2016-06-01

A Semi-Automated Algorithm for Segmenting the Hippocampus in Patient and Control Populations

Nathan McKay Muncy
Brigham Young University

A Semi-Automated Algorithm for Segmenting the Hippocampus in Patient and Control Populations

Nathan McKay Muncy

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science
Neuroscience

Christopher B. Kirwan, Chair
Michael D. Brown
Jonathan J. Wisco

Department of Physiology and Developmental Biology

Brigham Young University

June 2016

ABSTRACT

A Semi-Automated Algorithm for Segmenting the Hippocampus
in Control and Patient Populations

Nathan McKay Muncy
Department of Physiology and Developmental Biology
Master of Science
Neuroscience

Calculating hippocampal volume from Magnetic Resonance (MR) images is an essential task in many studies of neurocognition in healthy and diseased populations. The `gold standard' method involves hand tracing, which is accurate but laborious, requiring expertly trained researchers and significant amounts of time. As such, segmenting large datasets with the standard method is impractical. Current automated pipelines are inaccurate at hippocampal demarcation and volumetry. We developed a semi-automated hippocampal segmentation pipeline based on the Advanced Normalization Tools (ANTs) suite of programs to segment the hippocampus. We applied the semi-automated segmentation pipeline to 70 participant scans (26 female) from groups that included participants diagnosed with autism spectrum disorder, healthy older adults (mean age 74) and healthy younger controls. We found that hippocampal segmentations obtained with the semi-automated pipeline more closely matched the segmentations of an expert rater than those obtained using FreeSurfer or the segmentations of novice raters. Further, we found that the pipeline performed best when including manually-placed landmarks and when using a template generated from a heterogeneous sample (that included the full variability of group assignments) than a template generated from more homogeneous samples (using only individuals within a given age or with a specific neuropsychiatric diagnosis). Additionally, the semi-automated pipeline required much less time (5 minutes per brain) than manual segmentation (30-60 minutes per brain) or FreeSurfer (8 hours per brain).

Keywords: hippocampus, segmentation, algorithm, autism, advanced normalization tools (ANTs)

ACKNOWLEDGMENTS

I'd like to thank Dr. Kirwan for brainstorming this project and letting me work on it. He was so patient in teaching me the basic skills I needed and while we figured out the algorithm, and how to best represent the algorithm in the paper. I'd also like to thank Dr. Hunsaker, whose code I stalked heavily while learning how to use ANTs. Additionally, this paper would have been impossible if it were not for Chris Doxey, who answered all my scripting questions and spent so much time segmenting all the scans by hand. Finally, I'd like to thank Ariana Hedges for helping me to develop the repeated measures scripts needed, the analyses would have been impossible without her.

TABLE OF CONTENTS

TITLE PAGE i

ABSTRACT ii

ACKNOWLEDGMENTS iii

TABLE OF CONTENTS iv

LIST OF TABLES vi

LIST OF FIGURES vii

INTRODUCTION 1

METHODS 6

Participants 6

MRI Data Acquisition 6

Manual Hippocampal Segmentation 7

The Semi-Automated Pipeline 7

FreeSurfer 9

Variants of the SAP 9

RESULTS 10

Comparison of SAP to FreeSurfer and Manual Segmentation 10

Analysis of Variations on the Semi-Automated Pipeline 11

DISCUSSION 12

Overview 12

Researcher and Computational Time Considerations 12

SegAdapter 13

Application to Other Structures 14

Limitations 14

Conclusion 16

REFERENCES 24

APPENDIX A: Scripting Commands 29

CURRICULUM VITAE 30

LIST OF TABLES

Table 1: Various Segmentation Protocols..... 17

Table 2: Repeated Measures ANOVA of the Segmentation Methods 18

Table 3: Comparison of Variations of the SAP Method..... 18

LIST OF FIGURES

Figure 1: Distribution of Hippocampal Volumes per Segmentation Method 19

Figure 2: Distribution of DSCs 20

Figure 3: Comparison of Exp, SAP, SAPXL, and SAPH 21

Figure 4: Group-Specific Comparisons 22

Figure 5: Comparison of Segmentation Hippocampal Masks 23

INTRODUCTION

Cortical and subcortical segmentation is a useful morphometric tool used for both research and diagnostic purposes. Segmentation involves the process of labeling and separating voxels associated with regions-of-interest (ROIs) (Fischl et al., 2002; Pruessner et al., 2000; Yassa & Stark, 2009). This process, i.e. pipeline, is expedient for use in studying typical and atypical brain morphologies, as typical morphologic variations occur naturally and manifest when controlling for age and sex (Allen, Bruss, Brown, & Damasio, 2005; Allen, Damasio, & Grabowski, 2002; Avants, Yushkevich, et al., 2010; Avants, Cook, et al., 2010; Bartley, Jones, & Weinberger, 1997; Cox, 1996; Evans, 2006; Persson et al., 2014) and atypical morphologies are associated with trauma and a variety of neurodegenerative, developmental, and psychiatric disorders (Csernansky et al., 1998; Csernansky et al., 2002; Csernansky et al., 2005; Fischl et al., 2002; Sparks et al., 2002; Wang et al., 2006; Yassa et al., 2010). Volumetric analysis of structure is a valuable area of research as task performance and severity of disorder or disease are heavily correlated with cortical and subcortical volumes (Doxey & Kirwan, 2015; Jung et al., 2014; Turner, Furey, Drevets, Zarate Jr, & Nugent, 2012).

While various methods and pipelines for brain segmentation exist, two standard methods have emerged: manual segmentation and automated segmentation via FreeSurfer. These methods are widely used, but both have significant and different drawbacks. Manual segmentation is often considered the "gold standard" of segmentation (Avants, Epstein, Grossman, & Gee, 2008; Chupin et al., 2007; Pluta et al., 2009). A well-trained researcher manually segments a particular ROI, e.g. the hippocampus, by drawing a mask over the desired region. The mask assigns a label to the ROI, and segmentation occurs as labeled regions are separated one from another, and from non-labeled regions. The main benefit of manual segmentation is accuracy: a well-trained

researcher may accurately identify and distinguish both typical and atypical morphologic variations. Additionally, an expert researcher will not make the same systematic errors that more automated pipelines are known to make (Hunsaker & Amaral, 2014). There are significant drawbacks, however, to manual segmentation (Khan, Wang, & Beg, 2008; Wenger et al., 2014). First, the method is dependent on the expertise of the researcher. Becoming proficient in typical and atypical neuroanatomy is a difficult task, and the amount of time required to label data sets by hand is extensive (Avants et al., 2008). This proves costly in time, training, and computer resources. Second, as the task is researcher dependent, inter-rater reliability is an issue. As pointed out by Avants, raters may make systematic segmentation errors in some structures (Avants et al., 2011). It has been shown that researchers have biases to varying degrees, which result in left-right hippocampal asymmetries that are not present anatomically and significant inter-rater discrepancies in the head and tail; as Maltbie (2012) describes, such rater-caused volumetric variance is likely to greatly affect most neuroimaging studies since interrater variance may approach 11% for hippocampal volumes (Hasboun et al., 1996; Maltbie et al., 2012; Sparks et al., 2002). Third, as the manual segmentations are typically done slice-by-slice in one view, e.g. the sagittal view, the output of the manual segmentation method is often "boxy" and does not reflect the smooth nature of the underlying anatomy. Finally, the amount of time it takes to segment a dataset restricts the size of the dataset that can be analyzed in a reasonable time, restricting the size and scope of studies and necessitating smaller datasets for purely pragmatic reasons (Tustison et al., 2014). Such limitations in manual segmentation reduce the scientific power of the method.

Automated segmentation is most often accomplished via FreeSurfer (Fischl et al., 2002), which is a fully automated program that labels and segments cortical and subcortical structures.

