LDDMM-1	-0.4120	9.53e-15	0.4476	2.06e-17	325
	Pred-1	-0.3670	8.51e-12	0.4331	2.71e-16	
	Pred+Corr-1	-0.4109	1.15e-14	0.4632	1.09e-18	
	LDDMM-2	-0.4095	2.09e-14	0.4375	1.93e-16	321
	Pred-2	-0.3411	3.46e-10	0.3940	2.29e-13	
	Pred+Corr-2	-0.3943	2.20e-13	0.4336	3.79e-16	
36mo	LDDMM-1	-0.2474	0.55	0.2869	0.49	8
	Pred-1	-0.2474	0.55	0.2869	0.49	
	Pred+Corr-1	-0.2474	0.55	0.2869	0.49	
	LDDMM-2	0.0935	0.83	0.1695	0.69	8
	Pred-2	0.0935	0.83	0.1695	0.69	
	Pred+Corr-2	0.0935	0.83	0.1695	0.69	

Table 7: FPSGR-derived correlations with clinical variables, compared to correlations with clinical variables for SGR using optimization-based LDDMM. The #data column lists the number of data points analyzed. Green indicates that FPSGR using the prediction+correction network shows the strongest correlations; Yellow indicates that FPSGR using the prediction network alone shows the strongest correlations; Red indicates that LDDMM SGR shows the strongest correlations. The MMSE column lists correlations between atrophy scores and the mini-mental state exam scores; the DX column lists correlations between atrophy score and diagnostic category. Finally, the p-value column(s) list the p-values for the null-hypothesis that there is no correlation. Benjamini-Hochberg procedure was employed to reduce the false discovery rate and Purple highlight indicates statistically significant. FPSGR using the prediction+correction network generally improves performance over using the prediction network alone and frequently even performs slightly better than the SGR results obtained by optimization-based LDDMM.

Normality Test			
MMSE	LDDMM	Pred	Pred+Corr
LDDMM	N/A	0.1507	0.5361
Pred	0.1507	N/A	0.0183
Pred+Corr	0.5361	0.0183	N/A
Paired t-test			
MMSE	LDDMM	Pred	Pred+Corr
LDDMM	N/A	0.0005484	0.09469173
Pred	0.9994516	N/A	0.9999718
Pred+Corr	0.0530827	0.0000282	N/A
Normality Test			
DX	LDDMM	Pred	Pred+Corr
LDDMM	N/A	0.1963	0.2356
Pred	0.1963	N/A	0.3208
Pred+Corr	0.2356	0.3208	N/A
Paired T-test			
DX	LDDMM	Pred	Pred+Corr
LDDMM	N/A	0.0010944	0.9813582
Pred	0.9989056	N/A	0.9999869
Pred+Corr	0.0186418	0.0000131	N/A

Table 8: Results of a Shapiro-Wilk normality test and a paired t-test on MMSE and DX correlations among optimization-based LDDMM, FPSGR without prediction network and FPSGR with correction network. The null-hypothesis for the Shapiro-Wilk normality test is that the difference between column-method and row-method is normally distributed. The null-hypothesis for the paired t-test is that the column-method is statistically better than row-method. Green highlighted p-values indicate no rejection of the normality hypothesis (at 5% significance) and thus facilitate the paired t-test. p-values highlighted in red indicate a rejection of the normality null-hypothesis and consequently do not allow a paired t-test.

