Supplementary Material: Fast Predictive Simple Geodesic Regression

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

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

This supplementary material contains further details of the FPSGR approach. Specifically, Sec. 1 investigates estimation bias. Sec. 2 shows that using simply the slope and y-intercept of the linear fit to the estimated atrophy scores contains information useful for classification thereby supporting the use of atrophy measures to assess our results. Sec. 3 explores atrophy measures. Sec. 4 shows forecasting results not only for the prediction + correction method, but also for the prediction method *without* correction. Sec. 5 shows example graphs illustrating the numerical convergence of the deep learning model during training. Lastly, Sec. 6 reports the overall runtime and training time of our approach and discusses computational cost in relation to existing optimization-based approaches.

1. Bias

Table 1 is an extended version of Table 3 in the main manuscript, which also includes the prediction-only (i.e., without correction) results. As pairwise registration results were discussed in the main document and used there to justify SGR they are no longer reported here. Specifically, Table 1 shows the estimated slopes, intercepts, and 95% confidence intervals for optimization-based SGR LDDMM and for FPSGR predictions with and *without* correction for `ADNI-1` and `ADNI-2`, respectively. SGR LDDMM-1 and SGR LDDMM-2 denote the optimization-based results split into the same testing groups used for the Pred-1/2 and Pred+Corr-1/2 results to allow for a direct comparison. All of the results show intercepts that are near zero relative to the range of changes observed and all prediction intercept confidence intervals contain zero. Similar to our discussion in the main manuscript, we conclude that (1) neither optimization-based SGR LDDMM nor FPSGR produce deformations with significant bias to overestimate or underestimate volume change; (2) a linear model of atrophy scores generated by FPSGR can capture intrinsic volume change (i.e., slope) among different diagnostic change groups; and (3) The prediction+correction approach produces results which are more similar to SGR LDDMM than the prediction-only approach.

In Table 1, all slopes are positive, indicating average volume loss over time. This is consistent with expectations for an aging and neuro-degenerative population. The slopes capture increasing atrophy with disease severity. In `ADNI-1/ADNI-2`, we expect $\text{Slope}_{\text{NC-NC}} < \text{Slope}_{\text{NC-MCI}}$; $\text{Slope}_{\text{MCI-NC}} < \text{Slope}_{\text{MCI-MCI}} < \text{Slope}_{\text{MCI-AD}}$; and $\text{Slope}_{\text{NC-NC}} < \text{Slope}_{\text{MCI-MCI}} < \text{Slope}_{\text{AD-AD}}$. The experimental groups (SGR LDDMM-1/2, SGR Pred-1/2, and SGR Pred+Corr-1/2) are consistent with these expectations (and also consistent with results in Hua et al. (2013)). The slopes estimated from the prediction+correction results are generally larger than the for the prediction-only model and closer to the slopes obtained via optimization-based SGR LDDMM. This indicates that the correction network can generally improve prediction accuracy.

2. Classification

We performed classification experiments to assess if the trends estimated via FPSGR carry information indicative of the diagnostic category. Note that these experiments make use of very limited information (i.e., they use *only* the slopes and y-intercepts of the estimated atrophy trends obtained via linear regression from the FPSGR-estimated atrophy measures, as discussed in main paper Sec. 4.2). Hence, these experiments are not expected to yield state-of-the art classification results which can be obtained using much more sophisticated biomarkers (Davatzikos et al., 2011; Suk and Shen, 2013; Westman et al., 2012).

Nevertheless, we use a similar experimental setup and perform pair-wise classifications between the three diagnostic groups on both the `ADNI-1` and `ADNI-2` datasets. In particular, we compare NC vs. AD, NC vs. MCI, and MCI-converter vs. MCI-nonconverter. Here, MCI-converter indicates that a patient diagnosed with MCI will

Email address: `zp-ding@cs.unc.edu` (Zhipeng Ding)

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

Table 1: Slope and intercept values for linear regression of volume change over time. Our notation for *slope* and *intercept* indicates [lower bound of 95% C.I., **point estimate**, upper bound of 95% C.I.]. The intervals of intercept estimates all contain zero. The slope changes between the different diagnostic groups. The #data column lists the number of data points analyzed.