Although FreeSurfer requires significant computational time, it places much less demand on the researcher than manual segmentation. There are several limitations to FreeSurfer. First, the technique is dependent on previously rendered atlases, which may or may not be suitable for the studied group. For example, a subject pool consisting of children cannot be accurately segmented by a template derived from adults (Fonov et al., 2011). Nor would it make sense to use a healthy, young adult, male atlas for studying women, geriatric, or unhealthy groups, thereby necessitating several atlases for analysis within the same study. Second, FreeSurfer takes considerable computational time to perform a full segmentation, up to eight hours to segment a single subject. Without parallelization, large datasets may require weeks or months of processing time. Third, FreeSurfer subcortical segmentation is voxel-based, meaning that FreeSurfer compares voxel intensity to the atlas to guide segmentation. This raises issues of loss of fine detail and segmentation problems (Fischl et al., 2002; Khan et al., 2008). As an example, the alveus, a thin strip of white matter used to anatomically separate the amygdala from the head of the hippocampus, is only 1mm thick (Chupin et al., 2007) and with standard structural MRI scan resolution of about $1mm^3$, the intensity of the alveus is easily averaged with the nearby intensities of the amygdala and hippocampus through partial volume sampling. This leads to systematic errors in which FreeSurfer labels a portion of amygdala as hippocampus, errors that must be corrected manually (Figure 5). This issue does not happen as readily in other subcortical regions because FreeSurfer, in registering the voxel intensities, is able to account for the averaged differences in intensities between subcortical regions. The amygdala and hippocampus, however, have nearly identical image intensities (see: Fischl et al., 2002). Consequently, without the guidance of the alveus, FreeSurfer makes errors in the amygdaloid-hippocampal region.

Recently, the development of Advanced Normalization Tools software (ANTs) (Avants et al., 2008; Avants et al., 2011) has significantly addressed the issues associated with manual and automatic segmentation. Symmetric Normalizer (SyN), one of the toolkits in ANTs, is able to more accurately compare the template to the scan by utilizing two unique approaches, termed symmetric registration and cross-correlation (Avants et al., 2008). Symmetric registration refers to the fact that both scan and template are warped using an optimization function to a midpoint (so their movements are symmetric) and cross-correlation is the fact that the registration algorithm utilizes local intensity values, in addition to voxel intensity, from both the scan and template when finding the corresponding point in both images.

An additional advantage of ANTs is a decreased dependence on pre-labeled atlases. Instead, one is able to use the ANTs software to create a study-specific template derived from the very images that are to be studied, which yields a more sensitive and applicable form of comparison (B. B. Avants, P. Yushkevich, et al., 2010; Wilke, Holland, Altaye, & Gaser, 2008; Yoon, Fonov, Perusse, & Evans, 2009). These novel approaches by the ANTs program have been shown to outperform standard segmentation methods (Klein et al., 2010; Tustison et al., 2014).

Here it was our aim to develop a pipeline (dubbed the "semi-automated pipeline" or SAP) based around the ANTs software that was as accurate as manual segmentation, fast enough for large datasets, easy to use, and robust in atypical and aged populations. In addition to the pipeline, we supply the scripts and parameters used in the pipeline for the sake of algorithmic transparency (Kovacevic, 2006; Tustison et al., 2013). To assess the performance of the SAP, we compared hippocampal volumes obtained with the semi-automated pipeline (SAP) to those obtained with manual segmentation by both novice and expert researchers, and FreeSurfer (FS).

Furthermore, we considered several permutations of the SAP in order to establish the optimal procedure in terms of ease of use and robustness of results.

METHODS

Participants

Seventy volunteers (26 female) gave written informed consent prior to participating in MRI scanning. Participant groups included a group diagnosed with autism spectrum disorder (ASD) (n=19, 5 female, mean age 22.5), a group of healthy older adults (n=16, 8 female, mean age 74.3) and a group of healthy, young-adults which served as a control population (n=35, 13 female, mean age 22.2). All scans were obtained for the purpose of functional localization in functional MRI (fMRI) studies. The fMRI data are reported elsewhere. The Institutional Review Board at Brigham Young University approved all research.

MRI Data Acquisition

Data acquisition was performed at the BYU MRI Research Facility using a 3T Siemens TIM Trio MRI scanner. All structural MR images were acquired using one of two T1-weighted magnetization-prepared rapid acquisition with gradient echo (MP-RAGE) sequences. For the ASD and ASD-control groups (n=35), the sequence used the following parameters: TE = 2.26 ms, flip angle = 9°, slices = 176, slice thickness = 1.0 mm, matrix size = 256 × 224, field of view = 250 × 220.8 mm, voxel size = 1 × 1 × 1 mm. For the remaining participants in the Aged and Aged-control groups (n=35), the sequence used the following parameters: TE = 2.08 ms, flip angle = 8°, slices = 128, slice thickness = 1.20 mm, matrix size = 192 × 192, field of view = 220.8 × 220.8 mm, voxel size = 1.15 × 1.15 × 1.2 mm.

Initial spatial normalization of all individual MR images was accomplished using the Analysis of Functional NeuroImages (AFNI) suite of programs (Cox, 1996) to manually perform AC-PC alignment (Talairach, 1988).

Manual Hippocampal Segmentation

Manual tracings were performed by two researchers (one experienced, one novice) using a protocol based on the anatomical description of Insausti et al. (Cox, 1996; Insausti et al., 1998) and referencing common, published, anatomical guidelines (Hasboun et al., 1996). The novice researcher trained on a practice set of scans until they had obtained sufficiently high inter-rater reliability, calculated using the Dice Similarity Coefficient (DSC) (Dawant et al., 1999; Dice, 1945; Sparks et al., 2002) using the formula $2|A*B| / (|A| + |B|)$. Values for the DSC range from 0 to 1, with good scores ≥ 0.7, where a value of 1 indicates complete similarity (Bartko, 1991; Zijdenbos, Dawant, Margolin, & Palmer, 1994), and the novice researcher obtained DSCs ≥ 0.9 when compared with the experienced researcher. Manual segmentations were used as the standard against which we compared the semi-automated method described below (Avants et al., 2008; Chupin et al., 2007; Pluta et al., 2009). All scans (n=70) were segmented by both the novice and experienced researcher.

The Semi-Automated Pipeline

In order to address the limitations associated with manual and automatic segmentation, we adapted a segmentation pipeline (Table 1) that had previously been developed for use in rhesus macaques by Hunsaker and Amaral (2014), here dubbed the Semi-Automated Pipeline or SAP. The SAP progressed in the following steps: 1) Landmark the AC-PC oriented structural scans; 2) Render a *heterogeneous* study-specific template; 3) Landmark and manually segment the template; 4) Register the template to the structural scans.

Landmarks were manually placed within both hippocampi of all participants using the program Multi-image Analysis GUI (Mango; University of Texas Health Science Center; http://ric.uthscsa.edu/mango/) following the landmarking guidelines of Hunsaker and Amaral

(Hunsaker & Amaral, 2014) and referencing the locations specified by Pluta et al. (2009). Landmarks were used in order to guide the automated registration performed by the Advanced Normalization Tools (ANTs) software (Csernansky et al., 1998; Hunsaker & Amaral, 2014; Pluta et al., 2009; Tustison, Avants, & Gee, 2009) by biasing the registration of the fixed to the moved image (Klein et al., 2010) for certain regions of high inter-subject variability e.g. the boundaries of the hippocampus. All landmarks were placed by the novice researcher.

ANTs (Avants et al., 2011) was then used to render a study-specific template (SST) (Chupin et al., 2007; Evans, 2006) utilizing all participants in all groups reported above, plus an additional 104 scans from other healthy, young adults utilizing the same scanning protocols previously described (n=172, 59 female, mean age = 28.4 ± 16.7; see appendix A for specific command line options; Avants et al., 2010). That is, instead of rendering multiple homogenous templates to use with each different group as has previously been done (e.g., Hunsaker & Amaral, 2014; Klein et al., 2010), we rendered a heterogeneous template from all the scans of the various groups, thereby making an overall study-specific template rather than multiple group-specific templates. The additional 102 scans were included to improve overall generalizability of the template. We believe this method will prove superior to templates based on homogenous samples, standard atlases, or single scans (Fonov et al., 2011) as it both minimizes the deformation resulting from the requisite warping (i.e., the Jacobian) required for each registration and requires only a single manual segmentation in template space.

