Assessment of Reliability of Multi-site Neuroimaging Via Traveling Phantom Study*

Sylvain Gouttard[1], Martin Styner[2,3], Marcel Prastawa[1], Joseph Piven[3], and Guido Gerig[1]

[1] Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT 84112
gouttard@sci.utah.edu
[2] Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599
[3] Department of Psychiatry, University of North Carolina, Chapel Hill, NC 27599

Abstract. This paper describes a framework for quantitative analysis of neuroimaging data of traveling human phantoms used for cross-site validation. We focus on the analysis of magnetic resonance image data including intra- and inter-site comparison. Locations and magnitude of geometric deformation is studied via unbiased atlas building and metrics on deformation fields. Variability of tissue segmentation is analyzed by comparison of volumes, overlap of tissue maps, and a new Kullback-Leibler divergence on tissue probabilities, with emphasis on comparing probabilistic rather than binary segmentations. We show that results from this information theoretic measure are highly correlated with overlap. Reproducibility of automatic, atlas-based segmentation of subcortical structures is examined by comparison of volumes, shape overlap and surface distances. Variability among scanners of the same type but also differences to a different scanner type are discussed. The results demonstrate excellent reliability across multiple sites that can be achieved by the use of the today's scanner generation and powerful automatic analysis software. Knowledge about such variability is crucial for study design and power analysis in new multi-site clinical studies.

1 Introduction

Image data from patient populations acquired and pooled across multiple sites become necessary to collect larger number of samples for improved statistical power in investigating neuroanatomic correlates of disease. Examples are the Alzheimer's Disease Neuroimaging Initiative (ADNI) or the Autism Centers of Excellence (ACE) network projects. Such efforts require advanced concepts for multi-site imaging calibration by developing standardization of protocols, phantom calibration, and evaluation of cross-site differences and scanner stability. The NIH sponsored effort by the BIRN consortium aims at developing *"the ability to conduct clinical imaging studies across multiple sites ..."* [1]. Whereas a major effort of BIRN so far was dedicated to standardization and calibration of structural, functional and diffusion MRI, less knowledge is available on the

* This research was supported in part by the National Institutes of Health under Grant RO1 HD055741 (Autism Center of Excellence, project IBIS), and in part by the National Alliance for Medical Image Computing (NA-MIC), funded by the NIH through Grant U54 EB005149.

D. Metaxas et al. (Eds.): MICCAI 2008, Part II, LNCS 5242, pp. 263–270, 2008.

joint variability of scanning and quantitative analysis although this is the most crucial information for multi-site studies. Han et al. [2] proposed atlas renormalization to account for segmentation differences across scanners and demonstrated results with two scanners and 27 human phantoms. This paper focuses on a framework for evaluation of multi-site image data of human traveling phantoms using existing image analysis methods. We tested both intra- and inter-site variability but the following discussion focuses on the latter. Although applied to a limited study as a pilot for a new multi-site pediatric imaging study, our aim is a) to propose a framework for such analysis including a set of tests and b) to demonstrate reliability that can be achieved by using the latest generation of clinical scanners and well-established brain segmentation pipelines.

2 Evaluation of Traveling Phantom Image Data

This study does not focus on validity, i.e. closeness of results to an existing truth, but on reproducibility of image data and extracted quantitative measurements. We present three different analyses, deformation-based analysis to evaluate the amount of distortion across scanners, comparison of probabilistic brain tissue segmentations to judge the variability of such commonly applied tissue analysis, and finally assessment of the variability of subcortical structure volumes. Please note that the first evaluation is based on measurements on raw images, whereas the other two are metrics obtained through the *aperture* of a segmentation algorithm, thereby reflecting combined errors of the whole system from image acqusition at multiple sites to segmentation algorithms. This only will give the designers of such studies the necessary data on expected variability of measurements and will help to estimate the required sample size via power analysis.

