Multi-site validation of image analysis methods - Assessing intra and inter-site variability

Martin A. Styner[a,b], H. Cecil Charles[a], Jin Park[b], Guido Gerig[b]

[a]Duke Image Analysis Lab, DUMC, Durham, NC 27710, USA
[b]Dept. Computer Science, Dept. Psychiatry, UNC, Chapel Hill, NC 27599, USA

ABSTRACT

In this work, we present a unique set of 3D MRI brain data that is appropriate for testing the intra and inter-site variability of image analysis methods. A single subject was scanned two times within a 24 hour time window each at five different MR sites over a period of six weeks using GE and Phillips 1.5 T scanners. The imaging protocol included T1 weighted, Proton Density and T2 weighted images. We applied three quantitative image analysis methods and analyzed their results via the coefficients of variability (COV) and the intra correlation coefficient. The tested methods include two multi-channel tissue segmentation techniques based on an anatomically guided manual seeding and an atlas-based seeding. The third tested method was a single-channel semi-automatic segmentation of the hippocampus. The results show that the outcome of image analysis methods varies significantly for images from different sites and scanners. With the exception of total brain volume, which shows consistent low variability across all images, the COV's were clearly larger between sites than within sites. Also, the COV's between sites with different scanner types are slightly larger than between sites with the same scanner type. The presented existence of a significant inter-site variability requires adaptations in image methods to produce repeatable measurements. This is especially of importance in multi-site clinical research.

Keywords: Multi-site, Quantitative Image Analysis, Reliability, Validation, MRI

1. INTRODUCTION

Large scale clinical studies of MRI images often require the application of quantitative 3D MRI image analysis methods on datasets that were acquired by multiple sites. However, such methods are most often developed on datasets from a single MR source. In order to evaluate the variability in regard to different MR sources, the methods should be tested and validated on a dataset from multiple scanners with different properties using the same standard protocols. In this work, we present such a set of 3D MRI brain data that is appropriate for testing the intra and inter-site variability of image analysis methods.

In morphometric studies of psychiatric and neurological disorders such as schizophrenia[1] , Alzheimer's Disease[2] or Multiple Sclerosis[3] , it is often necessary to accurately segment brain MR images into their constituent tissue types prior to measuring structures of interest. Traditional manual and semi-automated tissue classification methods are prone to intra-rater and inter-rater variability, both of which can weaken the ability to discriminate subtle morphological differences. Many of these methods do not correct for the bias field created by MR scanner RF inhomogeneities and thus are prone to misclassifications[4] . These problems associated with traditional methods address the need for an automated, reliable and bias corrected tissue classification method in the analysis of brain MR images. In this paper, we investigate two methods (see also Park[5]): 1. an semi-automated, manual seeding based tissue segmentation without bias field correction. 2. an automated, atlas seeding based seeding segmentation[6] that includes bias field correction.

Following the tissue segmentation, regions of interest are segmented either based on the tissue segmentation or directly form the original brain MR images. In this paper we investigate a semi-automatic segmentation of the hippocampus from the original image. The hippocampus is a subcortical structure in the temporal lobe that is of special interest in schizophrenia[1, 7] and Alzheimer's disease[8] research.