ADNI-1		Slope	Intercept
NC-NC	SGR Pred-1	[0.37, **0.44**, 0.50]	[-0.21, **-0.08**, 0.05]
	Pairwise Pred-1	[0.44, **0.52**, 0.60]	[-0.46, **-0.30**, -0.14]
	SGR Pred-2	[0.43, **0.50**, 0.57]	[-0.16, **-0.02**, 0.11]
	Pairwise Pred-2	[0.48, **0.57**, 0.65]	[-0.34, **-0.18**, -0.01]
NC-MCI	SGR Pred-1	[0.39, **0.58**, 0.78]	[-0.43, **-0.05**, 0.33]
	Pairwise Pred-1	[0.39, **0.63**, 0.87]	[-0.63, **-0.16**, 0.30]
	SGR Pred-2	[0.72, **0.99**, 1.26]	[-0.68, **-0.16**, 0.36]
	Pairwise Pred-2	[0.65, **0.96**, 1.27]	[-0.69, **-0.10**, 0.50]
MCI-MCI	SGR Pred-1	[0.65, **0.80**, 0.96]	[-0.29, **-0.03**, 0.22]
	Pairwise Pred-1	[0.69, **0.86**, 1.03]	[-0.43, **-0.15**, 0.12]
	SGR Pred-2	[0.69, **0.82**, 0.96]	[-0.20, **0.02**, 0.24]
	Pairwise Pred-2	[0.70, **0.85**, 1.01]	[-0.29, **-0.04**, 0.21]
MCI-NC	SGR Pred-1	[0.26, **0.44**, 0.62]	[-0.61, **-0.29**, 0.03]
	Pairwise Pred-1	[0.21, **0.45**, 0.68]	[-0.74, **-0.31**, 0.12]
	SGR Pred-2	[0.40, **0.61**, 0.83]	[-0.62, **-0.24**, 0.14]
	Pairwise Pred-2	[0.29, **0.56**, 0.83]	[-0.61, **-0.14**, 0.34]
MCI-AD	SGR Pred-1	[1.28, **1.40**, 1.53]	[-0.24, **-0.02**, 0.20]
	Pairwise Pred-1	[1.28, **1.42**, 1.56]	[-0.31, **-0.06**, 0.19]
	SGR Pred-2	[1.42, **1.56**, 1.70]	[-0.11, **0.14**, 0.39]
	Pairwise Pred-2	[1.44, **1.60**, 1.75]	[-0.22, **0.06**, 0.33]
AD-AD	SGR Pred-1	[1.23, **1.50**, 1.77]	[-0.13, **0.21**, 0.54]
	Pairwise Pred-1	[1.25, **1.55**, 1.85]	[-0.23, **0.13**, 0.49]
	SGR Pred-2	[1.56, **1.85**, 2.15]	[-0.13, **0.22**, 0.57]
	Pairwise Pred-2	[1.53, **1.85**, 2.16]	[-0.15, **0.23**, 0.60]
ADNI-2		Slope	Intercept
NC-NC	SGR Pred-1	[0.41, **0.48**, 0.55]	[-0.03, **0.04**, 0.12]
	Pairwise Pred-1	[0.25, **0.33**, 0.41]	[0.15, **0.24**, 0.33]
	SGR Pred-2	[0.47, **0.55**, 0.62]	[-0.03, **0.05**, 0.13]
	Pairwise Pred-2	[0.26, **0.35**, 0.44]	[0.22, **0.32**, 0.43]
NC-MCI	SGR Pred-1	[0.53, **0.68**, 0.82]	[-0.14, **0.01**, 0.16]
	Pairwise Pred-1	[0.37, **0.57**, 0.77]	[-0.06, **0.14**, 0.33]
	SGR Pred-2	[0.58, **0.77**, 0.97]	[-0.19, **0.01**, 0.22]
	Pairwise Pred-2	[0.42, **0.65**, 0.88]	[-0.07, **0.18**, 0.42]
MCI-MCI	SGR Pred-1	[0.53, **0.61**, 0.68]	[-0.06, **0.02**, 0.10]
	Pairwise Pred-1	[0.43, **0.52**, 0.61]	[0.04, **0.14**, 0.23]
	SGR Pred-2	[0.58, **0.66**, 0.73]	[-0.05, **0.03**, 0.12]
	Pairwise Pred-2	[0.45, **0.54**, 0.63]	[0.09, **0.19**, 0.29]
MCI-NC	SGR Pred-1	[0.05, **0.29**, 0.52]	[-0.24, **0.05**, 0.33]
	Pairwise Pred-1	[-0.10, **0.17**, 0.45]	[-0.12, **0.21**, 0.53]
	SGR Pred-2	[0.24, **0.42**, 0.61]	[-0.17, **0.05**, 0.28]
	Pairwise Pred-2	[0.03, **0.26**, 0.49]	[0.02, **0.29**, 0.57]
MCI-AD	SGR Pred-1	[1.09, **1.27**, 1.46]	[-0.12, **0.09**, 0.30]
	Pairwise Pred-1	[0.88, **1.10**, 1.32]	[0.08, **0.33**, 0.58]
	SGR Pred-2	[1.15, **1.35**, 1.56]	[-0.09, **0.14**, 0.36]
	Pairwise Pred-2	[0.89, **1.13**, 1.37]	[0.18, **0.44**, 0.70]
AD-AD	SGR Pred-1	[1.74, **1.90**, 2.07]	[-0.09, **0.04**, 0.18]
	Pairwise Pred-1	[1.57, **1.77**, 1.96]	[0.01, **0.17**, 0.34]
	SGR Pred-2	[1.97, **2.14**, 2.31]	[-0.07, **0.07**, 0.21]
	Pairwise Pred-2	[1.79, **1.99**, 2.19]	[0.05, **0.21**, 0.37]

Table 9: SGR prediction model compared with a pairwise prediction model. Slope and intercept values for simple linear regression of volume change over time. The notation for slope and intercept columns indicates [Lower bound of 95% C.I., **point estimate**, Upper bound of 95% C.I.]. Green indicates that the intercept is closer to zero (also, zero is within the 95% confidence interval) for SGR prediction model; Yellow indicates that the intercept is closer to zero for pairwise prediction model; Red indicates that the point estimate is either biased to overestimate or underestimate volume change. The SGR prediction model performs better than the pairwise prediction model.

the deformation will in general be able to match the target image well. However, just as in linear regression, this may accentuate the effects of noise. In both setups, images can be interpolated or extrapolated based on the estimated geodesic.

