Longitudinal FreeSurfer for Reliable Imaging Biomarkers

Martin Reuter[13], H. Diana Rosas[1], and Bruce Fischl[23]

[1] Neurology, Massachusetts General Hospital, Harvard Medical School
[2] Radiology, Massachusetts General Hospital, Harvard Medical School
[3] Massachusetts Institute of Technology

Abstract. Longitudinal image processing algorithms aim to increase the reliability of automatic measurements in sequential imaging data by incorporating the knowledge that images come from the same subject. When transferring information across time great care needs to be taken to treat all time points the same in order to avoid introducing a systematic processing bias. We have presented an unbiased longitudinal processing framework previously. Here we discuss an extension to improve results in subjects with large ventricles. Furthermore, we demonstrate methods to investigate longitudinal data via scatter plots, as well as linear modles and non-linear flow lines. Using normalized brain volume to estimate disease progression, we find increased atrophy rates in several structures in advanced disease stages. We also highlight how linear fits into percent volume changes of different structures can help identify outliers and detect early disease effects. Finally we show improved surface placement accuracy when using longitudinal image processing in cases with low image quality relative to independent processing.

1 Introduction

Longitudinal study designs are integral to modeling and understanding disease trajectories with the goals of quantifying drug effects, predicting onset of symptoms and estimating disease staging and progression. With the growing availability of longitudinal MRI data, sophisticated image processing technologies and statistical analysis tools are being developed that aim to increase the reliability of sequential measurements. Under the assumption that accuracy is not biased, reduced variability positively impacts the statistical power to distinguish groups, e.g. when studying drug effects in clinical trials.

For example, the following mechanisms can help reduce the variability when quantifying repeated measures:

- a common space, achieved by co-registering the time points for each subject.
- transfer of information, such as the brain mask, Talairach transform, non-linear atlas registration, labels or surfaces across time to either directly determine results in, or initialize processing of follow-up time points.
- temporal regularization, to enforce smoothness across time.

These procedures, however, can easily bias longitudinal results by treating a single time point differently, usually the baseline scan. For example an algorithm that encourages follow-up results to be more similar to the baseline result certainly reduces variability (as can easily be proved in test-retest studies) but sacrifices accuracy and produces biased atrophy measures. Even worse, such a method is likely to affect a disease group with significant atrophy more severely than a control group with relatively less change. Such *over regularization* can be introduced by explicit temporal smoothness constrains, e.g. in the non-linear registration procedures.

Another source of bias is often introduced when mapping follow-up images to baseline, for a direct comparison in the same space. All follow-up images will thus be resampled while the baseline image remains untouched. These *interpolation asymmetries* can significantly bias a longitudinal study and can result in severe underestimation of sample sizes due to overestimation of effect sizes [11, 10]. As described in [6] and demonstrated in [9], interpolation asymmetries are not the only source of bias. Transferring information, such as labels or surfaces, from baseline to initialize processing in follow-up time points can be sufficient to bias results. In short, treating a single time point consistently different from others has strong potential to induce a bias and should be entirely avoided. While it is theoretically possible to detect processing bias simply by switching the order of time points, some types of bias can be small with respect to measurement noise and can therefore remain elusive. Furthermore, processing bias can be regional. For example, we showed in [9] that hippocampal volume was not affected by an induced bias, while measurements in other regions were clearly biased when initializing follow-up segmentation with labels from baseline. For these reasons we proposed in [8, 9] a novel longitudinal processing framework that employs an unbiased within-subject template to construct a common mid-space and average (median) image. Results from this unbiased template are subsequently used to initialize each time point with common information, increasing reliability while avoiding over-regularization by letting the algorithms evolve freely.

In this paper we describe an extension of our framework to improve the non-linear atlas registration in subjects with large ventricles. Furthermore, we investigate linear models and non-linear flow lines for longitudinal data. Using normalized brain volume to estimate disease progression, we find increased atrophy rates in several structures in advanced disease stages. We also demonstrate that linear fits into percent volume changes of different structures can help identify outliers, and detect early disease effects. Finally we show improved surface placement accuracy when using longitudinal image processing in cases with low image quality relative to independent processing.