Send correspondence to Cecil H. Charles: cecil.charles@duke.edu

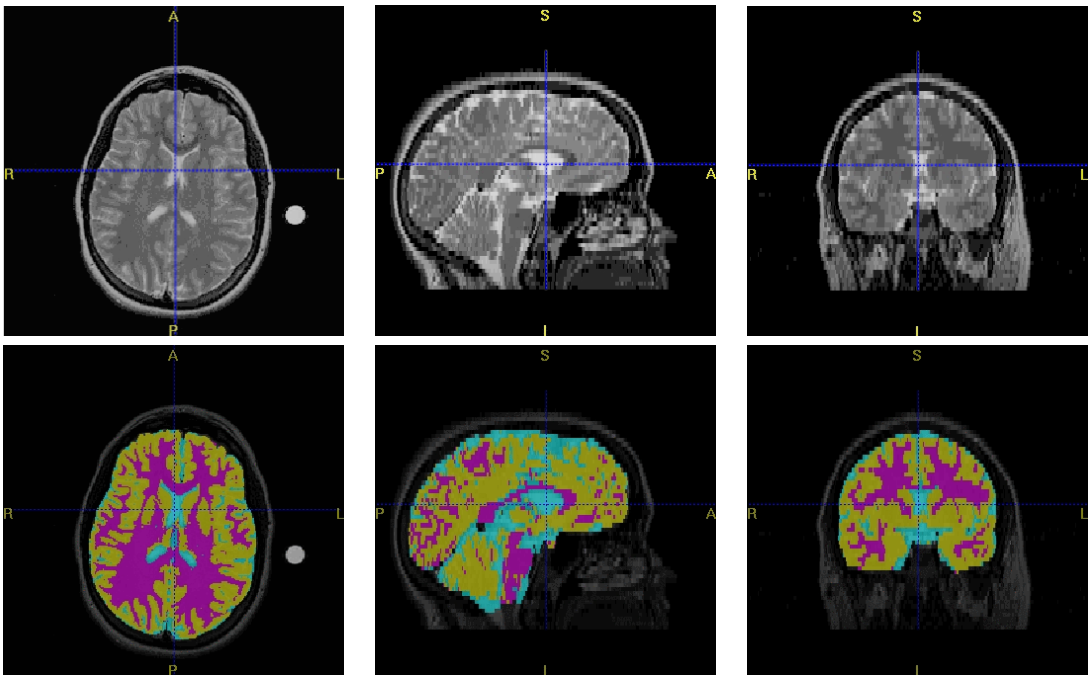


Figure 1. Three orthogonal views (axial, coronal, and sagittal) of the original T2-weighted MR scan (top row) and an overlay with the tissue classification label maps (bottom row, GM = yellow, WM = purple, CSF = cyan).

2. METHODS

2.1. Dataset for testing intra and inter-site variability

A single female subject (age 25 years) was scanned two times within a 24 hour time window each at five different MR sites over a period of six weeks. The age of the subject plus the absence of physical or mental illness suggests that the brain of the subject did not change in the six week period. Thus, the acquired images represent all the same brain.

The imaging protocol included 3D T1 weighted, Inversion Recovery prepped Spoiled Grass images (SPGR, 0.9375mm x 0.9375mm x 1.5mm, axial slicing direction) and contiguous Proton Density and T2 weighted fast spin echo images (FSE, 0.9375mm x 0.9375mm x 3.0mm, axial slicing direction). Four of the five sites acquired the images using a GE Signa 1.5 T scanner and one site was using a Phillips Gyro Scan NT 1.5 T scanner. The scanning protocols were provided by a central facility, the Duke Image Analysis Laboratory (DIAL). The same facility collected all data and supervised image quality control. Also, additional phantom data using a Hoffman brain phantom was acquired for each site. The phantom data allows for geometric fidelity between the images by normalizing the real field-of-view (FOV) to sub-milliter accuracy.

This dataset allows for an evaluation of reproducibility of image analysis methods within one site, between sites with the same type of scanner and between sites with different types of scanner. The main sources of variance in this dataset at the level of data acquisition are patient positioning, scanner geometry, scanner intensity variation and discrete image artifacts, e.g potion artifacts. At the level of image analysis variance mainly arises from procedures involving registration, interpolation, intensity bias field correction and manual interaction.

2.2. Brain tissue segmentation, manual seeding

The first tested image analysis method is a brain tissue segmentation based on the two FSE channels, T2 weighted and Proton Density (PD), which are acquired at the same time and thus no registration is needed. No

bias field correction is applied as pre-processing step. A semi-automatic extraction of the intra cranial cavity (ICC) is performed first. ICC is defined as the sum of the brain tissues white matter (WM), gray matter(GM) and cerebro-spinal fluid (CSF). Then, a human rater selects seed samples on several slices for the tissue classes WM, GM, CSF and background (BG). These seeds are used by a parzen-window classifier to segment all voxels of the images into one of the classes (see Figure 1).

The segmentation procedure was performed using the "Multi-spectral Segmentation" module within the Analyze 3.0 software package.[9] A standard operation procedure was established to standardize the ICC extraction and the locations of the seed samples for the tissue segmentation. Human raters performing the segmentation have to undergo a training phase, which is followed by a validation study to test for an appropriate inter-rater and intra-rater variability. This segmentation procedure has been tested and validated in several large scale multi-site trials with multiple raters.