develop into AD while MCI-nonconverter indicates that such a conversion did not occur throughout the imaging time-frame. We used a simple linear Support Vector Machine (SVM) to perform binary classifications for these three experiments. Table 2 shows the classification accuracies[1]. We only show results for ADNI-1 Pred-Corr-1 and ADNI-2 Pred-Corr-1. Similar results are obtained using the other models and are hence omitted here. For each experiment, we used two fold cross-validation on a dataset balanced with respect to diagnostic category. This is done to ensure that during training and testing of the SVM each class has the same number of samples (i.e., 186 NC vs 186 MCI, 100 NC vs 100 AD, 70 MCI-C vs 70 MIC-N, each separated into two folds). The reported accuracies are averages over the two folds. While our results are below state-of-the art (expected based on the simple two-dimensional slope/y-intercept feature we use) they clearly indicate that the estimated atrophy measures capture information discriminative for the different diagnostic groups. Example state-of-the-art results (Davatzikos et al., 2011; Suk and Shen, 2013; Westman et al., 2012) for the three diagnostic categories are NC vs AD: $\approx 96\%$, NC vs MCI: $\approx 90\%$, and MCI-C vs MCI-NC: $\approx 76\%$. However, these results were obtained using much more sophisticated biomarkers using complex models. Note also that the datasets for these results are not the same. Hence, this remains a qualitative comparison.

	NC vs. MCI	NC vs. AD	MCI-C vs MCI-NC
ADNI-1 Pred+Corr-1	66.22% (50%)	82.83% (50%)	68.22% (50%)
# samples	186 NC + 186 MCI	100 NC + 100 AD	70 MCI-C + 70 MCI-NC
ADNI-2 Pred+Corr-1	62.50% (50%)	83.77% (50%)	64.96% (50%)
# samples	178 NC + 178 MCI	142 NC + 142 AD	148 MCI-C + 148 MCI-NC

Table 2: Linear SVM classification results of three diagnostic groups with baseline accuracies in parentheses (i.e., random chance). MCI-C denotes MCI-converter; MCI-NC denotes MCI-nonconverter. Results show that the slope and y-intercept of the linear model of atrophy scores can be used as features to capture differences between different diagnostic groups.

3. Atrophy

This part extends the discussion of the atrophy measure in the main manuscript (Sec.4.2). Table 3 and Fig. 1 show the results of the Spearman rank-order correlation over all three diagnostic groups (NC, MCI and AD) combined. In detail, for ADNI-1/2, we randomly selected 200[2] cases from each diagnostic category at each month and calculated the Spearman rank-order correlation. Fig. 1 shows the results for 50 repetitions. Results are comparable with previous studies Fleishman and Thompson (2017b,a) as discussed in the main manuscript (Sec.4.2). Note that correlations at 18 month and 36 month do not follow an upward trend in Fig. 1. This is mainly, because diagnostic categories are not balanced for these two time-points. See Table 5 for details on data distribution. To study the dependence on diagnostic group, Fig. 2 therefore shows correlations separated by diagnostic groups, which do not exhibit the downward trend at 18 and 36 months.

Further, using the correction network, FPSGR achieves comparable and sometimes even slightly better performance compared to the optimization-based SGR LDDMM

[1]Note that we omitted training data of ADNI-1/2 Pred-Corr-1 in testing the classification accuracy.

[2]For ADNI-1 at 48 month, we selected 60, because there was not enough data; ADNI-2 36 month was omitted due to lack of data.