Tables 9 and 10 justify the use of SGR. Specifically, Table 9 shows linear regression results of atrophy measures over time as obtained via SGR (i.e., using an SGR fit over all time-points followed by atrophy computations based on the deformations of the regression geodesic) compared with atrophy measures obtained by pairwise registration. For both the `ADNI-1` and the `ADNI-2` datasets, SGR outperforms the pairwise registration approach in two aspects: (1) the estimated intercept of SGR is generally closer to zero than for the pairwise method and the intercept 95% confidence interval is narrower; (2) 11 out of 24 of the 95% confidence intervals of the pairwise methods show bias to either overestimate or underestimate volume change, while none of the SGR results show such significant bias. Table 10 compares the correlations between atrophy and clinical measures (MMSE and diagnostic category) of SGR and the pairwise approach. SGR performs better than the pairwise approach in 13 out of 18 cases for MMSE and in 15 out of 20 cases for the diagnostic category. Furthermore, when the pairwise method is better than SGR, the difference is much smaller compared to the differences observed for the cases where SGR is better than the pairwise method. Also note that the pairwise method shows better performance in later months compared to earlier months. This could, for example, be because the deformations are larger for later time-points and hence the registration result becomes more stable, or because SGR is also heavily influenced by the last time-point. To address the above observation, we used a Shapiro-Wilk normality test and a Wilcoxon signed-rank test. From Table 11 we see that we can reject the null-hypothesis of normality and hence, a paired t-test is not appropriate. As an

ADNI-1		MMSE	p-value	DX	p-value	#data
6mo	SGR Pred-1	-0.4642	8.09e-34	0.4754	1.30e-35	608
	Pairwise Pred-1	-0.3138	2.31e-15	0.3369	1.32e-17	
	SGR Pred-2	-0.4711	8.48e-35	0.4849	4.58e-37	606
	Pairwise Pred-2	-0.3431	3.51e-18	0.3680	7.24e-21	
12mo	SGR Pred-1	-0.5328	9.46e-43	0.4898	1.97e-35	565
	Pairwise Pred-1	-0.4393	4.67e-28	0.3996	4.51e-23	
	SGR Pred-2	-0.5351	9.79e-43	0.5120	1.11e-38	560
	Pairwise Pred-2	-0.4465	9.61e-29	0.4154	1.00e-24	
18mo	SGR Pred-1	-0.4659	3.18e-14	0.4313	3.37e-12	238
	Pairwise Pred-1	-0.4164	2.12e-11	0.3882	5.56e-10	
	SGR Pred-2	-0.4389	9.06e-13	0.3818	8.80e-10	241
	Pairwise Pred-2	-0.4078	4.52e-11	0.3356	9.38e-8	
24mo	SGR Pred-1	-0.5664	2.83e-38	0.5607	2.18e-37	435
	Pairwise Pred-1	-0.5805	1.51e-40	0.5791	2.55e-40	
	SGR Pred-2	-0.5911	1.41e-41	0.5714	2.26e-38	427
	Pairwise Pred-2	-0.5927	7.34e-42	0.5811	6.26e-40	
36mo	SGR Pred-1	-0.4731	7.38e-17	0.4926	2.42e-18	277
	Pairwise Pred-1	-0.4470	5.20e-15	0.4798	2.36e-17	
	SGR Pred-2	-0.4425	1.07e-13	0.4894	7.99e-17	256
	Pairwise Pred-2	-0.4538	2.08e-14	0.4990	1.59e-17	
48mo	SGR Pred-1	-0.7294	1.18e-12	0.6458	2.08e-9	69
	Pairwise Pred-1	-0.7100	8.43e-12	0.6168	1.67e-8	
	SGR Pred-2	-0.6995	9.08e-11	0.6048	9.49e-8	65
	Pairwise Pred-2	-0.6709	9.65e-10	0.5924	2.01e-7	
ADNI-2		MMSE	p-value	DX	p-value	#data
3mo	SGR Pred-1	N/A	N/A	0.4142	4.72e-23	522
	Pairwise Pred-1	N/A	N/A	0.1744	6.17e-5	
	SGR Pred-2	N/A	N/A	0.4280	1.05e-24	523
	Pairwise Pred-2	N/A	N/A	0.1503	5.64e-4	
6mo	SGR Pred-1	-0.4768	6.22e-28	0.4625	3.47e-26	468
	Pairwise Pred-1	-0.3378	5.93e-14	0.2633	7.29e-9	
	SGR Pred-2	-0.4718	2.02e-27	0.4742	9.96e-28	470
	Pairwise Pred-2	-0.3312	1.70e-13	0.2849	3.14e-10	
12mo	SGR Pred-1	-0.4530	7.32e-25	0.4771	9.39e-28	464
	Pairwise Pred-1	-0.4305	2.34e-22	0.4472	3.40e-24	
	SGR Pred-2	-0.4626	7.94e-26	0.4913	2.21e-29	461
	Pairwise Pred-2	-0.4223	2.30e-21	0.4374	5.72e-23	
24mo	SGR Pred-1	-0.3670	8.51e-12	0.4331	2.71e-16	325
	Pairwise Pred-1	-0.3772	3.06e-12	0.4515	9.99e-18	
	SGR Pred-2	-0.3411	3.46e-10	0.3940	2.29e-13	321
	Pairwise Pred-2	-0.3517	8.89e-11	0.4239	1.99e-15	
36mo	SGR Pred-1	-0.2474	0.55	0.4536	0.26	8
	Pairwise Pred-1	-0.1650	0.70	0.2869	0.49	
	SGR Pred-2	0.0935	0.83	0.1695	0.69	8
	Pairwise Pred-2	0.0935	0.83	0.2608	0.53	

Table 10: SGR prediction model compared with pairwise prediction model. Results show correlations with clinical variables. The #data column lists the number of data points analyzed. Green indicates a stronger correlation for the SGR prediction method; Yellow indicates a stronger correlation for the pairwise model. The p-value column lists p-values for the null-hypothesis that there is no correlation. The Benjamini-Hochberg procedure was employed to reduce the false discovery rate (FDR). The Purple highlight indicates statistically significant results after correction for multiple comparisons. In general, SGR prediction performs better than pairwise prediction demonstrating that regression stabilizes the correlation results. ADNI-2 36mo only has 8 data points and the p-value is greater than 0.1, thus we ignore this month in our comparison.

