Medical Image Analysis 56 (2019) 193–209

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

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,1], The Alzheimer's Disease Neuroimaging Initiative

[a] Department of Computer Science, University of North Carolina at Chapel Hill, USA 201 S. Columbia St., Chapel Hill, NC 27599, USA
[b] Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, USA 125 Mason Farm Road, Chapel Hill, NC 27599, USA
[c] Imaging Genetics Center, University of Southern California, USA 2001 N. Soto Street, SSB1-102, Los Angeles, CA 90032, USA
[d] Department of Radiology, University of Pennsylvania, USA 3400 Civic Center Boulevard Atrium, Ground Floor, Philadelphia, PA 19104, USA
[e] Department of Computer Science, University of Salzburg, Austria Jakob Haringer Strasse 2, 5020 Salzburg, Austria

ARTICLE INFO

Article history:
Received 23 March 2018
Revised 31 May 2019
Accepted 11 June 2019
Available online 12 June 2019

Keywords:
Fast prediction
Image regression
ADNI dataset
Longitudinal data

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.

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 (Jack et al., 2015) and the Rotterdam study (Ikram et al., 2015), involve analyzing thousands of images. In fact, even larger studies will be available in the near future. For example, the UK Biobank bio 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 (Niethammer et al., 2011; Hong et al., 2012b; 2012a; Singh et al., 2013; Fletcher, 2013; Hong et al., 2014b; Singh and Niethammer, 2014; Hong et al., 2014a; Singh et al., 2015; Hong et al., 2016) have been developed to analyze deformation trends in longitudinal imaging studies. One such approach is geodesic regression (GR) (Niethammer et al., 2011; Singh et al., 2013; Fletcher, 2013) which (for images) builds on the large displacement diffeomorphic metric mapping model (LDDMM) (Beg et al., 2005). 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 (Hong et al., 2012c) (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

* Corresponding author.
E-mail addresses: zp-ding@cs.unc.edu (Z. Ding), greg.nli10me@gmail.com (G. Fleishman), xy@cs.unc.edu (X. Yang), pthomp@usc.edu (P. Thompson), roland.kwitt@gmail.com (R. Kwitt), mn@cs.unc.edu (M. Niethammer).

[1] 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.

https://doi.org/10.1016/j.media.2019.06.003
1361-8415/

a single graphics processing unit (GPU) to process thousands of 3D image registrations for large-scale imaging studies such as ADNI (Jack et al., 2015). The primary computational bottleneck for SGR is the optimization required to compute pair-wise registrations.

Recently, efficient approaches have been proposed for deformable image registration (Cao et al., 2017; Miao et al., 2016; Sokooti et al., 2017; Yang et al., 2016; 2017; Zhang et al., 2017). 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 (Zhang and Fletcher, 2015) and frequency diffeomorphisms (Zhang et al., 2017), or by fast predictive image registration (FPIR) (Yang et al., 2016; 2017). FPIR predicts the initial conditions (specifically, the initial momentum) of LDDMM, which fully characterize the geodesic and the spatial transformation using a *learned* patch-based deep regression model. Because numerical optimization of standard LDDMM registration is replaced by a *single* prediction step, followed by optional correction steps (Yang et al., 2017), FPIR is dramatically faster than optimization-based LDDMM without compromising registration accuracy, as measured on several registration benchmarks (Klein et al., 2009).

Besides FPIR, other predictive image registration approaches have been proposed. Dosovitskiy et al. (Dosovitskiy et al., 2015) use a convolutional neural network (CNN) to directly predict optical flow. Liu et al. (Liu et al., 2017) use an encoder-decoder network to synthesize video frames. Schuster et al. (Schuster et al., 2016) investigate strategies to improve optical flow prediction via a CNN. Cao et al. (Cao et al., 2017) 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. (Miao et al., 2016) use CNN regression for 2D/3D rigid registration. Sokooti et al.(Sokooti et al., 2017) use CNNs to directly predict a 3D displacement vector field from input image pairs. An end-to-end approach for image registration was proposed by de Vos et al. (de Vos et al., 2017); 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. Hong et al. (2017) employ a low-dimensional band-limited representation of velocity fields in Fourier space (Zhang and Fletcher, 2015) to speed up SGR (Hong et al., 2012c) 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 and a geodesic path, 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 from (Yang et al., 2017), 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 (FPSGR) 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 the use of 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 $>6,000$ 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. Due to imaging protocol differences in ADNI-1 and ADNI-2, we separately analyze these two datasets.

This work is an extension of a recent conference paper (Ding et al., 2017). 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) a comparison with pairwise registration (Section 4.2); (b) a more in-depth analysis of atrophy scores correlated with clinical variables (Section 4.2); (c) correlations within diagnostic groups (Section 4.2); (d) an example to visualize the performance of regression models and associated quantitative comparisons (Section 4.3); (e) experiments on extrapolation on unseen data (Section 4.4, Section 4.3) ; (f) and more detailed atrophy assessments (Section 4.5).

Organization. The remainder of this article is organized as follows: Section 2 describes FPSGR, Section 3 discusses the experimental setup and the training of the prediction models. In Section 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 variant of LDDMM (Singh et al., 2013); Fig. 1 illustrates our overall approach. The individual components are described in the following.

2.1. LDDMM

Shooting-based LDDMM and geodesic regression minimize

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

$$s.t. \quad m_t + \mathrm{ad}^*_v m = 0, I_t + \nabla I^T v = 0, m - Lv = 0, \\ m(t_0) = m_0, \quad I(t_0) = I_0, \tag{2}$$

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

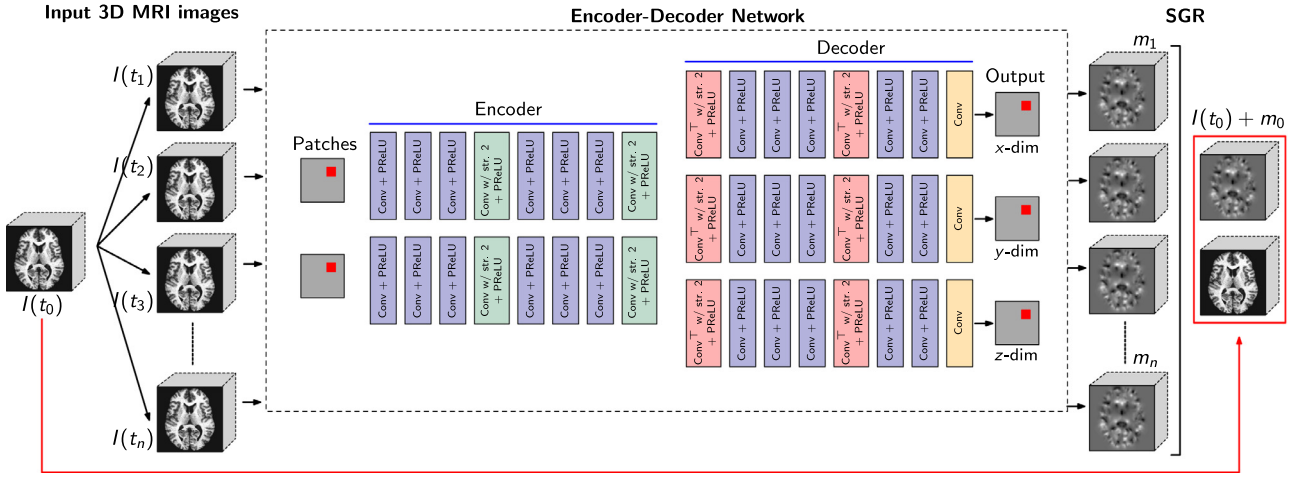

Fig. 1. Principle of fast predictive simple geodesic regression (FPSGR). In the encoder-decoder network (middle), the inputs are patches from the moving image and the target image at the *same* spatial location; the outputs are the predicted initial momenta (i.e., $m_1, \ldots, m_n$) of the corresponding 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. (3) to produce the initial momentum of the regression geodesic (marked `red`). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

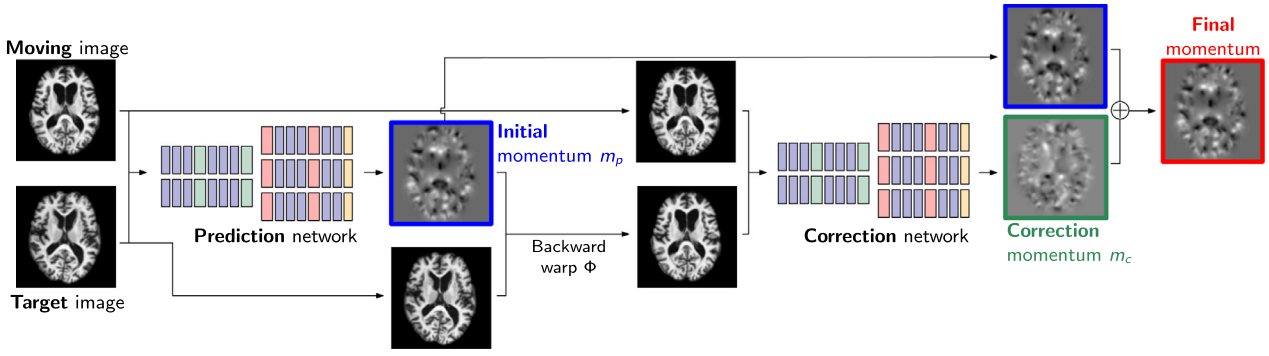

Fig. 2. Architecture of the prediction + correction network. Here, we use 2D images and the momentum in the x-direction for illustration. All images are 3D in our experiments. (1) Predict the initial momentum m_p and the corresponding backward deformation, Φ; (2) Predict a correction of the initial momentum, 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.

denotes the Jacobian. The deformation of the source image $I_0 \circ \Phi^{-1}$ can be computed by solving $\Phi_t^{-1} + D\Phi^{-1}v = 0,\ \Phi^{-1}(t_0) = \mathrm{id}$, where id denotes the identity map.

2.2. FPIR

Fast predictive image registration (Yang et al., 2016; 2017) 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 learned features are then concatenated to form the input to the decoder. The output is the predicted initial momentum in the x, y and z directions (obtained by numerical optimization on the training samples). We use an l_1 loss to train the network. 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 Yang et al. (2017) 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. Quantitative statistical results about deformation errors for such networks (with and without correction) can be found in Yang et al. (2017). Specifically, comparisons between deformations from the prediction models and the ones derived via

optimization showed good performance of the prediction models for diffeomorphic image registration on four different datasets.