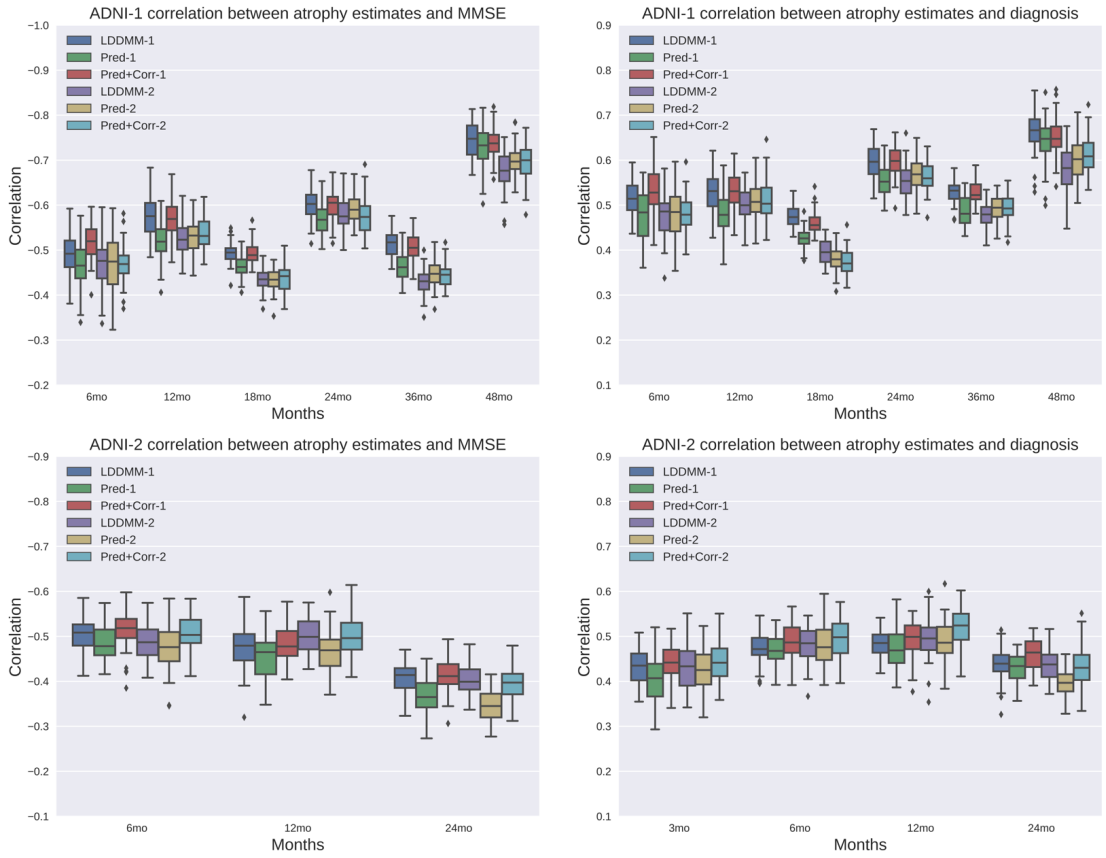

Figure 1: Boxplot of FPSGR-derived correlations with clinical variables in `ADNI-1` and `ADNI-2`. Prediction + correction results are comparable with optimization-based LDDMM. Adding a correction network generally improves over the prediction-only results.