The SST was then manually segmented by the experienced researcher and landmarked, according to the same protocols previously described, and registration of each scan to the SST was calculated using the ANTs software (Hunsaker & Amaral, 2014). Upon registration, segmentation occurred by warping the template hippocampal mask into subject space for all

scans. The hippocampal masks were then split according to hemisphere, and converted to binary masks.

FreeSurfer

In order to better evaluate the output of ANTs hippocampal segmentation was also performed using FreeSurfer 5.0 on all scans. In this way we were able to assess the performance of the SAP with regard to the commonly accepted methods of manual segmentation and FreeSurfer.

Variants of the SAP

Additionally, we examined two variants of the SAP as outlined above. The SAPH variant consists of utilizing multiple homogeneous templates instead of a single, heterogeneous template, and the SAPXL variant does not utilize landmarks in either the structural scans or the template, and does not include the Point-Set Expectation (PSE) option during the registration process. SAPH and SAPXL represent common methods of ANTs usage, and were included in order to assess whether or not a heterogeneous template and the addition of landmarks increased the reliability and robustness of SAP when compared to an experienced rater and the more standard segmentation protocols. Accordingly, all scans were processed using SAP, SAPH, SAPXL, FS, and manual segmentation (Table 1). Furthermore, when considering the hippocampal volume measurements resulting from the various segmentation protocols, hippocampal volumes from our experienced researcher were used as a standard against which we compared the multiple pipelines.

RESULTS

Comparison of SAP to FreeSurfer and Manual Segmentation

Hippocampal volumes derived from each segmentation method (SAP, FS, and novice and experienced manual segmentation) for all participants (n=70) were analyzed with a repeated-measures ANOVA, accounting for the lack of independence in the two hemispheres and reported using the multivariate counterpart of the Student's t-test, the Hotelling's T2. A significant main effect of segmentation method on hippocampal volumes was found for each method (alpha=0.05, Hotelling's T2 (3,68) = 4997.343, two-sided p-value < 0.000001). All methods were found to differ significantly one from another (Table 2), except for the experienced-SAP comparison (Hotelling's T2 (1, 68) = 0.905, two-sided p-value = 0.345), indicating that the SAP produced statistically identical volumes to the experienced researcher. Likewise, the distribution of hippocampal volumes produced by SAP were nearly identical to the experienced researcher, and unlike those produced by the novice researcher or FreeSurfer (Figure 1).

Dice similarity coefficients (DSCs), the ratio of overlap to non-overlap, were used in determining the extent to which the same space was segmented by each method. The SAP method had a higher agreement with the experienced (Exp) researcher (mean DSC = 0.85) than either the novice (mean DSC = 0.78) or FreeSurfer (mean DSC = 0.60), and the SAP-E DSC was revealed to have the highest degree of agreement between all DSCs (Figure 2). According to a repeated measures ANOVA, this agreement was found to be significantly different from the DSCs resulting from the other comparisons of segmentation methods (Hotelling's T^2 (1, 68) = 184.368, two-sided p-value < 0.00001).

Analysis of Variations on the Semi-Automated Pipeline

In order to justify both the landmarks and the heterogeneous template, two variations of the SAP method were used to segment the same groups of participants. The semi-automated pipeline without landmarks (SAPXL) does not include landmarks in either the participant scans or the template, and does not use the PSE option during ANTs registration. The semi-automated pipeline with homogeneous templates (SAPH) utilizes a unique template built for each group (ASD, ASD-control, Aged, Aged-control), that was then manually segmented and landmarked (see Table 1).

Repeated-measures analysis of variance revealed that SAPXL produced volumes that were not significantly different from the experienced researcher (Hotelling's T^2 (1, 68) = 0.955, two-sided p-value = 0.332), while a main effect was found for the SAPH method (Table 3). This suggests that Exp, SAP, and SAPXL produce nearly identical volumes in all 70 participants, but that SAPH does not. This is further evidenced in both the distribution of volumes and DSCs (Figure 3).

Finally, repeated-measure analyses conducted on the specific, individual groups (ASD, Aged, and control) revealed the same results, that is, that SAP and SAPXL produce volumes statistically equal to Exp in all individual groups while SAPH and FS do not, and that differences in DSCs between SAP_Exp and SAPXL_Exp appear to only be statistically, but not practically, significant. Furthermore, SAPH had better agreement in the control groups with E than it had in the ASD and Aged groups, which is likely a template effect resulting of the different hippocampal masks used in each of the four homogeneous templates (Figure 4).

DISCUSSION

Overview

The purpose of this study was to develop a protocol for hippocampal segmentation that was as accurate as standard methods but was faster, more easily used, required minimal training, and proved robust in atypical populations. The Semi-Automated Pipeline protocol accomplished this, performing equally well to our experienced rater and outperforming both FreeSurfer (Klein et al., 2010) and novice raters, while minimizing researcher bias and interrater reliability concerns inherent in manual segmentation (Figure 5).

The standard pipeline (SAP) and a variations of the pipeline that did not employ landmarks on the individual subjects or template (SAPXL) performed equally to manual segmentations performed by an experienced rater (Exp). A variation of the pipeline that used templates specific to the homogenous groups studied rather than a heterogeneous template based on subjects from each group under investigation did not perform as well. Similarly, manual segmentation performed by a novice rater (Nov) and FreeSurfer both resulted in less accurate hippocampal segmentation. As SAP takes more processing time than SAPXL due to the inclusion of landmarks, we therefore recommend considering the SAPXL method of segmentation, in addition to SAP, for hippocampal volumetric studies. We anticipate the SAPXL method to decrease processing time while increasing reliability both between groups and studies. Additionally, we recommend not using the SAPH pipeline, as is commonly done, in order to reduce template-mask confounds when comparing different groups.

Researcher and Computational Time Considerations

Manual segmentation required 20 hours of instruction and training for the novice researcher, and over 100 man-hours to manually segment 70 scans, which unfortunately resulted

in less-than-desirable inter-rater Dice similarity coefficients (DSCs). FreeSurfer required approximately eight hours of processing time per participant scan. In our case, we were able to take advantage of a supercomputing cluster to perform FS on all scans in parallel, but this option is not always widely available. In contrast, the novice researcher was able to learn, become proficient in, and then landmark all scans for both this study and another (n=172) within ten hours, and moreover, achieve hippocampal volumes similar to those of an expert rater, and not including landmarks would further reduce this processing time. The Study-specific Template took 147 hours of computation time to render, utilizing 4 cores and 8gb of RAM, and the ANTs registration required only one hour of processing time when performed in parallel on all subjects on a supercomputing cluster.

Template rendering time may be reduced by incorporating fewer scans in the template, as it is not necessary to include all scans from the study, but only a relevant percentage from each group; a template later rendered with only 20 scans required a mere five hours of computation time. Indeed, it likely the better option to build a template from a smaller, representative sample as our template was not particularly sensitive to the hippocampal sulcus.

SegAdapter

SegAdapter (https://www.nitrc.org/projects/segadapter/) is a learning-based software that can be used to correct automatic, consistent errors resulting from segmentation algorithms (Wang, Das et al. 2011). The application in this paper would have been to help correct the output of SAP to the manual segmentations. We decided not to employ SegAdapter (SA) for a number of reasons. First, a pipeline utilizing SA is still dependent on manual segmentation, a protocol that the SAP is meant to replace. Second, SA is heavily dependent on the quality of the manual segmentation. In our hands, the post-SA volumes had very high DSCs with both researchers

(>0.9), novice and experienced alike. We felt that these results could not be trusted since we previously determined that Nov had inferior segmentations when compared to Exp. As such, the SA correction was adversely affected by the novice's segmentations, thereby resulting in high DSCs. Indeed, post-SA volumes differed from both SAP and SAPXL volumes. As such, only SA volumes guided by Exp could be reasonably trusted, but again this involves either segmenting a group twice (manually, SA-SAP), or segmenting a training set for the SA software, which only adds to the post-processing workload. If a high level of sensitivity is needed, however, the SA-SAP method could prove very useful, once the SegAdapter algorithm has been properly trained on expert segmentations.