Four trained human raters segmented each image of the dataset twice in a randomized, blinded study design. The average coefficients of variance (COV) for all tissue *volume* measurements were determined for each rater individually and then averaged over all raters. These COV's were computed for within one site, between sites with the same type of scanner and between sites with different types of scanner. The intra-class correlation was also computed using site, timepoint and rater as nested effects. The results are presented and compared with the atlas-based tissue segmentation in section 3.2.

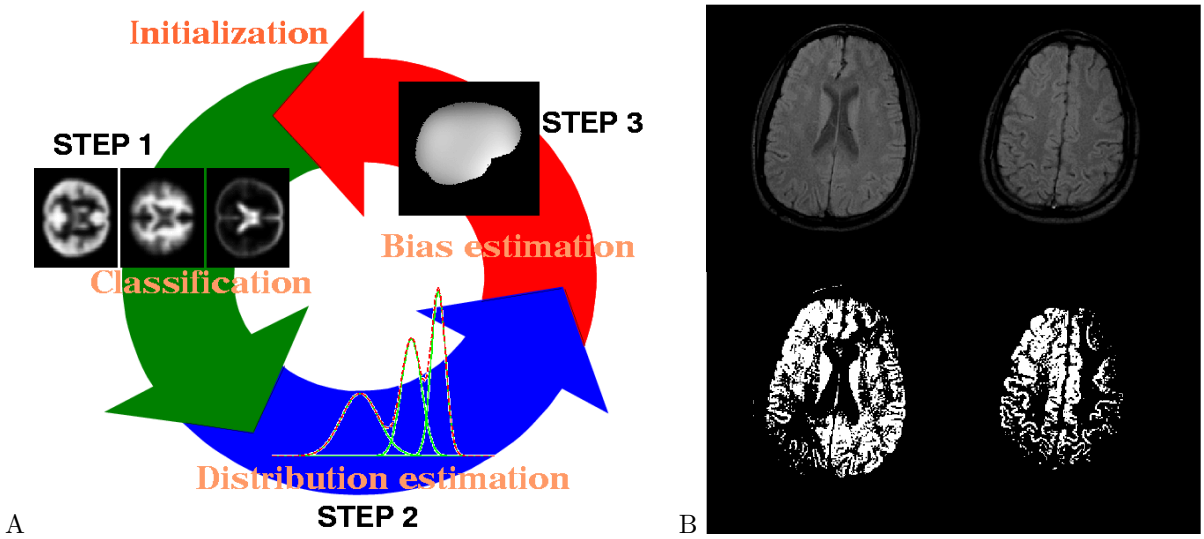


Figure 2. A: Three-step iterative algorithm for atlas-based tissue segmentation (illustration courtesy of K. Van Leemput). B: Tissue classification without bias field correction of PD MR images (top) show misclassifications of gray matter (bottom) due to a visually imperceptible bias field in the original data.

2.3. Brain tissue segmentation, atlas-based seeding

The second multi-channel brain tissue segmentation technique was proposed and developed by Leemput at al[6] . It is an *automated* model-based method using a digital probability atlas containing *a-priori* expectations about the spatial location of brain tissue classes. The method interleaves a mixture-model Bayesian classification with an estimation of tissue class parameters and with an estimation of the parameters for a polynomial bias field correction. Multiple images can be used for the segmentation, after these images have been registered to each other. Since the T1 images in our dataset have a higher resolution, the T2 and PD images are up-interpolated to the T1-image resolution using a windowed-sync interpolation. The scheme of the method and an illustration of the the need for bias field correction is shown in Figure 2.

The digital atlas is an average probability atlas from 155 normal subjects and is distributed with the SPM96[10] package. In order to apply the atlas to the image, it is first registered with a T1 template image provided by SPM 96 that is already coregistered with the atlas. The registration algorithm is based on maximizing mutual information.

The circular, iterative segmentation procedure starts with the bias field parameters set to zero, and a rough estimation of the classification for each pixel using the atlas probabilities. The distribution parameters for the tissue classes are then estimated from the classification. After this step, the bias field can be estimated from the classification and the class distributions. Subsequently the class distributions and the bias field lead to a new estimation of the classification. These steps are repeated for several cycles. The computation of the final classification is fully deterministic and repeatable.

