Segmentation and registration are two categories of image processing methods that are often required in the analysis of neuroimaging data. A variety of automated segmentation and registration methods exist, and new versions and new methods are being created continually. However, performance characteristics of these tools are rarely quantified. Furthermore, they are rarely compared with one another in a quantitative and consistent manner. Even when segmentation and registration software tools are associated with published validation studies, these studies may not necessarily be comparable with one another. Thus, many end users are not provided with sufficient information to make well-informed choices when they use or design an image analysis sequence.

This project proposes to address the above problem through the development of an online validation resource available to the public. Briefly, this resource will provide a set of manually delineated test data (Aim 1), automated tools for evaluating performance of segmentation or registration methods using those data (Aim 2) and an online repository to facilitate the application of the automated tools and the archival of their results, which will presented and compared via a web interface (Aim 3). Users will first download data appropriate for a given segmentation or registration task and will then perform that task using the method that they have chosen choose to evaluate. The user will then upload the resulting data to the resource website, and the web server will compute a series of measures on the data to assess performance. These measures and the uploaded data will be archived on the resource server as a validation study. An interactive web-based interface will then provide convenient access to the validation results for the study. Visitors to the resource website will be able to explore the different studies, allowing for detailed comparison of different approaches. We will develop this resource through the following three Specific Aims.

Aim 1: Evaluation Data. We will develop a collection of test data to be used within the validation framework.

We will produce a test set composed of 80 human brain MRI selected from the OASIS database. Each brain will be labeled independently by 4 trained raters following a delineation protocol. To ensure that evaluations may be conducted repeatedly without overfitting to the test data, 40 of these sets of neuroanatomical labels will not be made public. We will also make use of additional training and test data produced external to this project. These will include (a) 68 manually delineated brain MRI volumes that are currently available to the public and have been used in published studies; (b) 100 MRI volumes that are currently being delineated for a related project. Users of the resource will download these data as sets of individual brains for segmentation tasks or as sets of paired brains for registration tasks.

Aim 2: Evaluation Measures We will develop a series of analytic tools for evaluating MRI brain segmentation and registration methods.

Evaluating user segmentations of MRI brain data will be based on comparisons with manually labeled neuroanatomical structures. We will create the tools to make these comparisons using well-established measures of set agreement. Segmentation will be evaluated by comparing the agreement between the sets of labels produced by the user with sets of the same labels delineated by human raters. Evaluating the quality of registration between source and target images will also be based on set agreement; in this case, the transformed source label sets will be compared with the manually labeled target data. In addition to overlap and distance errors used for segmentation evaluation, measures specific to registration comparison will include measures of image similarity and measures of distortion computed from the displacement field used to match the source brain image to its target. Aim 2 also devises the registration or segmentation tasks that will be performed during the evaluation.

Aim 3: Online Framework We will develop a public online resource that will provide a system for providing test and training data, for performing evaluations, for archiving results, and for disseminating information.

This Aim develops the web interface, control code, and data repository for the online resource. The web interface will (a) allow a user to download the validation data identified in Aim 1, (b) accept user-uploaded segmented or registered image data, (c) compute the measures described in Aim 2, and (d) archive all data and results. The web interface will present these data in numeric and graphical form to site visitors to enable comparisons across different methods. The website will also be supported by a forum that will allow users to comment on their own results and to discuss issues relevant to segmentation, registration, and validation.

A. Significance

Neuroimaging studies typically make use of automated approaches to perform tasks such as registration, skull stripping, tissue-class segmentation, and anatomical labeling. Investigators use automated methods because they may be fast, convenient, sophisticated, and consistent, but these methods are often treated as black-box solutions. The appropriate settings, accuracy, and robustness of these automated algorithms is often unknown, especially to the end user.

Interpretation of scientific results from imaging studies that use segmentation or registration methods usually assumes that each step in the processing has produced accurate results. In many instances, this assumption is not valid. It is therefore vital that the neuroimaging community understands the performance capabilities and limitations of the different techniques available. To find out what tools researchers are using for registering brain MRI data, we conducted a review of the 593 articles published in NeuroImage in the first eight months of 2008. We found that 82 studies used nonlinear inter-subject registration; 74 of these used SPM Normalize (Ashburner and Friston, 1999). In our recent comparison of registration methods (Klein et al., 2009), we found that SPM Normalize produced among the worst results. From our evaluation and literature search, it appears that numerous researchers are employing substandard tools for registration and labeling, and they have yet to be convinced that there exist compelling alternatives.

The traditional place for investigators to receive information about the capabilities and performance of algorithms is in the published literature, where one may find several brain image registration and segmentation studies. Early work by Strother et al. (1994) compared 4 different registration approaches. West et al. (1997) performed a large-scale study in which developers of several algorithms were invited to perform the registration of test data, and their results were then evaluated independently. Boesen et al. (2004) evaluated 4 different skull-stripping methods and also introduced a web-based evaluation service, termed Brain Extraction Evaluation (BEE); users could download a set of 15 test MRI volumes used in the paper, submit their brain extraction results, and have their evaluation results e-mailed to them. A series of publications by Hellier et al. (2001, 2002, 2003) evaluated five different fully automated nonlinear brain image registration software programs using global measures comparing 17 deformed MRI source images and one target image. Fennema-Notestine et al. (2006) conducted a comparison of 4 publicly available methods for skull-stripping applied to 16 data volumes. Authors of each algorithm were invited to participate by suggesting parameters based on their own tests on practice data (3 of the 4 elected to participate). The data processing was then performed without further input from the developers. Yassa and Stark (2009) compared 4 nonlinear registration methods applied to regions in the medial temporal lobe by apply these methods to manually labeled brain regions, to weighted masks for these regions, or to the original unlabeled brains. Evaluation was performed on pairs of images using an overlap measure and a measure of blur in a group average of co-registered images. An extensive comparison of 14 algorithms was performed by Dr. Klein (Subcontract PI) (Klein et al., 2009). Principal architects of the respective algorithms were coauthors on that publication, bit their participation was limited to providing guidance to Dr. Klein, who applied the algorithms to the data (see **Sec. C.1.**).

Many of the validation studies in the neuroimaging field make use of publicly available test data. The Internet Brain Segmentation Repository (IBSR) provides datasets that have been delineated using manually-guided processes (http://www.nitrc.org/projects/ibsr/). IBSR also provides similarity metrics computed on results produced by several algorithms, giving developers of new algorithms a basis for comparison. The BrainWeb digital phantom (http://www.bic.mni.mcgill.ca/brainweb/) has also emerged as a standard in brain imaging (Collins et al., 1998). The phantom was constructed based on a high-resolution image from one individual. The BrainWeb site provides several versions of the scan at different resolutions, noise levels, and with various degrees of RF nonuniformity. The site was later extended to provide templates produced from 20 additional individuals (Aubert-Broche et al., 2006). Both the IBSR and BrainWeb datasets have been used in several publications and have played a key role in the validation and improvement of image processing algorithms.

At the 2007 conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), two segmentation competitions were held as part of a workshop (van Ginneken et al., 2009). Participants were able to segment test data for one of two problems: caudate segmentation or liver segmentation. For both segmentation problems, a training set and a test set were provided to interested participants prior to the conference. A second test set was provided on the day of the workshop and had to be segmented within 3 hours after it was made available. The results were ranked using a set of metrics. Two websites (http://sliver07.org and http://www.cause07.org) were established to provide access to the results from the competition, and also allow existing and new participants to submit new results for evaluation. Additional 3D segmentation competitions with grand challenges were held as part of the 2008 and 2009 MICCAI conferences (http://grand-challenge2009.bigr.nl/) and are again being conducted for MICCAI 2010.