	Shapiro-Wilk normality test	Wilcoxon signed-rank test
MMSE	0.01943	0.0005226
DX	0.03286	0.0005083

Table 11: p-values for a Shapiro-Wilk normality test and Wilcoxon signed-rank test on MMSE and DX correlations between the SGR prediction model and the pairwise prediction model. The null-hypothesis for the Shapiro-Wilk normality test is that the difference of two methods is normally distributed (at a significance level of 5%). The null-hypothesis for the Wilcoxon signed-rank test is that the pairwise prediction method is statistically better than the SGR prediction method (at a significance level of 5%).

alternative, we conducted a Wilcoxon signed-rank test to compare the SGR prediction model and the pairwise prediction model. Table 11 shows that the SGR prediction model is statistically significantly better than the pairwise prediction model. Based on the above points, we conclude that SGR is more stable over time than the pairwise

ADNI-1			MMSE	p-value	DX	p-value	#data
60mo	Forecast	LDDMM-1	-0.5242	1.34e-13	0.5157	3.85e-13	173
		Pred-1	-0.4727	5.16e-11	0.4816	1.98e-11	
		Pred+Corr-1	-0.5193	2.48e-13	0.5240	1.38e-13	
		LDDMM-2	-0.4501	2.32e-10	0.4761	1.43e-11	180
		Pred-2	-0.4527	1.77e-10	0.4620	6.63e-11	
		Pred+Corr-2	-0.4582	9.97e-11	0.4652	4.73e-11	
72mo	Forecast	LDDMM-1	-0.4607	1.60e-10	0.4507	4.37e-10	174
		Pred-1	-0.4132	1.45e-8	0.4364	1.75e-9	
		Pred+Corr-1	-0.4615	1.47e-10	0.4667	8.52e-11	
		LDDMM-2	-0.3662	3.18e-7	0.4233	2.15e-9	184
		Pred-2	-0.3793	1.09e-7	0.4273	1.46e-9	
		Pred+Corr-2	-0.3793	1.09e-7	0.4259	1.67e-9	
84mo	Forecast	LDDMM-1	-0.3986	1.40e-6	0.4108	6.17e-7	137
		Pred-1	-0.3495	2.84e-5	0.4018	1.13e-6	
		Pred+Corr-1	-0.3946	1.83e-6	0.4211	2.98e-7	
		LDDMM-2	-0.3293	4.65e-5	0.3622	6.53e-6	147
		Pred-2	-0.3199	7.81e-5	0.3629	6.25e-6	
		Pred+Corr-2	-0.3187	8.35e-5	0.3609	7.12e-6	

Table 12: Correlations of forecasting results. The #data column lists the number of data points analyzed. Green indicates that FPSGR using the prediction+correction network shows the strongest correlations; Yellow indicates that FPSGR using the prediction network alone shows the strongest correlations; Red indicates that LDDMM SGR shows the strongest correlations. The Benjamini-Hochberg procedure was employed to reduce the false discovery rate (FDR). The Purple highlight indicates statistically significant results after correction for multiple comparisons.

ADNI-1			MMSE	p-value	DX	p-value	#data
36mo	Original	LDDMM-1	-0.5142	4.29e-20	0.5300	1.81e-21	277
		Pred-1	-0.4731	7.38e-17	0.4926	2.42e-18	
		Pred+Corr-1	-0.5069	1.71e-19	0.5296	1.99e-21	
	Forecast	Pred-1	-0.4583	1.09e-15	0.4825	1.93e-17	
		Pred+Corr-1	-0.4708	1.42e-16	0.4980	1.21e-18	
	Replace	Pred-1	-0.4923	3.43e-18	0.5104	1.21e-19	
		Pred+Corr-1	-0.5097	1.37e-19	0.5375	5.47e-22	
	Original	LDDMM-2	-0.4334	3.79e-13	0.4815	2.93e-16	256
		Pred-2	-0.4425	1.07e-13	0.4894	7.99e-17	
		Pred+Corr-2	-0.4393	1.67e-13	0.4863	1.34e-16	
	Forecast	Pred-2	-0.4078	1.36e-11	0.4398	1.95e-13	
		Pred+Corr-2	-0.4005	3.34e-11	0.4301	7.40e-13	
	Replace	Pred-2	-0.4202	2.75e-12	0.4635	6.27e-15	
		Pred+Corr-2	-0.4164	4.51e-12	0.4582	1.38e-14	
48mo	Original	LDDMM-1	-0.7456	2.01e-13	0.6635	5.20e-10	69
		Pred-1	-0.7294	1.18e-12	0.6458	2.08e-9	
		Pred+Corr-1	-0.7443	2.30e-13	0.6575	8.43e-10	
	Forecast	Pred-1	-0.6332	5.29e-9	0.6165	1.70e-8	
		Pred+Corr-1	-0.6541	1.10e-9	0.6317	5.86e-9	
	Replace	Pred-1	-0.6446	2.27e-9	0.6478	1.78e-9	
		Pred+Corr-1	-0.6668	3.98e-10	0.6800	1.31e-10	
	Original	LDDMM-2	-0.6889	2.25e-10	0.5927	1.98e-7	65
		Pred-2	-0.6995	9.08e-11	0.6048	9.49e-8	
		Pred+Corr-2	-0.7005	8.31e-11	0.6067	8.49e-8	
	Forecast	Pred-2	-0.6528	3.79e-9	0.5568	1.46e-6	
		Pred+Corr-2	-0.6403	9.25e-9	0.5460	2.55e-6	
	Replace	Pred-2	-0.6334	1.49e-8	0.5970	1.53e-7	
		Pred+Corr-2	-0.6307	1.79e-8	0.5973	1.50e-7	