The segmentation procedure works on single and multi-channel data. We investigated which combination of the T1, T2 and PD channels gives the best performance for tissue classification (see section 3.1). The best channel combination is then used for the computation of the intra and inter-site variability. The COV's and intra-class correlation were computed in the same way as for the manual-seed based tissue classification. The results are presented in section 3.2.

2.4. Hippocampus segmentation

Segmentation of the hippocampus is known to be a difficult problem in medical image analysis. No fully automated algorithm for hippocampus segmentation is evident in the literature. The existing algorithms require either varying amounts of landmark selection and/or delineation, or extensive post-processing. Our method is a histogram based morphometric region growing algorithm that requires seeding of the hippocampus on multiple slices. Additional manual interaction involves restricting the region growing to prevent leakage, as well as small manual editing. The segmentation was performed using the "Region of Interest" module within the Analyze 3.0 software package[9] .

Four trained human raters segmented each image of the dataset twice in a randomized, blinded study design. The COV's for the volume measurements were determined for each rater individually and then averaged over all raters. These COV's were computed for within one site, between sites with the same type of scanner and between sites with different types of scanner. The intra-class correlation was computed using site, time point and rater as nested effects. The results are presented in section 3.3.

3. RESULTS

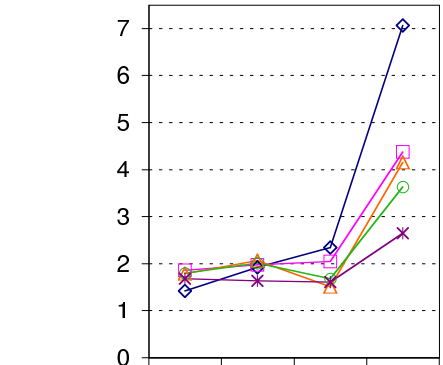
3.1. Optimal channel combination in multi-channel tissue segmentation

We evaluated several combinations of channels for the atlas-based tissue segmentation in order to determine which combination has best performance. The COV's were computed over the whole dataset. A low COV means that the segmentation procedure produces similar volume measurements over all images independent of site or scanner type. The COV's for different channel combinations are shown in Figure 3A. The COV's vary the most for the CSF volumes, whereas all other tissue classes are segmented equally stable by all combinations. From the comparison of COV's alone, it would seem as if the combination of the T2 and PD channels is performing best. The lower COV's for the (T2,PD) combination are caused by the inexistence of variance due to registration and also by the higher signal-to-noise ratio in FSE images compared to SPGR images. The (T2,PD) combination is also used in our manual-seeding based tissue segmentation.

Besides the stability of the segmentation addressed in studying the COV's, the segmentation precision has to be studied as well in order to decide which combination of channels performs best. As can be seen in Figure 3B, the precision of a tissue segmentation using the higher resolution T1 channel is better than the segmentation using the low resolution images. Considering that the tissue segmentation is often used to measure small subparts of the brain such as lateral ventricles or deep gray matter structures, we decided that the (T1,T2,PD) combination performs best. In the remainder of this work, we used this combination for the atlas-based tissue segmentation.

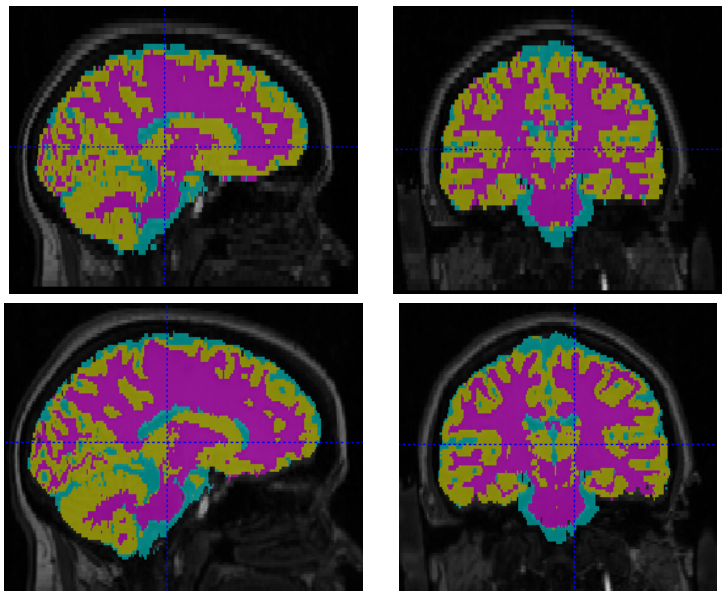
A

COV for channel combinations



	ICC	GM	WM	CSF
T1	1.41	1.92	2.34	7.06
T1T2	1.86	1.97	2.04	4.37
T1PD	1.78	2.06	1.50	4.15
T1T2PD	1.79	2.01	1.68	3.63
T2PD	1.67	1.63	1.60	2.65