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 (left-most 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 with respect to the first image. The approximated optimal m_0 of the energy functional in Eq. (1) for a fixed I_0 is then

$$\overline{m} \approx \frac{\sum_i (t_i - t_0)^2 m_i}{\sum_i (t_i - t_0)^2} = \frac{\sum_i (t_i - t_0)\widetilde{m}_i}{\sum_i (t_i - t_0)^2}, \tag{3}$$

where $\widetilde{m}_i$ is obtained by registering I_0 to Y^i in unit time followed by a rescaling of the momentum to account for the original time duration: $m_i = \frac{1}{t_i - t_0}\widetilde{m}_i$. See Appendix A for details.

3. Setup / training

All experiments use 3D images from the ADNI dataset[2] which consists of 6471 3D MR brain images of size $220 \times 220 \times 220$ voxels (a voxel is of size $1\,mm \times 1\,mm \times 1\,mm$). 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 preprocessed all images. Specifically, images are a) brain extracted using ROBEX (Iglesias et al., 2011) and b) affinely registered using FSL-FLIRT (Jenkinson et al., 2002; Jenkinson and Smith, 2001; Greve and Fischl, 2009) to a common atlas (ICBM brain template (Mazziotta et al., 1995)). Further, their c) intensities are normalized to a mean intensity of 1.0 within the brain. Since the patients in ADNI−1 / ADNI−2 were imaged at different time points (see Appendix B for details) and acquired with different acquisition protocols, we treat them as *two separate datasets*. Consequently, we evaluate them separately in what follows.

In particular, we perform two types of studies:

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

Regression. 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.

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

Table 1
Overview of the trained prediction models.

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

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-subject 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[3] (which serves as the "ground-truth").

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. The baseline images of each subject were registered to *all* the later timepoints within the same subject. To maintain the diagnostic ratio, we randomly 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 randomly 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 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. Training sets within ADNI−1 and ADNI−2, resp., were not overlapping. We trained four prediction models (i.e., two prediction models for each dataset in a two-fold cross-validation setup; denoted as Pred−1/2, respectively) and their four corresponding correction models, leading to eight prediction models overall (Table 1).

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

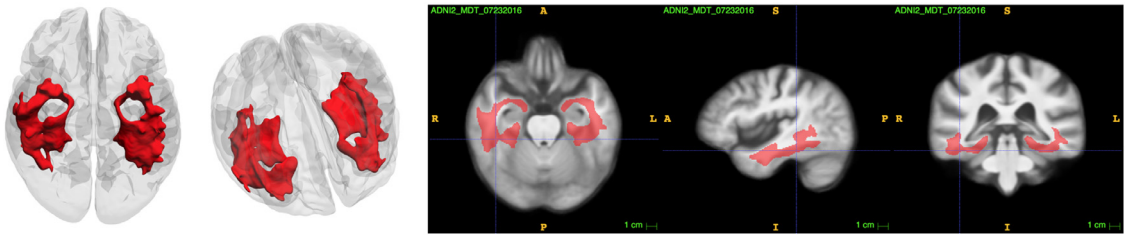

Fig. 3. Region of Interest (ROI) significantly associated with atrophy in AD used to compute atrophy scores.

For our experiments, we created 10 different (dataset / registration approach) combinations, each combination specifically designed to assess certain properties of our proposed method. The combinations are as follows:

(1) All subjects from the ADNI−1 dataset in combination with optimization-based LDDMM (to which we refer as SGR LDDMM when used for regression).
(2) Two subgroups of ADNI−1 (i.e., different cross-validation folds) in combination with FPSGR *without* a correction network. Denoted as SGR Pred or Pred in short.
(3) The same two subgroups as in 2), but in combination with FPSGR *with* a correction network. Denoted as SGR Pred+Corr or Pred+Corr in short.
(4) The same five groups of 1–3, but for ADNI−2.

With an additional correction network, the results are generally better than using the prediction network alone. Hence, to simplify the presentation of our results, we only show the prediction + correction results in the main manuscript. Selected results obtained when using the prediction network only (combination 2 above) can be found in the supplementary material.

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 (Kingma and Ba, 2014) over 10 epochs with a learning rate of 0.0001. Additional training and convergence details can be found in the supplementary material.

3.3. Efficiency

Once trained, the prediction models allow fast computations of registrations. We use a Nvidia TITAN X (Pascal) GPU and PyTorch[4] 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. SGR LDDMM[5] would require ≈ 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 recent fast geodesic regression model by Hong et al. (2017), FPSGR is approximately twice as fast, though this comparison is only qualitative as FPSGR is implemented and run on a GPU, whereas Hong et al. (2017) used a CPU compute cluster. We therefore base our qualitative comparisons on the obtained speed-ups. Specifically, the model by Hong et al. (2017) achieves ≈16 times speed-up compared with optimization-based SGR for the same CPU setting (parallel computing with the same number of cores). In our case, we achieve more than 40 times speed-up compared with SGR for the same GPU setting (using a single Nvidia GTX1080 Ti GPU). Note also that an interesting future direction could be to combine the approach by Hong et al. (2017) with our prediction approach. This would likely yield extremely fast prediction methods for registration and regression. Additional details on computation times for trainining and testing can be found in the supplementary material.

4. Experimental results for 3D ADNI data

Our general hypothesis is that the prediction models (for ADNI−1/2) show similar performance to SGR LDDMM and that using the correction network for the predictions improves results. As using a correction network indeed improved results only these results are presented in the main document. See the supplementary material for results when not using the correction network. To assess differences, we compare differences in deformations. Specifically, for every deformation produced by the different approaches, we compute its local Jacobian determinants (JD). The JDs are then warped to a common coordinate system for the entire ADNI dataset using existing deformations from Fleishman and Thompson (2017b,a) obtained via LDDMM registration. 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). This region was determined in Fleishman and Thompson (2017b) and Fleishman and Thompson (2017a) using a training dataset of AD subjects and controls and optimization-based LDDMM registration[6]. Only voxels whose atrophy measurements were significantly associated with the AD disease group (vs. controls) after a Bonferroni correction on the number of voxels, were retained. We prefer this data-driven approach to defining the area impacted by AD to using anatomical boundaries. As can be seen from Fig. 3, our statistically derived ROI reassuringly overlaps with the hippocampus and surrounding grey matter. There are likely deformations over time outside this region, either due to local tissue loss or elastic deformation from non-local tissue loss, but we (conservatively) only study voxels that passed our statistical test. Specifically, we quantify atrophy as

$$s(\phi) := \left(1 - \frac{1}{|\omega|}\int_{\omega} \det(D\phi(x))\, dx\right) \times 100, \tag{4}$$

where $\det(\cdot)$ denotes the determinant and $|\cdot|$ the cardinality/size of a set; ω is the stat-ROI region described above. The resulting scalar value is an estimate of the relative volume change experienced by that region between the baseline and a follow-up image.

[4] http://pytorch.org.

[5] Here, we used 300 fixed iterations for each registration. Empirically, 300 iterations were sufficient for convergence. Note that the optimization-based LDDMM also uses a GPU implementation.

[6] Less than 5% of the images of the ADNI−1 dataset were used to define this statistical region of interest. This may result in some analysis bias for the ADNI−1 dataset. The ADNI−2 results are *not* subject to this possible analysis bias as we use the same ADNI−1-derived stat-ROI for the analysis of the ADNI−2 data.

Hence, its sign is positive when the region has lost volume over time and is negative if the region has gained volume over time. To estimate atrophy trends for longitudinal data, we compute atrophy measurements according to Eq. (4) at all measurement timepoints and then fit them via a linear regression model. For the regression formulations the measurements are the ones based on the deformations of the regression geodesic at these timepoints. Instead, for pairwise registrations (LDDMM), atrophy measurements are computed *independently* for each timepoint.

We limited our experiments to the applications in Hua et al. (2013, 2016), 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[7]), and normal ageing (NC: normal control) (0) in an elderly population. These are the diagnostic groups for `ADNI-1`. For `ADNI-2`, we use the following three diagnostic categories[8]: normal ageing (0) (including SMC), mild cognitive impairment (including EMCI and LMCI) (1), and AD (2).

Specifically, we investigate the following *five* aspects:

S1 Prediction Models for Longitudinal Data (Section 4.1)
Can we learn models for *longitudinal* image data which predict optimization-based registration results to high accuracy?
We show that this is possible. Hence it is appropriate to use our training and prediction strategy as a component of SGR.

S2 Quantitative Validation (Section 4.2)
(a) Are regression results more stable and hence capture trends better than pairwise registrations?
(b) Are FPSGR atrophy measurements consistent with those derived from deformations via numerical optimization (SGR LDDMM) which produced the training dataset?
Our experiments show that SGR is indeed more stable than pairwise registration and FPSGR results are consistent with results obtained via numerical optimization. Hence, our prediction approach can reliable replace costly numerical optimization.

S3 Visual Validation (Section 4.3)
Can the prediction models for regression visually capture similar trends to the regression model obtained by numerical optimization?
Our visual results show that FPSGR approximates longitudinal image data well, providing visual confirmation for our quantitative validation results (**S2**).

S4 Forecasting (Section 4.4)
Is the predictive power of the regression models strong enough to forecast deformations for unseen future timepoints?
We show that FPSGR can capture correlation trends for future (unseen) images. This is evidence that FPSGR captures trends which allow for extrapolation in time.

S5 Atrophy Assessment via Transitivity Analysis and Sample Size Estimates (Section 4.5)
Does transitivity hold for our atrophy regression results, i.e., do regression results from $A \rightarrow C$ agree with results obtained by regressing from $A \rightarrow B$ and $B \rightarrow C$? Furthermore, what sample sizes are required to show differences based on the regressed atrophy measures?
Our results show that FPSGR a) shows limited saturation effects when analyzing transitivity, and b) shows consistent sample size estimates with SGR LDDMM.

Aspects **S1**-**S5** justify the use of FPSGR. In turn, the substantially improved computational efficiency of FPSGR justifies its use for large-scale imaging studies. Appendix B shows the distributions of the prediction cases per time-point and the diagnostic groups in `ADNI-1`/`ADNI-2`, respectively.

4.1. ***S1***: *Prediction models for longitudinal data*

A key aspect to the success of FPSGR for the analysis of longitudinal imaging data is to verify that the predictive registration component of FPSGR can reliably predict longitudinal registration results. In particular, this question also relates to how one should go about training such longitudinal models. Our hypothesis, based on the prior work in Yang et al. (2017), was that highly accurate prediction models can be obtained. Going beyond these results, we further hypothesized that training a prediction model on longitudinal data yields higher accuracies than training on cross-subject data, as the models can then become more data-specific, because they are trained on deformations that are *expected* for longitudinal registrations. To test these hypotheses, we trained two different prediction models and tested them on longitudinal and cross-subject registration tasks. Our training strategy for the different prediction models is detailed in Section 3.1. In brief, we trained our prediction models on data of 100 subjects of the `ADNI-1` dataset and tested on 20. We trained models only using longitudinal pairs between baseline and the 24 month follow-up images, as well as using the same time-points but across subjects. Testing was done on data for a separate set of 20 subjects and compared with respect to results obtained via numerical optimization of LDDMM.