2 Methods

Our longitudinal processing framework is not based on pairwise comparisons of time points to quantify atrophy, nor does it compare each time point to the subject template directly. Instead we compute a full set of subcortical and cortical

volume and thickness measurements for each time point: labels and volumes for subcortical gray matter structures, white matter, corpus callosum, ventricles, cerebellum, as well as white matter and pial surfaces, local cortical thickness measurements and cortical parcellations. Reliability of these measures is obtained by working in the *common template space* and by *initializing* processing in each time point with common results from the template.

First all time points are processed independently to obtain intensity normalized and skull stripped images. Then the within-subject template image is created by an iterative inverse-consistent robust registration [7] of each time point to an average image. The average is based on the intensity median at each voxel, instead of the mean, to remove the influence of outliers (e.g. scans with strong motion artifacts). See [9, 5] for details of the procedure. Following the co-registration and template creation, the template image is processed with FreeSurfer to obtain initial results for segmentations and surfaces, basically a robust estimate of the subject anatomy.

Finally each time point is processed by initializing several procedures with results from the template, for instance, the non-linear atlas registration or white matter and pial surfaces are initialized from the subject template and then allowed to evolve freely. The brain mask is computed once for each subject and remains fixed across time, as does the affine Talairach transformation, which is meaningful under the assumption of fixed head sizes[4]. See [9] for results demonstrating the improved reliability and discrimination power in synthetic examples and different group studies. For the results presented here, FreeSurfer 5.1 was used with a modification for some subjects with large ventricles as described below.

The automated labeling in FreeSurfer depends on the calculation of a transformation T that maps the individual subject into a probabilistic atlas coordinate system. This procedure (documented in [1, 2]) employs a number of terms in an energy functional, the minimization of which results in the desired warp field. These terms can be grouped into data matching term (e.g. finding the warp that maximizes the probability of the observed image given the atlas parameters) and smoothness terms (e.g. priors on the space of allowable warp fields). The data matching terms depend on the subject anatomy being drawn from the same distribution as those used to create the atlas, which currently consists of a set of 40 subjects, distributed in age and Alzheimer's pathology. However, when individual subject anatomies diverge significantly from those used to construct the atlas, the procedure can fail. Specifically, if the ventricles in the subject are considerably larger than those seen in the atlas, the gradient of the energy functional, used in the numerical minimization, will not point in the correct direction.

In order to resolve this issue and make the atlas warp robust to the presence of significantly enlarged ventricles we designed a pre-processing step to specifically handle the enlarged ventricle case. The procedure begins with the calculation

[4] It is possible to relax these constrains for data sets where head size changes, e.g. pediatric data

of a distance transform in the atlas coordinates that results in a scalar field over the image that specifies the distance to the borders of the atlas ventricles. We then employ the identical energy functional used in our standard warp, but instead of computing the gradient of the data matching terms we substitute the gradient of the distance transform, thus moving the warp field in the direction of ventricular expansion. The result is a procedure that expands the ventricles in the atlas as much as necessary in order to optimally match the subject data. An example of this technique is given in Figure 1.

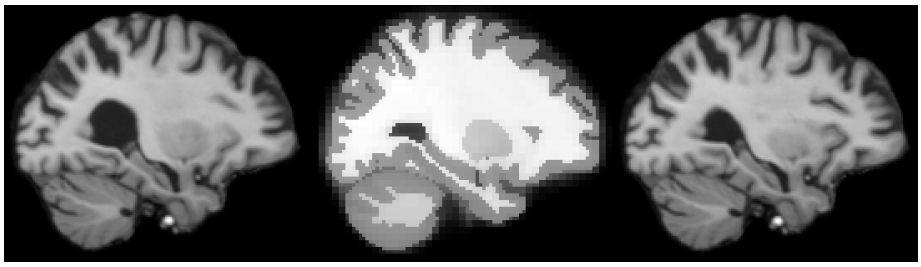

Fig. 1. Example of large ventricle pre-processing. Left: individual subject data with significantly enlarged ventricles affine transformed to the atlas. Center: the target atlas. Right: individual subject data warped by the transform resulting from the large ventricle pre-processing. Note the significant reduction in ventricular size, a decrease that is sufficient for our standard nonlinear warp to result in good alignment and hence accurate segmentation of the ventricular system.