Table 13: Forecast results compared with real data results. The #data column lists the number of data points analyzed. The Benjamini-Hochberg procedure was employed to reduce the false discovery rate (FDR). Purple highlight indicates statistically significant results after corrections for multiple comparisons. Forecast results are calculated by using SGR excluding 36mo and 48mo data points and then predicting 36mo and 48mo correlations. Results are compared based on the same dataset except for two invalid data points for the 36mo data.

method and in general also results in stronger correlations.

4.5. Forecasting

Another interesting question for SGR and geodesic regression in general is the suitability of the model for the data. To address this question, we evaluate if SGR can *forecast* unseen future time-points. Specifically we consider this question in two different scenarios:

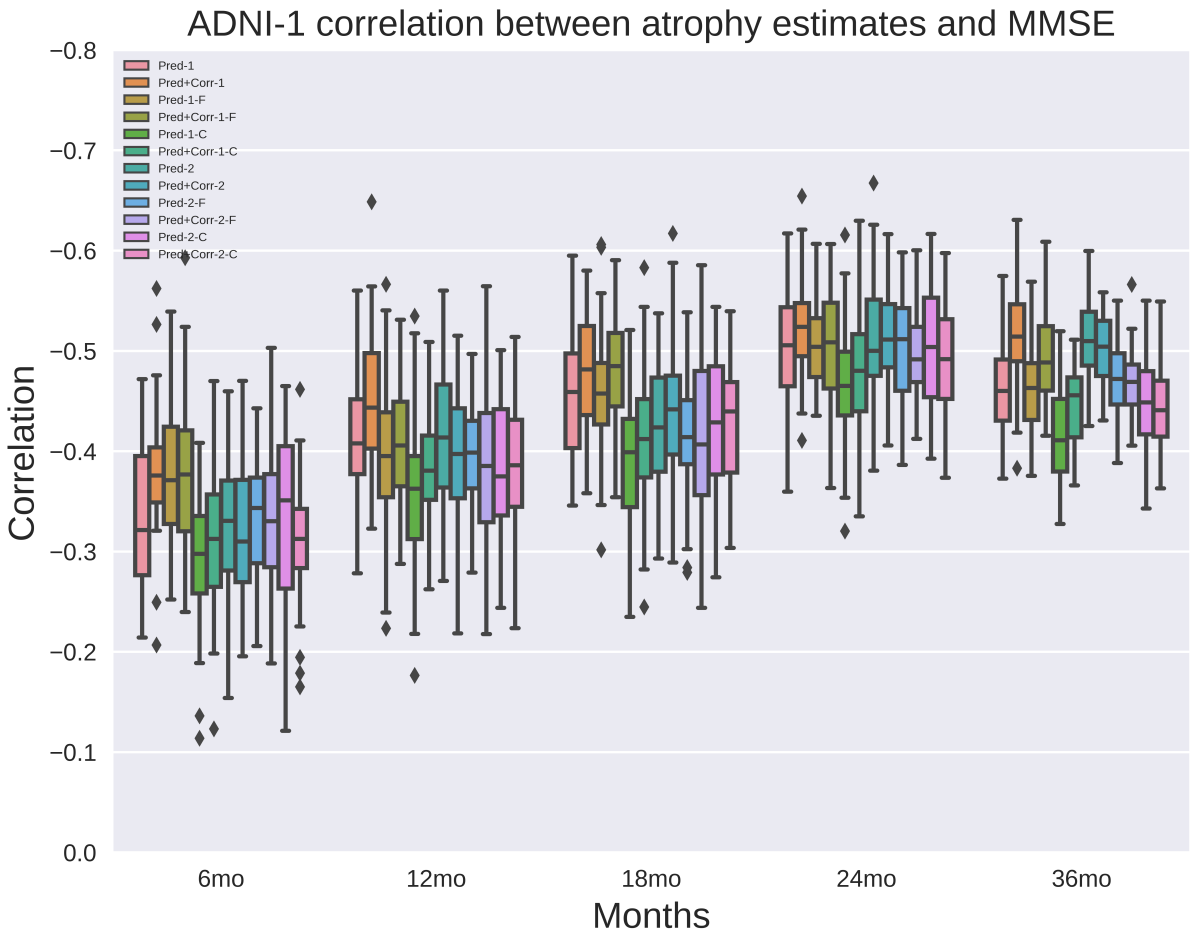

Figure 9: Comparison of correlations among prediction results, **Forecast** results (-F) and **Replace** results (-C) in MCI converter groups (MCI-NC, MCI-MCI and MCI-AD). In Pred-1, Forecast results outperform Replace results; In Pred-2, Forecast results and Replace results are comparable.