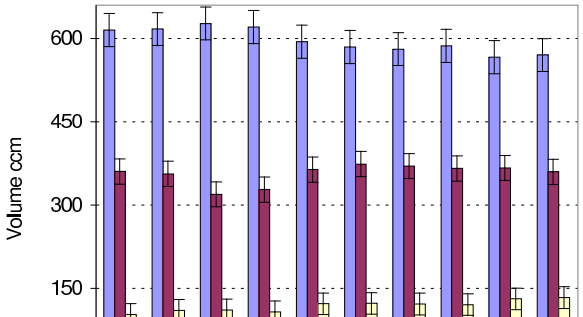
B



Saggittal Coronal

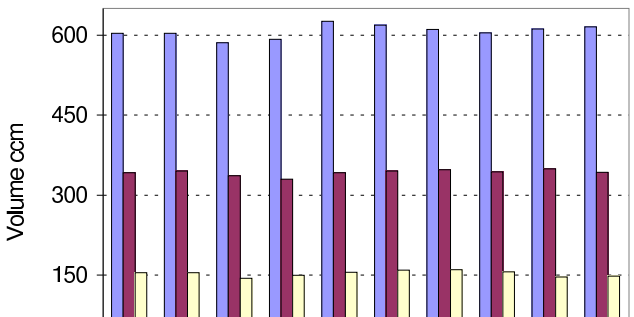
Figure 3. Left: Coefficients of Variance (COV) for different tissue classes over the whole dataset computed for different combinations of channels. The combination of T2 and PD yields the lowest COV's. Right: Visual comparison of a segmentation from low resolution (3.0mm) T2,PD channels and high resolution (1.5mm) T1,T2,PD channels (T2,PD up-interpolated). The segmentation is shown in a sagittal and coronal slice (data has axial slicing direction) overlaid with the registered T1 image. The precision of the segmentation is clearly improved for the higher resolution channel combination.

WM/GM/CSF volumes for seed based segmentation



	GE1	GE1	GE2	GE2	GE3	GE3	GE4	GE4	PH	PH
GM	615	617	627	620	594	585	581	586	566	570
WM	360	356	319	328	364	374	370	366	367	360
CSF	103	110	111	107	122	123	122	121	131	133

WM/GM/CSF volumes for atlas based segmentation



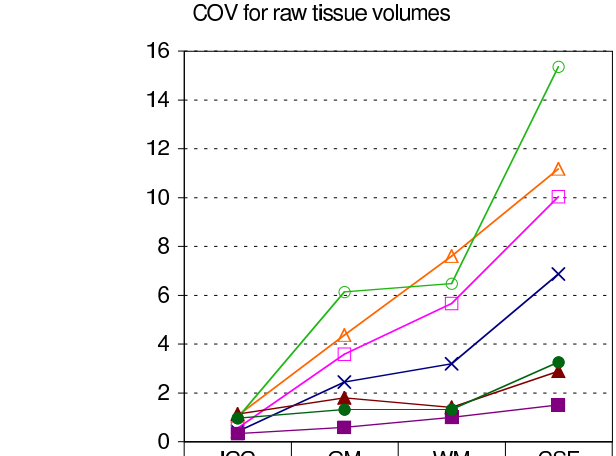
	GE1	GE1	GE2	GE2	GE3	GE3	GE4	GE4	PH	PH
GM	603	603	585	592	626	619	611	604	611	616
WM	342	345	336	329	342	345	347	343	349	343
CSF	154	154	143	149	154	159	159	155	146	147

Figure 4. Volume measurements for WM, GM and CSF tissue classes over the whole dataset (GE = GE scanner site, PH = Phillips scanner site). Left: Average volumes from segmentations using manual seeding on (T2,PD) images, the bar indicates the range of the measurements over all raters. Right: Volumes from segmentations using atlas-based seeding on (T1,T2,PD) images.