Related works that can be used for evaluating segmentation or registration results include the Valmet software tool (Gerig et al., 2001) and the Non-rigid Image Registration Evaluation Project (NIREP) (http://www.nirep.org). The Windows-based Valmet provides tools for measuring and visualizing the differences between corresponding image segmentations. It uses several measures to compare segmentations: overlap ratio, Hausdorff distance, surface distance, and probabilistic overlap. The NIREP project "has been started to develop, establish, maintain, and endorse a standardized set of relevant benchmarks and metrics for performance evaluation of nonrigid image registration algorithms." The initial phase of the project will include 16 manually labeled brain images (32 labeled regions in 8 men and 8 women) and four evaluation metrics: (a) relative overlap (b) variance of the registered intensity images for an image population, (c) inverse consistency error between a forward and reverse transformation between two images, and (d) transitivity.

Assessment of accuracy is an essential part of the algorithm development process. Developers need to understand how their algorithms perform so that they can observe where current methods need to be improved and so that they can see how new developments compare with existing ones. Still, researchers often struggle with identifying ways to test their methods, and they do not always select methods or data that others have used. Validation studies that are published in separate publications are not always comparable, as the testing procedures and assumptions may differ even when the same assessment measures and input data are used. Furthermore, the publication of results by an algorithm developer will usually result in that algorithm performing the best in the evaluation study. This could be an indication of bias on the part of the group performing the comparison, a reluctance to publish methods that do not perform as well as the state of the art, or a greater familiarity with their own algorithms and data than they have with existing methods.

In this application, we propose the development a web-based resource that will provide an automated framework where users can evaluate different brain MRI segmentation and registration methods. This addresses problems faced by both the developers and end users of brain segmentation and registration algorithms. Users will be able to download test sets of imaging data, perform tasks using the algorithms they are evaluating, and upload the resulting data to the server. The server will then compute a set of evaluation measures based on comparison with expert-delineated data. It will archive the uploaded data and evaluation measures and then provide the evaluation results to the user in a convenient web-browser interface. The server will also provide access to these validation studies to visitors of the site, allowing them to make comparisons across methods.

B. Innovation

This application proposes several novel components. First, it proposes a system of evaluation that will be completely automatic with results computed and disseminated online. Though a few sites associated with the MICCAI Grand Challenges allow users to upload new data for evaluation of single structures, we are unaware of any that provide fully automated facilities for evaluation of segmenting the brain into multiple regions or for assessing brain registration algorithms. Second, the test data used will be labeled four times; each delineation performed by a different trained human rater. We are unaware of any publicly available brain data that has been labeled this comprehensively. The multiple rater data will allow us to provide better evaluation of how machine algorithms compare with human observers by providing confidence bounds based on interrater comparisons. An additional aspect of this work is that we will produce both private and public evaluation data. The public data will provide users with data for algorithm training and testing; the private data will ensure the integrity of the test results. Importantly, the results of the evaluations will be open, which will encourage greater transparency in validation studies and provide better opportunities for verification of results.

There are several benefits offered by these innovations that will impact the brain imaging community. First, algorithm developers who need to evaluate their own methods during development or in preparation for publication will be provided with an easy-to-use common framework. As this will aid developers in improving their own approaches, it should also have a broader impact on the neuroimaging community by producing more accurate and more automated approaches. Furthermore, the public online display of comparative results will alert and inform the larger developer and user communities of these alternatives. They will be able to make more informed decisions about which algorithms to use in different settings. By improving the quality and understanding of the software packages used in neuroimaging analysis, this proposal has the potential to improve the scientific products that are derived from their application. In this proposal, we focus our evaluation on segmentation and registration tasks performed on human brain MRI data. The framework may also be extended to other imaging modalities and other segmentation or registration problems.

C. Approach

The approach we adopt for validating segmentation and registration algorithms will reply upon data that have been delineated manually by trained raters following an established protocol. Under **Aim 1**, the raters will delineate sets of MRI data multiple times. This repeat delineation by multiple raters is critical to establish confidence bounds on the boundaries of anatomical regions. During the initial phase of development of our system, we will also use data that were previously delineated and released publicly; only a single label set is available for most of these volumes, which limits their utility somewhat. In **Aim 2**, we develop a testing protocol and related software to perform evaluations that will allow us to rank and compare image processing algorithms. Finally, in **Aim 3**, we integrate the data, measures, and testing protocols into a web-based resource that will guide users through the process of validating their algorithms and will also provide the results to the public at large so that the neuroimaging community may benefit from the evaluations that are performed.

C.1. Preliminary Studies

C.1.a. LONI Probabilistic Brain Atlas Projects

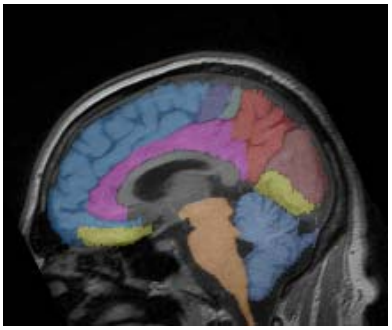

Figure C.1. A labeled subject MRI from the LPBA40 data collection. Each subject brain MRI was manually delineated to identify 56 neuroanatomical regions.

Dr. Shattuck (PI) supervised the development of the publicly released 40 Subject LONI Probabilistic Brain Atlas (LPBA40) (Shattuck et al., 2008). MRI data were acquired from 40 normal healthy research volunteers; each MRI was delineated by trained raters using our BrainSuite software (Shattuck and Leahy, 2002). The delineations consisted of 56 structures per brain (see **Fig. C.1.**) and were performed according to protocols that we developed for that project. The protocols provided detailed written instructions and figures showing surface renderings and 3D cross-sectional views of reference anatomy. Each protocol also provided a general description of the structure to be delineated and specified a single plane of section (coronal, sagittal, or transaxial) in which that structure should be labeled. Additionally, each protocol defined where the rater should initiate and complete the labeling. The protocol was published as supplemental material for the manuscript describing the LPBA40 (Shattuck et al., 2008) and is also available (http://www.loni.ucla.edu/Protocols/LPBA40). All data resulting from the LPBA40 were made publicly available (http://www.loni.ucla.edu/Atlases/LPBA40), allowing users to apply the labels to machine learning approaches (e.g., Tu et al., 2008), to construct their own atlases, or to use them in validation studies (see **Aim 3**). Dr. Shattuck (PI) is currently supervising an effort at the Laboratory of Neuro Imaging to label MRI from 100 ICBM subjects. The brains are being labeled using the LPBA40 protocol. The demographics include 10 males and 10 females in each of the age ranges: 20-29, 30-39, 40-49, 50-59, 60-69. These data will also be available to the present project during its development.

C.1.b. Registration Comparisons

Dr. Klein (subcontract PI) recently conducted a very large evaluation study of brain image registration algorithms (Klein et al., 2009). Fourteen methods produced by laboratories around the world were evaluated using 8 different error measures. More than 45,000 registrations between 80 manually labeled brains were performed by algorithms including AIR (Woods et al., 1998), ANIMAL (Collins et al., 2005), ART (Ardekani et al., 2005), Diffeomorphic Demons (Vercauteren et al., 2007), FNIRT (Andersson et al., 2008), IRTK (Rueckert et al., 2006), JRD-fluid (Chiang et al., 2008), ROMEO (Hellier et al., 2001B), SICLE (Christensen, 1999; Christensen and Johnson, 2001; Johnson and Christensen, 2002), SyN (Avants et al., 2008), and four different SPM5 algorithms -- “SPM2-type” (Friston et al., 1995) and regular Normalization (Ashburner and Friston, 1999), Unified Segmentation (Ashburner and Friston, 2005), and the DARTEL Toolbox (Ashburner, 2007). All of these registrations were preceded by linear registration between the same image pairs using FLIRT (Smith et al., 2004). One of the most significant findings of this study was that the relative performances of the registration methods under comparison appear to be little affected by the choice of subject population, labeling protocol, and type of overlap measure. This is important because it suggests that the findings are generalizable to new subject populations that are labeled or evaluated using different labeling protocols. Furthermore, we ranked the 14 methods according to three completely independent analyses (permutation tests, one-way ANOVA tests, and indifference-zone ranking) and derived three almost identical top rankings of the methods.