Q1) **Extrapolate-clinical:** Can we extrapolate the SGR results into the future (to time-points that do not exist in the ADNI image dataset, but for the clinical data) while still obtaining strong correlations.

Q2) **Extrapolate-image:** How well can correlations between atrophy and clinical measures be predicted for time-points when we do or do not use image data at that very time-point. We artificially leave out image measurements so that we can compare prediction results to results when we have the image measurement.

For both scenarios we use two different forecasting approaches. In the first approach (**Forecast**) we simply compute SGR results with the available image time-points and then extrapolate using the resulting regression geodesic to the desired time-point in the future. In the second approach (**Replace**), we artificially impute the missing image time-points by simply replacing them by the image at the closest measured time-point. For example, if we have images at 6, 12, and 18 month, but we want to forecast at 24 month, we use the 18 month image as the imputed 24 month image and then perform SGR on the 6, 12, 18, and the imputed 24 month images. We then obtain the deformation at 24 months from the SGR result.

ad Q1. Table 12 shows correlations between atrophy and the clinical measures for the **Forecast** results for 60 month, 72 month and 84 month.The resulting correlations of atrophy with diagnostic category are all above 0.3 (or below -0.3). Furthermore, the **Forecast** correlations show a downward trend with respect to time, which means that the prediction of "far-away" points is not as accurate as for the "near" future. On the other hand, SGR using the 6 month to 48 month time points results in correlation around -0.5 for MMSE and 0.5 for DX on average. Hence the correlation with the diagnostic category is consistent for that of 60 months. In other words, using 6 month to 48 month data, our prediction model can predict accu-

rately up to 60 month. Our prediction+correction network performs as well as and even slightly better than SGR using optimization-based LDDMM. Fig. 10 shows that these forecasting results capture the trends of the changes in the temporal lobes near the hippocampus and changes in the ventricles.

ad Q2. Table 13 and Fig. 9 show **Forecast** and **Replace** results for correlations between atrophy and clinical measures in comparison to using all images. Specifically, for the **Forecast** and **Replace** results we did not use the available images at 36 and 48 month so we could compare against the results obtained when using these images. If FPSGR is a good model, it should results in correlation results as close to the correlation results using all images as possible. The **Forecast** correlations are only slightly weaker (0.02 to 0.05 lower) than the original correlations using all images illustrating that FPSGR can approximately forecast future changes.

The overall correlations in Table 13 show that the **Replace** group performs better than the **Forecast** group. In particular, we are also interested in the prediction of MCI converters, namely, MCI to NC, MCI to MCI, and MCI to AD. The boxplots in Fig. 9 show the correlations for such predictions. The **Replace** group in Fig. 9 show relatively worse correlation performance than the **Forecast** group in ADNI-1 Pred-1 and consistent performance in ADNI-1 Pred-2. Hence SGR on a longitudinal image data can achieves good forecasting result for MCI converters.

Thus, both **Extrapolate-clinical** and **Extrapolate-image** experiments justify the use of FPSGR in predicting near future longitudinal trends especially for MCI converters.

4.6. Jacobian Determinant (JD)

The average JD images qualitatively agree with prior results [33, 34]: severity of volume change increases with severity of diagnosis and time. Change is most substantial in the temporal lobes near the hippocampus (see Fig. 10). In Fig. 10, 6 month to 48 month are existing data points; 60 month to 84 month are forecast results. Blue indicates volume loss. Red indicates expansion. Results are consistent with expectations: volume loss increases with time and severity of diagnosis in temporal lobes; volume expansion increases with respect to time and severity of diagnosis around the ventricles / cerebrospinal fluid. The forecast results capture visually sensible volume loss or expansion over time, qualitatively illustrating the performance of our method.