2.1 Study Design

This traveling phantom study has been initiated as pilot calibration work for collecting image data in a new pediatric multi-site autism study. The purpose was two-fold, measuring the variability of imaging across three sites equipped with the *same* scanner system and upgrade (Siemens 3T Tim Trio), and evaluation if a *different* scanner (Siemens 3T Allegra head-only) could be used for this study. Two human phantoms (male, age 26 and 27), visited the four different sites within one week and got two repeated scans within 24 hours at each site. Pulse sequences included MPRage with $1 \times 1 \times 1mm^3$, high-resolution T2 (TSE) with $1 \times 1 \times 1mm^3$, and also DTI which is not discussed here. In the following, image data from the three sites with same scanner type will be named $\mathcal{A}_itj$, with $i = \{1, 2, 3\}$ for scan site and $j = \{1, 2\}$ for the two time points. The site with a different scanner will be named $\mathcal{B}tj$. The set of image data has been acquired during one week. It is therefore safe to assume that there are no major brain changes over this short time period.

2.2 Evaluation of Cross-Site Image Deformations

Assessment of deformations between images requires the use of high-dimensional nonlinear registration that is sensitive to even localized distortions at sub-voxel scale. Candidates of such registration procedures are, among others, large-deformation fluid

registration [3,4] and elastic registration [5]. Such registrations result in volumetric deformation fields $h(x)$, and quantitative measurements commonly derived from these vector fields are a) $\log|Dh(x)|$, the log determinant of the Jacobian and b) $\|h(x)\|_2$, the L^2-norm of the deformation vector field.

We expect $\log|Dh|$ to be very close to zero in regions of no change, and positive and negative in local areas of extraction and contraction. The L^2-norm $\|h(x)\|_2$ helps to evaluate the magnitude of local voxel shifts. Image maps created with the two metrics were used for qualitative assessment of locality and magnitude of changes. Histograms of both the $log|Dh|$ and $\|h(x)\|_2$ and measures derived from their cumulative histograms are used for comparison (Fig. 1). Image data available to this study are the raw DICOM data and have not been pre-corrected by phantom calibration. We are interested in measuring the magnitude of eventual geometric distortions across scanners since such deformations and voxel-scaling differences might significantly contribute to variability of quantitative measurements. Here, we use the large-deformation registration framework extended to image populations by [4].

The volumetric image scans are preprocessed for bias correction and intensity normalization, which both are part of the EM tissue segmentation method [6] described in a later paragraph. Intensity normalization is necessary since the matching functional of our registration is the L^2 norm of image intensities. After co-registration by rigid transformation, we created an unbiased atlas $\bar{\mathcal{A}}$ given the set of six T1 images from scanner type $\mathcal{A}$. The image data of scanner $\mathcal{B}$ were deformed into the atlas $\bar{\mathcal{A}}$ using the same fluid deformation method. The resulting diffeomorphic registration fields $h_k(x)$ describe distortions between each image and the unbiased atlas $\bar{\mathcal{A}}$, used here as the best estimate of truth. Quantitative analysis of deformation is calculated within a brain mask which was automatically obtained by the EM tissue segmentation (see Fig. 2).

2.3 Evaluation of Tissue Segmentation

Automatic Tissue Segmentation: Several methods have been developed for automatically segmenting healthy adult brain MRI, mostly variations of multi-variate statistical classification techniques [7,6,8] and most recent work by Pohl et al. [9] that augments tissue class segmentation by a detailed segmentation of neuro-anatomical structures. Here, we use a modified version of [6] written in ITK [10], which takes a set of multimodal MRI as input and performs probabilistic atlas registration, bias correction, brain stripping, user-selected nonlinear filtering, and multi-variate classification in one integrated tool [1]. As results, we get tissue probability maps $p(class|x)$ for the categories white matter (wm), gray matter (gm), cerebrospinal fluid (csf), and background (bg) and also binary label maps of the maximum posterior classification.

Overlap Measures on Segmented Structures: Commonly used overlap measures for binary segmentations are the Dice [11] coefficient $DSC(A,B) = \frac{A \cap B}{(|A|+|B|)}$ and the Jaccard [12] similarity coefficient $JSC(A,B) = \frac{A \cap b}{A \cup B}$. By definition, JSC is more stringent and therefore results in lower values: $JSC(A,B) = \frac{2|A \cap B|}{2(|A|+|B|-|A \cap B|)} = \frac{DSC(A,B)}{2-DSC(A,B)}$.