An extension of that study was recently completed by Dr. Klein (Klein et al., 2010). This new work directly compared some of the most accurate of the volume-based registration methods with surface-based registration methods, and also compared registrations of whole-head and brain-only images, and custom template vs. direct brain-to-brain registrations. More than 16,000 registrations between 80 manually labeled brain images

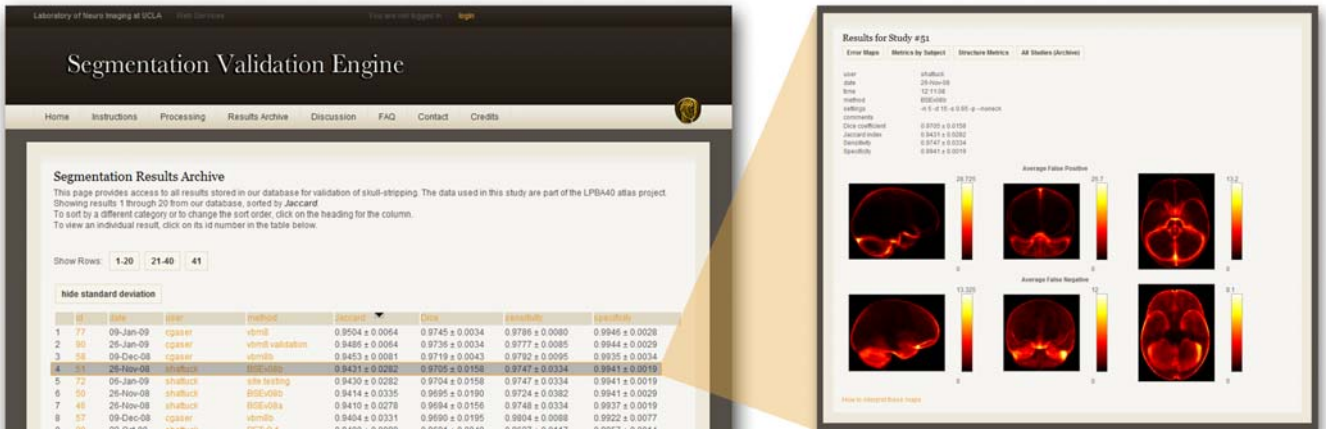

Figure C.2. Segmentation Validation Engine Results Page. Results from several algorithms have been uploaded by site users. Site visitors can rank the methods according to different measures and explore the details of the results. When a user clicks on a particular row, the server presents a new page that details the validation results for that particular entry (left). Parameter settings, user comments, and the various measures are shown. Also shown are projection maps indicating the location of false positive and false negative errors. The user can drill down further to examine scores for individual structures or for individual subject data.

were performed by the volume-based algorithms ART and SyN and the surface-based algorithms FreeSurfer (Fischl et al., 1999) and Spherical Demons (Yeo et al., 2009). Overlap performance was computed for every combination of volume-based and surface-based: labeling, registration, and evaluation.

C.1.c. Segmentation Validation Engine

We developed a web-based resource (http://sve.loni.ucla.edu) for evaluating the performance of brain extraction in T1-weighted MRI (Shattuck et al., 2009). The resource provides both the MRI data to be segmented and an online tool that performs an evaluation of the resulting segmented data against manual segmentations of the data. Registered users may download the set of 40 LPBA40 whole-head MRI volumes, segment each volume into brain/non-brain with their method of choice, and upload their results to our SVE server. The server then archives the data, along with user-specified information regarding the method and parameters used, and any other notes the user may wish to share. The uploaded segmentation data are processed by the server, typically in less than two minutes, to produce a set of agreement measures based on comparison with manually delineated brain volumes. The measures include Dice coefficient, Jaccard index, sensitivity, and specificity. These evaluation results are archived on the server, and the user is presented with a detailed report of the results that can be explored at several levels.

The archived results are also visible to any visitor to the SVE site. Visitors are presented with a table that can be sorted by any one of these metrics (see **Fig. C.2**). Additional features allow a visitor to examine results uploaded by a particular user or results attributed to a particular algorithm. This provides a convenient method to compare different approaches rapidly. The interface allows visitors to drill down further into the data by selecting a particular result. Visual displays of the segmentation results are presented to the user, where pre-computed spatial warps map the errors in each of the 40 subjects to a common atlas space. The errors are averaged at each spatial point and summed along each cardinal axis. This provides axial, sagittal, and coronal displays of the average false positives and average false negatives in the image. The user can drill down even further and view the measures for each brain. We are also able to compute error rates in individual structures in the brain. These different views of the data provide users with many ways to examine which combinations of algorithms and settings perform best on the data. We applied this framework to the evaluation of three popular skull-stripping algorithms — the Brain Extraction Tool (Smith, 2002), the Hybrid Watershed Algorithm (Ségonne et al., 2004), and the Brain Surface Extractor (Shattuck et al., 2001) under several different program settings (Shattuck et al., 2009). Additional methods have been evaluated by other users of the website.

C.2. Aim 1: Evaluation Data

Aim 1 will produce evaluation data by manually labeling 80 MRI, termed the LONI Validation Engine 80 (**LVE80**). Each of image in **LVE80** will be labeled separately by 4 raters following a delineation protocol. During the initial development of this project, we will make use of existing labeled MRI data (see **Table C.1.**) that have been used previously in several validation studies. These existing sets have some limitations, such as using lab-specific protocols and having delineations produced by only single raters. Though the production of the LVE80 data will be labor-intensive, having inter-rater validation data for each image will establish confidence bounds in our results which is not available for the existing labeled data.

Dataset	Subjects	Raters/ brain	Anatomic Regions	Ages μ=mean (years)	Volume dimensions (voxels)	Voxel size (mm)
LVE80**	80 (40 ♂, 40 ♀)	**4**	98+	to be determined	256x256x128	1x1x1.25
LPBA100*	100 (50 ♂, 50 ♀)	1	56	μ= 43.64	160x256x256	1x1x1
LPBA40	40 (20 ♂, 20 ♀)	1	56	μ=29.20	256×256×124	38: 0.86×0.86×1.5 2: 0.78×0.78×1.5
IBSR18	18 (14 ♂, 4 ♀)	1	84	7–71 μ=38.4 +4 "juveniles"	256×256×128	8: 0.94×0.94×1.5 6: 0.84×0.84×1.5 4: 1×1×1.5
MGH10	10 (4 ♂, 6 ♀)	1	74	range 22–29	256×256×128	1×1×1.33

Table C2. Validation data collections. (*in progress; **proposed) A total of 248 delineated brain MRI will be available for testing or training purposes. This proposal will produce LVE80, a new collection of 80 delineated MRI that will serve as ground truth for testing. All data collections consist of T1-weighted scans.