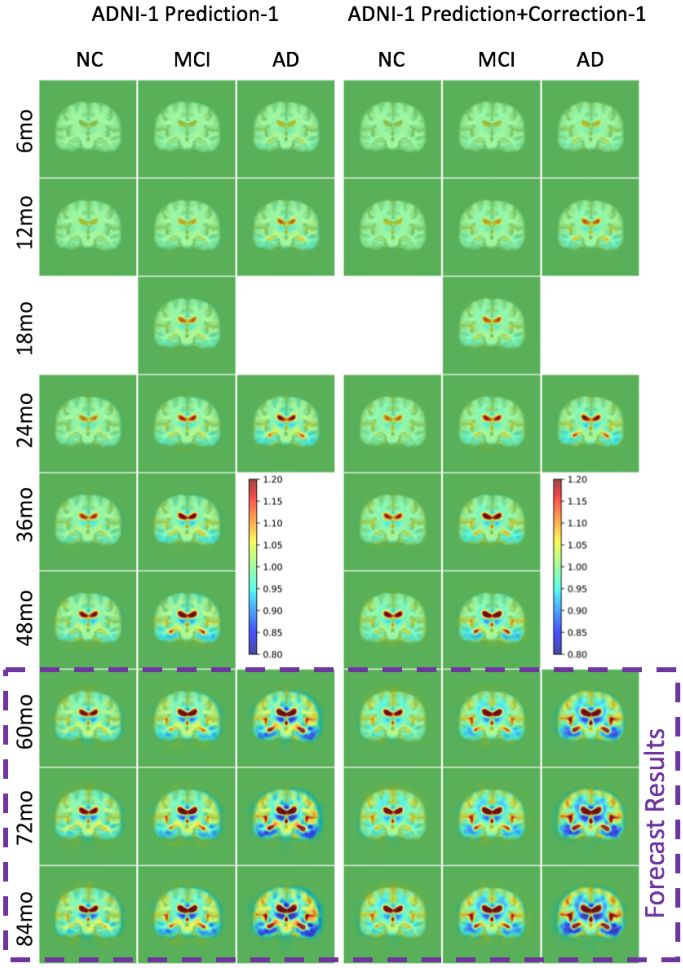

Figure 10: Average Jacobian determinant over time and diagnostic category for ADNI-1 Prediction-1 and ADNI-1 Prediction+Correction-1 (experiments in ADNI-2 show similar results). A value < 1 means shrinkage and value > 1 means expansion. The 60 month - 84 month results contained in the purple rectangle are forecasts using the data from 6 month - 48 month. Results show consistent volume loss over time near the temporal lobes and expansion over time near the ventricles/cerebrospinal fluid.

5. Conclusion and future work

In this work, we proposed a fast approach for geodesic regression (FPSGR) to study longitudinal image data. FPSGR incorporates the recently proposed FPIR [19, 20] into the SGR [15] framework, thus leading to a computationally efficient solution to geodesic regression. Since FPSGR replaces the computationally intensive intermediate step of computing pairwise initial momenta via a deep-learning prediction method, it is orders of magnitude faster than existing approaches [15, 28], without compromising accuracy. Consequently, FPSGR facilitates the analysis of large-scale imaging studies. Experiments on the `ADNI-1/ADNI-2` datasets demonstrate that FPSGR captures expected atrophy trends of normal aging, MCI and AD. It further (1) exhibits negligible bias towards volume changes within stat-ROIs, (2)

shows high correlations with clinical variables (MMSE and diagnosis) and (3) produces consistent forecasting results on unseen data.

In future work, it will be be interesting to explore FPSGR for the task of *classifying* stable Mild Cognitive Impairment (sMCI) and progressive Mild Cognitive Impairment (pMCI). Currently, FPSGR only shows modest accuracy for distinguishing these types within MCI. Extending our approach to time-warped geodesic regression models [9] might improve the accuracy in this context. Furthermore, end-to-end prediction of averaged initial momenta would be an interesting future direction, as this would allow *learning* representations that characterize the geodesic path among multiple time-series images, not only based on averages of momenta for two images as in FPIR [19, 20].

Acknowledgements

Research reported in this publication was supported by the National Institutes of Health (NIH) and the National Science Foundation (NSF) under award numbers NIH R01AR072013, NSF ECCS-1148870, and EECS-1711776. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the NSF. We also thank Nvidia for the donation of a TitanX GPU.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] C. R. Jack, J. Barnes, M. A. Bernstein, B. J. Borowski, J. Brewer, S. Clegg, A. M. Dale, O. Carmichael, C. Ching, C. DeCarli, et al., Magnetic resonance imaging in ADNI 2, Alzheimer's & Dementia 11 (7) (2015) 740–756.

[2] M. A. Ikram, A. van der Lugt, W. J. Niessen, P. J. Koudstaal, G. P. Krestin, A. Hofman, D. Bos, M. W. Vernooij, The Rotterdam scan study: design update 2016 and main findings, European journal of epidemiology 30 (12) (2015) 1299–1315.

[3] Biobank website: www.ukbiobank.ac.uk.

[4] M. Niethammer, Y. Huang, F.-X. Vialard, Geodesic regression for image time-series, in: MICCAI, 2011, pp. 655–662.

[5] Y. Hong, Y. Shi, M. Styner, M. Sanchez, M. Niethammer, Simple geodesic regression for image time-series., in: WBIR, Vol. 12, Springer, 2012, pp. 11–20.

[6] Y. Hong, S. Joshi, M. Sanchez, M. Styner, M. Niethammer, Metamorphic geodesic regression, Medical Image Computing and Computer-Assisted Intervention–MICCAI (2012) 197–205.

[7] N. Singh, J. Hinkle, S. Joshi, P. T. Fletcher, A vector momenta formulation of diffeomorphisms for improved geodesic regression and atlas construction, in: ISBI, 2013, pp. 1219–1222.

[8] P. T. Fletcher, Geodesic regression and the theory of least squares on Riemannian manifolds, IJCV 105 (2) (2013) 171–185.

[9] Y. Hong, N. Singh, R. Kwitt, M. Niethammer, Time-warped geodesic regression, in: Medical image computing and computer-assisted intervention–MICCAI, Vol. 17, 2014, p. 105.