[1] Free download at http://www.ia.unc.edu/dev/download/itkems/index.htm

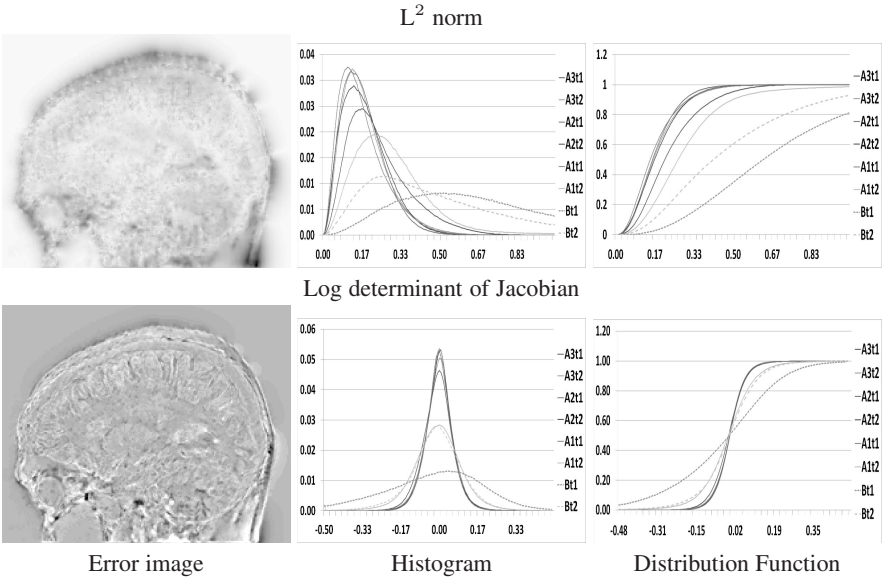


Fig. 1. Example images and distributions of the L^2 norm and the log determinant of the Jacobian for the eight scans of one phantom

	90 percentile	
	Phantom1	Phantom2
A1t1	0.30	0.33
A1t2	0.48	0.48
A2t1	0.30	0.27
A2t2	0.30	0.27
A3t1	0.42	0.36
A3t2	0.30	0.27
Bt1	1.17	0.90
Bt2	0.90	0.78

	5 percentile		95 percentile	
	Phantom1	Phantom2	Phantom1	Phantom2
A1t1	-0.12	-0.16	0.08	0.12
A1t2	-0.12	-0.24	0.16	0.20
A2t1	-0.12	-0.12	0.08	0.08
A2t2	-0.12	-0.12	0.08	0.08
A3t1	-0.12	-0.12	0.08	0.12
A3t2	-0.12	-0.12	0.08	0.04
Bt1	-0.44	-0.20	0.28	0.12
Bt2	-0.20	-0.20	0.16	0.12

Fig. 2. Quantitative evaluation of image deformation. Left: 90 percentiles for the L2 norm (in mm). Right: 5 and 95 percentiles for the log det Jacobian (log volume change).

Given probabilistic segmentations, i.e. segmentations with class probabilities at each voxel, it is preferable to use a probabilistic overlap measure POV [13,4], a metric derived from the normalized $L1$ distance between two probability distributions

$$POV(A,B) = 1 - \frac{\sum_x |P_A - P_B|}{2\sum_x P_{AB}} \,. \tag{1}$$

For POV, P_A and P_B are the probability maps representing the two fuzzy segmentations and P_{AB} is the joint probability, calculated by integrating the pair of probabilistic

segmentations over the image space and appropriate normalization. The numerator expresses the probabilities of primarily non-intersecting regions.

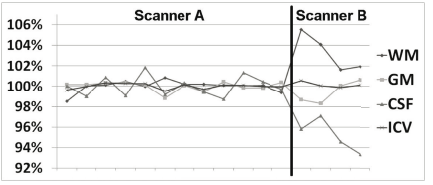
Reliability of Segmentation Via Kullback-Leibler Divergence: The problem of comparing a pair of segmentations presented as sets of posterior probability maps $p(c|x)$ can be rephrased so that we ask for the divergence between probability densities for locations as a function of classes $p(x|c)$. In the following, the class-specific measure is called $D_{KL}^{(c)}$, and the total divergence D_{KL} is obtained as the sum over all categories c:

$$D_{KL}(p\|q) = \sum_c D_{KL}^{(c)}(p\|q) \;=\; \sum_c \left(\sum_x p(x|c) \log \frac{p(x|c)}{q(x|c)} \right) . \qquad (2)$$