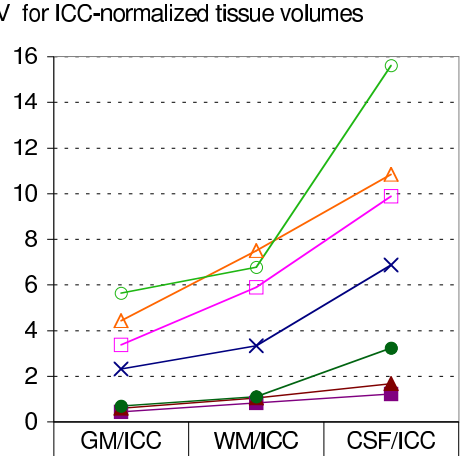
3.2. Inter/intra-site variability of brain tissue segmentation

We compared the segmentations of the two proposed tissue segmentation procedures (see Figure 4). The volume measurements seem to be more stable across the dataset for the atlas-based segmentation. The CSF volumes are in average 20% smaller in the manual-seed based segmentation. This is mainly due to the different channel combinations since the T1 image is an additional channel in the atlas-based segmentation. This suggests that the T2 and PD channels are either underestimating the CSF volume or that the T1 channel overestimates it or that the truth lies somewhere in-between. The atlas-based segmentation also produces in average 16% smaller CSF volumes when only the T2 and PD channels are used. Since we do not have ground-truth information, this matter cannot be solved for this dataset.

Since the atlas-based segmentation is deterministic and thus fully repeatable, the same image will always be segmented with the same result. This is not the case for the manual-seed based segmentation, which can be seen from the COV's values for segmenting the same image in Figure 5. These COV's are the minimal values for the manual-seed based segmentation; this method cannot perform better. Not surprisingly, the intra and inter-site COV's are all larger than the intra-image COV's. We also see that these intra-image COV's are larger than the COV's of the atlas-based segmentation. It is thus evident that the atlas-based segmentation is considerably more stable than the manual based segmentation.



	ICC	GM	WM	CSF
M same img	0.41	2.43	3.18	6.87
M same site	0.54	3.58	5.65	10.02
M same type	1.06	4.36	7.58	11.16
M diff type	0.98	6.13	6.46	15.35
A same site	0.33	0.57	0.98	1.51
A same type	1.13	1.78	1.40	2.88
A diff type	0.95	1.31	1.32	3.25



	GM/ICC	WM/ICC	CSF/ICC
M same img	2.32	3.32	6.87
M same site	3.37	5.90	9.87
M same type	4.42	7.48	10.83
M diff type	5.63	6.78	15.61
A same site	0.44	0.83	1.22
A same type	0.59	1.03	1.66
A diff type	0.68	1.09	3.23

Figure 5. Average COV for tissue segmentation methods: Manual-seed based segmentation = "M", Atlas-based segmentation = "A"; inter-image COV = "same img", COV within site = "same site", COV between sites with same scanner type = "same type", COV between sites with different scanner type = "diff type". Left: Raw tissue volumes, Right: Tissue volumes normalized with the ICC volume. The atlas-based segmentation is more stable than the manual-seed based segmentation. The intra-site variability is lower than the inter-site variability.

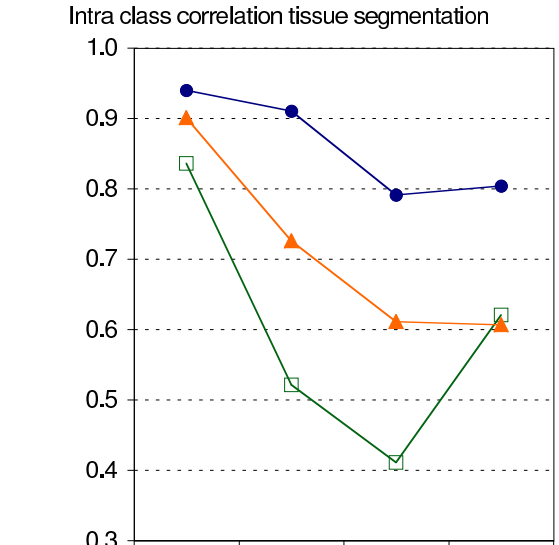
From the intra class correlation analysis shown in Figure 6A we can conclude the same findings as above. Additionally, the intra class correlation was computed for the average volume measurements over all raters in the manual-seed based segmentation. Also the average volumes have a lower intra class correlation than the volumes from the atlas-based segmentation, and thus are less stable.