MRI Data The LVE80 data will be produced from the NIH-funded Open Access Series of Imaging Studies (OASIS) project (http://www.oasis-brains.org), which allows redistribution of data under its data usage agreement. Images will selected to include 40 male and 40 female normal subjects covering a range of ages with an approximately uniform distribution. The age ranges in the OASIS database are 18-96.

Labeling Protocol An essential part of this project will be the use of a protocol that is not specific to an individual lab. We have been involved in the development of a new, neuroimaging community-oriented brain labeling protocol as part of an effort termed BrainCOLOR (Collaborative Open Labeling Online Resource), The BrainCOLOR protocol has been being developed by Dr. Jason Tourville (Consultant) as a consultant to NeuroMorphometrics, Inc. Dr. Klein (subcontract PI) has catalyzed this effort in several ways, including (a) forming an advisory group, called the "Brain Color Crowd (BCC)" that has held meetings and online discussions; (b) created the www.brainCOLOR.org website for the free dissemination of related labeled and shape analysis data; (c) promoted the labeling protocol derived from some of the best and most frequently used parcellation methods established by other research groups; (d) conducted two surveys on the BrainCOLOR website to gather information and opinions from neuroscientists and neuroanatomists regarding the labeling protocol and to determine what image data they would want to have labeled using this protocol. Formation of the BCC has created a small community of people interested in brain labeling and has led to very active discussions among investigators at different institutions (Columbia, UCLA, MIT, Harvard, Vanderbilt University, Neuromorphometrics, U. Mass).

The BrainCOLOR divides the cerebral cortex is into 49 macro-anatomically defined regions in each hemisphere. Region of interest (ROI) boundary definitions were derived from a number of cortical parcellation methods currently in use. Protocols from the Laboratory of Neuroimaging at UCLA (LONI; Shattuck et al., 2008), the University of Iowa Mental Health Clinical Research Center (IOWA; Crespo-Facorro et al., 2000; Kim et al., 2000), the Center for Morphometric Analysis at Massachusetts General Hospital (MGH-CMA; Caviness et al., 1996), a collaboration between the Freesurfer group at MGH and Boston University School of Medicine (MGH-Desikan; Desikan et al., 2006), and UC San Diego (Carper & Courchesne, 2000; Carper & Courchesne, 2005; Carper et al., 2002) are specifically referenced in the protocol. Methods developed at Boston University (Tourville & Guenther, 2003), Brigham and Women's Hospital (McCarley & Shenton, 2008), Stanford (Allan Reiss lab), the University of Maryland (Buchanan et al., 2004), and the University of Toyoma (Zhou et al., 2007) were also consulted. The development of the protocol was also guided by the Ono, Kubik, and Abernathy (1990), Duvernoy (1999), and Mai, Paxinos, and Voss (Mai et al., 2008) neuroanatomical atlases. The subcortical protocol is currently in development.

Labeling Team: The labeling of the brain MRI data will be performed by 4 trained raters – 2 at Columbia University supervised by Dr. Klein (subcontract PI) and 2 at UCLA supervised by Dr. Shattuck (PI). These raters will be trained on the BrainCOLOR protocol by Dr. Tourville. Dr. Tourville was the lead developer of the BrainCOLOR protocol and has experience training individuals on it (see support letter from Dr. Tourville).

Training Procedure Dr. Tourville will train the 4 raters to label brain MRI according to the BrainCOLOR protocol. In Year 1, Dr. Tourville will visit the UCLA and Columbia sites to perform in-person training of the raters; year 1 training will be focused on the proper delineation of the cortical areas (49 structures / hemisphere). In Year 2, Dr. Tourville will again visit both labs to train the raters on the subcortical delineation protocol. After each visit, the raters will be responsible for applying the protocol to four brains. Additional training in both years will be conducted remotely, with feedback provided on the delineation of the training

data. During the training phase, the four raters will be tested for interrater consistency using intraclass correlation methods; we note that we have used this approach previously in training raters to delineate data such as the LPBA40 atlas (Shattuck et al., 2008).

Labeling Procedure During the labeling phase, each brain MRI will be independently labeled by the 4 trained raters. The MRI brain labels will be reviewed by Dr. Tourville. Because we are interested in inter-rater variation measures, this review will be focused primarily on identifying major errors in labeling. We will use Subversion version control repository to store the label data during the labeling procedure. This will guard against potential problems such as file system corruption or the accidental overwriting of files. It will also facilitate the sharing of the current version of the label data among the members of the research team.

C.2 Aim 2: Evaluation Measures

Anatomical labels delineated manually by an expert rater are often used as ground truth data for assessing performance in image segmentation and image registration tasks. Performance of automated segmentation algorithms can be measured by how closely the label sets they produce match the corresponding manual labels delineated on the same data. Similarly, automated image registration algorithms may also be evaluated using expert-labeled data. After registering a source image to a target image, the resulting transform can be used to transform the manual labels from the source image space to the target image space. If the registration has produced a suitable alignment of the labeled anatomical regions, then the transformed source labels should be in agreement with the target labels.

Under **Aim 2**, we develop the validation framework that will allow us to evaluate (a) segmentation results and (b) registration results. In the case of segmentation, our method will be to compare the newly segmented data against the **LVE80** label data that were manually delineated in **Aim 1**. This will be performed using a series of set similarity and distance measures. For the assessment of registration algorithms, we also make use of the manually delineated data. We will evaluate a registration algorithm's performance by examining its ability to transform labeled anatomical areas in a source brain to match labels in a target brain. This evaluation is performed using set similarity and distance measures; measures relevant to spatial transformations will also be computed to examine characteristics such as the smoothness of the spatial transformation. Evaluation studies for registration and segmentation will be performed for sets of specified tasks, which we describe below. The key components requiring effort in this aim will be (a) the development of software to compute, organize, and aggregate the validation measures on the data, and (b) the application of these algorithms to establish baseline data for the framework.

C.2.a. Evaluation Tasks

Evaluation Data The evaluation studies will be performed using the **LVE80** data set produced under **Aim 1**. To provide users with exemplar data, we will release the label volumes for 40 of the images in the **LVE80**. This will include 4 label sets (one for each rater) per MRI. These data may be used by algorithm developers and evaluators for such purposes as training data for machine learning algorithms, basis data for constructing anatomical priors, or test data for algorithm development. To preserve the integrity of the evaluation test, the label data for the remaining 40 MRI will not be released publicly. As the production of all 80 brains will span the course of this Project, we will provide incremental releases at three points in the Project. The first will occur when 20 brains have been labeled, the second will occur when 40 brains have been labeled, and the final release will occur when 80 brains have been labeled. At each interval, half of the brains will be released as training data along with their delineated labels, and half will be reserved as test data.

Segmentation Tasks We will evaluate the ability of segmentation algorithms to identify brain regions that are part of the BrainCOLOR protocol. A limitation of this type of approach is that algorithms may be developed with different interpretations of neuroanatomical structures than is specified in the test protocol. We note here that the BrainCOLOR protocol was based on an analysis of several existing protocols. Furthermore, it has been developed with openness in mind and with input from the neuroimaging community obtained via surveys available from the NITRC website. Additionally, users of the validation framework will be provided with training data that may be used during their algorithm development and tuning.

For the test segmentation, the users will be provided with N MRI volumes, where N=10, 20 or 40 depending on the phase of the Project. An equivalent number of training data with labels will also be provided. For the test phase of the evaluation, the users will be instructed to download the test data volumes, process them with the algorithm they are evaluating, and upload the generated labels for all images. We do not expect every segmentation algorithm to label the brain comprehensively. For example, specialized algorithms are often developed for labeling specific structures (e.g., hippocampus). Thus, each algorithm will be evaluated on the

structures that it does label. In **Aim 3**, users will be able to document the details of the method under evaluation, including specifying the subset of the BrainCOLOR structures that were labeled.