Here, the class-specific divergences $D_{KL}^{(c)}$ describe differences between probability densities $p(x|c)$ and $q(x|c)$ of locations x spread across the whole image volume. This gives us the probability for location x given a specific category c. The $p(x|c)$ can be calculated from the probabilistic EM tissue segmentations $p(c|x)$ since $p(c|x)/\sum_x p(c|x)$ is exactly $p(x|c)$ if we assume that the prior $p(x)$ is uniform within the image volume: $\frac{p(c|x)}{\sum_x p(c|x)} = \frac{p(c,x)}{p(x)} \times \frac{1}{\frac{1}{p(x)}\sum_x p(c,x)} = \frac{p(c,x)}{\sum_x p(c,x)} = \frac{p(c,x)}{p(c)} = p(x|c)$. In our implementation, we use the symmetrized Jensen-Shannon divergence D_{JS} betweeen distributions p and $m = \frac{(p+q)}{2}$ to account for $p(x)$ and $q(x)$ drawn from different images.

Results of Tissue Segmentation Comparison: Fig. 3 displays the relative tissue volumes across all scanners, where volumes were normalized by the average volume of scanner type $\mathcal{A}$ for each phantom, used as estimated truth. This figure and the table to the right illustrate that the coefficient of variation for measurements on scanners type $\mathcal{A}$ is in the range of 0.5%, a value that might very well be in the range of normal brain variability. Tissue segmentation from scanner type $\mathcal{B}$, shows significantly larger differences, in particular for white matter and csf, which might be attributed to sensitivity of the EM tissue segmentation to a slightly different contrast mechanism. The probabilistic overlap measure (see Fig. 4) reflects a similar picture, with overlap for white and gray matter for scanner $\mathcal{A}$ over 97% but for scanner $\mathcal{B}$ lower then 95%. Again here, overlap is measured relative to an estimated truth, which are the probabilistic tissue segmentation maps of the unbiased atlas. Please note that we cannot judge *which scanner and segmentation is right* but only that the two types are different.

The Kullback-Leibler distance was calculated based on the same tissue probability maps $p(c|x)$. Table 4 right lists the total D_{KL} and the class-specific $D_{KL}^{(c)}$ values. The values reflect again that scanner $\mathcal{B}$ differs much more for all three tissue categories. The relationship between the overlap POV and the information-theoretic D_{KL} is not straightforward, but a correlation between the two sets of values over all tissue categories is -0.992 across both type scanners and -0.980 for scanner type $\mathcal{A}_k$ only, which indicates that they describe the pairwise differences in a similar way. The total D_{KL} can be calculated as the sum of all the class-specific measures and can be used as an overall difference. With overlap measures, such a combination is less obvious.



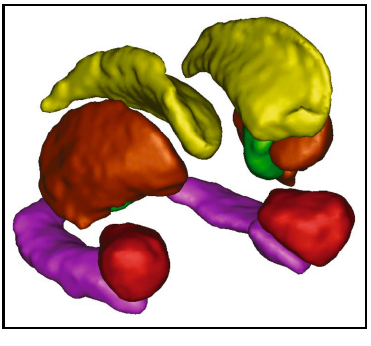
Scanner type $\mathcal{A}$	
tissue class	COV (%)
wm	0.573
gm	0.471
csf	0.988
icv	0.267

Fig. 3. Left: Tissue volumes for all the scanners normalized relative to scanner $\mathcal{A}$ averages for each phantom. Right: Coefficient of variation for tissue volumes.