Reasons for the higher stability of the atlas-based segmentation are most likely the additional bias field correction, the consistent seeding and the reproducibility of the computation. The remaining sources of variance due to the processing are registration and interpolation errors. The other sources of variance are at the level of data acquisition, such as patient positioning, scanner geometry, scanner intensity variation and discrete image artifacts.

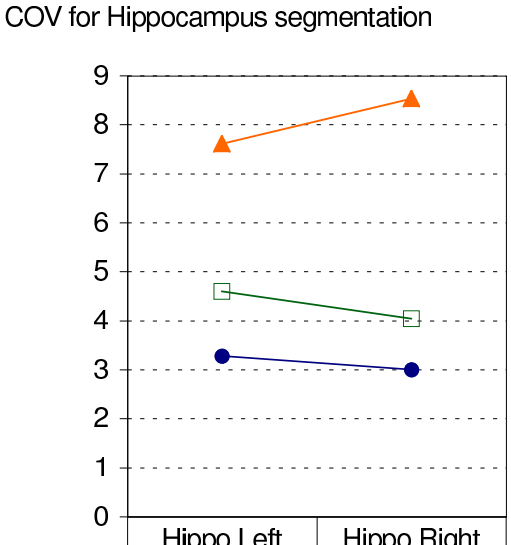
It is common when studying brain tissue volumes to normalize the tissue volumes with the ICC volume. This normalization usually helps in dealing with gender and age effects in the studied population since it leads to a higher degree of similarity between volume measurements of different subjects. The COV's for the normalized tissue volumes are also displayed in Figure 5 and exhibit the same patterns as the non-normalized volume measurements.

The analysis of the COV's also shows that the ICC volumes are most stable across the dataset, followed by GM, WM and CSF. The CSF volumes are clearly the least stable. The same pattern can be seen in the analysis of the ICC normalized volumes. This COV pattern correlates with the volume size, i.e. a larger tissue volume leads to a lower COV. This suggests that segmentations of large objects have a smaller relative error than segmentations of smaller objects.

With the exception of the ICC volume, which shows low variability across all tests, the COV's are clearly larger between sites than within sites. Thus, the intra-site variability is, as expected, lower than the inter-site variability. The picture is less clear when comparing inter-site variability between sites with same and different scanner types. The variabilities are quite close and due to the small sample size no conclusive findings can be drawn. However, a trend is visible that the inter-site variability between sites with the same scanner type is slightly lower.



	ICC	GM	WM	CSF
Atlas	0.94	0.91	0.79	0.80
M all	0.84	0.52	0.41	0.62
M Avg	0.90	0.73	0.61	0.61



	Hippo Left	Hippo Right
same site	3.27	3.00
same type	4.59	4.03
diff type	7.61	8.53

Figure 6. Left: Intraclass correlation for the volumes from the tissue segmentation : Manual-seed based segmentation = "M", Atlas-based segmentation = "Atlas"; correlation based on all measurements = "all", correlation based on averaged measurements = "Avg". The atlas-based segmentation is more stable. Right: Average COV for hippocampus segmentation: COV within site = "same site", COV between sites with same scanner type = "same type", COV between sites with different scanner type = "diff type". The intra-site variability is lower than the inter-site variability. The inter-site variability for sites with same scanner type is lower than for sites with different scanner type.

3.3. Inter/intra-site variability of hippocampus segmentation

The analysis of the COV's for the segmentation of the left and right hippocampus present a similar pattern similar as the tissue segmentations. The intra-site variability is lower than the inter-site variability. In contrast to the tissue segmentation analysis, the inter-site variability between sites with same scanner type is considerably lower than for sites with different scanner type. This disparity could be partially due to the fact that the hippocampus segmentation is based purely on the SPGR T1 images. The SPGR protocol seems to lead to images with an increased variability between scanner types compared to the FSE protocol. A single channel T1 tissue segmentation also shows an increased variability between sites with different scanner types.

We also computed the intra class correlation for the hippocampus segmentation, which is for the right hippocampus 0.93 and for the left 0.92. These values and the COV analysis suggest that our method allows a stable segmentation of the hippocampus.