Users will also be able to segment the existing LPBA40, IBSR18, and MGH10 data, as well as the LPBA100 data when it is completed. We note that these evaluations have some limitations, including the public availability of the label data and the lack of repeat rater scores.

Registration Tasks The 80 labeled MRI volumes that will comprise the **LVE80** provide 6,320 possible non-trivial pairings of images (i.e., excluding registering an MRI to itself) that could be used in the registration evaluation. This is far more than were used in our previous study (Klein et al., 2009). It is not feasible investigators with an interest in registration validation to perform an exhaustive analysis of all possible pairs. We propose instead to provide evaluation based on subsets of the possible pairs. To provide subsets that represent the differing degrees of difficulty present in the pairs of images, we will first evaluate the registration difficulty by performing an exhaustive comparison with these new data. Based on our previous study (Klein et al., 2009) we have chosen to evaluate two methods, ART (Ardekani et al., 2005) and SyN (Avants et al., 2008). These will serve as initial benchmarks for our registration validation. We will use the Mean Overlap measure (see below), averaged over the ART and SyN results, to rank order the pairings. We will then draw random distributions from these to produce our evaluation sets.

We restrict our selection of pairs to prevent repeats in a test set. Using the existing MGH10, IBSR18, and LPBA40 labeled data, we can generate 5, 9, or 20 registration problems for the sets, respectively. Similarly, the LPBA100 data will provide 50 additional pairs drawn from drawn from 9,900 non-trivial pairings. During the initial period of the project, we will develop the system to use these data for registration evaluation. Users will be able to download the full sets of MRI data and request a set of evaluation pairings.

As the label data from the LVE80 become available, we will establish registration tasks based on them. This will again provide 3 phases of evaluation as 20, 40, and 80 brains become available, with half of the label data public during each phase. We will evaluate the degree of difficulty in rating each possible pairing as described above. For training and development purposes, users will be able to request a randomized set of data from the public component of the LVE80. With access to the label data for both source and target, the users will be able to observe the anatomical overlap directly. When algorithms are being evaluated, we will form each registration pair with one MRI that is in the training set and one MRI that is in the private set. Thus, users will have access to labels for only one image in the pair. With each release, we will establish one of the random pairings as a standard for comparison across algorithms.

Bandwidth An important consideration in this system will be the bandwidth required for uploading the registration results data for evaluation. There are three components to this, each of which can be used in evaluation: (1) the transformation field that maps from source space to target space; (2) the resampled source labels in target space; and (3) the resampled source image in target space. Formats for storing deformation fields are not standard. However, we can instead provide users with spatially indexed volumes (e.g., I(x,y,z) = x) that, when resampled into a new space, provide a mapping of coordinates from the target space T to the source space S. Because the initial images are linear gradients and the applied distortions are expected to be locally smooth, the encoded warp fields should compress well and not require prohibitive network transmit times. We could also use these fields to transform the label data, but we note that the labels are expected to be locally contiguous and will also compress very well. As a final consideration, we will be computing measures of image similarity between co-registered pairs of images. Even with compression, this would require that several hundred megabytes of data be uploaded for each evaluation study during final phase of the Project. If this is prohibitive to our users, we will explore the alternative approach of using the encoded spatial warp fields to resample the image data.

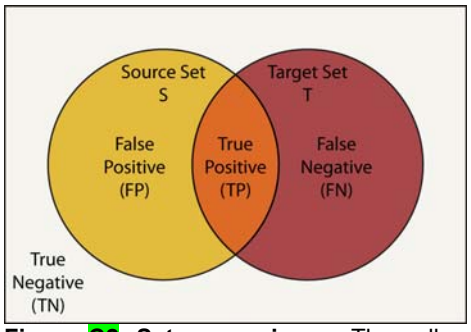

Figure C2. Set comparisons. The yellow circle represents the S, the results set or source set. The red circle represents T, the truth set or target set. The orange region represents their overlap, and is the set of correctly labeled voxels.

C.2.b. Evaluating Label Agreement

The segmentation and registration tasks described in **Sec. C.2.a.** produce label sets that must be examined for correspondence. We will evaluate this correspondence using set similarity measures and set distance functions. We note that these are well-established standards for inter-rater agreement. More novel measures could also be adopted. Notably, because the set data will be stored in a database, new approaches can be added and computed retrospectively.

Set Agreement Measure	Formula (binary case)	Formula (multivalue case)
union overlap (UO) a.k.a. Jaccard index	$UO = \frac{\lvert S \cap T \rvert}{\lvert S \cup T \rvert} = \frac{\lvert TP \rvert}{\lvert FP \rvert + \lvert TP \rvert + \lvert FN \rvert}$	$UO = \frac{\sum_r \lvert S_r \cap T_r \rvert}{\sum_r \lvert S_r \cup T_r \rvert}$
mean overlap (MO) a.k.a. Dice coefficient	$MO = \frac{\lvert S \cap T \rvert}{\frac{1}{2}(\lvert S \rvert + \lvert T \rvert)} = 2\frac{\lvert TP \rvert}{(\lvert FP \rvert + \lvert TP \rvert) + (\lvert TP \rvert + \lvert FN \rvert)}$	$MO = 2\frac{\sum_r \lvert S_r \cap T_r \rvert}{\sum_r (\lvert S_r \rvert + \lvert T_r \rvert)}$
target overlap (TO) a.k.a. Sensitivity a.k.a. Recall	$TO = \frac{\lvert S \cap T \rvert}{\lvert T \rvert} = \frac{\lvert TP \rvert}{\lvert TP \rvert + \lvert FN \rvert}$	$TO = \frac{\sum_r \lvert S_r \cap T_r \rvert}{\sum_r \lvert T_r \rvert}$
source overlap (SO) a.k.a. precision	$SO = \frac{\lvert S \cap T \rvert}{\lvert S \rvert} = \frac{\lvert TP \rvert}{\lvert TP \rvert + \lvert FP \rvert}$	$SO = \frac{\sum_r \lvert S_r \cap T_r \rvert}{\sum_r \lvert S_r \rvert}$
specificity	$specificity(\mathrm{S},\mathrm{T}) = \frac{\lvert (S \cup T)^C \rvert}{\lvert T^C \rvert} = \frac{\lvert TN \rvert}{\lvert TN \rvert + \lvert FP \rvert}$	

Table C2. Set agreement measures used in this project. We will compute these standard measures of set overlap to assess agreement between: (a) segmented data and ground truth delineations; (b) registered source label data and target label data. Notation: $A \setminus B$ is the set difference, i.e., the subset of voxels in A that are not members of B. $A \cap B$ is the intersection of sets A and B. The size, i.e., number of voxels, in set A is indicated by $|A|$. A^C is the set complement of A.