POV %	wm	gm	csf
A1t1	97.45	97.38	95.04
A1t2	97.33	97.13	94.56
A2t1	97.47	97.40	95.12
A2t2	97.42	97.34	94.87
A3t1	97.38	97.32	94.38
A3t2	97.54	97.37	94.43
Bt1	94.66	94.14	88.58
Bt2	95.08	94.48	88.90

	D_{KL}	$D_{KL}^{(wm)}$	$D_{KL}^{(gm)}$	$D_{KL}^{(csf)}$
A1t1	0.0612	0.0141	0.0137	0.0331
A1t2	0.0681	0.0153	0.0158	0.0369
A2t1	0.0605	0.0139	0.0140	0.0321
A2t2	0.0641	0.0143	0.0143	0.0349
A3t1	0.0674	0.0141	0.0142	0.0390
A3t2	0.0661	0.0132	0.0142	0.0385
Bt1	0.2034	0.0467	0.0456	0.1099
Bt2	0.1909	0.0411	0.0424	0.1057

Fig. 4. Probabilistic overlap measure POV for the tissue segmentation of each case compared to the tissue segmentation of the atlas. Right: KL divergence measure of reliability of tissue probability maps. Total D_{KL} and class-specific $D_{KL}^{(c)}$ for each scan are listed. Correlation between POV and D_{KL} is -0.992 for wm, gm and csf combined (see text).



Structures	COV
Amygdala L	1.05%
Amygdala R	0.75%
Caudate L	0.58%
Caudate R	0.95%
Hippocampus L	0.75%
Hippocampus R	0.55%
Pallidus L	1.49%
Pallidus R	1.71%
Putamen L	0.47%
Putamen R	0.50%

Dice (%)	JSC (%)
97.95	95.99
98.13	96.34
98.14	96.36
98.17	96.40
98.26	96.57
98.21	96.48
98.38	96.81
98.63	97.29
98.55	97.15
98.90	97.82

Fig. 5. Segmentation of subcortical structures averaged over scanner type $\mathcal{A}$ and both phantoms. Left: 3D display of binarized probabilistic template. Middle: Coefficient of variation for resulting volumes. Right: Overlap of binary segmentations.

2.4 Evaluation of Subcortical Structure Segmentation

Subcortical structures were segmented by high-dimensional deformation of an unbiased population average MRI carrying probabilistic segmentations of the set of subcortical structures (see [14] for details). Figure 5 lists the coefficient of variation for resulting volumes as averages over all scanner type $\mathcal{A}$ images and both phantoms. In addition,

POV	Scanners	
Structures	A (%)	B (%)
Amygdala L	98.96	95.81
Amygdala R	99.12	96.79
Caudate L	99.12	97.87
Caudate R	99.12	97.78
Hippocampus L	99.11	97.30
Hippocampus R	99.15	97.13
Pallidus L	99.18	97.65
Pallidus R	99.29	97.20
Putamen L	99.28	97.61
Putamen R	99.40	97.13

	MAD (in mm)		Hausdorff (in mm)	
Structures	A	B	A	B
Amygdala L	0.10	0.34	1.15	2.33
Amygdala R	0.08	0.30	1.03	2.31
Caudate L	0.07	0.17	0.95	1.04
Caudate R	0.07	0.17	1.27	1.44
Hippocampus L	0.06	0.20	0.98	1.62
Hippocampus R	0.06	0.22	1.00	1.56
Pallidus L	0.06	0.18	0.83	1.10
Pallidus R	0.05	0.21	0.88	1.00
Putamen L	0.06	0.22	0.98	1.57
Putamen R	0.05	0.26	0.95	1.09

Fig. 6. Reliability of segmentation of subcortical structures. Left: Probabilistic overlap coefficient. Right: Surface distances (in mm) per structure for scanners $\mathcal{A}$ and $\mathcal{B}$ relative to estimated truth.

Dice and Jaccard overlap coefficients are calculated for the same list of structures. These coefficients are calculated relative to an estimated truth, which is obtained by averaging the probabilistic segmentations of scanner type $\mathcal{A}$ structures and extracting the level-set at probability 0.5 to represent a hard segmentation, a process similarly to [4]. Tables 5 and 6 list the coefficient of variation (COV), Dice, JSC, probabilistic overlap (POV), and the mean absolute and Hausdorff surface distances. These tables clearly demonstrate the excellent reproducibility of scanner $\mathcal{A}$ results and the significantly larger differences of scanner $\mathcal{B}$, with lower POV and larger surface distances.