To analyze changes in thickness on the cortical surfaces, slopes of thickness with respect to time (or any other variable) can be estimated on a per-subject basis. The longitudinal processing framework ensures that surfaces are in vertex correspondence across time. Smoothing and linear regression can therefore be easily performed on the thickness maps for each subject. Results are then mapped via a spherical registration [3] to a template subject to establish correspondences for a group analysis. In all steps (linear fitting and smoothing) it is ensured that values from outside the cortex cannot influence any results by applying a common cortex label as a mask in both the within-subject processing and later in the group analysis steps.

3 Results

3.1 Data

In this section we present results on the data from the MICCAI 2012 atrophy challenge to assess measurement reliability and bias using structural MRI in Alzheimer's disease. The data consists of 46 patients fulfilling NINCDS-ADRDA criteria for probable AD and 23 age-matched elderly controls, scanned at 0, 2,

6, 12, 26, 38 and 52 weeks (and a subset additionally at 18 and 24 months). All subjects also had 2 back-to-back scans at 3 of the time points. All participants of the challenge were blinded with respect to age, gender, group membership and time (only the baseline time point was known). All inputs are T1-weighted images with voxel sizes: $0.9375, 0.9375, 1.5$ mm and dimension: $256 \times 256 \times 124$ (except scan 192 F with only 123 slices).

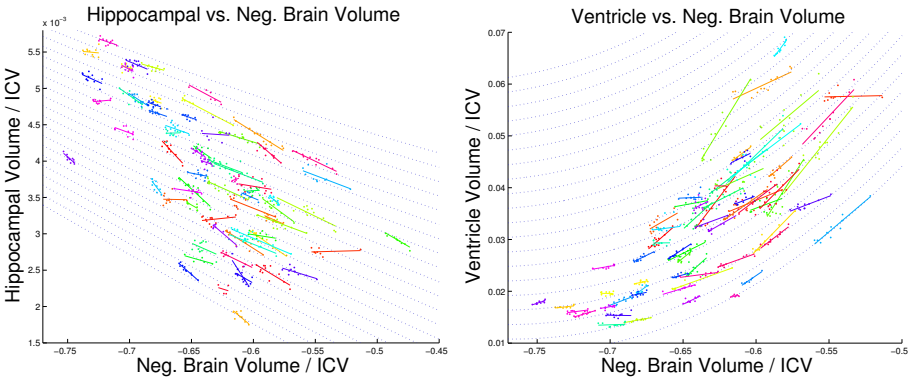

Fig. 2. Scatter plots of ICV normalized hippocampal and ventricle volume as a function of negative brain volume (without ventricles). Smaller brain volumes to the right of the x-axis indicate progressed disease or age. Colors denote individual subjects. The dotted blue lines are field lines satisfying a linear relation between subject averages and longitudinally-derived slopes.

For longitudinal data it is recommended to first look at a scatter plot of the data. This can help one to understand how the data behaves and whether outliers are present. Since the time of each scan is unknown, we decided to use the negative brain volume divided by intracranial volume (ICV) as a measure of time. Furthermore, we will split the subjects into two groups based on mean normalized brain volume below. Figure 2 depicts the ICV normalized hippocampal volume (sum of left and right hippocampal volume) and the normalized ventricle volume each as a function of the negative normalized brain volume. Within each subject, longitudinal slopes were obtained via linear fits. It can be seen that both hippocampal and ventricle volume shows a correlation, cross-sectionally and longitudinally, with brain volume. While the longitudinal slopes of the hippocampus are independent of the actual brain volume (and approximately the same as the cross-sectional slope), the relation between ventricles and brain volume seems to be more complex and indicates faster ventricle enlargement for subjects with smaller brain volumes (i.e., larger negative brain volume, to the right of the plot). To test these hypotheses, we first obtain the slopes y_i' from the robust linear fits of the data for each subject. Then we fit the subjects mean volumes x_i and y_i into these slopes using robust regression (with bisquare weight