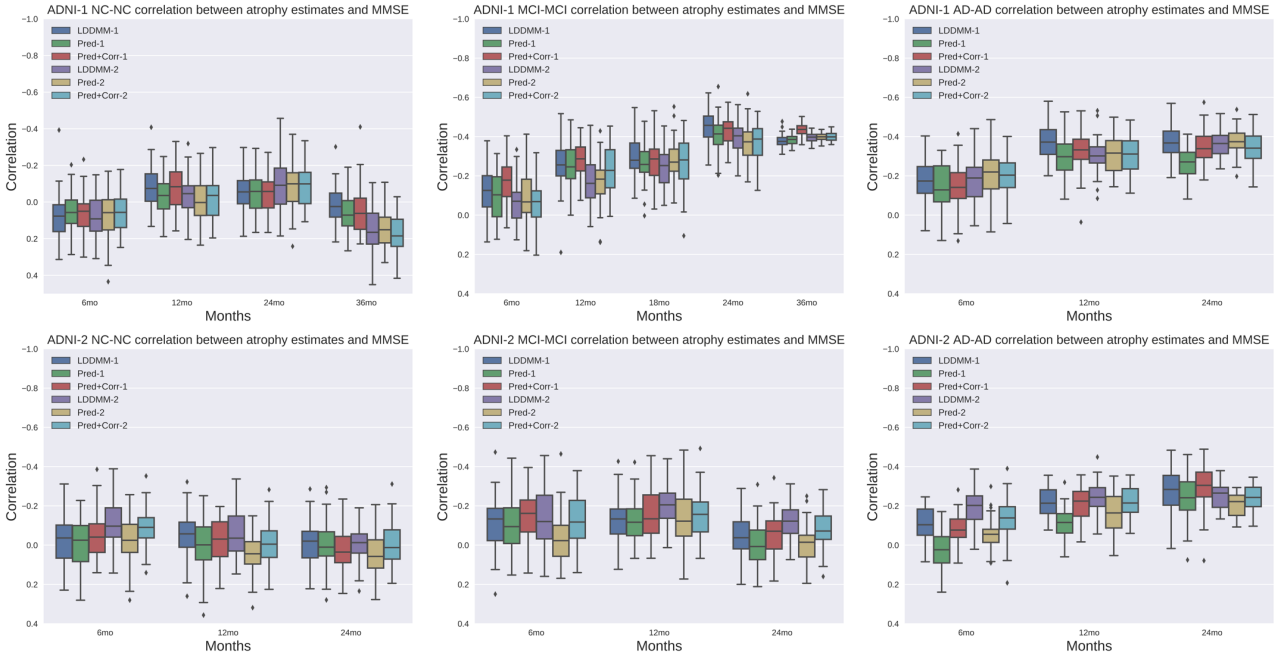

Figure 2: Boxplot of Spearman rank-order correlations between atrophy measures and MMSE with respect to time in `ADNI-1` and `ADNI-2`. **Top row:** `ADNI-1` NC-NC group (left), `ADNI-1` MCI-MCI group (middle), `ADNI-1` AD-AD group (right). **Bottom row:** `ADNI-2` NC-NC group (left), `ADNI-2` MCI-MCI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

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

Table 3: FPSGR-derived correlations with clinical variables, compared to correlations with clinical variables for SGR using optimization-based LDDMM. The #data column lists the number of data points analyzed. Green indicates that FPSGR using the prediction+correction network shows the strongest correlations; Yellow indicates that FPSGR using the prediction network alone shows the strongest correlations; Red indicates that SGR LDDMM shows the strongest correlations. The MMSE column lists correlations between atrophy scores and the mini-mental state exam scores; the DX column lists correlations between atrophy score and diagnostic category. Finally, the p-value column(s) list the p-values for the null-hypothesis that there is no correlation. Benjamini-Hochberg procedure was employed to reduce the false discovery rate and Purple highlight indicates statistically significant. FPSGR using the prediction+correction network generally improves performance over using the prediction network alone and frequently even performs slightly better than the SGR results obtained by optimization-based LDDMM.

method; see Table 3 for additional quantitative results. Specifically, FPSGR using the prediction+correction network performs best in 10 out of 18 comparisons for

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

Table 4: Results of a Shapiro-Wilk normality test and a paired t-test on MMSE and DX correlations among optimization-based LDDMM, FPSGR without prediction network and FPSGR with correction network. The null-hypothesis for the Shapiro-Wilk normality test is that the difference between column-method and row-method is normally distributed. The null-hypothesis for the paired t-test is that the correlation of the column-method is greater than that of the row-method, i.e. the column-method is statistically better than row-method (at a significance level of 5%). Green highlighted p-values indicate no rejection of the normality hypothesis (at 5% significance) and thus facilitate the paired t-test. p-values highlighted in red indicate a rejection of the normality null-hypothesis and consequently do not allow a paired t-test. Specifically, green highlighted p-values in paired t-test indicate that SGR Pred+Corr $\geq$ SGR LDDMM $>$ SGR Pred. Hence the FPSGR with correction network works best in terms of correlation with MMSE and DX.

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 5: Distribution of Pred/Corr-1 and Pred/Corr-2 cases in ADNI-1. MCI* is the combination of the MCI and LMCI diagnostic groups.

MMSE and in 14 out of 20 comparisons for diagnostic group. In the cases where FPSGR with prediction+correction network does not perform best its difference to the optimization-based method is generally very small. In general FPSGR using the correction network performs better than FPSGR without the correction network. To check for statistical differences in the performance of FPSGR, we use a paired t-test. Table 4 shows the resulting p-values for the three methods: optimization-

based SGR LDDMM, FPSGR without correction network (i.e., Pred) and FPSGR with correction network (i.e., Pred+Corr). For both correlation with MMSE and DX, FPSGR with correction network shows significantly better performance than FPSGR without correction network and slightly better performance than SGR LDDMM, which justifies using FPSGR with correction network. In summary, FPSGR captures correlations between atrophy and clinical measures well.