Application to Other Structures

Practically, this protocol is not limited to the hippocampus. As it is dependent on only a template and ROI mask, any MRI scan utilizing a T1 or T2 image could be segmented with SAP, both within and without the CNS. Also, the increase in sensitivity and robustness and reduction in confounds may yield more consistent results and make SAP a useful tool for studying neurodegenerative disorders like Alzheimer's and Parkinson's disease (Tustison, et. al., 2014). Finally, with a reduced time required for the segmentation of each scan by SAP, less training, and less demand on the researcher and resources, large datasets may be segmented in a very reasonable amount of time.

Limitations

There are several limitations to this study. First, as this study used images that had previously been acquired, two different sets of scanning parameters were used to acquire the MR images. When controlling for scan parameters, however, no significant difference was detected

between control scans that were acquired using different parameters. Had a difference been detected, this study would not necessarily be invalidated as the comparisons were not performed between the various groups and scanning protocols, but between the segmentation methods within each participant. A second limitation in this study is its dependency on the "standard" hippocampal segmentation procedure: manual segmentation. As reported, significant differences were found between the researchers who participated in manual segmentation. While the novice researcher had trained until they had achieved DSCs > 0.9 with the experienced researcher, it seems that the novice researcher only improved on the training scans and underperformed on the actual dataset. More important to the study, however, than the inter-rater reliability issue (which was not unexpected) is the inherent relativity of the analysis. In the absence of a known standard against which we could compare the various segmentation pipelines, we were forced to assume that our experienced rater produced the most accurate hippocampal masks. As such, all of the analyses done in this study were inherently contrastive and limited by the skill of the experienced researcher. Indeed, our study could say nothing of the true performance of the various segmentation methods, but speaks only of their performance relative to Exp, and while it is not uncommon nor unwarranted to use an experienced rater as a standard (Avants, Epstein, Grossman, & Gee, 2008; Chupin et al. 2007), we were nevertheless unable to assess how well the segmentation protocols performed in actuality. For this reason, we cannot know for certain whether the SAP, SAPXL, or SAPH variation performed most accurately. This, however, highlights the core issues which this paper attempts to address: the lack of a reliable and consistent subcortical segmentation pipeline, that is sensitive to atypical groups, and the utter dependency on manual segmentation (with its known various issues) that exists in the field. A potential solution could be found in repeating this study on cadavers: upon segmenting the scans

the hippocampi could then be extracted and its true volume calculated. This may give a more definitive answer as to which segmentation method, if any, performs most accurately. One anticipated issue with this approach is that fixing the brain, as is commonly done post mortem, would change the MR signal, and that an unfixed, unsuspended brain is likely to collapse somewhat during the scanning procedure.

Conclusion

In sum, we recommend using ANTs for hippocampal segmentation, over other common methods such as FreeSurfer or manual segmentation. When using ANTs for volumetric segmentation, we recommend rendering a single, heterogeneous template; landmarks may or may not be beneficial. This will allow for faster processing speeds, more sensitive segmentations, increase the number of groups that may be studied (ASD, AD), and reduce rater-reliability concerns. Furthermore, utilizing a single template mask will reduce template-mask effects that are likely to confound sensitive studies.

Table 1: Various Segmentation Protocols. Manual Segmentation (MS) was performed on scans that were AC-PC aligned, and then the binary hippocampal masks were smoothed. The semi-automated pipeline (SAP) involves rendering a heterogeneous template, manually segmenting this template and smoothing the segmentation masks, landmarking both the template and scans, and using ANTs to register each scan to the template and then warp the segmentation mask from template to scan space., and then splitting, thresholding, and smoothing the output. SAP differs from SAPH (the semi-automated pipeline utilizing homogeneous templates) in that different templates, and template masks are used. SAP differs from SAPXL (the semi-automated pipeline without utilizing landmarks) in that SAPXL did not landmark the scans nor the template, and the Point Set Expectation (PSE) option could not be used during the registration process. All output of the various segmentation pipelines based on ANTs were split, thresholded, and smoothed. FreeSurfer (FS) is entirely self-contained, and no additional pre-or post-processing was used. All methods (MS, SAP, SAPH, SAPXL, FS) were used to segment all scans (n=70).

	Methods				
Steps	MS	SAP	SAPH	SAPXL	FS
AC-PC align	**X**	**X**	**X**	**X**	
Landmark scans		**X**	**X**		
Landmark template		**X**	**X**		
Heterogeneous template		**X**		**X**	
Homogeneous template			**X**		
Segment template(s)		**X**	**X**	**X**	
Smooth template masks		**X**	**X**	**X**	
ANTs registration with PSE		**X**	**X**		
Split output masks		**X**	**X**	**X**	
Thresh output masks		**X**	**X**	**X**	
Smooth output masks	**X**	**X**	**X**	**X**	

Table 2: Repeated Measures ANOVA of the Segmentation Methods. Hotelling's T-squared value and their associated p-values for each comparison of segmentation methods: novice researcher (Nov), experienced researcher (Exp), the semi-automated pipeline (SAP), and FreeSurfer (FS). Every method differs significantly one from another, save Exp and SAP, indicating that SAP and Exp do not perform significantly different.

Comparison	T^2 **(1,68)**	**P-value**
Nov-Exp	667.3428	< .00001
Nov-SAP	429.1486	< .00001
Nov-FS	2481.4046	< .00001
Exp-SAP	**0.9058**	**0.34462**
Exp-FS	4710.4845	< .00001
SAP-FS	5060.1502	< .00001

Table 3: Comparison of Variations of the SAP Method. Hippocampal volumes produced by the experienced researcher are equivalent to both the SAP and SAPXL methods, but SAPH differs significantly from Exp, SAP, and SAPXL.

Comparison	T^2 **(1,68)**	**P-value**
Exp-SAP	0.9058	0.34462
Exp-SAPXL	0.9559	0.33173
Exp-SAPH	176.3029	< .00001
SAP-SAPXL	0.1017	0.7508
SAP-SAPH	202.6593	< .00001
SAPXL-SAPH	245.4571	< .00001

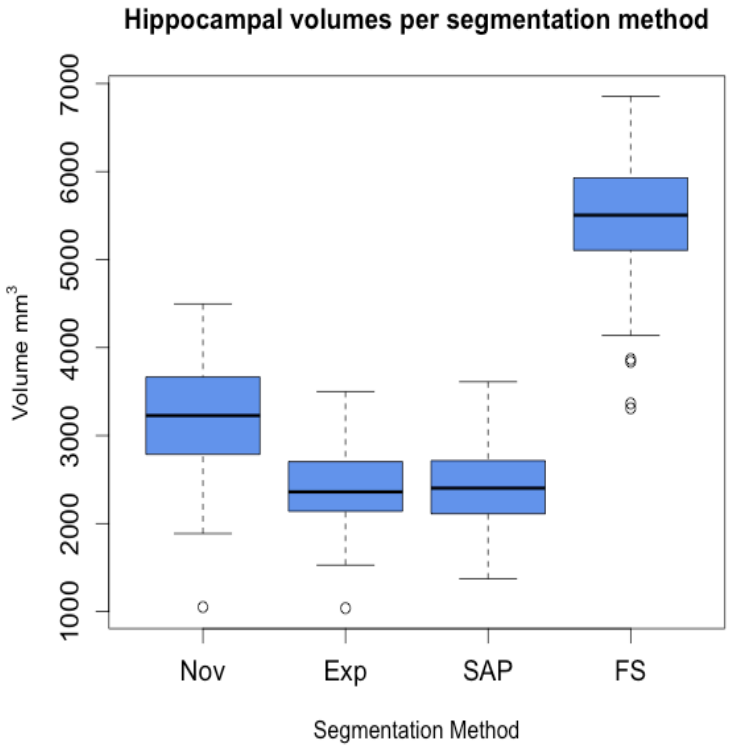

Figure 1: Distribution of Hippocampal Volumes per Segmentation Method. Hippocampal volumes as derived via the various segmentation methods for all 70 participants: Nov, Exp, SAP, and FS. The performance of Exp is more closely approximated by SAP than by Nov or FS.