Table 2 shows the resulting deformation errors and confirms our hypotheses. Results with respect to optimization-based LDDMM are highly accurate with a median deformation error substantially below a millimeter for the longitudinal registration task which is relevant for SGR/FPSGR. Furthermore, training on longitudinal image registration pairs is clearly beneficial. Hence, we conclude that a prediction model trained on longitudinal data works well while allowing much faster computations than optimization-based LDDMM. Hence, we use such models for all our following experiments.

4.2. ***S2***: *Quantitative validation*

Now that we justified that highly accurate prediction models for longitudinal data can be trained (see Section 4.1), it is important to validate the performance of FPSGR. Specifically, we investigate if (1) regression is beneficial for the analysis of longitudinal data with more than two timepoints and (2) if FPSGR can perform as well as SGR LDDMM (i.e., simple geodesic regression via numerical optimization).

For simple geodesic regression to be a useful model it should outperform pairwise image registration. The 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 considered a trivial case of geodesic regression with two images only) the deformation will in general be able to match the target image better. However, just as in linear regression, this may accentuate the effects of noise.

We assess the performance of our models by evaluating *bias* of regressed atrophy scores and strength of *correlation* with respect to

[7] 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. (2013), 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`

[8] 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 *ADNIDiagnosis Summary* available on the IDA website. Hereafter, in all discussions of `ADNI-2`, NC includes NC and SMC and MCI includes EMCI and LMCI.

Table 2
Deformation error of longitudinal and cross-subject models tested on longitudinal and cross-subject data. 2-norm deformation errors in millimeters 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-subject registration is preferred for cross-subject registrations.

3D Longitudinal Test Case Deformation Error [mm]							
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-subject Training	0.0544	0.1424	0.2641	0.3723	0.5067	0.7502	0.8425
3D Cross-subject Test Case Deformation Error [mm]							
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-subject Training	**0.1123**	**0.3024**	**0.5863**	**0.8737**	**1.2743**	**2.2659**	**2.7836**

clinical measures. A successful model should not exhibit bias and is expected to result in high correlations comparable to the correlation levels achieved via numerical optimization.

Bias. Estimates of atrophy are susceptible to bias (Yushkevich et al., 2010; Fox et al., 2011). We use two bias measures: regression intercept of the atrophy score and the transitivity of the regression results. In this section, we only assess bias via the atrophy regression intercept, as it is a direct assessment of bias when there is no expected change. We leave the more detailed transitivity analysis and sample size estimates for Section 4.5. 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)[9], (3) MCI for all time points (MCI-MCI), (4) starting with MCI and reverting 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)[10] In particular, we follow Hua et al. (2013) 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 3 lists the slopes, intercepts, and 95% confidence intervals for optimization and prediction results on `ADNI-1` and `ADNI-2`, respectively. Specifically, it shows linear regression results of atrophy measures over time as obtained via (1) FPSGR (i.e., using an FPSGR fit over all time-points followed by atrophy computations based on the deformations of the regression geodesic) compared with atrophy measures obtained by (2) pairwise predictive registration and (3) SGR LDDMM. The different cross-validation testing folds are indicated with suffix -1 and -2, e.g., SGR LDDMM-1 and SGR LDDMM-2. Comparisons between approaches should therefore be within folds.

As shown in Table 3, FPSGR (i.e., SGR Pred+Corr-1/2) outperforms the pairwise registration approach in two aspects: (1) the estimated intercept of FPSGR is generally closer to zero than for the pairwise method and the intercept 95% confidence interval is narrower; (2) 8 out of 24 of the 95% confidence intervals of the pairwise methods show bias to either overestimate or underestimate volume change. None of the FPSGR results show such significant bias. Both SGR LDDMM and FPSGR show intercepts that are near zero relative to the range of changes observed and both intercept confidence intervals contain zero. For all diagnostic change groups, FPSGR results are more stable than the results for the SGR LDDMM method, as indicated by the tighter confidence intervals. A possible explanation for the tighter confidence intervals is that the prediction method at the core of FPSGR learns a relatively conservative mapping from images to initial momentum. Hence, it will avoid, for example, large outliers (as also observed in the original Quicksilver work of Yang et al. (2017) for image-to-image registration). Methods based on optimization-based image registration (such as SGR LDDMM) are more sensitive to misregistrations and imperfections in image pre-processing (e.g., imperfect brain extraction results, which can be tolerated much more gracefully by a deep-learning-based registration approach; see Yang et al. (2017)). Appendix E visually 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. Additional discussions about disease severity and the linear regression slope as well as a more in-depth analysis of the estimation bias can be found in the supplementary material. We conclude that (1) neither SGR LDDMM optimization nor FPSGR produced deformations with significant bias to overestimate or underestimate volume change; (2) the pairwise prediction model suffers from bias while the regression prediction model (FPSGR) shows little bias. Hence, from the perspective of bias, **S2** has been validated.

Correlation. Atrophy estimates are shown to correlate[11] with clinical variables (Fleishman and Thompson, 2017b). To quantify this effect, we computed the Spearman rank-order correlation[12] 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 computed these measures for FPSGR (SGR Pred+Corr-1/2), for optimization-based SGR (SGR LDDMM-1/2) and for pairwise predicted registrations (Pairwise Pred+Corr-1/2). We applied the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) for all the

[9] Very few cases convert from NC to AD in the imaged time-frame.

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

[11] The true correlation between ideal noiseless atrophy measures and clinical variables is unknown. Validation of dense volumetric image registration is known to be a challenging task. In segmentation, automatic segmentations are often compared to manual segmentations. However, obtaining a manual gold standard for *dense* volumetric registration is infeasible. Alternatively, segmentations are frequently used as an indirect way of validating registration accuracy. When assessing atrophy over time in a region of interest, manual and automated segmentation errors are often comparable in magnitude to the expected atrophy (Shen et al., 2010). For example, expected annual tissue loss in the hippocampus for an AD subject is around 6%, which can be produced by shifting the hippocampal segment boundary by only a few voxels along its extent. Hence, segment boundaries are not reliable forms of ground truth in this setting either. In the absence of reliable alternatives, we therefore hypothesize that tissue loss will correlate with disease severity and then trust the measurements that clinicians utilize to measure disease severity, such as cognitive scores.

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

Table 3
Slope and intercept values for 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. C.I. Length is the average 95% confidence interval length of this linear regression over time. The #data column lists the number of data points analyzed. **Green** indicates that the intercept is closer to zero (also, zero is within the 95% confidence interval) for SGR Pred+Corr model; *Yellow* indicates that the intercept is closer to zero for pairwise Pred+Corr model; Red indicates that the point estimate is either biased to overestimate or underestimate volume change. FPSGR (SGR Pred+Corr) model performs better than the pairwise Pred+Corr model.