Set Similarity Measures Label agreement can be measured using several set similarity metrics. For this project, we will develop software for computing several of these measures (see **Table C2**). **Figure C2** illustrates set agreement for a binary classification comparison. An example of this would be brain extraction, where each voxel in an MRI is assigned one of two labels: brain or non-brain. The target set, T, represents the voxels identified as brain by an expert. For image segmentation, the source set S represents voxels identified as brain by another rater, e.g., a brain extraction algorithm. In the case of image registration, S represents the set of resampled voxels that were labeled as brain in the original source image and transformed to the target space. The four regions in Fig C2 are basic measures of accuracy and error: correctly labeled brain voxels are the True Positive set, $TP(T,S) = T \cap S$; correctly labeled non-brain voxels are the True Negative set, $TN(T,S) = (T \cap S)^C$; voxels that are incorrectly labeled as foreground are the false positive set, $FP(T,S) = S \setminus T$; and voxels that are falsely labeled as background are the false negative set, $FN(T,S) = T \setminus S$. The latter two measures are also known as Type I and Type II errors, respectively. Several widely-used measures of set agreement can be formulated from the relative sizes of combinations of these regions (see **Table C.2.**). When multiple label classes are used, e.g., for our multi-label protocol, the source set S and target set T are each partitioned into R sets of structures, S_r and T_r, where $r \in \{1,2,\ldots,R\}$ is the index of a given structure. The set similarity measures typically have extensions to quantify multi-label set agreement. More details on multi-label evaluation appear in (Crum et al., 2005).

Specificity measures the fraction of voxels that are correctly rejected as not belonging to the target set. Sensitivity, or Target overlap (TO), measures the fraction of target voxels that are correctly detected; this measure can be extended when multiple labels are present. Source overlap (SO), or precision, measures what fraction of the voxels of S are in T; it penalizes for discrepancies in overlap but not in size. It also can be extended for multiple label sets. Overlap measures have been devised that weigh sensitivity and specificity. The Jaccard index (Jaccard, 1912) is specified as the size of the intersection of S and T divided by the size of their union and is thus a measure of union overlap (UO). Perfect agreement results in a score of 1; completely disjoint sets produce a score of 0. A decrease in sensitivity or specificity produces a decrease in the Jaccard index. This measure of union overlap can be extended to multiple sets (see **Table C.2.**). The mean overlap (MO) or Dice coefficient (Dice, 1945), is the size of the intersection divided by the mean size of the two regions, and may again be summed over multiple regions. Decreases in sensitivity or specificity will cause decreases in the Dice coefficient.

Set Distance Measures The above overlap measures do not explicitly account for boundary discrepancies between corresponding source and target regions. Thus, we will develop software to compute distance measures that assess the distance between the boundary points on the resampled source labels to the boundary points on the target labels. Given a pair of source and target regions, S_r and T_r respectively, we

identify for each of these the subset of boundary voxels, ∂S_r and ∂T_r, where a boundary voxel is identified by having an assigned label that is different from at least one of its neighbors. We compute a directed distance measure, DE, between these boundary regions as the average minimum distance from each source region boundary point in ∂S_r to the set of points making up the target region boundary ∂T_r:

$$DE(T_r, S_r) = \frac{1}{|\partial S_r|} \sum_{p \in \partial S_r} \min_{q \in \partial T_r} d(p,q)$$

where $d(p,q)$ is the Euclidean distance between points and $|\partial S_r|$ is the number of points in ∂S_r. We make this measure symmetric by averaging the distance error in both directions:

$$DE_2(T_r, S_r) = \frac{1}{2} DE(T_r, S_r) + \frac{1}{2} DE(S_r, T_r) = \frac{1}{2|\partial S_r|} \sum_{p \in \partial S_r} \min_{q \in \partial T_r} d(p,q) + \frac{1}{2|\partial T_r|} \sum_{p \in \partial T_r} \min_{q \in \partial S_r} d(p,q)$$

which produces one of the modified Hausdorff distance measures described by Dubuisson and Jain (1994). We note that we can also compute the variance in the minimum set distances as a measure of similarity.

C.2.c. Comparison with multi-rater data

The **LVE80** data will include 4 sets of rater-delineated labels for each MRI. This creates opportunities to compute validation measures to assess how an algorithm compares to a group of human raters. We will process the rater-labeled data to produce inter-rater scores for the measures described in **Sec. C.2. b**. This will be performed for each brain on the sets of labels produced by each pair of raters. For each measure, we will then compute the inter-rater mean and variance across the brains used in the evaluation study. These measures will establish a baseline for assessing new labeling results that are uploaded to the server.

To evaluate a new set of label data that is uploaded to the server, we will compute these measures for each pairing of the uploaded data with the 4 manually delineated label sets. We will then compute the mean and variance across the 4 sets and across the 40 brains in the test set. These statistics will provide us with a measure of how the computer algorithm compares to our expert delineations and will enable us to rank the uploaded data according to their agreement with the human raters. Assuming normality, we can apply an F-ratio test to determine if the algorithm results are distinguishable from the human raters. This will also enable us to test if the results of two algorithms are within the confidence limits of the human raters, and thus not distinguishable by this comparison.

In the case of image registration validation, we will be evaluating how well the rater label data that have been resampled from the space of a source brain to that of a target brain compare with the rater labels that were manually delineated on the target brain. There are multiple (16) possible pairings given data from 4 raters; we will compute each of these to enable comprehensive analysis of the algorithms being tested. We anticipate that users will be most interested in seeing how resampled data from the same rater compare.

STAPLE The Simultaneous Truth and Performance Level Estimation (STAPLE) method proposed by Warfield et al. (2004) produces an estimate of a ground truth labeling based on multiple labeled images. For each of the 80 brains in the **LVE80**, we will use STAPLE to create a basis image from the 4 sets of rater-delineated labels. For evaluating new segmentation results with the existing manual delineations, we will compute the set similarity and distance metrics (**Sec. C.2.b**) for the new results and the STAPLE ground truth. Summary statistics (mean, variance) will be computed across the test set of brains for each delineated region. For registration between source and target, the source STAPLE ground truth label data will be resampled from source to target space and compared against the target STAPLE ground truth label data.

C.2.d. Registration-specific Measures

Image Similarity: We will compute a series of image intensity similarity measures (e.g., mutual information, cross-correlation, and ratio image). As these measures often drive the image registration process, it will be important during the presentation of results (**Aim 3**) that this is made clear. While these evaluation metrics are not independent of the registration methods, they may still provide valuable insight into the functioning of a particular method. Definitions of these measures appear in many references (e.g., Hill et al., 2001).

Warp Field Distortion: An important aspect of image registration is the amount of deformation or distortion introduced by the algorithm. In extreme cases, images may be torn in order to push similar intensities into alignment. We will compute measures on the image warp transforms that are produced by the registration algorithms. The image transforms will be represented as 3D volumes of displacement vectors. To measure distortion, we will compute Jacobian maps from the displacement vector fields using a central difference scheme. By taking the determinant of the Jacobian matrices at each voxel we will produce a map of the image

distortion. This will allow us to detect and quantify instances where the determinant is non-positive, which indicates folding or tearing in the displacement vector field. We will also compute bending energy for each transformation as a global measure of image distortion.

C.3. Aim 3: Online Resource Framework

The intent of this proposal is to develop a validation system that is broadly available to the neuroimaging community. The data amd measures produced by **Aims 1** and **2** are the first steps to this. In **Aim 3**, we develop them into a publicly available online validation web server, termed the LONI Validation Engine Server (LVES). LVES will provide convenient access via the Internet to the validation process and archived results. The typical usage of someone performing a validation study is illustrated in **Fig. C3**.

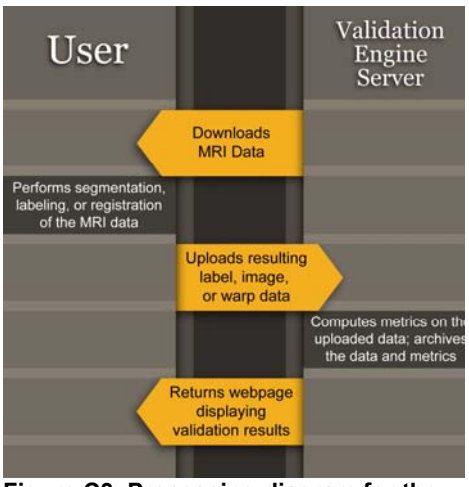

Figure C3. **Processing diagram for the web server**. The user downloads the MRI and related data, processes them, and uploads the resulting files to the server. The server then computes the validation measures, and returns a page detailing the results of the validation study.