4. DISCUSSIONS

In this paper, we have presented a unique dataset that allows the assessment of intra-site and inter-site variability. We have applied three image analysis methods to the dataset and analyzed its results. The results show that the outcome of image analysis methods varies significantly for images from different sites and scanners. With the exception of total Brain volume, which shows consistent low variance, the COV's were clearly larger between sites than within sites. Also, the COV's between sites of different scanner type are larger than between sites with the same scanner type. The low sample size of the dataset slightly hampers the generalization of the latter finding.

The presented existence of a significant inter-site variability suggest a need for adaptations in image analysis methods to produce repeatable measurements across sites. This is especially of importance in multi-site clinical research.

Our dataset can also be used to find for the optimal combination of image channels in a multi-channel tissue segmentation. Our results suggest that a (T2,PD) channel combination produces the most stable results. Adding a higher resolution T1 channel and up-interpolating the T2 and PD channels leads to a higher precision of the segmentation while maintaining good stability. We suggest the use of this (T1,T2,PD) channel combination since small structures, such as lateral ventricles or deep gray matter structures, are often measured on tissue segmentations. These structures demand a higher precision than the FSE's 3.0mm slice thickness.

We also used the dataset to compare two different tissue segmentations. The outcome showed that the atlas-based method showed better stability than the manual-seed based method. The stability of measuring whole brain volume was very good in both methods. The stability of measuring CSF volume was decreased in both methods, but to a larger extend for the manual-seed based method.

We plan to use this dataset to assess intra and inter-site variability of all our image analysis algorithm that are used in multi-site studies of brain morphometry. Since we think our dataset represents an important mean to test image analysis methods, we plan to make access to it publicly available via internet.

ACKNOWLEDGMENTS

We would like to acknowledge D. Vandermeulen, F. Maes and K. Van Leemput, Catholic University of Leuven, Belgium, for providing the atlas-based brain tissue segmentation software. ACKNOWLEDGE Lilly.

REFERENCES

1. R. McCarley, C. Wible, M. Frumin, Y. Hirayasu, J. Levitt, I. Fischer, and M. Shenton, "MRI anatomy of schizophrenia," *Biological psychiatry* **45**, pp. 1099–1119, 1999.
2. J. Tanabe, D. Amend, N. Schuff, V. DiSclafani, F. Ezekiel, D. Norman, G. Fein, and M. Weiner, "Tissue segmentation of the brain in alzheimer disease.," *AJNR Am J Neuroradiol* **18**(1), pp. 115–123, 1997.

3. K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection," *IEEE Transactions on Medical Imaging* **20**, pp. 677 – 688, Aug 2000.
4. A. Simmons, P. Tofts, G. Barker, and S. Arridge, "Sources of intensity nonuniformity in spin echo images," *Magnetic Resonance in Medicine* **32**, pp. 121–128, 1994.
5. J. Park, G. Gerig, M. Chakos, D. Vandermeulen, and J. Lieberman, "Structural neuroimaging of psychiatric disease: A reliable and efficient method for auotmated tissue classification," in *Schizophrenia Researc*, **49**, p. 167, Elsevier, April 2001.
6. K. V. Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classication of mr images of the brain," *IEEE Transactions on Medical Imaging* **18**(10), pp. 897–908, 1999.
7. J. Csernansky, S. Joshi, L. Wang, J. Haller, M. Gado, J. Miller, U. Grenander, and M. Miller, "Hippocampal morphometry in schizophrenia via high dimensional brain mapping," *Proc. Natl. Acad. Sci. USA* **95**, pp. 11406–11411, September 1998.
8. B. Dickerson, I. Goncharova, M. Sullivan, C. Forchetti, R. Wilson, D. Bennett, L. Beckett, and L. deToledo Morrell, "MRI-derived entorhinal and hippocampal atrophy in incipient and very mild alzheimer's disease," *Neurobiol Aging* **22**(5), pp. 747–754, 2001.
9. Biomedial Imaging Resource Mayo Foundation, *Analyze AVW 3.1 Manual*, Online Documentation, 2001.
10. J. Ashburner, K. Friston, A. Holmes, and J. Poline, "Statistical parametric mapping." The Wellcome Dept. Cognitive Neurology, Univ. College London, UK. Avialable at http://www/fil.ion.ucl.ac.uk/spm.