ADNI-1		Slope	Intercept	C.I. Length	#data
NC-NC	SGR LDDMM-1	[0.62, **0.70**, 0.78]	[-0.25,**-0.08**, 0.09]	0.2941	154
	SGR Pred+Corr-1	[0.61, **0.68**, 0.75]	[-0.15, **-0.01**, 0.13]	0.2478	
	Pairwise Pred+Corr-1	[0.46, **0.55**, 0.63]	[0.07, **0.24**, 0.40]	0.2899	
	SGR LDDMM-2	[0.57, **0.66**, 0.75]	[-0.21, **-0.04**, 0.14]	0.3137	156
	SGR Pred+Corr-2	[0.51, **0.58**, 0.65]	[-0.12, **0.01**, 0.15]	0.2421	
	Pairwise Pred+Corr-2	[0.49, **0.56**, 0.64]	[-0.10, **0.05**, 0.20]	0.2632	
NC-MCI	SGR LDDMM-1	[0.72, **0.94**, 1.16]	[-0.45, **-0.03**, 0.39]	0.7363	24
	SGR Pred+Corr-1	[0.71, **0.90**, 1.10]	[-0.40, **-0.01**, 0.37]	0.6737	
	Pairwise Pred+Corr-1	[0.60, **0.88**, 1.15]	[-0.15, **0.38**, 0.92]	0.9373	
	SGR LDDMM-2	[0.88, **1.19**, 1.50]	[-0.65, **-0.05**, 0.55]	1.0657	22
	SGR Pred+Corr-2	[0.80, **1.07**, 1.34]	[-0.66, **-0.14**, 0.38]	0.9236	
	Pairwise Pred+Corr-2	[0.65, **0.91**, 1.17]	[-0.39, **0.11**, 0.61]	0.8875	
MCI-MCI	SGR LDDMM-1	[0.97, **1.17**, 1.38]	[-0.28, **0.05**, 0.39]	0.6683	146
	SGR Pred+Corr-1	[0.92, **1.09**, 1.26]	[-0.14, **0.14**, 0.42]	0.5612	
	Pairwise Pred+Corr-1	[0.75, **0.91**, 1.08]	[0.08, **0.35**, 0.63]	0.5416	
	SGR LDDMM-2	[0.83, **1.00**, 1.17]	[-0.21, **0.06**, 0.33]	0.5365	148
	SGR Pred+Corr-2	[0.77, **0.90**, 1.04]	[-0.15, **0.07**, 0.29]	0.4406	
	Pairwise Pred+Corr-2	[0.78, **0.94**, 1.11]	[-0.11, **0.16**, 0.42]	0.5332	
MCI-NC	SGR LDDMM-1	[0.48, **0.72**, 0.96]	[-0.85, **-0.42**, 0.01]	0.7873	16
	SGR Pred+Corr-1	[0.51, **0.68**, 0.86]	[-0.52, **-0.20**, 0.13]	0.5951	
	Pairwise Pred+Corr-1	[0.27, **0.52**, 0.76]	[-0.18, **0.27**, 0.71]	0.8187	
	SGR LDDMM-2	[0.54, **0.79**, 1.03]	[-0.79, **-0.36**, 0.07]	0.8087	17
	SGR Pred+Corr-2	[0.49, **0.70**, 0.91]	[-0.59, **-0.21**, 0.17]	0.7020	
	Pairwise Pred+Corr-2	[0.28, **0.54**, 0.80]	[-0.39, **0.07**, 0.53]	0.8577	
MCI-AD	SGR LDDMM-1	[1.94, **2.10**, 2.27]	[-0.28, **0.02**, 0.31]	0.5484	148
	SGR Pred+Corr-1	[1.70, **1.84**, 1.98]	[-0.17, **0.08**, 0.33]	0.4601	
	Pairwise Pred+Corr-1	[1.46, **1.60**, 1.74]	[0.18, **0.43**, 0.67]	0.4516	
	SGR LDDMM-2	[1.75, **1.92**, 2.09]	[-0.16, **0.14**, 0.44]	0.5595	147
	SGR Pred+Corr-2	[1.49, **1.64**, 1.78]	[-0.08, **0.17**, 0.43]	0.4708	
	Pairwise Pred+Corr-2	[1.50, **1.64**, 1.77]	[0.00, **0.24**, 0.48]	0.4415	
AD-AD	SGR LDDMM-1	[1.97, **2.33**, 2.69]	[-0.17, **0.27**, 0.70]	0.8878	143
	SGR Pred+Corr-1	[1.74, **2.05**, 2.35]	[-0.04, **0.33**, 0.70]	0.7486	
	Pairwise Pred+Corr-1	[1.62, **1.91**, 2.21]	[0.15, **0.51**, 0.87]	0.7328	
	SGR LDDMM-2	[1.92, **2.28**, 2.65]	[-0.20, **0.24**, 0.68]	0.9067	140
	SGR Pred+Corr-2	[1.65, **1.95**, 2.24]	[-0.10, **0.25**, 0.60]	0.7244	
	Pairwise Pred+Corr-2	[1.70, **1.99**, 2.27]	[-0.03, **0.32**, 0.66]	0.7030	
ADNI-2		Slope	Intercept		
NC-NC	SGR LDDMM-1	[0.55, **0.65**, 0.75]	[-0.08, **0.03**, 0.13]	0.2635	170
	SGR Pred+Corr-1	[0.50, **0.57**, 0.65]	[-0.04, **0.05**, 0.13]	0.2040	
	Pairwise Pred+Corr-1	[0.37, **0.46**, 0.54]	[0.10, **0.19**, 0.29]	0.2259	
	SGR LDDMM-2	[0.51, **0.62**, 0.72]	[-0.10, **0.01**, 0.12]	0.2700	175
	SGR Pred+Corr-2	[0.35, **0.44**, 0.52]	[-0.09, **+0.00**, 0.08]	0.2134	
	Pairwise Pred+Corr-2	[0.35, **0.44**, 0.53]	[-0.09, **0.00**, 0.10]	0.2357	
NC-MCI	SGR LDDMM-1	[0.56, **0.79**, 1.02]	[-0.22, **0.01**, 0.25]	0.4624	16
	SGR Pred+Corr-1	[0.63, **0.80**, 0.97]	[-0.16, **0.02**, 0.19]	0.3431	
	Pairwise Pred+Corr-1	[0.59, **0.82**, 1.05]	[-0.20, **0.04**, 0.27]	0.4620	
	SGR LDDMM-2	[0.62, **0.90**, 1.18]	[-0.32, **-0.02**, 0.28]	0.5691	17
	SGR Pred+Corr-2	[0.46, **0.68**, 0.91]	[-0.25, **+0.02**, 0.22]	0.4554	
	Pairwise Pred+Corr-2	[0.53, **0.77**, 1.02]	[-0.40, **-0.15**, 0.10]	0.4927	
MCI-MCI	SGR LDDMM-1	[0.71, **0.83**, 0.94]	[-0.13, **-0.00**, 0.12]	0.3008	184
	SGR Pred+Corr-1	[0.64, **0.73**, 0.82]	[-0.08, **0.02**, 0.11]	0.2384	
	Pairwise Pred+Corr-1	[0.59, **0.69**, 0.79]	[-0.01, **0.09**, 0.20]	0.2577	
	SGR LDDMM-2	[0.71, **0.82**, 0.92]	[-0.14, **-0.02**, 0.09]	0.2844	183
	SGR Pred+Corr-2	[0.50, **0.59**, 0.67]	[-0.12, **+0.02**, 0.07]	0.2253	
	Pairwise Pred+Corr-2	[0.59, **0.69**, 0.78]	[-0.23, **-0.13**, -0.02]	0.2559	
MCI-NC	SGR LDDMM-1	[0.03, **0.39**, 0.74]	[-0.38, **0.05**, 0.47]	0.9551	16
	SGR Pred+Corr-1	[0.08, **0.36**, 0.64]	[-0.28, **0.05**, 0.38]	0.7421	
	Pairwise Pred+Corr-1	[0.00, **0.25**, 0.51]	[-0.11, **0.18**, 0.48]	0.6702	
	SGR LDDMM-2	[0.14, **0.40**, 0.67]	[-0.28, **0.04**, 0.35]	0.7109	21
	SGR Pred+Corr-2	[0.05, **0.26**, 0.48]	[-0.22, **0.03**, 0.29]	0.5744	
	Pairwise Pred+Corr-2	[-0.04, **0.23**, 0.50]	[-0.32, **-0.01**, 0.31]	0.7148	
MCI-AD	SGR LDDMM-1	[1.65, **1.95**, 2.25]	[-0.21, **0.13**, 0.47]	0.7908	70
	SGR Pred+Corr-1	[1.39, **1.62**, 1.85]	[-0.15, **0.11**, 0.37]	0.6060	
	Pairwise Pred+Corr-1	[1.25, **1.48**, 1.72]	[0.09, **0.35**, 0.60]	0.6122	
	SGR LDDMM-2	[1.59, **1.91**, 2.23]	[-0.16, **0.19**, 0.53]	0.8447	65
	SGR Pred+Corr-2	[1.20, **1.45**, 1.69]	[-0.13, **0.14**, 0.41]	0.6498	
	Pairwise Pred+Corr-2	[1.27, **1.50**, 1.74]	[-0.14, **0.12**, 0.38]	0.6193	
AD-AD	SGR LDDMM-1	[2.49, **2.76**, 3.04]	[-0.15, **0.07**, 0.30]	0.5128	101
	SGR Pred+Corr-1	[2.14, **2.34**, 2.54]	[-0.09, **0.08**, 0.24]	0.3810	
	Pairwise Pred+Corr-1	[2.12, **2.34**, 2.57]	[-0.02, **0.17**, 0.35]	0.4223	
	SGR LDDMM-2	[2.72, **2.99**, 3.27]	[-0.15, **0.07**, 0.29]	0.5124	103
	SGR Pred+Corr-2	[2.16, **2.36**, 2.56]	[-0.15, **0.02**, 0.18]	0.3796	
	Pairwise Pred+Corr-2	[2.14, **2.37**, 2.59]	[-0.24, **-0.05**, 0.13]	0.4226	

correlation results to account for multiple comparisons. The overall false discovery rate was set to 0.01, which resulted in an effective significance level of $\alpha \approx 0.0093$. Detailed results can be found in Table 4. FPSGR performs better than the pairwise approach in 14 out of 18 cases for MMSE and in 17 out of 20 cases for the diagnostic category. Furthermore, when the pairwise method is better than FPSGR, the difference is much smaller compared to the differences observed for the cases where FPSGR 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 FPSGR is also heavily influenced by the last time-point. Furthermore, FPSGR shows statistically significant improved (higher in magnitude) correlations over the pairwise approach. Specifically, we tested if the two approaches show different means based on the correlations reported in Table 4. Details on the statistical tests can be found in the Appendix C.

We observe median correlations for all four FPSGR prediction + correction models (ADNI1/2 Pred+Corr-1/2) 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 (Fleishman and Thompson, 2017b; 2017a). Our two SGR 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 SGR prediction+correction models.

In fact, FPSGR with correction network shows similar correlations between atrophy and MMSE/DX to optimization-based SGR (SGR LDDMM), justifying the use of FPSGR. Statistical testing details are given in Appendix D.

Table 4
SGR prediction models (FPSGR and SGR LDDMM) 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 FPSGR (SGR prediction+correction) method; *Yellow* indicates a stronger correlation for the pairwise prediction+correlation 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, FPSGR performs better than the pairwise prediction+correction model 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 timepoint in our comparison.