3 Discussion

We present methodology and results on validation of MRI data of human traveling phantoms. The methods include analysis of image deformations, comparison of tissue segmentation and automatic segmentation of sets of subcortical structures. Analysis of deformation demonstrates that most voxels show shifts below $0.4mm$ for scanners of the same type but much higher values for a different scanner. The log determinant of the Jacobian is known to be sensitive to noise through the building of derivatives and the determinant, and values within the 90 percentile range can be as high as 20% volume change. However, scanner $\mathcal{B}$ again shows distinctly different values. Tissue volume analysis shows volume variation close to 0.5% for white and gray matter of scanners of type $\mathcal{A}$ but much higher values for scanner $\mathcal{B}$. The probabilistic overlap POV shows a similar pattern and is in the range of 97.5% for same scanner types and 94.0% for the different scanner, all measures calculated relative to the truth which is the unbiased average image $\bar{\mathcal{A}}$. For automatic subcortical structure segmentation, the COV in the range 0.5% to 1.0% and POV of mostly above 99% for scanner type $\mathcal{A}$ represent a level of multi-site reliability which to our knowledge has not been reported yet.

This paper primarily proposes comparison metrics for *probabilistic segmentations*, taking into account that today's segmentation algorithms can increasingly provide such data which more robustly cope with partial voluming and avoid discretization artifacts.

The Kullback-Leibler divergence D_{KL} as presented here seems a viable alternative to commonly used overlap measures. The information-theoretic KL divergence is based on a well-researched theory and has a known interpretation for hypothesis testing, a property that will be explored more extensively in our future research.

The results shown here might serve as a benchmark for other research groups but will also be useful for clinical researchers involved in multi-site imaging studies to get information on expected variability of measurements. Upon completion of the phantom acquisition study, we will make the MRI and DTI image data publicly available.

References

1. Jovicich, J., Beg, M.F., Pieper, S., Priebe, C., Miller, M., Buckner, R., Rosen, B.: Biomedical informatics research network: integrating multi-site neuroimaging data acquisition, data sharing and brain morphometric processing. In: 18th IEEE Symposium on Computer-Based Medical Systems, pp. 288–293. IEEE, Los Alamitos (2005)
2. Han, X., Fischl, B., Sch, H., Hosp, M., Charlestown, M.: Atlas Renormalization for Improved Brain MR Image Segmentation Across Scanner Platforms. IEEE Transactions on Medical Imaging 26, 479–486 (2007)
3. Christensen, G., Rabbitt, R., Miller, M.: Deformable templates using large deformation kinematics. IEEE TMI 5, 1435–1447 (1996)
4. Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased diffeomorphic atlas construction for computational anatomy. Neuro. Image 23 (Suppl. 1), 151–160 (2004)
5. Shen, D., Davatzikos, C.: Hammer: hierarchical attribute matching mechanism for elastic registration. IEEE TMI 21, 1421–1439 (2002)
6. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based tissue classification of MR images of the brain. IEEE Trans. Medical Imaging 18, 897–908 (1999)
7. Wells, W.M., Kikinis, R., Grimson, W.E.L., Jolesz, F.: Adaptive segmentation of MRI data. IEEE Trans. Medical Imaging 15, 429–442 (1996)
8. Cocosco, C.A., Zijdenbos, A.P., Evans, A.C.: A fully automatic and robust brain MRI tissue classification method. Medical Image Analysis 7, 513–527 (2003)
9. Pohl, K., Bouix, S., Nakamura, M., Rohlfing, T., McCarley, R., Kikinis, R., Grimson, W., Shenton, M., Wells, W.: A hierarchical algorithm for mr brain image parcellation. IEEE Transactions on Medical Imaging 26, 1201–1212 (2007)
10. Insight Consortium: Insight Toolkit (2004)
11. Dice, L.R.: Measure of the amount of ecological association between species. Ecology 26, 297–302 (1945)
12. Jaccard, P.: The distribution of flora in the alpine zone. New Phytologist 11 (1912)
13. Gerig, G., Jomier, M., Chakos, M.: VALMET: a new validation tool for assessing and improving 3D object segmentation. In: Niessen, W.J., Viergever, M.A. (eds.) MICCAI 2001. LNCS, vol. 2208, pp. 516–523. Springer, Heidelberg (2001)
14. Gouttard, S., Styner, M., Joshi, S., Smith, R., Cody Hazlett, H., Gerig, G.: Subcortical structure segmentation using probabilistic atlas priors. In: SPIE Medical Imaging, vol. 6512, p. 65122J–1 – 65122J–11 (2007)