[10] N. Singh, M. Niethammer, Splines for diffeomorphic image regression, in: Medical image computing and computer-assisted intervention – MICCAI, Vol. 17, 2014, p. 121.

[11] Y. Hong, R. Kwitt, N. Singh, B. Davis, N. Vasconcelos, M. Niethammer, Geodesic regression on the Grassmannian, in: European Conference on Computer Vision, Springer, 2014, pp. 632–646.

[12] N. Singh, F.-X. Vialard, M. Niethammer, Splines for diffeomorphisms, Medical image analysis 25 (1) (2015) 56–71.

[13] Y. Hong, R. Kwitt, N. Singh, N. Vasconcelos, M. Niethammer, Parametric regression on the Grassmannian, IEEE transactions on pattern analysis and machine intelligence 38 (11) (2016) 2284–2297.

[14] M. F. Beg, M. I. Miller, A. Trouvé, L. Younes, Computing large deformation metric mappings via geodesic flows of diffeomorphisms, International journal of computer vision 61 (2) (2005) 139–157.

[15] Y. Hong, Y. Shi, M. Styner, M. Sanchez, M. Niethammer, Simple geodesic regression for image time-series, in: WBIR, 2012, pp. 11–20.

[16] X. Cao, J. Yang, J. Zhang, D. Nie, M. Kim, Q. Wang, D. Shen, Deformable image registration based on similarity-steered CNN regression, in: MICCAI, Springer, 2017.

[17] S. Miao, Z. J. Wang, Y. Zheng, R. Liao, Real-time 2D/3D registration via CNN regression, in: Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on, IEEE, 2016, pp. 1430–1434.

[18] H. Sokooti, B. d. Vos, F. Berendsen, B. P. Lelieveldt, I. Isgum, M. Staring, Nonrigid image registration using multi-scale 3D convolutional neural networks, in: MICCAI, Springer, 2017.

[19] X. Yang, R. Kwitt, M. Niethammer, Fast predictive image registration, in: International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Springer, 2016, pp. 48–57.

[20] X. Yang, R. Kwitt, M. Styner, M. Niethammer, Quicksilver: Fast predictive image registration–a deep learning approach, NeuroImage 158 (2017) 378–396.

[21] M. Zhang, R. Liao, A. V. Dalca, E. A. Turk, J. Luo, P. E. Grant, P. Golland, Frequency diffeomorphisms for efficient image regis-

tration, in: International Conference on Information Processing in Medical Imaging, Springer, 2017, pp. 559–570.
[22] M. Zhang, P. T. Fletcher, Finite-dimensional Lie algebras for fast diffeomorphic image registration, in: IPMI, 2015, pp. 249–260.
[23] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, et al., Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration, Neuroimage 46 (3) (2009) 786–802.
[24] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, T. Brox, Flownet: Learning optical flow with convolutional networks, in: ICCV, 2015, pp. 2758–2766.
[25] Z. Liu, R. Yeh, X. Tang, Y. Liu, A. Agarwala, Video frame synthesis using deep voxel flow, arXiv preprint arXiv:1702.02463.
[26] T. Schuster, L. Wolf, D. Gadot, Optical flow requires multiple strategies (but only one network), arXiv preprint arXiv:1611.05607.
[27] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, I. Išgum, End-to-end unsupervised deformable image registration with a convolutional neural network, arXiv preprint arXiv:1704.06065.
[28] Y. Hong, P. Golland, M. Zhang, Fast geodesic regression for population-based image analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 317–325.
[29] Z. Ding, G. Fleishman, X. Yang, P. Thompson, R. Kwitt, M. Niethammer, A. D. N. Initiative, et al., Fast predictive simple geodesic regression, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2017, pp. 267–275.
[30] D. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.
[31] G. Fleishman, P. M. Thompson, The impact of matching functional on atrophy measurement from geodesic shooting in diffeomorphisms, in: ISBI, 2017.
[32] G. Fleishman, P. M. Thompson, Adaptive gradient descent optimization of initial momenta for geodesic shooting in diffeomorphisms, in: ISBI, 2017.
[33] X. Hua, D. P. Hibar, C. R. K. Ching, C. P. Boyle, P. Rajagopalan, B. Gutman, A. D. Leow, A. W. Toga, C. R. J. Jr., D. J. Harvey, M. W. Weiner, P. M. Thompson, Unbiased tensor-based morphometry:improved robustness & sample size estimates for Alzheimer's disease clinical trials, NeuroImage 66 (2013) 648–661.
[34] X. Hua, C. R. K. Ching, A. Mezher, B. Gutman, D. P. Hibar, P. Bhatt, A. D. Leow, C. R. J. Jr., M. Bernstein, M. W. Weiner, P. M. Thompson, MRI-based brain atrophy rates in ADNI phase 2: acceleration and enrichment considerations for clinical trials, Neurobiology of Aging 37 (2016) 26–37.
[35] P. A. Yushkevich, B. B. Avants, S. R. Das, J. Pluta, M. Altinay, C. Craige, Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: An illustration in ADNI 3T MRI data, NeuroImage 50 (2) (2010) 434 – 445.
[36] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the royal statistical society. Series B (Methodological) (1995) 289–300.