ADNI-1		MMSE	*p*-value	DX	*p*-value	#data
6mo	SGR LDDMM-1	-0.4957	5.17e-39	0.5140	2.66e-42	608
	SGR Pred+Corr-1	-0.5104	1.22e-41	0.5259	1.53e-44	
	Pairwise Pred+Corr-1	-0.3216	4.28e-16	0.2695	1.42e-11	
	SGR LDDMM-2	-0.4667	4.17e-34	0.4814	1.75e-36	606
	SGR Pred+Corr-2	-0.4734	3.54e-35	0.4890	9.67e-38	
	Pairwise Pred+Corr-2	-0.3289	9.34e-17	0.3041	1.97e-14	
12mo	SGR LDDMM-1	-0.5749	5.23e-51	0.5313	1.81e-42	565
	SGR Pred+Corr-1	-0.5799	4.39e-52	0.5406	3.44e-44	
	Pairwise Pred+Corr-1	-0.4605	5.22e-31	0.3773	1.49e-20	
	SGR LDDMM-2	-0.5301	6.81e-42	0.5055	1.17e-37	560
	SGR Pred+Corr-2	-0.5374	3.73e-43	0.5155	2.89e-39	
	Pairwise Pred+Corr-2	-0.4377	1.46e-27	0.3602	1.44e-18	
18mo	SGR LDDMM-1	-0.4939	4.86e-16	0.4776	5.76e-15	238
	SGR Pred+Corr-1	-0.4924	6.16e-16	0.4643	3.98e-14	
	Pairwise Pred+Corr-1	-0.4324	2.90e-12	0.3851	7.82e-10	
	SGR LDDMM-2	-0.4385	9.50e-13	0.4000	1.12e-10	241
	SGR Pred+Corr-2	-0.4384	9.75e-13	0.3790	1.19e-9	
	Pairwise Pred+Corr-2	-0.4057	5.79e-11	0.3191	4.16e-7	
24mo	SGR LDDMM-1	-0.6064	5.01e-45	0.5978	1.69e-43	435
	SGR Pred+Corr-1	-0.6001	6.55e-44	0.5943	6.82e-43	
	Pairwise Pred+Corr-1	-0.6005	5.57e-44	0.5445	6.12e-35	
	SGR LDDMM-2	-0.5822	4.11e-40	0.5534	1.24e-35	427
	SGR Pred+Corr-2	-0.5898	2.28e-41	0.5709	2.65e-38	
	Pairwise Pred+Corr-2	-0.5881	4.36e-41	0.5443	2.64e-34	
36mo	SGR LDDMM-1	-0.5142	4.29e-20	0.5300	1.81e-21	277
	SGR Pred+Corr-1	-0.5069	1.71e-19	0.5296	1.99e-21	
	Pairwise Pred+Corr-1	-0.4759	4.61e-17	0.4726	8.13e-17	
	SGR LDDMM-2	-0.4334	3.79e-13	0.4815	2.93e-16	256
	SGR Pred+Corr-2	-0.4393	1.67e-13	0.4863	1.34e-16	
	Pairwise Pred+Corr-2	-0.4526	2.49e-14	0.4801	3.64e-16	
48mo	SGR LDDMM-1	-0.7456	2.01e-13	0.6635	5.20e-10	69
	SGR Pred+Corr-1	-0.7443	2.30e-13	0.6575	8.43e-10	
	Pairwise Pred+Corr-1	-0.7124	6.65e-12	0.6592	7.34e-10	
	SGR LDDMM-2	-0.6889	2.25e-10	0.5927	1.98e-7	65
	SGR Pred+Corr-2	-0.7005	8.31e-11	0.6067	8.49e-8	
	Pairwise Pred+Corr-2	-0.6686	1.16e-9	0.5840	3.28e-7	
ADNI-2		MMSE	p-value	DX	p-value	#data
3mo	SGR LDDMM-1	N/A	N/A	0.4254	2.34e-24	522
	SGR Pred+Corr-1	N/A	N/A	0.4353	1.52e-25	
	Pairwise Pred+Corr-1	N/A	N/A	0.1915	1.06e-5	
	SGR LDDMM-2	N/A	N/A	0.4409	2.77e-26	523
	SGR Pred+Corr-2	N/A	N/A	0.4445	9.64e-27	
	Pairwise Pred+Corr-2	N/A	N/A	0.1951	7.01e-6	
6mo	SGR LDDMM-1	-0.4989	8.01e-31	0.4688	6.09e-27	468
	SGR Pred+Corr-1	-0.5128	9.64e-33	0.4846	6.19e-29	
	Pairwise Pred+Corr-1	-0.3721	8.20e-17	0.2830	4.58e-10	
	SGR LDDMM-2	-0.5072	4.29e-32	0.4883	1.58e-29	470
	SGR Pred+Corr-2	-0.5066	5.25e-32	0.4913	6.33e-30	
	Pairwise Pred+Corr-2	-0.3890	1.97e-18	0.3273	3.37e-13	
12mo	SGR LDDMM-1	-0.4756	1.43e-27	0.4859	7.22e-29	464
	SGR Pred+Corr-1	-0.4908	1.67e-29	0.5064	1.37e-31	
	Pairwise Pred+Corr-1	-0.4623	5.98e-26	0.4762	1.22e-27	
	SGR LDDMM-2	-0.4937	1.07e-29	0.5026	7.05e-31	461
	SGR Pred+Corr-2	-0.4987	2.35e-30	0.5149	1.44e-32	
	Pairwise Pred+Corr-2	-0.4647	5.04e-26	0.4780	1.22e-27	
24mo	SGR LDDMM-1	-0.4120	9.53e-15	0.4476	2.06e-17	325
	SGR Pred+Corr-1	-0.4109	1.15e-14	0.4632	1.09e-18	
	Pairwise Pred+Corr-1	-0.4178	3.66e-15	0.4796	4.22e-20	
	SGR LDDMM-2	-0.4095	2.09e-14	0.4375	1.93e-16	321
	SGR Pred+Corr-2	-0.3943	2.20e-13	0.4336	3.79e-16	
	Pairwise Pred+Corr-2	-0.4045	4.60e-14	0.4607	2.85e-18	
36mo	SGR LDDMM-1	-0.2474	0.55	0.2869	0.49	8
	SGR Pred+Corr-1	-0.2474	0.55	0.2869	0.49	
	Pairwise Pred+Corr-1	-0.2887	0.49	0.4434	0.27	
	SGR LDDMM-2	0.0935	0.83	0.1695	0.69	8
	SGR Pred+Corr-2	0.0935	0.83	0.1695	0.69	
	Pairwise Pred+Corr-2	0.3429	0.41	0.1956	0.64	

In summary, based on the discussions above, FPSGR shows excellent performance. It shows negligible bias, works better than the pair-wise approach, shows strong correlations with clinical variables, and works as well, or better, than simple geodesic regression via numerical optimization (SGR LDDMM).

4.3. ***S3**: Visual validation*

Section 4.1 quantitatively assessed the ability of FPIR to predict longitudinal deformations. Section 4.2 quantitatively demonstrated the good performance of FPSGR in relation to a pair-wise prediction approach and SGR via numerical optimization. Here, we qualitatively illustrate the behavior of FPSGR via the visualization of intensity differences (for completeness these are also quantified in Table 5) and Jacobian determinants (JD) for some example image data.

Intensity. Fig. 4 shows an example regression result. In this specific case, large changes can be observed around the ventricles. To illustrate differences between the methods, Fig. 4 visualizes regression results based on optimization-based SGR LDDMM and for FPSGR *with* a correction network. Both methods successfully capture the expanding ventricles and generally capture the image changes. The difference between SGR LDDMM and FPSGR is barely noticeable in the 5th row of Fig. 4. To further illustrate

Table 5
Mean+standard deviation of the overlay errors, see Eq. (5), over 100 patients in ADNI−1 dataset. Prediction + correction model exhibits performance comparable to optimization-based regression results (SGR LDDMM).

	$E_{overlay}(I_0 \circ \Phi_{t_i}^{-1}, Y_i)$					
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
SGR 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
SGR 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

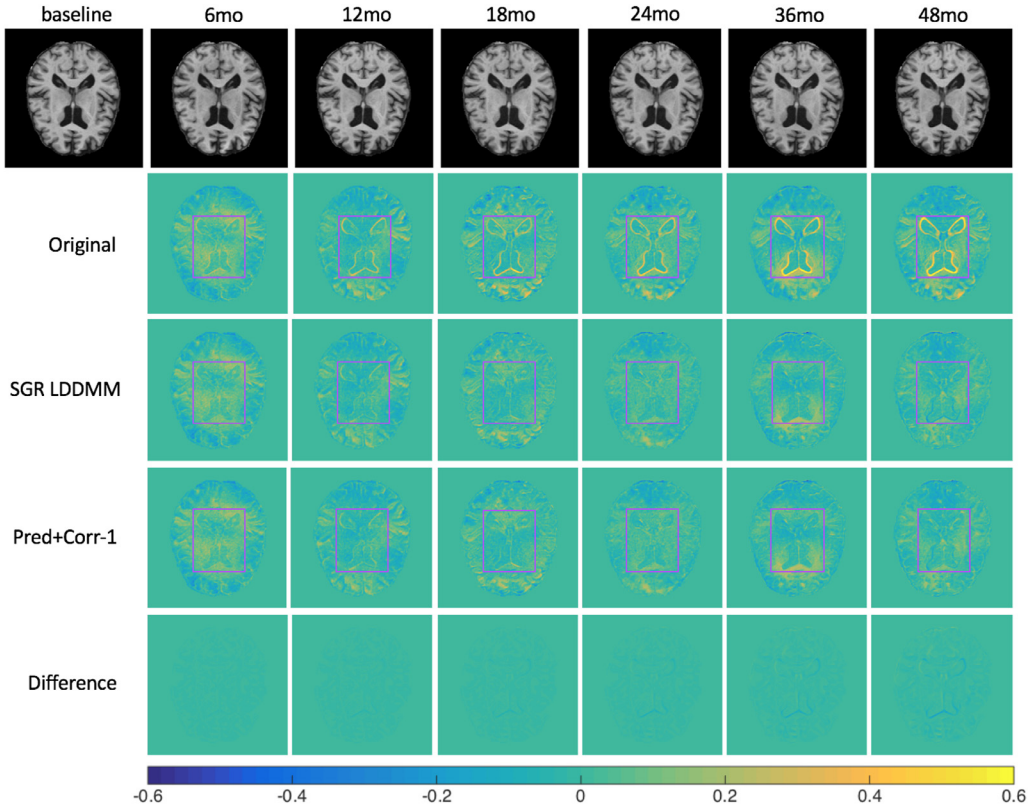

Fig. 4. Example regression result: one subject with 6 follow-up images from the ADNI−1 dataset. Image intensity range is [0, 2.49]. **Top row:** Axial slices extracted from the 3D MR images at the same axial location for different months. **Original:** intensity differences between the baseline image and its 6-month, 12-month, etc. follow-up images. **SGR LDDMM:** intensity differences between the acquired images in the top row and optimization-based regression results at each follow-up month(s). **Pred+Corr-1:** intensity differences between the acquired images in the top row and the Pred+Corr-1 regression results at each follow-up month(s). **Difference:** intensity differences between **SGR LDDMM** and **Pred+Corr-1** at each follow-up month(s). 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 model* (Pred+Corr-1) give very similar results to the regression results obtained by numerical optimization (SGR LDDMM).

the regression results, we compute the overlay error between measured images and the images on the geodesic as

$$E_{overlay}(I_0 \circ \Phi_{t_i}^{-1}, Y_i) = \frac{1}{|\Omega|} \| I_0 \circ \Phi_{t_i}^{-1} - Y_i \|_{L_1} \tag{5}$$

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 a randomly selected population of 100 subjects of the ADNI−1 dataset. This random set includes all diagnostic groups. FPSGR obtains results comparable with optimization-based SGR LDDMM. This justifies the use of the proposed method.

Jacobian Determinant (JD). The average JD images qualitatively agree with prior results (Hua et al., 2013; 2016): severity of volume change increases with severity of diagnosis and time. Change is most substantial in the temporal lobes near the hippocampus. In Fig. 5, 6 month to 48 month are existing data points; 60 month to 84 month are forecasting results (i.e., results obtained via extrapolation of the estimated regression geodesic; see upcoming Section 4.4 for a detailed discussion on how these forecasting results were computed). 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.

4.4. ***S4***: *Forecasting*

Another interesting question for SGR and geodesic regression in general is if SGR is able to *forecast* unseen future time-points. Specifically we consider two scenarios:

(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 maintaining strong correlations.