function [4]):

$$y_i' = \alpha x_i + \beta y_i + \gamma + \epsilon_i \tag{1}$$

After estimating the parameters α, β and γ, the corresponding differential equation can be solved analytically (yielding an exponential plus linear term). Selected flow lines are shown in Fig. 2 as dotted blue lines and highlight the overall behavior of both the means and derivatives. According to the p values of the fit, subject slopes significantly depend on the position of the subject means for the ventricles only, not for the hippocampus. This can also be seen in Fig. 3 where box plots show the difference in median slopes between subjects stratified at mean normalized brain volume for the ventricle slopes and slopes into perirhinal thickness (average of both hemispheres, normalized by $\sqrt[3]{ICV}$).

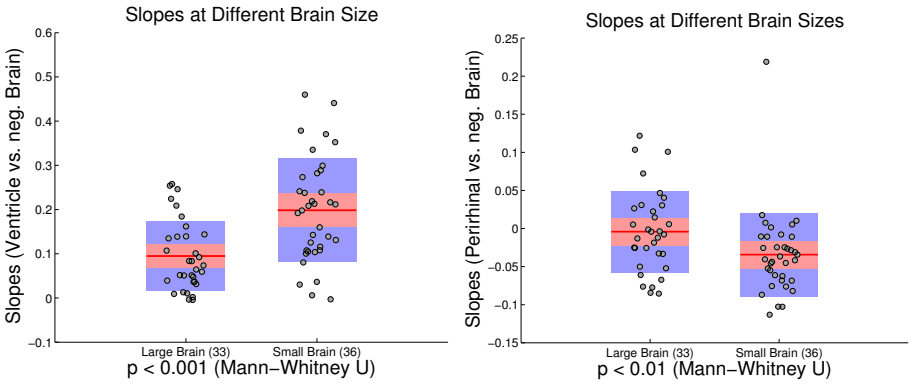

Fig. 3. Box plots showing subject slopes of venctricle volume (left) and perirhinal thickness (right) for subjects with small and large brain volumes. The medians differ significantly in both cases according to the Mann-Whitney U test (also called Wilcoxon Rank-Sum test).

As indicated in Fig. 3 (right), perirhinal cortical thickness is also affected by the disease, presumably in later stages. In order to analyze atrophic behavior in the full cortex, we computed vertex-wise within-subject slopes of thickness as a function of negative normalized brain volume (smoothed at full width half maximum 15). The subjects were then split into two groups according to mean normalized brain volume and the assumption was tested that slopes are steeper in the group with smaller brain volumes. Figure 4 shows the regions where this hypothesis is likely true (red: $p < 0.05$, yellow: $p < 0.001$, two sided test, the opposite assumption of flatter slopes does not yield significant results anywhere).

A faster ventricle enlargement and cortical thinning with respect to brain volume loss in subjects with advanced brain atrophy can have several causes. The majority of these subjects is likely diseased and further progressed, compared to the subjects with small ventricles and large brain volume, who probably consist, to a larger extend, of controls (group memberships are unknown to the authors).