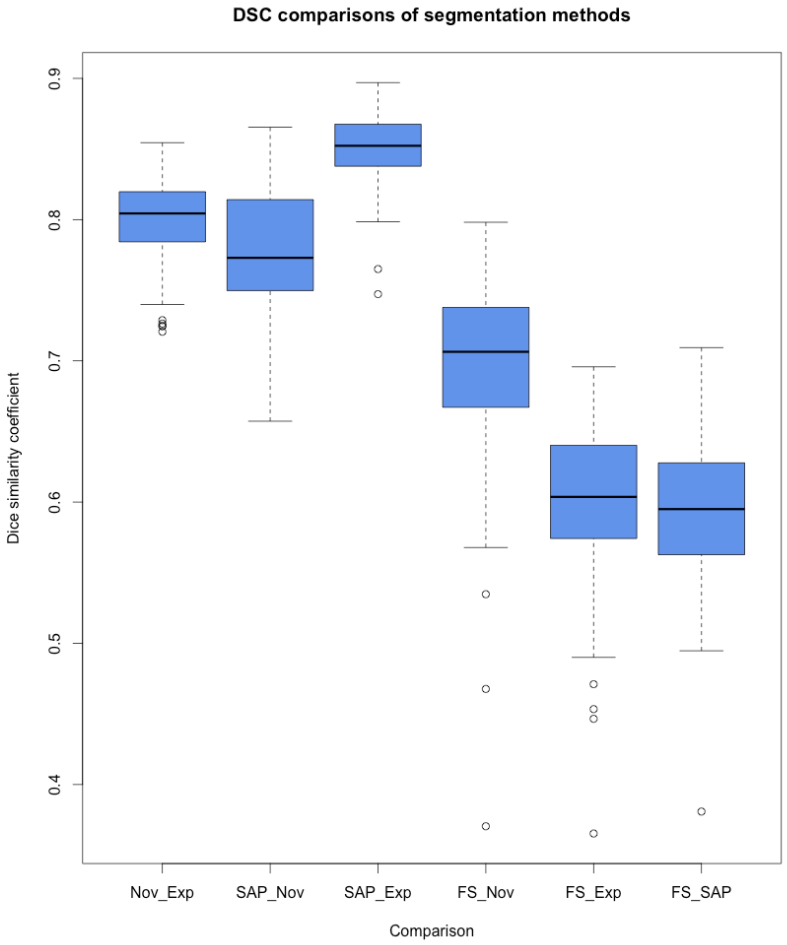

Figure 2: Distribution of DSCs. SAP is shown to have the highest degree of agreement with Exp (SAP_Exp), when compared to the other segmentation methods. Additionally, SAP_Exp has a higher degree of similarity than any of the other comparisons.

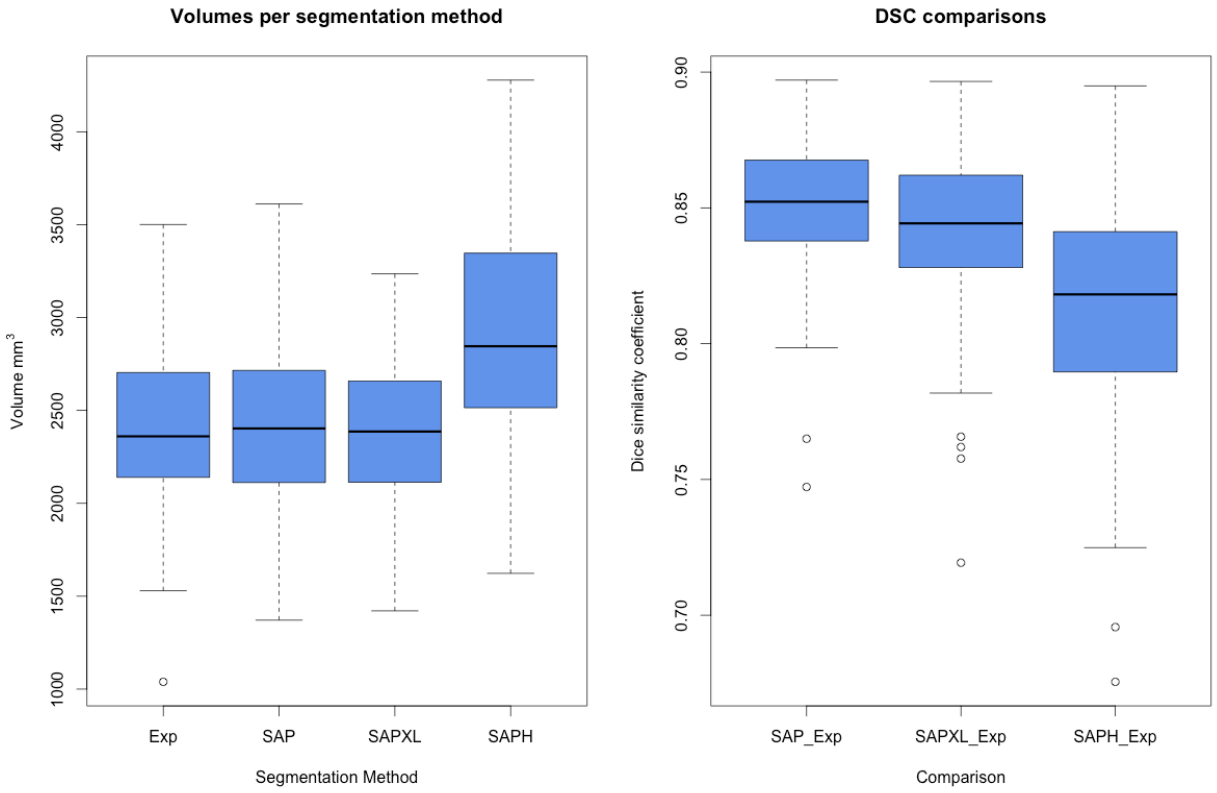

Figure 3: Comparison of Exp, SAP, SAPXL, and SAPH. Left, Distributions of hippocampal volumes per segmentation method. While both SAP and SAPXL produce distributions of volumes that are equivalent to Exp, SAPH produces volumes that are unique. Right, DSCs between Exp and each SAP variation. While all distributions are significantly different one from another (Hotelling's T2 (1, 68) > 3.978, two-sided p-value < 0.0001), the difference between SAP_Exp and SAPXL_Exp DSCs appears to be statistically but not practically significant.