To further explore the correlations of atrophy with MMSE scores, we visualize them separated by diagnostic group where diagnosis did not change (i.e., NC-NC, MCI-MCI, AD-AD) in Fig. 2. For the `ADNI-1` dataset, we observe (as expected) very low correlations for the normal diagnostic group (with no clear trend), and much stronger correlations for the MCI and AD groups. MCI and AD also exhibit increasingly stronger correlations with time. In case of `ADNI-2`, the MCI group shows modest correlations, which remain consistent across time. Correlations are relatively low for the normal groups. The AD groups show increasingly strong correlations over time. In contrast to `ADNI-1`, `ADNI-2` focuses mainly on earlier stages of the diagnostic groups (Hua et al., 2016). Hence, the deformations in `ADNI-2` are generally smaller than in `ADNI-1`. This may explain why the NC and MCI diagnostic groups show consistent correlation values over time (instead of stronger correlations as for AD in `ADNI-2` or the MCI and AD groups in `ADNI-1`).

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

$$s(\phi)(x) := (1 - \det(D\phi(x))) \times 100 \ . \qquad (1)$$

I.e., each voxel in a stat-ROI has an associated atrophy score. Fig. 3 shows kernel density estimates of the highest 10% local correlations in a violin plot. For the `ADNI-1` MCI and AD groups, a clear shift toward stronger correlations can be observed over time, similar to the boxplots of Fig. 2. This indicates the progression of the disease. Correlations for the normal groups in `ADNI 1/2` are mostly centered around a modest correlation (as expected). In `ADNI-2`, only the AD diagnostic group shows a shift towards stronger correlations over time. All the other diagnostic groups show a relatively consistent distribution over time. This is also similar to Fig. 2.

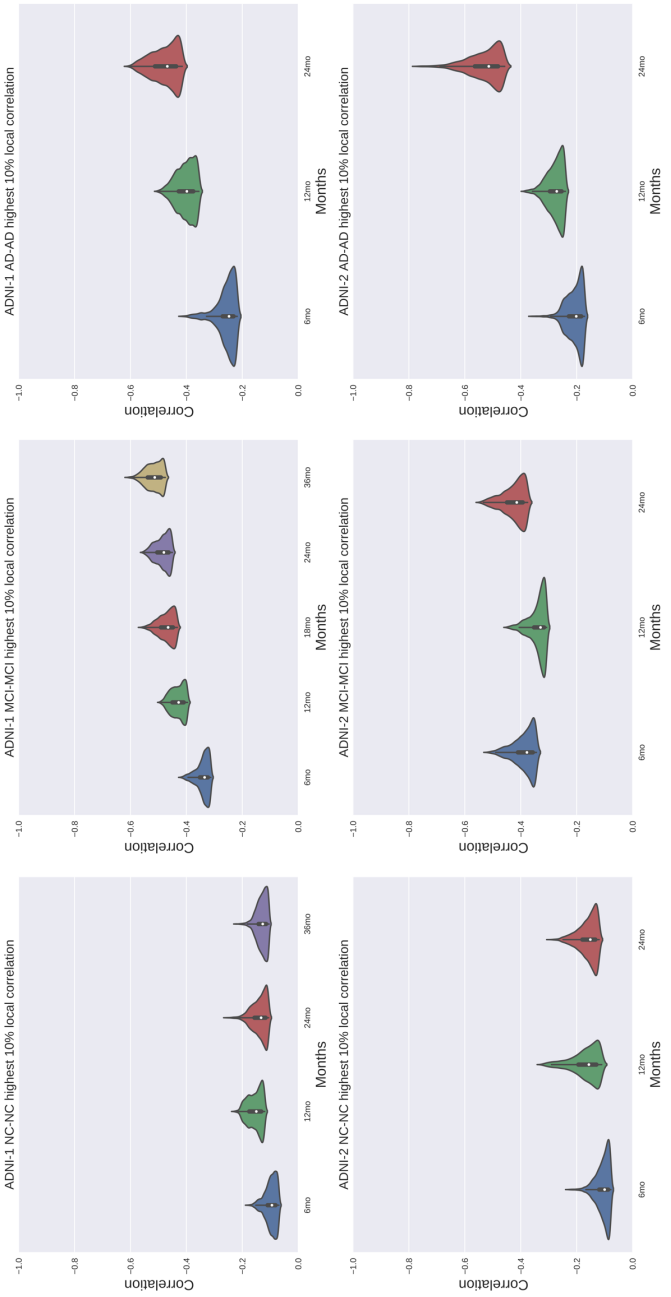

Figure 3: Kernel density estimates of highest 10% local correlations of atrophy with MMSE within the ROI depicted in the main paper. **Top row:** results of NC group, MCI group and AD group from `ADNI-1`. **Bottom row:** results of NC group, MCI group and AD group from `ADNI-2`. Results show a shifting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