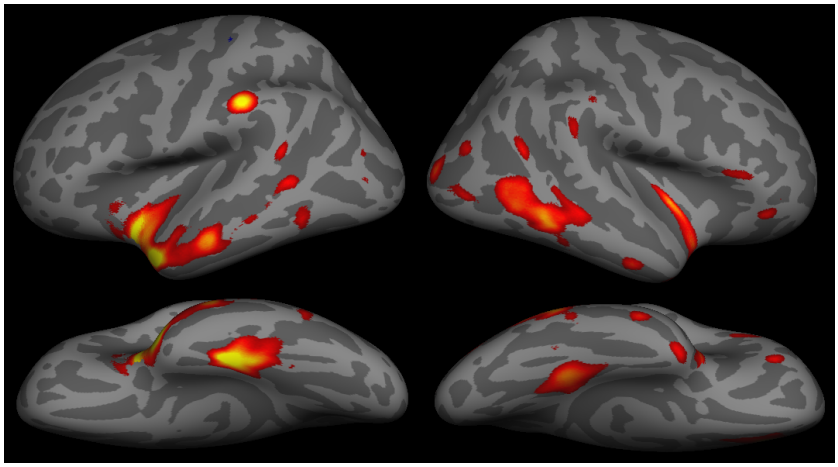

Fig. 4. Cortical regions where the within-subject slopes of thickness vs. negative normalized brain volume are significantly steeper in subjects with small brain volume (red: $p < 0.05$, yellow: $p < 0.001$).

Note that the ventricle sizes can also be affected by head size, gender, as well as age. To remove head size differences, we already normalized all volume measures by ICV. The original volume measures also demonstrate a very similar behavior as the ICV-normalized results presented above (not shown).

3.2 Percent Volume Change

In addition to the confounding effect of different head sizes, gender and especially age, absolute volume differences may not be very informative: a 0.5ml hippocampal volume change may not be considered much in a 9ml hippocampus, but seems large in a 3ml one. For these reasons we analyze percent volume changes. If the time variable was known, it would be possible to compute yearly percent changes or develop more sophisticated longitudinal mixed effects models. Here, however, it is possible to analyze the relationship between the structures after dividing measurements by the mean structure volume for each subject. Figure 5 shows plots of the normalized volumes for 2 subjects with a linear fit. The baseline measurement is denoted by a green circle. For a short period of time (1-2 years), longitudinal atrophy measures can be assumed to be approximately linear, as can be seen by the good linear fit.

The first subject on the top (blue line) shows around 8% volume change in the ventricle and approx. 5% in the hippocampus. The second subject (bottom) covers a wider range (20% ventricle volume increase and 10% hippocampal decrease). Both subjects are potentially diseased as the yearly hippocampal volume loss of 4% to 5% is relatively large (assuming the time points cover the maximum time frame of 24 months). The former subject already has large ventricles (121ml) and may be in a later stage of the disease (with decreasing growth of

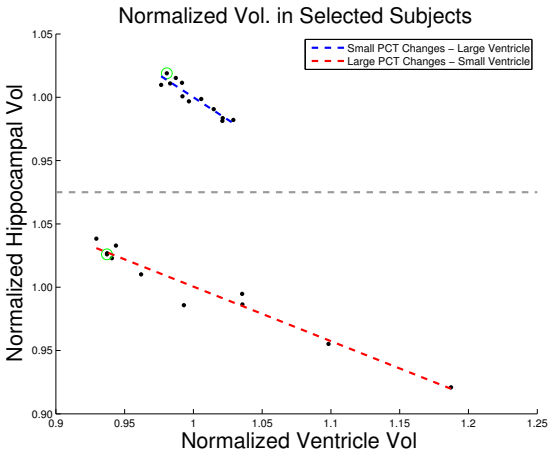

Fig. 5. Two subjects with different percent changes in ventricle volume (due to different ventricle sizes).

the ventricle) while the second subject may be in an early stage with rather large percent changes (ventricle volume 43ml). Controls can be expected to have less correlation among the variables, due to smaller percent changes and thus a larger influence of measurement noise.

3.3 Outlier

Images with poor quality had been previously removed from this dataset, causing subjects to have a variable number of time points. Still we detected several cases with low image quality based on percent change scatter plots. These cases were *not* removed in the submission of our results to the atrophy challenge. A thorough quality check and removal of low quality images from the longitudinal processing stream can therefore be expected to further improve results. Figure 6, for example, shows a subject with 10% hippocampal volume increase with respect to baseline. Inspection of the input image indicates strong motion artifacts.