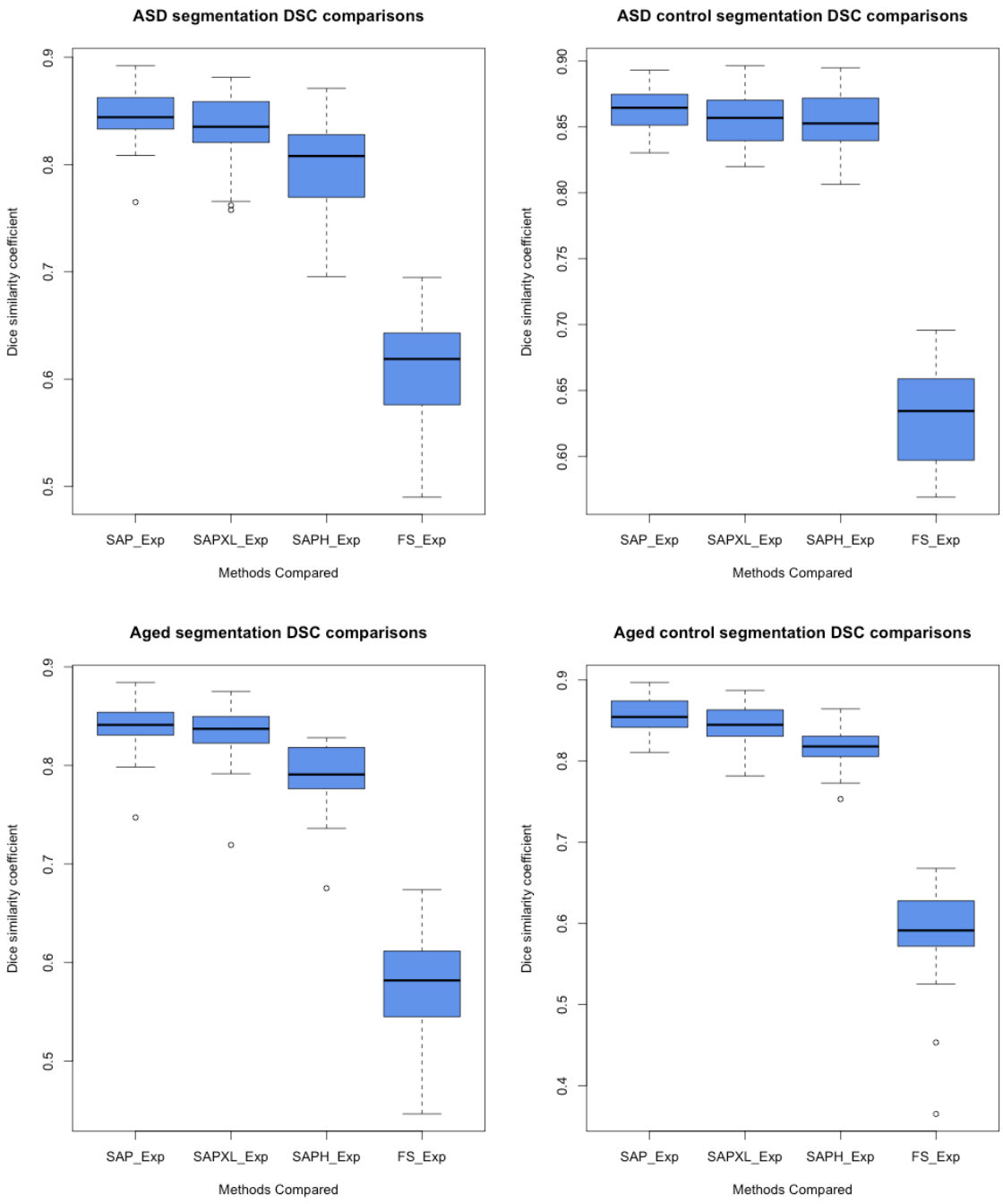

Figure 4: Group-Specific Comparisons. While SAP, SAPXL, and FS appear to have consistent DSCs with Exp in each group, SAPH DSCs with Exp range from being nearly equal to those of SAP and SAPXL (ASD control, top right) to being quite different (Aged, bottom left). This difference in performance, both in comparison to itself and to SAP, SAPXL, is most likely the result of minute differences in segmentation masks for each of the homogeneous templates.

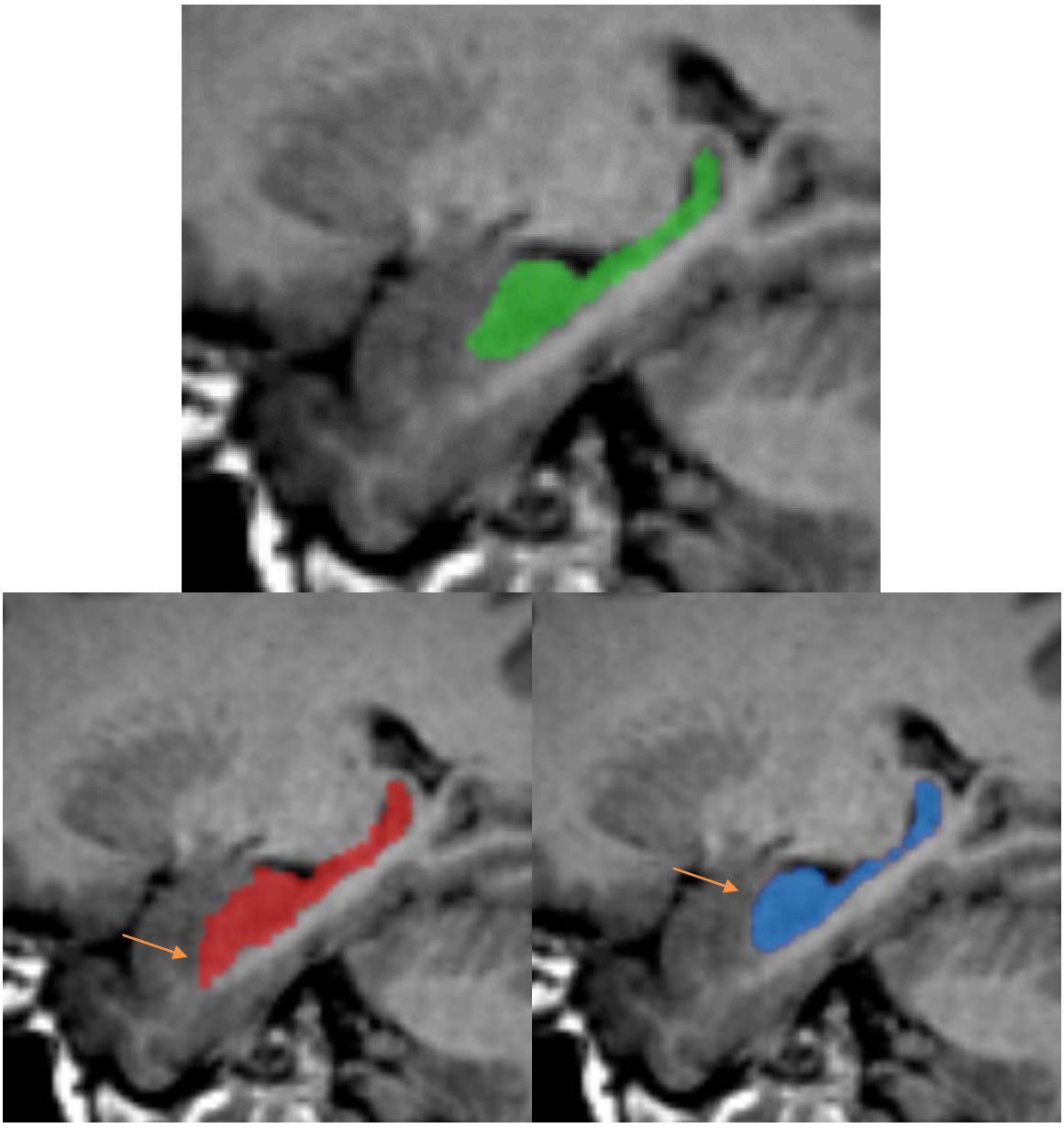

Figure 5: Comparison of Segmentation Hippocampal Masks. SAP (top), FS (bottom left), and Nov (bottom right) hippocampal masks superimposed over an ASD structural scan. The SAP protocol based on the ANTs software is sensitive to the alveus and gives a more accurate segmentation in typical and atypical groups than FS. Additionally, ANTs outputs a probabilistic mask with values that can be constrained more liberally or conservatively.

REFERENCES

Aljabar, P., Heckemann, R. A., Hammers, A., Hajnal, J. V., & Rueckert, D. (2009). Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage, 46*(3), 726-738.