4. Forecasting

The forecasting results shown here correspond to the results of Sec. 4 of the main manuscript, but include the results for the prediction-only models. Specifically, Table 6 corresponds to Table 7 in the main manuscript with the prediction-only results included. Prediction + correlation results are similar to prediction-only results.

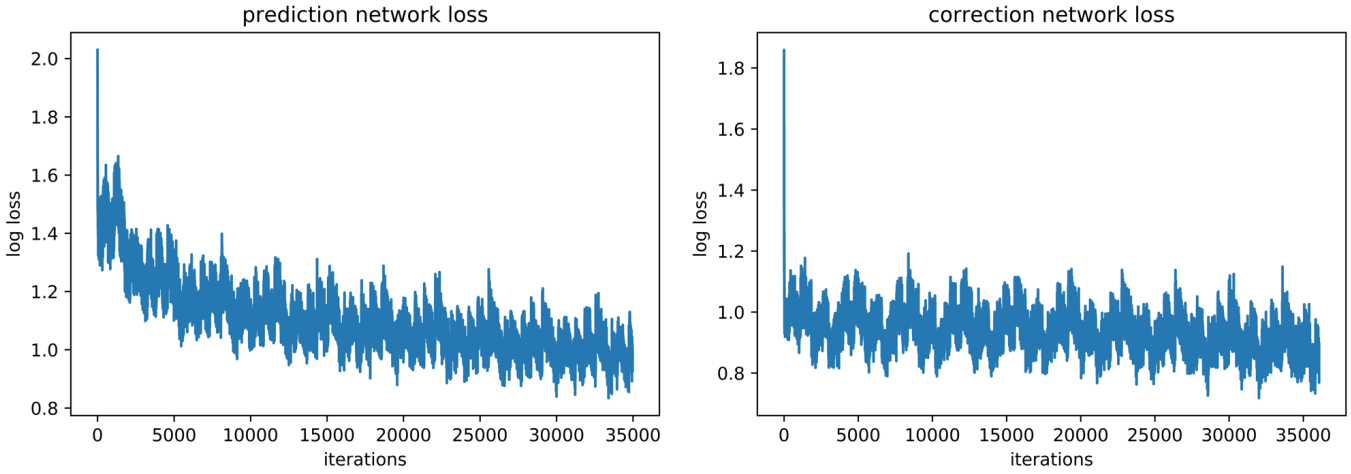

Figure 4: Logarithmic loss curves for prediction and correction network. Left: log base 10 loss for `ADNI-1` Pred-2 model; Right: log base 10 loss for `ADNI-1` Corr-2 model. Loss curves for the other models are similar.

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-1	-0.4731	7.38e-17	0.4926	2.42e-18	
		SGR Pred+Corr-1	-0.5069	1.71e-19	0.5296	1.99e-21	
	Forecast	SGR Pred-1	-0.4583	1.09e-15	0.4825	1.93e-17	
		SGR Pred+Corr-1	-0.4708	1.42e-16	0.4980	1.21e-18	
	Replace	SGR Pred-1	-0.4923	3.43e-18	0.5104	1.21e-19	
		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-2	-0.4425	1.07e-13	0.4894	7.99e-17	
		SGR Pred+Corr-2	-0.4393	1.67e-13	0.4863	1.34e-16	
	Forecast	SGR Pred-2	-0.4078	1.36e-11	0.4398	1.95e-13	
		SGR Pred+Corr-2	-0.4005	3.34e-11	0.4301	7.40e-13	
	Replace	SGR Pred-2	-0.4202	2.75e-12	0.4635	6.27e-15	
		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-1	-0.7294	1.18e-12	0.6458	2.08e-9	
		SGR Pred+Corr-1	-0.7443	2.30e-13	0.6575	8.43e-10	
	Forecast	SGR Pred-1	-0.6332	5.29e-9	0.6165	1.70e-8	
		SGR Pred+Corr-1	-0.6541	1.10e-9	0.6317	5.86e-9	
	Replace	SGR Pred-1	-0.6446	2.27e-9	0.6478	1.78e-9	
		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-2	-0.6995	9.08e-11	0.6048	9.49e-8	
		SGR Pred+Corr-2	-0.7005	8.31e-11	0.6067	8.49e-8	
	Forecast	SGR Pred-2	-0.6528	3.79e-9	0.5568	1.46e-6	
		SGR Pred+Corr-2	-0.6403	9.25e-9	0.5460	2.55e-6	
	Replace	SGR Pred-2	-0.6334	1.49e-8	0.5970	1.53e-7	
		SGR Pred+Corr-2	-0.6307	1.79e-8	0.5973	1.50e-7	