In order to analyze the effect of outliers on the subject template, we removed time points E and H from subject 237 and time point G from subject 217. Each subject template was recomputed and the remaining time points processed with the new template. Figure 7 shows the effect of time point removal on the volumes and slopes. Results are relatively stable, but changes in the template can affect the longitudinal results, especially in subjects with few time points. We therefore recommend to inspect the data and remove outliers if possible before processing.

Time points with low quality images can however profit from longitudinal processing as information from all time points (via the template) is used to improve individual results. An example can be seen in Figure 8, showing 237 E (left) and 217 G (right) with the pial surfaces overlayed. When processing the images directly surface quality is low; the red surface is not accurately placed

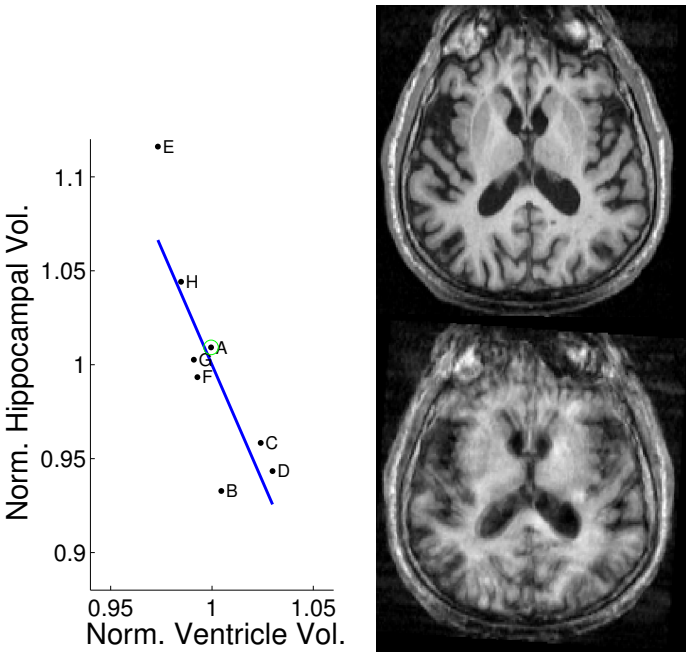

Fig. 6. Left: Normalized hippocampal volume vs. normalized ventricular volume with linear fit. Right top: baseline time point A. Right bottom: time point E. It can be seen that time point E (also H) are outliers in the scatter plot. Inspection of the image shows strong motion artifacts in E (some motion also in H, not shown).

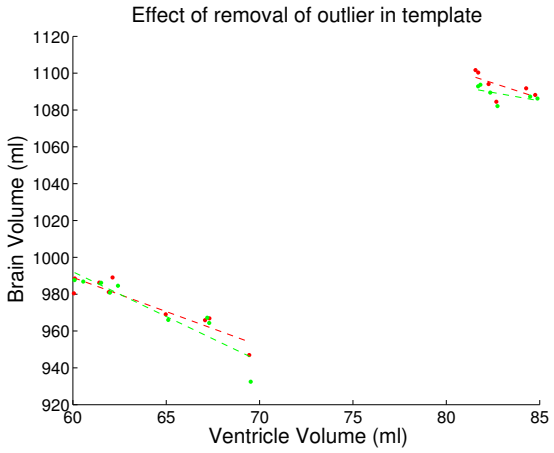

Fig. 7. Outlier images were removed in these two subjects to investigate the effect. Red: Initial volumes and fit estimated using a template constructed from all time points. Green: new results after removing the outlier images from the template construction.

along the gray matter boundary. Our longitudinal framework, however, is capable of accurately placing the pial surface, based on the good initialization from the subject template.