Hosting The LVES will reside at the UCLA Laboratory of Neuro Imaging. The physical hardware will consist of a dedicated server with dual Intel Xeon processors (8 cores total) connected to an 11TB RAID storage system. This storage will be backed up by the existing LONI data archive system. The machine will use VMware vSphere Enterprise and the VMware ESX operating system. These will host virtual machines running RedHat Enterprise Edition, with Apache, MySQL, PHP, and Tomcat to provide the necessary web resources. The VMware product will allow us to migrate new versions of the website rapidly from a development environment to the production environment. This will also allow us to maintain snapshots of working versions of the server. In the event of a primary hardware failure, the webserver may be redeployed rapidly to other physical hardware hosting VMware servers provided by LONI. This will allow us to maintain a highly-available resource. We note that the LONI system administration team is experienced with this methodology. We have employed this strategy for our existing Segmentation Validation Engine site, as well as several other web servers provided by the Laboratory of Neuro Imaging.

Data Selection User registration will be required before site visitors may download or upload data. This will require that they agree to the terms of service for the site, which will provide basic guidelines to encourage fair use of the resources and the validation process. Registered users wishing to perform validation studies may then download the data defined in **Aim 1** from the LVES or, in the case of the IBSR data, from the NITRC website. For segmentation evaluation, there will be prepackaged downloadable files for each level of evaluation. For registration evaluation, the LVES will offer the user with a selection of choices. Some of the datasets for registration testing will use random draws from the possible registration pairings. We will develop server-side software that performs the selection of image registration pairings to provide varying levels of difficulty according to the error distribution rankings computed in **Aim 2**. When a user requests a randomized trial, the LVES will invoke the selection software and assign a unique validation study identifier to the resultant collection of brain pairings. These will be stored in a MySQL database and will allow the LVES to properly process the results files when the user returns to upload them. The unique ID will also allow users to repeat trials on the same registration pairs that were used previously. For both segmentation and registration validation, the user will be provided with instructions specifying the task to perform, the file format and naming conventions to follow, and the procedure for uploading data back to the LVES.

Data Processing and Archival After selecting the appropriate task and acquiring the data, the user will process the data using the method being evaluated. The user will then upload the results files to the website, with names following conventions described on the site to ensure the data are handled appropriately. The user will also be prompted to provide descriptions of the methods used to process the data, including any settings, preprocessing, or other relevant information that would affect the results. The webserver will first check to see if the uploaded data match the expected conventions. If not, an error message will be returned to the user. Assuming the data are properly named and formatted, processing will proceed. All uploaded data will be archived on the server. A sequence of evaluation steps will be invoked to compute the software to compute the measures described in **Aim 1**. These results will be archived in a MySQL database. The LVES will also invoke commands to produce graphical representations of errors, similar in spirit to the false positive and false negative maps shown in **Fig. C1**, and archive those on the server. The archiving of the uploaded data is an important component of making the server extensible and robust. In the event that we develop software to compute new performance measures, we can apply these retrospectively to the archived data.

Results Presentation Visitors to the site will be able to explore the validation results. Users who have uploaded tests and who are logged in will be able to navigate directly to their own results. **Figure C1** shows the existing SVE results interface. A table displays validation results from numerous algorithms that have been uploaded by several users. The table is generated on the fly from data stored in a MySQL database. The interface is interactive, and the user can re-order the results based on a particular measure. Individual results may be further explored using a drill-down approach. For example, maps of segmentation errors may be examined by selecting an individual result (see **Fig. C2**). Users can examine averages across subjects, averages across structures, or explore the individual numbers. This ability to drill down into the validation results data can provide algorithm developers with insights into where an algorithm may be succeeding or failing. We will extend this interface to provide access to the measures described in **Aim 1**. Users will be able to select a subset of algorithms of interest to be displayed. The interface will provide sort facilities, and the tables will be exportable into widely-used formats, e.g., comma-separated-value (CSV) files, that may be used for further analysis or other formats that are suitable for publication. We will develop functionality on the server to produce visual representations of segmentation and registration errors, similar to those for skull stripping shown in **Fig. C.2.** Users will be provided with functionality allowing them to group sets of methods, such as those that use fluid registration or those that segment a particular set of structures, and produce comparisons within these groups. As larger trial and test data sets are incorporated (e.g., the LPBA100, LVE80), users will also be able to view results stratified according to various demographics, such as the age of the subject at the time of MRI acquisition.

In addition to the tabular data, we will develop enhanced navigation using schematic representations of the brain. In these views, a representation of the brain will be shown in a flattened form, separated into anatomical areas. As the user hovers over a particular entry in the tabular display, that corresponding region of the schematic brain will be highlighted. A similar action will be performed when the user hovers over areas of the schematic brain. Traversing into a structure will update both the schematic and tabular views. This type of navigation can assist the user in sifting through the many metrics and many results in order to find the fine-grained detail for a particular region, a particular metric, or a particular algorithm.

An additional area of research will involve the development of software that analyzes the various metrics to find trends or interesting results. As a simple example, the software might compare an analyzed set of data to a set of benchmarks and find areas where the measures are significantly different. Curious anomalies in the data might be flagged or otherwise expanded in the results table to identify them to a user.

Community We will host a forum on the LVES website. This will allow the discussion of results, trouble-shooting of problems encountered with the server, or general discussions on segmentation and registration. Users will be able to add comments to their own results to indicate the method used or the intent of the evaluation.

C.4 Infrastructure

LONI has a long history of providing computational resources to the neuroimaging community. These computational resources are maintained by a staff of system administrators. Under the supervision of Dr. Arthur Toga, the founder and director of LONI, the system administration team will provide all necessary support for the computational infrastructure required by this proposal. The services offered by the web-based validation engine will be further augmented by the availability of the LONI Grid and Pipeline. These provide an infrastructure that may be used to perform large scale studies, e.g., the data processing required to process data for validation. Additionally, the LONI infrastructure will ensure that these validation resources continue to provide utility to the neuroimaging community after the funding period of the current proposal.

The personnel involved in this proposal are well-suited for carrying out this project. Dr. Shattuck has developed publicly released segmentation and visualization software (e.g., BrainSuite; Shattuck and Leahy, 2002), multi-subject atlases (Shattuck et al., 2008), and also produced an online validation resource for evaluating skull stripping algorithms (Shattuck et al., 2009). Dr. Klein has performed an extensive brain image registration evaluation (Klein et al., 2009) that provides a basis for registration evaluation. He also produced publicly available software for labeling of structures in brain MRI (MindBoggle; Klein et al., 2005). Drs. Toga and Thompson have led numerous brain registration and segmentation efforts. In addition to our expertise in segmentation and registration, the personnel can draw on significant professional design experience. Amanda Hammond, the project's web designer, leads a team of artists, web designers, and web developers at LONI who are available for discussion of design ideas and testing of the products in this proposal. She developed the look-and-feel of the current Segmentation Validation Engine site. Dr. Klein has previous experience as an Information Synthesis Theorist and Program Analyst at the Parsons Institute for Information Mapping.

C.5. Development Process

The resource developed within this proposal has the potential to serve the neuroimaging community for many years beyond the funding period. It may also have utility in other image analysis problems. The software development process we adopt will be essential to both the maintainability of the code and its ability to be adopted in other domains. We will follow several software engineering principles to ensure maintainability.