(Q2) **Extrapolate-image:** How well can correlations between atrophy and clinical measures be predicted for time-points

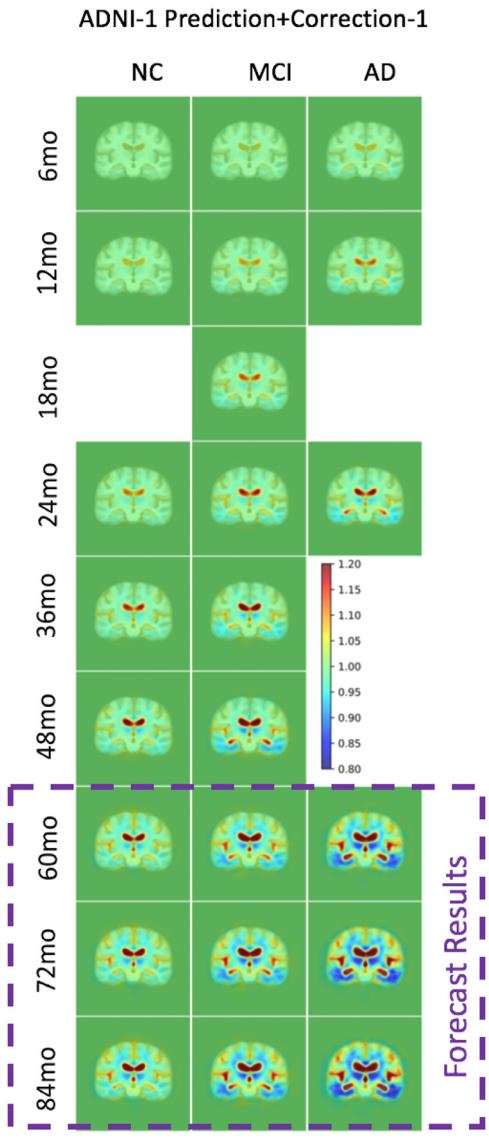

Fig. 5. Average Jacobian determinant over time and diagnostic category for `ADNI-1` Prediction+Correction-1 (experiments in `ADNI-2` show similar results). A value <1 indicates shrinkage and value >1 indicates expansion. The 60 month - 84 month results contained in the purple rectangle are forecasts 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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 access to the image measurement.

We use different forecasting approaches for the two scenarios. 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 6 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. This indicates that relative volume change within the ROI may not be accurately described by a regression geodesic for later time points. Note that SGR using the 6 month to 48 month time points only results in correlations around -0.5 for MMSE and 0.5 for DX on average. The correlations for the 60 months forecast in Table 6 show similar magnitudes. This suggests that the model successfully predicts into the near-future. Overall, our prediction+correction network performs as well as and sometimes even slightly better than SGR using optimization-based LDDMM. Fig. 5 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 7 shows **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 result in correlations close to the correlations when using all images. The **Forecast** correlations are only slightly weaker (0.02 to 0.05 lower in absolute values) than the original correlations using all images illustrating that FPSGR can approximately forecast future changes. The overall correlations in Table 7 show that the **Replace** approach performs better than the **Forecast** approach. Thus, both **Extrapolate-clinical** and **Extrapolate-image** experiments justify the use of FPSGR in predicting near future longitudinal trends. Besides, Fig. 5 shows the forecast results for 60 month, 72 month and 84 month. Results illustrate a clear trend extending the existing 6 month to 48 month deformations.

4.5. ***S5***: *Atrophy assessment via transitivitiy analysis and sample size estimates*

Section 4.2 used atrophy score regression results to establish that FPSGR does not produce deformations with significant bias to overestimate or underestimate volume change in the analyzed `ADNI` data. In the following, we investigate bias in more detail via a transitivity analysis. We also provide sample size estimates.

Transitivity analysis (Fox et al., 2011) is a common approach to test registration bias for atrophy measures derived from images. As our work is concerned with image *regression* for multiple time-points, rather than pair-wise image registration, we modify the approach of Fox et al. (2011) for transitivity analysis. Specifically, in Fox et al. (2011) three sequential scans *A, B,* and *C* are used. Transitivity is then assessed by measuring atrophy when directly registering $A \to C$ versus composing the two registration results for $A \to B$ and $B \to C$. Instead, to assess transitivity effects for regression, we selected a group of 115 patients from `ADNI-1` with longitudinal data for five time-points: I_0, I_1, I_2, I_3 and I_4, with $t_0 < t_1 < t_2 < t_3 < t_4$. We then perform three different regressions all using FPSGR with a correction network: (1) full FPSGR over all images, G: $I_0 \to I_1 \to I_2 \to I_3 \to I_4$; (2) FPSGR only over the first three images, $F1$: $I_0 \to I_1 \to I_2$; and (3) FPSGR over the last three images $F2$: $I_2 \to I_3 \to I_4$. To compare with the result of FPSGR over all time-points we compose the transformations obtained from F_1 and F_2 to obtain atrophy measures over the entire time range. We assess atrophy with respect to the baseline at the last timepoint, t_4. Fig. 6 shows the resulting atrophy differences. Specifically, we

Table 6
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; Red indicates that SGR LDDMM 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
60mo	Forecast	SGR LDDMM-1	-0.5242	1.34e-13	0.5157	3.85e-13	173
		SGR Pred+Corr-1	-0.5193	2.48e-13	0.5240	1.38e-13	
		SGR LDDMM-2	-0.4501	2.32e-10	0.4761	1.43e-11	180
		SGR Pred+Corr-2	-0.4582	9.97e-11	0.4652	4.73e-11	
72mo	Forecast	SGR LDDMM-1	-0.4607	1.60e-10	0.4507	4.37e-10	174
		SGR Pred+Corr-1	-0.4615	1.47e-10	0.4667	8.52e-11	
		SGR LDDMM-2	-0.3662	3.18e-7	0.4233	2.15e-9	184
		SGR Pred+Corr-2	-0.3793	1.09e-7	0.4259	1.67e-9	
84mo	Forecast	SGR LDDMM-1	-0.3986	1.40e-6	0.4108	6.17e-7	137
		SGR Pred+Corr-1	-0.3946	1.83e-6	0.4211	2.98e-7	
		SGR LDDMM-2	-0.3293	4.65e-5	0.3622	6.53e-6	147
		SGR Pred+Corr-2	-0.3187	8.35e-5	0.3609	7.12e-6	

Table 7
Forecast results which are based on the 6mo and 24mo images compared with results obtained when using all available time-points. 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 correcting 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.

ADNI-1			MMSE	p-value	DX	p-value	#data
36mo	All months	SGR LDDMM-1	-0.5142	4.29e-20	0.5300	1.81e-21	277
		SGR Pred+Corr-1	-0.5069	1.71e-19	0.5296	1.99e-21	
	Forecast	SGR Pred+Corr-1	-0.4708	1.42e-16	0.4980	1.21e-18	
	Replace	SGR Pred+Corr-1	-0.5097	1.37e-19	0.5375	5.47e-22	
	All months	SGR LDDMM-2	-0.4334	3.79e-13	0.4815	2.93e-16	256
		SGR Pred+Corr-2	-0.4393	1.67e-13	0.4863	1.34e-16	
	Forecast	SGR Pred+Corr-2	-0.4005	3.34e-11	0.4301	7.40e-13	
	Replace	SGR Pred+Corr-2	-0.4164	4.51e-12	0.4582	1.38e-14	
48mo	All months	SGR LDDMM-1	-0.7456	2.01e-13	0.6635	5.20e-10	69
		SGR Pred+Corr-1	-0.7443	2.30e-13	0.6575	8.43e-10	
	Forecast	SGR Pred+Corr-1	-0.6541	1.10e-9	0.6317	5.86e-9	
	Replace	SGR Pred+Corr-1	-0.6668	3.98e-10	0.6800	1.31e-10	
	All months	SGR LDDMM-2	-0.6889	2.25e-10	0.5927	1.98e-7	65
		SGR Pred+Corr-2	-0.7005	8.31e-11	0.6067	8.49e-8	
	Forecast	SGR Pred+Corr-2	-0.6403	9.25e-9	0.5460	2.55e-6	
	Replace	SGR Pred+Corr-2	-0.6307	1.79e-8	0.5973	1.50e-7	

calculated a relative atrophy score difference $\frac{s_{F2\circ F1}-s_G}{s_G}$. Results are mostly centered at zero with a slight shift (a median value of -7.4%[13] in the violin plot) observable towards negative values, suggesting saturation effects with time. Overall, the mean shift and hence the transitivity errors of the approach are relatively small. This shift might be mitigated by exploring more advanced regression models in the future (for example, a time-warped variant of FPSGR as discussed in the future work Section 5). Next, we illustrate with a simple toy example why saturations may cause such a negative shift.

Fig. 7 illustrates our toy atrophy example. Here, atrophy is large at the beginning and then starts to saturate. Consequentially, from the perspective of a transitivity check, regressing over all timepoints will overestimate atrophy at the last timepoint (where saturation has already set in). Breaking the regression into two parts will result in a model that is more faithful to the data and can better model the saturation. Hence, the atrophy measure at the last timepoint will be smaller. Consequentially a negative relative atrophy error will result, consistent with what we observed for real data in Fig. 6.

As suggested by Fox et al. (2011); Fleishman and Thompson (2017b), sample size is a good measure to assess the distribution of atrophy scores within diagnostic groups. Specifically, we used N80 sample size. N80 sample size is the estimated number

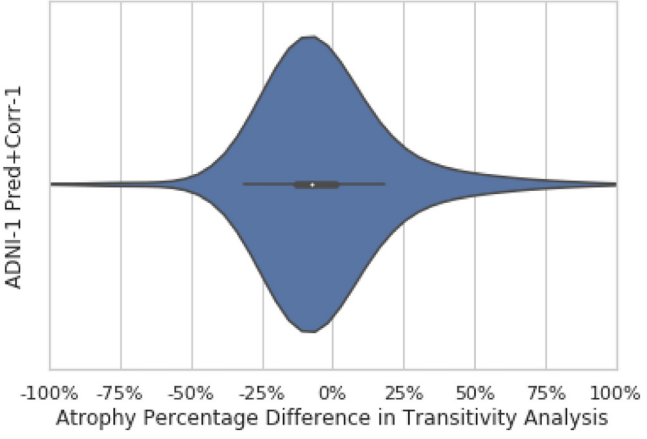

Fig. 6. Violin plot of the transitivity analysis.

of individuals required to detect a 25% reduction in the mean rate of atrophy, with 80% power, and with 95% confidence in the result. The formula to calculate N80 is as follows:

$$N80 = \frac{2\sigma^2(z_{1-0.05/2} + z_{0.8})^2}{(0.25\mu)^2}. \quad (6)$$

Here, μ is the average atrophy score for a prediction, σ is the standard deviation, and z_α is the value at which the cumulative standard normal distribution equals α. Numerically evaluating z_α re-

[13] The atrophy values range from -4.06 to 18.31 with a median difference value of -0.3932.