Table 6: Forecast results compared with real data results. The #data column lists the number of data points analyzed. The Benjamini-Hochberg procedure was employed to reduce the false discovery rate (FDR). Purple highlight indicates statistically significant results after corrections for multiple comparisons. Forecast results are calculated by using SGR excluding 36mo and 48mo data points and then predicting 36mo and 48mo correlations. Results are compared based on the same dataset except for two invalid data points for the 36mo data.

5. Numerical Convergence During Training

For completeness, Fig. 4 shows some convergence curves for the training of a prediction and a correction network. We use a batch size of 50, a learning rate of 0.0001, and 10 epochs to train both the prediction and correction network. We use approximately 180,000 patches to train each model. At a batch size of 50 this corresponds to 3,600 iterations/epoch. We observe that 10 epochs are sufficient for convergence of the models. No overfitting was noticed in our experiment, nor in the experiments of Yang et al. (2017).

6. Efficiency

Training a network takes about 20 hours on one NVIDIA GTX1080Ti. Thus, the prediction+correction model takes about 40 hours of training time, because two networks are trained. To prepare the training dataset, it took about 50 hours to use optimization-based LDDMM to obtain the initial momenta. For the prediction+correction model an additional 0.5 hours are required since a correction step is used to generate differences of the initial momentum and the predicted initial momentum; hence, the initial momentum needs to be predicted, based on which the transform from the target to the source domain can be computed, which is then used to spatially warp the target image back to the source image domain. The analysis of the `ADNI-1` dataset (2,646 pairwise image registrations) using FPSGR took about 24 hours. Hence the total time for such a large-scale analysis was about 114.5 hours, which is less than 5 days. Using optimization-based LDDMM to process the same dataset took over 40 days using a single GPU. The computation time is similar for `ADNI-2`. Hence, the net improvement for the large-scale image analysis of `ADNI-1` and `ADNI-2` is roughly 70 days including the training time of the models.

References

Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ. Prediction of mci to ad conversion, via mri, csf biomarkers, and pattern classification. Neurobiology of aging 2011;32(12):2322–e19.

Fleishman G, Thompson PM. Adaptive gradient descent optimization of initial momenta for geodesic shooting in diffeomorphisms. In: ISBI. 2017a. .

Fleishman G, Thompson PM. The impact of matching functional on atrophy measurement from geodesic shooting in diffeomorphisms. In: ISBI. 2017b. .

Hua X, Ching CRK, Mezher A, Gutman B, Hibar DP, Bhatt P, Leow AD, Jr. CRJ, Bernstein M, Weiner MW, Thompson PM. MRI-based brain atrophy rates in ADNI phase 2: acceleration and enrichment considerations for clinical trials. Neurobiology of Aging 2016;37:26–37.

Hua X, Hibar DP, Ching CRK, Boyle CP, Rajagopalan P, Gutman B, Leow AD, Toga AW, Jr. CRJ, Harvey DJ, Weiner MW, Thompson PM. Unbiased tensor-based morphometry:improved robustness & sample size estimates for Alzheimer's disease clinical trials. NeuroImage 2013;66:648–61.

Suk HI, Shen D. Deep learning-based feature representation for ad/mci classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2013. p. 583–90.

Westman E, Muehlboeck JS, Simmons A. Combining mri and csf measures for classification of alzheimer's disease and prediction of mild cognitive impairment conversion. Neuroimage 2012;62(1):229–38.

Yang X, Kwitt R, Styner M, Niethammer M. Quicksilver: Fast predictive image registration–a deep learning approach. NeuroImage 2017;158:378–96.