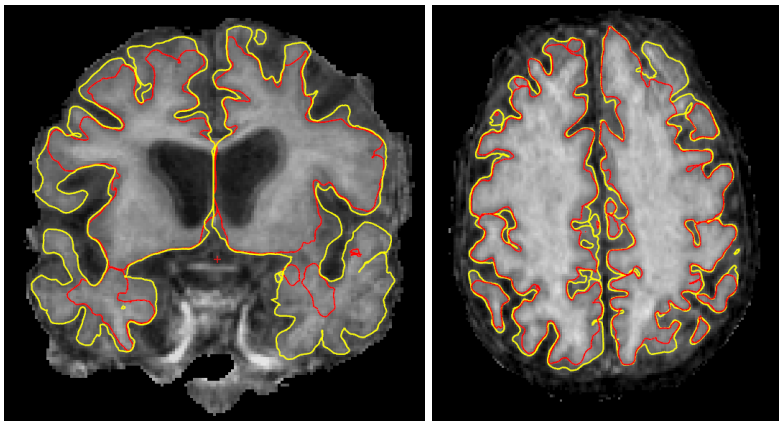

Fig. 8. Two low quality images showing the pial surface (overlayed as a 2D curve intersecting this slice) estimated from the image alone (red) and estimated using the subject template and longitudinal processing (yellow). The original surface placement (red) is very inaccurate and improves significantly when using the longitudinal processing stream (yellow).

3.4 Big Ventricles

Several subjects with enlarged ventricles have been identified and pre-processing in the non-linear registration step has been applied, as described above, to improve the atlas registration in these cases. An example can be seen in Figure 9 where the linear fit into the normalized brain and ventricle volumes improves the correlation of measurements drastically ($\rho = -0.74$ to $\rho = -0.97$).

4 Conclusion

In spite of the fact that important variables such as time, age, gender or group membership are hidden, we demonstrated methods to investigate the data via scatter plots, as well as linear and non-linear models. It may be possible to distinguish diseased subjects from controls by analyzing the volumes together with the slope of the within-subject linear fits of one structure as a function of a different structure as indicated by the results presented in Fig. 2 and Fig. 3. It is unlikely that diseases affect the whole brain uniformly. This can also be seen from the cortical thickness analysis in Fig. 4, potentially indicating increased cortical atrophy rates in advanced disease stages. Furthermore, percent changes can be

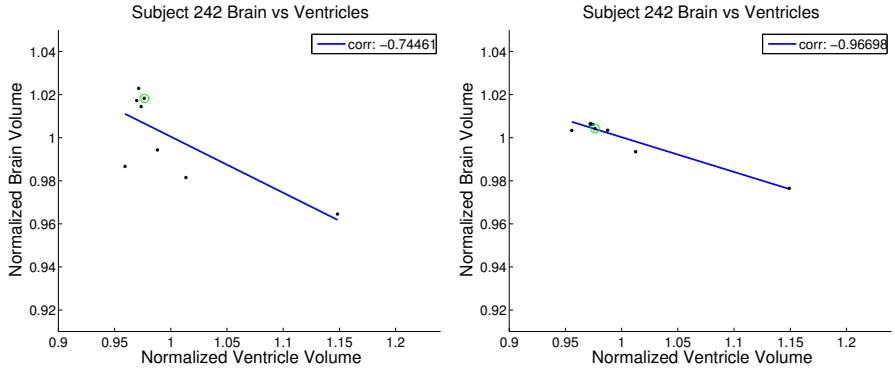

Fig. 9. Improvements of linear fit when switching from regular processing (left) to special treatment for large ventricles (right).

helpful to distinguish early disease stages from controls, even pre-symptomatic stages (see [9] for an example in Huntington's disease). While controls may still have similar volumes as early diseased or pre-symptomatic subjects, percent changes can reveal disease effects in structures that get affected early (see Fig. 5).