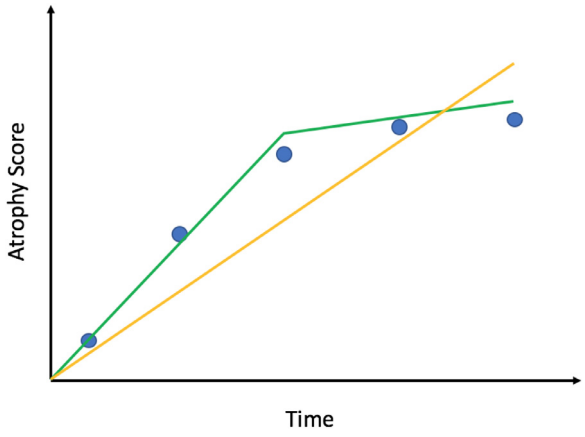

Fig. 7. Toy example to illustrate the transitivity analysis results of Fig. 6. The green line illustrate the two separate regression results (F_1, F_2) which are composed to obtain the deformation and from it the atrophy measure at the last timepoint. The yellow line indicates the regression results when using all timepoints at once (G). Because the deformation is fast at the beginning and slows down later, G will overestimate atrophy at the last timepoint. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sults in $N80 \approx 250.88 \times (\sigma/\mu)^2$. Table 8 shows the results for the ADNI−1 and ADNI−2 datasets for the three diagnostic categories: NC, MCI, and AD. A lower N80 score indicates a lower variance of the atrophy score, a higher average atrophy score or both. The more severe the disease (NC < MCI < AD), the lower the sample size estimation. Table 8 shows that in general FPSGR (with correction network) has either similar or lower N80 sample size estimates than optimization-based SGR LDDMM. This is likely due to the lower variance of FPSGR as supported by the tighter confidence intervals of FPSGR shown in Section 4.2.

Our results for regression of atrophy score, transitivity analysis, and sample size estimates indicate that FPSGR is an effective method to learn a general mapping from images to the initial momentum of an approximate geodesic.

5. Conclusion & future work

We proposed a fast approach for geodesic regression (FPSGR) to study longitudinal image data. FPSGR incorporates the recently proposed FPIR (Yang et al., 2016; 2017) into the SGR (Hong et al., 2012c) 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 (Hong et al., 2012c; 2017), without compromising accuracy. Consequently, FPSGR facilitates the analysis of large-scale imaging studies. Experiments on the ADNI−1 and 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.

Several limitations should be acknowledged:

Firstly, the model is relatively simple and attempts to model longitudinal changes via an approximated geodesic, combined with a linear regression model on the estimated atrophy scores. While such simple models are a desirable first step (as they simplify estimations) they, of course, may be too simplistic to model, for example, atrophies saturating over time (where large changes can initially be observed, but changes diminish later on). Such saturation effects may explain decreases in correlations for predicted months when predicting further ahead (see Table 6). Additional evidence for such saturation effects is given by Table 7, where the correlations of the **Replace** approach are higher than for the **Forecast** approach, indicating that stat-ROI deformations show less change between later time points than between earlier time points.

Secondly, correlations with clinical variables are moderate. This could, for example, be the case because the stat-ROI we choose only provides a very spatially limited view of the development of AD, or because the specific clinical variables we test are not strongly correlated with this particular stat-ROI. Though previous studies (Fleishman and Thompson, 2017b; 2017a; Hua et al., 2013; 2016) have shown the usefulness of such a statistically determined ROI and the studied clinical variables, it would be interesting to expand our study to other areas within the brain and to additional clinical variables.

Thirdly, the proposed framework requires the training of a deep neural network. Hence, what it captures will depend on the training data. Specifically, the testing images are required to have the same characteristics as the ones during training. Encouraging results have been obtained in Yang et al. (2017) for cross-dataset applications of models (using image intensity normalization), but our work only investigated dataset-specific models. Furthermore, changing registration parameters would require re-training the

Table 8
Estimated N80 sample size for ADNI−1 and ADNI−2. Results for ADNI−1 Pred+Corr-2 and ADNI−2 Pred+Corr-2 are similar and are omitted here for brevity. FPSGR shows similar and for ADNI−2 often smaller sample size estimates compared to optimization-based SGR LDDMM.

SGR LDDMM-1	6mo	12mo	18mo	24mo	36mo	48mo
NC	758	246		204	197	140
MCI	203	161	154	146	138	86
AD	125	101		101		
ADNI-1 Pred+Corr-1	6mo	12mo	18mo	24mo	36mo	48mo
NC	783	222		198	185	120
MCI	207	162	153	145	137	84
AD	127	104		101		
SGR LDDMM-2	3mo	6mo	12mo	24mo	36mo	
NC	844	418	310	253	87	
MCI	418	336	304	282	82	
AD	98	60	68	27		
ADNI-2 Pred+Corr-1	6mo	12mo	18mo	24mo	36mo	
NC	688	361	271	231	67	
MCI	384	311	288	268	80	
AD	92	59	67	30		

network. Note however that the trained model in some sense goes beyond the original registration model: it captures the *statistics* of *all* the registrations in the training set and hence becomes, for example, less susceptible to outliers.

There are several possible avenues for future work. To address the possible saturation effects, it would be interesting to explore alternative models and extensions to FPSGR. A straightforward and easy to compute extension would be to develop an FPSGR variant to allow for dynamic time-warping, similar to what has been proposed in Hong et al. (2014b); Durrleman et al. (2013). As the underlying registrations for FPSGR can be computed very fast, such a time-warped variant could likely also be optimized very quickly and could address saturations while keeping model complexity at a minimum. More ambitiously, FPSGR could be extended to a hierarchical model (in the spirit of Singh et al. (2016)) to jointly model longitudinal data across patients. The resulting model would be significantly more complex than FPSGR or its envisioned time-warped variant, but would be expected to also greatly benefit computationally from replacing costly numerical optimizations to compute registration by approximate regression models. Combinations with spline models (Singh et al., 2015) to capture an overall population trend are also conceivable, though significantly more complex.

Finally, as we currently use separate models for ADNI−1 and ADNI−2 (as these datasets use different image acquisition protocols), it would also be interesting to explore more generic models that are trained on a set of different datasets and hence can be applied across a wider range of datasets without retraining them. As registration settings influence the registration results, it would also be of great interest to investigate approaches that allow estimating these parameters from data. 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 across multiple time-points, instead of focusing on pair-wise image registrations, as done in FPIR (Yang et al., 2016; 2017).

Acknowledgements

Research reported in this publication was supported by the National Institutes of Health (NIH) and the National Science Foundation (NSF) (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 (http://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.

Overview of Appendix

The following sections provide additional analysis results and explanations for our proposed method. Specifically, Appendix A provides details on the mathematical formulation of FPSGR. Appendix B details the data distributions in ADNI−1 and ADNI−2. Appendix C shows statistical testing results for the differences in correlation strength between atrophy measures with clinical variables between SGR and pairwise registration. Appendix D lists the corresponding statistical results when comparing SGR LDDMM and FPSGR (SGR Prediction+Correction). Lastly, Appendix E contains additional visualizations for the estimated linear regressions for the atrophy scores and highlights their consistency with disease severity.

Appendix A. Estimating the initial momentum of FPSGR

This section describes the mathematical formulation of simple geodesic regression and how it is used for FPSGR. We start by defining the quadratic distance d^2 in Eq. (1) between two images A and B as

$$d^2(A,B) = \frac{1}{2}\int_0^1 \|v^*\|_L^2 dt, \tag{A.1}$$

$$\text{where } v^* = \arg\min_v \frac{1}{2}\int_0^1 \|v\|_L^2 dt + \frac{1}{\sigma^2}\|Q(1) - B\|_2^2,$$

$$\text{s.t.} \quad Q_t + \nabla Q^T v = 0, \text{ and } Q(0) = A \ .$$

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

$$\Phi = \text{Exp}_{\text{Id}}((t_i - t_0)v^B) \circ \text{Exp}_{\text{Id}}(-(t_i - t_0)v^A), \tag{A.2}$$

which can be approximated to first order as

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

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

$$d^2(A(t_i), B(t_i)) \approx \frac{1}{2}(t_i - t_0)^2 \langle K(m^B - m^A), m^B - m^A \rangle, \tag{A.4}$$

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

$$E(\bar{I}, \overline{m}) = \frac{1}{2}\langle \overline{m}, K\overline{m} \rangle + \frac{1}{2\sigma^2}\sum_i (t_i - t_0)^2 \langle K(\overline{m} - m_i), \overline{m} - m_i \rangle, \tag{A.5}$$

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. (A.5) w.r.t. $\overline{m}$ results in

$$\nabla_{\overline{m}} E = K[\overline{m} + \frac{1}{\sigma^2}\sum_i (t_i - t_0)^2(\overline{m} - m_i)] \stackrel{!}{=} 0. \tag{A.6}$$

Thus,

$$\overline{m} = \frac{\sum_i (t_i - t_0)^2 m_i}{\sigma^2 + \sum_i (t_i - t_0)^2}. \tag{A.7}$$

In practice, σ^2 is very small and can thus be omitted. Furthermore, m_i is obtained by either registering $\bar{I}$ 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 Eq. A.7 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 $\bar{I} = I_0$ as

$$\overline{m} \approx \frac{\sum_i (t_i - t_0)\widetilde{m}_i}{\sum_i (t_i - t_0)^2}. \tag{A.8}$$

Appendix B. Distribution of diagnostic groups in ADNI-1/2 for predictions

For completeness and to be able to better appreciate the data we used, this section details the distributions of the diagnostic groups we used for our prediction experiments. Tables B.1 and B.2 show these distributions for the ADNI−1 and the ADNI−2 datasets respectively. Diagnostic groups are based on the information on the ADNI website http://www.adni.loni.usc.edu. We combine MCI and LMCI in ADNI−1, Normal and SMC in ADNI−2, and EMCI and LMCI in ADNI−2, because such detailed diagnoses are only available for the baseline images. Images at later time points are only labeled as NC, MCI, and AD. This has already been noticed in Section 4. Each case is reflected as a blue point in the visualizations of Appendix E.

Table B.1
Distribution of Pred/Corr−1 and Pred/Corr−2 cases in ADNI−1. MCI* is the combination of the MCI and LMCI diagnostic groups.

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 B.2
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.

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

Appendix C. Statistical correlation difference between the regression model (FPSGR) and the pairwise model.

This section relates to Section 4.2 and provides statistical testing details to show that FPSGR shows stronger correlations than a pairwise registration approach for the clinical measures MMSE and DX. The statistical results are based on the correlations reported in Table 4 and tests are for differences in mean. Specifically, we first checked the normality of the distribution using a Shapiro-Wilk normality test. As can be seen from Table C.1 normality can be rejected at a significance level of 5%. Hence, using a paired t-test would be inappropriate. We therefore used a paired Wilcoxon signed-rank test to compare these correlations. Results are statistically significant at a significance level of 5% suggesting that FPSGR indeed improves correlation measures over pairwise registrations.

Table C.1
One-sided p-values for a Shapiro-Wilk normality test and Wilcoxon signed-rank test on MMSE and DX correlations between the FPSGR 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 correlation of pairwise prediction method is greater than that of FPSGR, i.e. the pairwise prediction method is statistically better than the FPSGR prediction method (at a significance level of 5%).