Allen, J. S., Bruss, J., Brown, C. K., & Damasio, H. (2005). Normal neuroanatomical variation due to age: The major lobes and a parcellation of the temporal region. *Neurobiology of Aging, 26*(9), 1245-1260.

Allen, J. S., Damasio, H., & Grabowski, T. J. (2002). Normal neuroanatomical variation in the human brain: An MRI-volumetric study. *American Journal of Physical Anthropology, 118*(4), 341-358.

Ashburner, J., & Friston, K. J. (2000). Voxel-Based Morphometry—The Methods. *NeuroImage, 11*(6), 805-821.

Avants, B.B., Cook, P. A., McMillan, C., Grossman, M., Tustison, N. J., Zheng, Y., & Gee, J. C. (2010). Sparse unbiased analysis of anatomical variance in longitudinal imaging. *Med Image Comput Comput Assist Interv, 13*(Pt 1), 324-331.

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis, 12*(1), 26-41.

Avants, B. B., Schoenemann, P. T., & Gee, J. C. (2006). Lagrangian frame diffeomorphic image registration: Morphometric comparison of human and chimpanzee cortex. *Special Issue on The Second International Workshop on Biomedical Image Registration (WBIR'03), 10*(3), 397-412.

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., & Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage, 54*(3), 2033-2044.

Avants, B. B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., & Gee, J. C. (2010). The optimal template effect in hippocampus studies of diseased populations. *NeuroImage, 49*(3), 2457-2466.

Bartko, J. J. (1991). Measurement and reliability: Statistical thinking considerations. *Schizophrenia Bulletin, 17*(3), 483-489.

Bartley, A. J., Jones, D. W., & Weinberger, D. R. (1997). Genetic variability of human brain size and cortical gyral patterns. *Brain, 120 (Pt 2)*, 257-269.

Chupin, M., Mukuna-Bantumbakulu, A. R., Hasboun, D., Bardinet, E., Baillet, S., Kinkingnéhun, S., . . . Garnero, L. (2007). Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: Method and validation on controls and patients with Alzheimer's disease. *NeuroImage, 34*(3), 996-1019.

Cox, R. W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and biomedical research, an international journal, 29*(3), 162-173.

Csernansky, J. G., Joshi, S., Wang, L., Haller, J. W., Gado, M., Miller, J. P., . . . Miller, M. I. (1998). Hippocampal Morphometry in Schizophrenia by High Dimensional Brain Mapping. *Proceedings of the National Academy of Sciences of the United States of America, 95*(19), 11406-11411.

Csernansky, J. G., Wang, L., Jones, D., Rastogi-Cruz, D., Posener, J. A., Heydebrand, G., . . . Miller, M. I. (2002). Hippocampal Deformities in Schizophrenia Characterized by High Dimensional Brain Mapping. *American Journal of Psychiatry, 159*(12), 2000-2006. doi: doi:10.1176/appi.ajp.159.12.2000

Csernansky, J. G., Wang, L., Swank, J., Miller, J. P., Gado, M., McKeel, D., . . . Morris, J. C. (2005). Preclinical detection of Alzheimer's disease: hippocampal shape and volume predict dementia onset in the elderly. *NeuroImage, 25*(3), 783-792.

Dawant, B., Hartmann, S., Thirion, J. P., Maes, F., Vandermeulen, D., & Demaerel, P. (1999). Automatic 3-D segmentation of internal structures of the head and MR images using a combination of similarity and free-form transformations, part II: methodology and validation on severely atrophied brains. *IEEE TRans. Med. Imaging, 18*, 926-971.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology, 26*(3), 297-302.

Doxey, C. R., & Kirwan, C. B. (2015). Structural and functional correlates of behavioral pattern separation in the hippocampus and medial temporal lobe. *Hippocampus, 25*(4), 524-533.

Evans, A. C. (2006). The NIH MRI study of normal brain development. *NeuroImage, 30*(1), 184-202.

Evans, A. C., Janke, A. L., Collins, D. L., & Baillet, S. (2012). Brain templates and atlases. *NeuroImage, 62*(2), 911-922.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., . . . Dale, A. M. (2002). Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron, 33*(3), 341-355.

Fischl, B., Salat, D. H., van der Kouwe, A. J. W., Makris, N., Ségonne, F., Quinn, B. T., & Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *NeuroImage, 23, Supplement 1*(0), S69-S84.

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical Surface-Based Analysis: II: Inflation, Flattening, and a Surface-Based Coordinate System. *NeuroImage, 9*(2), 195-207.

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage, 54*(1), 313-327.

Hasboun, D., Chantome, M., Zouaoui, A., Sahel, M., Deladoeuille, M., Sourour, N., . . . Dormont, D. (1996). MR determination of hippocampal volume: Comparison of three methods. *American journal of neuroradiology, 17*(6), 1091-1098.

Hunsaker, M. R., & Amaral, D. G. (2014). A Semi-Automated Pipeline for the Segmentation of Rhesus Macaque Hippocampus: Validation across a Wide Age Range. *Plos One, 9*(2). doi: 10.1371/journal.pone.0089456

Hunsaker, N. (2014). BYU MRI Guide / semiautomated_hippocampus_tracing. 2015, from https://bitbucket.org/njhunsaker/byu-mri-guide/wiki/semiautomated_hippocampus_tracing

Insausti, R., Juottonen, K., Soininen, H., Insausti, A. M., Partanen, K., Vainio, P., . . . Pitkanen, A. (1998). MR volumetric analysis of the human entorhinal, perirhinal, and temporopolar cortices. *American journal of neuroradiology, 19*(4), 659-671.

Jung, R. E., Ryman, S. G., Vakhtin, A. A., Carrasco, J., Wertz, C., & Flores, R. A. (2014). Subcortical correlates of individual differences in aptitude. *Plos One, 9*(2), 1.

Khan, A. R., Wang, L., & Beg, M. F. (2008). FreeSurfer-initiated fully-automated subcortical brain segmentation in MRI using Large Deformation Diffeomorphic Metric Mapping. *NeuroImage, 41*(3), 735-746.

Klein, A., Ghosh, S. S., Avants, B., Yeo, B. T. T., Fischl, B., Ardekani, B. A., . . . Parsey, R. V. (2010). Evaluation of volume-based and surface-based brain image registration methods. *NeuroImage, 51*(1), 214-220.

Kovacevic, J. (2006). From the Editor-in-Chief. *Image Processing, IEEE Transactions on, 15*(12), 3625-3626.

Maltbie, E., Bhatt, K., Paniagua, B., Smith, R. G., Graves, M. M., Mosconi, M. W., . . . Styner, M. A. (2012). Asymmetric bias in user guided segmentations of brain structures. *NeuroImage, 59*(2), 1315-1323.

Persson, J., Spreng, R. N., Turner, G., Herlitz, A., Morell, A., Stening, E., . . . Söderlund, H. (2014). Sex differences in volume and structural covariance of the anterior and posterior hippocampus. *NeuroImage, 99*(0), 215-225.

Pluta, J., Avants, B. B., Glynn, S., Awate, S., Gee, J. C., & Detre, J. A. (2009). Appearance and incomplete label matching for diffeomorphic template based hippocampus segmentation. *Hippocampus, 19*(6), 565-571.

Pruessner, J. C., Li, L. M., Serles, W., Pruessner, M., Collins, D. L., Kabani, N., . . . Evans, A. C. (2000). Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. *Cerebral cortex, 10*(4), 433-442.

Sparks, B. F., Friedman, S. D., Shaw, D. W., Aylward, E. H., Echelard, D., Artru, A. A., . . . Dager, S. R. (2002). Brain structural abnormalities in young children with autism spectrum disorder. *Neurology, 59*(2), 184-192.

Talairach, J., Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System—An Approach to Cerebral Imaging*. New York: Thieme Medical Publishers.