Moreover, linear fits into percent change plots can help detect outliers for further investigation. Motion artifacts, for example, can produce large outliers in the computed measurements and therefore badly influence longitudinal analyses, as the assumption of normally distributed data is violated. This is especially problematic if one group (usually the diseased) is more likely to move in the scanner, e.g. in Huntington's disease or in older subjects. Methods with strong temporal regularization may reduce the effect of such outliers, at the cost of introducing a bias and reducing their ability to detect true large changes. Removing outliers manually can be expected to improve results of the remaining data. Therefore, for a fair comparison of different processing methods, it will be essential to fix the dataset to a common subset that does not contain missing values for any individual method.

We were also able to demonstrate the robustness of the longitudinal stream when dealing with low quality images, where surface placement accuracy increased significantly. Furthermore, the template image remains relatively stable due to the robustness of the median when removing (or adding) time points. However, small changes still affect the results and propagate into individual time points. A thorough analysis of stability will be necessary in the future. Finally, we observed an increased correlation of measurements in cases with large ventricles when performing a pre-processing step for the non-linear atlas registration.

Acknowledgments

Support for this research was provided in part by the National Center for Research Resources (P41RR14075, and the NCRR BIRN Morphometric Project

BIRN002, U24RR021382), the National Institute for Biomedical Imaging and Bioengineering (R01EB006758), the National Institute on Aging (R01AG022381, U01AG024904), the National Institute for Neurological Disorders and Stroke (R01NS052585, R01NS042861, P01NS058793, R21NS072652, R01NS070963). Additional support was provided by *The Autism & Dyslexia Project* funded by the Ellison Medical Foundation, the National Center for Alternative Medicine (RC1AT005728), and was made possible by the resources provided by Shared Instrumentation Grants (S10RR023401, S10RR019307, S10RR023043). The authors would also like to thank Louis Vinke and Iman Aganj for help with processing the data and valuable comments.

References

1. Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M.: Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. Neuron 33(3), 341–355 (2002)
2. Fischl, B., Salat, D.H., van der Kouwe, A., Makris, N., Ségonne, F., Quinn, B.T., Dale, A.M.: Sequence-independent segmentation of magnetic resonance images. NeuroImage 23(Supplement 1), 69 – 84 (2004)
3. Fischl, B., Sereno, M.I., Tootell, R.B., Dale, A.M.: High-resolution intersubject averaging and a coordinate system for the cortical surface. Human Brain Mapping 8(4), 272–284 (1999)
4. Holland, P.W., Welsch, R.E.: Robust regression using iteratively reweighted least-squares. Communications in Statistics - Theory and Methods 6(9), 813–827 (1977)
5. Reuter, M.: Longitudinal processing in freesurfer. http://freesurfer.net/fswiki/LongitudinalProcessing (2009)
6. Reuter, M., Fischl, B.: Avoiding asymmetry-induced bias in longitudinal image processing. NeuroImage 57(1), 19–21 (2011), http://dx.doi.org/10.1016/j.neuroimage.2011.02.076
7. Reuter, M., Rosas, H.D., Fischl, B.: Highly accurate inverse consistent registration: A robust approach. NeuroImage 53(4), 1181–1196 (2010), http://dx.doi.org/10.1016/j.neuroimage.2010.07.020
8. Reuter, M., Rosas, H.D., Fischl, B.: Unbiased robust template estimation for longitudinal analysis in freesurfer. In: 16th Annual Meeting of the Organization for Human Brain Mapping (2010)
9. Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B.: Within-subject template estimation for unbiased longitudinal image analysis. NeuroImage 61(4), 1402–1418 (2012), http://dx.doi.org/10.1016/j.neuroimage.2012.02.084
10. Thompson, W.K., Holland, D.: Bias in tensor based morphometry Stat-ROI measures may result in unrealistic power estimates. NeuroImage 57(1), 1–4 (2011)
11. Yushkevich, P.A., Avants, B., Das, S.R., Pluta, J., Altinay, M., Craige, C., ADNI: Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: An illustration in ADNI 3 tesla MRI data. NeuroImage 50(2), 434–445 (2010)