	Shapiro-Wilk normality test	Wilcoxon signed-rank test
MMSE	0.03425	0.0007959
DX	0.03596	0.0001951

Appendix D. Statistical correlation differences between optimization-based SGR and FPSGR

This section relates to Section 4.2 and shows statistical testing results for differences in correlations obtained via optimization-based SGR (i.e., SGR LDDMM) and FPSGR (i.e., SGR Pred+Corr). Specifically, we use a paired t-test to compare the correlations between atrophy and clinical variables for all the months for the ADNI−1 and ADNI−2 datasets in Table 4. Table D.1 shows the resulting p-values. Note that a t-test was appropriate based on the results of a Shapiro-Wilk normality test. We conclude that FPSGR using the correction approach works as well as, or better than, SGR via LDDMM optimization. This justifies the use of FPSGR for image regression.

Table D.1
Results of a Shapiro-Wilk normality test and a paired t-test on MMSE and DX correlations among SGR LDDMM, and FPSGR with correction network. The null-hypothesis for the Shapiro-Wilk normality test is that the difference between the two methods is normally distributed (at a significance level of 5%). The null-hypothesis for the paired t-test is that the correlation of SGR LDDMM is greater than that of FPSGR, i.e. the optimization based SGR method is statistically better than the FPSGR method (at a significance level 5%).

	Shapiro-Wilk normality test	Paired t-test
MMSE	0.5361	0.0530827
DX	0.2356	0.0186418

Appendix E. Linear regression of atrophy scores

Here, we show graphical illustrations of the linear regression results over the atrophy scores as presented in Table 3 of Section 4.2. Specifically, Fig. E.1 visually shows the linear regression results of the atrophy scores in ADNI−1 Pred+Corr-1 and

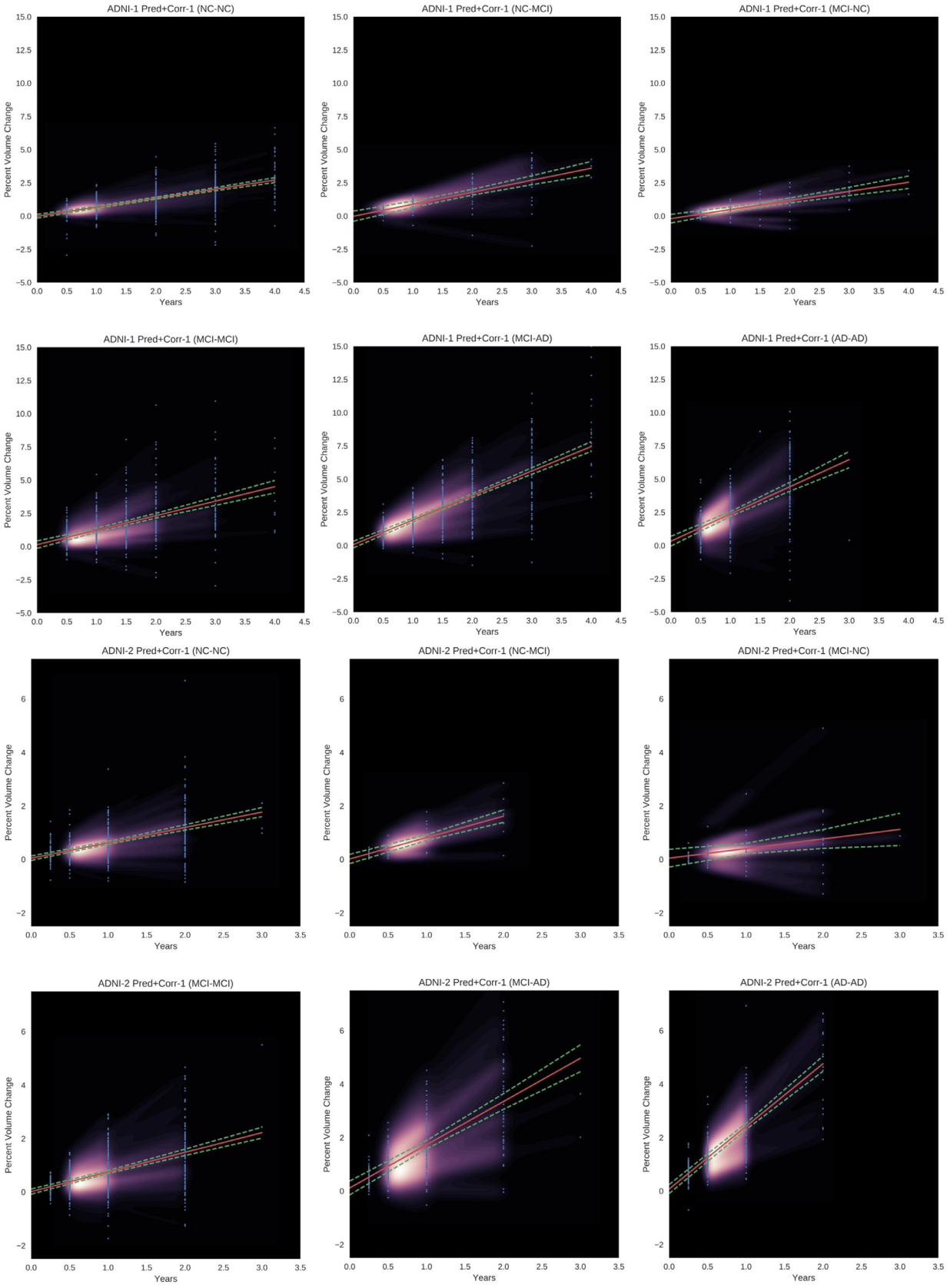

Fig. E.1. Linear regression of atrophy scores with respect to time for different diagnostic changes of `ADNI-1 Pred+Corr-1` and `ADNI-2 Pred+Corr-1`. Red line is the estimated regression line. green curves are the lower and upper bounds of the 95% confidence interval. Blue dots indicate actual data points. Bright white / purple images indicate kernel density estimations for all real data points illustrating dominant longitudinal trends in the data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ADNI−2 Pred+Corr-1. The slopes of the linear regressions are consistent with disease severity, i.e. NC-NC < NC-MCI, MCI-NC < MCI-MCI < MCI-AD, and NC-NC < MCI-MCI < AD-AD. All 95% confidence intervals contain zero, which indicates that FPSGR with correction did not produce deformations with significant bias to over- or underestimate volume changes.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2019.06.003.

References

Biobank website: http://www.ukbiobank.ac.uk.

Beg, M.F., Miller, M.I., Trouvé, A., Younes, L., 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. 61 (2), 139–157.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B (Methodological) 289–300.

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

Ding, Z., Fleishman, G., Yang, X., Thompson, P., Kwitt, R., Niethammer, M., Initiative, A.D.N., et al., 2017. Fast Predictive Simple Geodesic Regression. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 267–275.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: Learning optical flow with convolutional networks. In: ICCV, pp. 2758–2766.

Durrleman, S., Pennec, X., Trouvé, A., Braga, J., Gerig, G., Ayache, N., 2013. Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. Int. J. Comput. Vis. 103 (1), 22–59.

Fleishman, G., Thompson, P.M., 2017. Adaptive gradient descent optimization of initial momenta for geodesic shooting in diffeomorphisms. ISBI.

Fleishman, G., Thompson, P.M., 2017. The impact of matching functional on atrophy measurement from geodesic shooting in diffeomorphisms. ISBI.

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

Fox, N.C., Ridgway, G.R., Schott, J.M., 2011. Algorithms, atrophy and alzheimer's disease: cautionary tales for clinical trials. Neuroimage 57 (1), 15–18.

Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48 (1), 63–72.

Hong, Y., Golland, P., Zhang, M., 2017. Fast geodesic regression for population-based image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 317–325.

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

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

Hong, Y., Kwitt, R., Singh, N., Vasconcelos, N., Niethammer, M., 2016. Parametric regression on the grassmannian. IEEE Trans. Pattern Anal. Mach. Intell. 38 (11), 2284–2297.

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

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

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

Hua, X., Ching, C.R.K., Mezher, A., Gutman, B., Hibar, D.P., Bhatt, P., Leow, A.D., Jr., C.R.J., Bernstein, M., Weiner, M.W., Thompson, P.M., 2016. MRI-Based brain atrophy rates in ADNI phase 2: acceleration and enrichment considerations for clinical trials. Neurobiol. Aging 37, 26–37.

Hua, X., Hibar, D.P., Ching, C.R.K., Boyle, C.P., Rajagopalan, P., Gutman, B., Leow, A.D., Toga, A.W., Jr., C.R.J., Harvey, D.J., Weiner, M.W., Thompson, P.M., 2013. Unbiased tensor-based morphometry:improved robustness & sample size estimates for Alzheimer's disease clinical trials. Neuroimage 66, 648–661.

Iglesias, J.E., Liu, C.-Y., Thompson, P.M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging 30 (9), 1617–1634.

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

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

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17 (2), 825–841.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv: 1412.6980.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., et al., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage 46 (3), 786–802.

Liu, Z., Yeh, R., Tang, X., Liu, Y., Agarwala, A., 2017. Video frame synthesis using deep voxel flow. arXiv: 1702.02463.

Mazziotta, J.C., Toga, A.W., Evans, A., Fox, P., Lancaster, J., 1995. A probabilistic atlas of the human brain: theory and rationale for its development: the international consortium for brain mapping (icbm). Neuroimage 2 (2), 89–101.

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

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

Schuster, T., Wolf, L., Gadot, D., 2016. Optical flow requires multiple strategies (but only one network). arXiv: 1611.05607.

Shen, L., Saykin, A.J., Kim, S., Firpi, H.A., West, J.D., Risacher, S.L., McDonald, B.C., McHugh, T.L., Wishart, H.A., Flashman, L.A., 2010. Comparison of manual and automated determination of hippocampal volumes in mci and early ad. Brain Imaging Behav. 4 (1), 86–95.

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

Singh, N., Hinkle, J., Joshi, S., Fletcher, P.T., 2016. Hierarchical geodesic models in diffeomorphisms. Int. J. Comput. Vis. 117 (1), 70–92.

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

Singh, N., Vialard, F.-X., Niethammer, M., 2015. Splines for diffeomorphisms. Med. Image Anal. 25 (1), 56–71.

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

de Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I., 2017. End-to-end unsupervised deformable image registration with a convolutional neural network. arXiv: 1704.06065.

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

Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Quicksilver: fast predictive image registration–a deep learning approach. Neuroimage 158, 378–396.

Yushkevich, P.A., Avants, B.B., Das, S.R., Pluta, J., Altinay, M., Craige, C., 2010. 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), 434–445.

Zhang, M., Fletcher, P.T., 2015. Finite-dimensional Lie algebras for fast diffeomorphic image registration. In: IPMI, pp. 249–260.

Zhang, M., Liao, R., Dalca, A.V., Turk, E.A., Luo, J., Grant, P.E., Golland, P., 2017. Frequency diffeomorphisms for efficient image registration. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 559–570.