Turner, A. D., Furey, M. L., Drevets, W. C., Zarate Jr, C., & Nugent, A. C. (2012). Association between subcortical volumes and verbal memory in unmedicated depressed patients and healthy controls. *Neuropsychologia, 50*(9), 2348-2355.

Tustison, N. J., Avants, B. B., & Gee, J. C. (2009). Directly Manipulated Free-Form Deformation Image Registration. *Image Processing, IEEE Transactions on, 18*(3), 624-635.

Tustison, N. J., Cook, P. A., Klein, A., Song, G., Das, S. R., Duda, J. T., . . . Avants, B. B. (2014). Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *NeuroImage, 99*(0), 166-179.

Tustison, N. J., Johnson, H. J., Rohlfing, T., Klein, A., Ghosh, S. S., Ibanez, L., & Avants, B. B. (2013). Instrumentation bias in the use and evaluation of scientific software: recommendations for reproducible practices in the computational sciences. *Front Neurosci, 7*, 162. doi: 10.3389/fnins.2013.00162

Wang, L., Miller, J. P., Gado, M. H., McKeel, D. W., Rothermich, M., Miller, M. I., . . . Csernansky, J. G. (2006). Abnormalities of hippocampal surface structure in very mild dementia of the Alzheimer type. *NeuroImage, 30*(1), 52-60.

Weisstein, E. W. (1999). Euler-Lagrange Differential Equation. Retrieved 06/09, 2015, from http://mathworld.wolfram.com/Euler-LagrangeDifferentialEquation.html

Wenger, E., Mårtensson, J., Noack, H., Bodammer, N. C., Kühn, S., Schaefer, S., . . . Lövdén, M. (2014). Comparing manual and automatic segmentation of hippocampal volumes: Reliability and validity issues in younger and older brains. *Human Brain Mapping, 35*(8), 4236-4248.

Wilke, M., Holland, S. K., Altaye, M., & Gaser, C. (2008). Template-O-Matic: A toolbox for creating customized pediatric templates. *NeuroImage, 41*(3), 903-913.

Yassa, M. A., & Stark, C. E. L. (2009). A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe. *NeuroImage, 44*(2), 319-327.

Yassa, M. A., Stark, S. M., Bakker, A., Albert, M. S., Gallagher, M., & Stark, C. E. L. (2010). High-resolution structural and functional MRI of hippocampal CA3 and dentate gyrus in patients with amnestic Mild Cognitive Impairment. *NeuroImage, 51*(3), 1242-1252.

Yoon, U., Fonov, V. S., Perusse, D., & Evans, A. C. (2009). The effect of template choice on morphometric analysis of pediatric brain data. *NeuroImage, 45*(3), 769-777.

Zijdenbos, A. P., Dawant, B. M., Margolin, R. A., & Palmer, A. C. (1994). Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Transactions on Medical Imaging, 13*(4), 716-724.

APPENDIX A: Scripting Commands

This appendix only contains a small portion of the scripts used. A comprehensive list can be found at: https://github.com/nmuncy/HippSeg_Pipeline

1. buildtemplateparallel.sh –d 3 –o <prefix> -c 2 –r 1 –j 4 <input_struct.nii.gz>

2. ANTS 3 -o <prefix> -i 100x100x100x20 -t SyN[0.1] -r Gauss[3,0.] –m CC[<template.nii.gz>,<struct.nii.gz>,<4>,4] –m PSE[<template.nii.gz>,<struct.nii.gz>,<template_mask.nii.gz>,<struct_mask.nii.gz>,<4>, <0.8>,<100>,0,25,10000]

3. WarpImageMultiTransform 3 <struct.nii.gz> <prefix>ParticipantToTemplate.nii.gz <prefix>Warp.nii.gz <prefix>Affine.txt -R <template.nii.gz>
   WarpImageMultiTransform 3 <template.nii.gz> <prefix>TemplateToParticipant.nii.gz -i <prefix>Affine.txt <prefix>InverseWarp.nii.gz –R <struct.nii.gz>
   WarpImageMultiTransform 3 <template_mask.nii.gz> <prefix>auto.nii.gz –i <prefix>Affine.txt <prefix>InverseWarp.nii.gz –R <struct.nii.gz>

4. c3d <input.nii.gz> -as SEG –cmv –pop –pop –thresh 50% inf 1 0 –as MASK –push SEG – times –o <L_output.nii.gz> -push MASK –replace 1 0 0 1 –push SEG –times –o <R_output.nii.gz>

5. c3d <input_L.nii.gz> -thresh 0.25 1 1 0 –o <output_L_thresh.nii.gz>; c3d <input_R.nii.gz – thresh 1.25 2 2 0 –o <output_R_thres.nii.gz>

6. recon-all –all –subjid <subjectID> -sd <subjectDirectory> -notal-check –hippo-subfields

CURRICULUM VITAE

Nathan Muncy
Nathan.muncy1@gmail.com
13405 Garfield Ct.
Thornton, CO. 80241
(720) 280-9344

Education

Brigham Young University Expected Graduation August 2016
Provo, UT

- MS degree in Neuroscience
- Worked under Dr. Kirwan
- Thesis work in MRI segmentation algorithms
- Researched test-retest affects on memory specificity
- Studied and implemented MRI processing pipelines
- Worked in Unix, Linux, Bash
- Used AFNI, ANTs, ITK-snap, FSL, FreeSurfer, MANGO, SPSS, R, e-prime for research
- 3.87 G.P.A.

Brigham Young University–Idaho Graduated July 2013
Rexburg, ID

- BS degree in Psychology
- Minor in Philosophy
- Conducted research on Parkinson's Disease in collaboration with UCLA under Dr. Eckersell
- Member of the Neuroscience Academic Society
- 3.92 G.P.A.

Work Experience

Teacher's Assistant July 2015 – April 2016
BYU Provo, UT

- Worked under Dr. Kirwan and Dr. Hedges for Cognitive Neuroscience
- Held office hours, test reviews, tutoring sessions
- Wrote quizzes and test questions, graded assignments, maintained gradebook

Teacher's Assistant September 2014 – July 2015
BYU Provo, UT

- Worked under Dr. Woods, for the classes of Physiology and Advanced Physiology
- Lectured on physiology topics
- Taught lab techniques and etiquette

Tutor, Teacher's Assistant January 2010 – July 2013
BYU–Idaho Rexburg, ID

- Worked under Dr. Lowry, for the classes of Behavioral Neurobiology, Sensation and Perception, and general Psychology
- Tutored students in Behavioral Neurobiology, Sensation and Perception
- Substitute taught for the Professor
- Helped create and graded homework, quizzes, and tests
- Assisted in neurobiology anatomy labs

Awards

- BYU, Graduate Student Scholarship: half tuition — September 2014 – May 2016
- BYUI, Graduated with Honors — August 2013
- BYUI, Academic Scholarship: full tuition — January 2010 – August 2013

Published Presentations

A Semi-automated pipeline for hippocampal segmentation in patient and control populations
Cognitive Neuroscience Conference — San Francisco, 2015

Posters Presented

The effects of diet on hippocampal volume
Mary Lou Fulton Conference, BYU — Provo, UT 2016

The effects of testing encoding on pattern separation
Mary Lou Fulton Conference, BYU — Provo, UT 2016

A Semi-automated pipeline for hippocampal segmentation in patient and control populations
Cognitive Neuroscience Conference — San Francisco, 2015

Volunteer Work

Volunteer Theology Teacher — September 2008 – September 2014
Rexburg YSA Branch, Denver YSA Branch — The Church of Jesus Christ of Latter-day Saints
- Trained in methods of teaching
- Taught weekly classes of 60 students

Official Religious Representative in Portugal — March 2006 – April 2008
Lisbon Mission — Portugal
- Fluent in Portuguese, proficient in Spanish
- Conducted over 1000 religious meetings and discussions over two years
- Trained new Religious Representatives
- Managed Groups of Religious Representatives
- Served as one of three members of a Religious Executive Board