The source code for evaluating the volumetric data will be written in C++, which will provide high performance for rapid computation of results. For this code, we will adopt a basic set of style guidelines based on the Google C++ Style Guide (http://code.google.com/p/google-styleguide/) with some slight modifications based on our own local conventions. Web interface code will be written in PHP, JSP, Javascript, and Flash as appropriate for different sections of the server. Visualization components will make use of AJAX techniques to reduce visual clutter and to enhance navigation and drill down through the data. This will be implemented in a cross-platform and cross-browser compliant manner with testing on various major web browsers. We will make use of a model-view-controller architecture, which separates the software into data access, business logic, and user interface layers. This separation improves maintainability as each layer can be modified with fewer repercussions in the other layers than if they were tightly coupled. For example, if we wish to develop a new visualization interface that uses the same data as an existing one, we can make modifications that are confined to the view layer.

Source code be developed using Subversion revision control system (SVN) and documented using Doxygen (http://doxygen.org), which extracts documentation from inline comments in source code. We will produce additional documentation in the form of manuals and tutorials. Bug-tracking will be handled via LONI's Mantis server (http://www.mantisbt.org/). The validation engine source code will be released through the Laboratory of Neuro Imaging's software download website under an open source license. We will also maintain a presence on the Neuroimaging Informatics Tools and Resources Clearinghouse website (http://www.nitrc.org). This will allow us to have greater interaction with the neuroimaging community.

C.6. Schedule

The planned timeline for the four years of the proposed project is shown in **Fig. C.4**. We will have 3 major release phases to coincide with the completion under **Aim 1** of delineation of (1) the first 20 LVE80 images (2) the second 20 LVE80 images and (3) all LVE80 images. These will be completed in Years 2, 3, and 4. **Aim 2** will be heavily focused on developing evaluation software modules during the early part of the project; this focus will transition into applying these measures to the test data to establish baselines for comparison during the later phases. Under **Aim 3**, we will launch the initial version of the website in Year 2, with subsequent updates in Years 3 and 4 as more data become available.

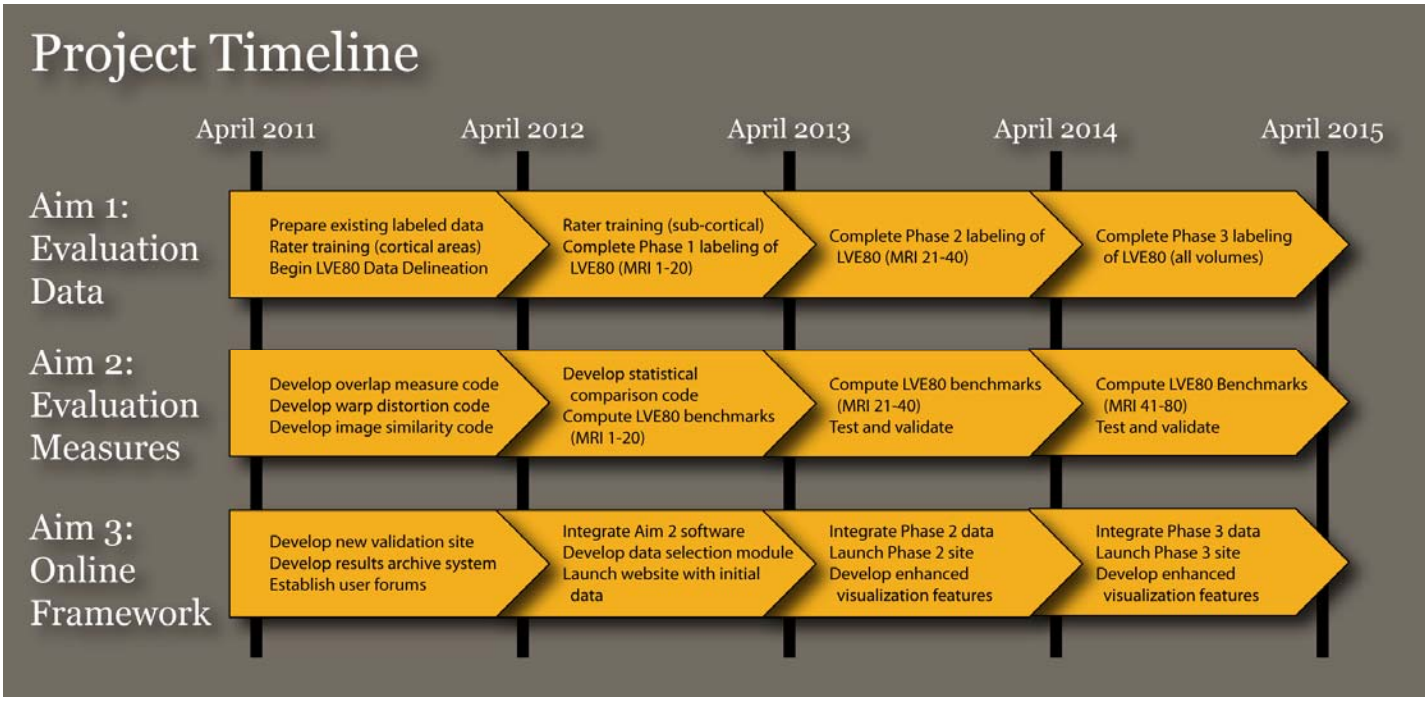

Figure C.4. Project Timeline.

References Cited

Ashburner J, Friston K (1999) Nonlinear spatial normalization using basis functions Hum Brain Map 7:254–66

Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang M-C, Christensen GE, Collins LD, Gee J, Hellier P, Song JH, Jenkinson M, Lepage C, Rueckert D, Thompson P, Vercauteren T, Woods RP, Mann JJ, Parsey RV (2009) Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration NeuroImage 46(3):786-802 PMCID: PMC2747506

URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2747506

Strother SC, Anderson JR, Xu XL, Liow JS, Bonar DC, Rottenberg DA (1994) Quantitative comparisons of image registration techniques based on high resolution MRI of the brain J Comput Assist Tomogr 18 (6), 954–962

West J, Fitzpatrick JM, Wang MY, Dawant BM, Maurer CR, Kessler RM, Maciunas RJ, Barillot C, Lemoine D, Collignon A, Maes F, Suetens P, Vandermeulen D, van den Elsen PA, Napel S, Sumanaweera TS, Harkness B, Hemler PF, Hill DL, Hawkes DJ, Studholme C, Maintz JB, Viergever MA, Malandain G, Woods RP (1997) Comparison and evaluation of retrospective intermodality brain image registration techniques J Comput Assist Tomogr 21 (4), 554–566

Boesen K, Rehm K, Schaper K, Stoltzner S, Woods R, Lüders E, Rottenberg D (2004) Quantitative comparison of four brain extraction algorithms NeuroImage 22(3):1255-61.

Hellier P, Barillot C, Corouge I, Gibaud B, Goualher GL, Collins L, Evans A, Malandain G, Ayache N (2001A) Retrospective evaluation of inter-subject brain registration Med Image Comput Comput-Assist Interv: MICCAI 2001 LNCS2208:258–266

Hellier P, Ashburner J, Corouge I, Barillot C, Friston K (2002) Inter subject registration of functional and anatomical data using SPM Med Image Comput Comput-Assist Interv: MICCAI 2002 LNCS2489:590–597

Hellier P, Barillot C, Corouge I, Gibaud B, Goualher GL, Collins DL, Evans A, Malandain G, Ayache N, Christensen GE, Johnson HJ (2003) Retrospective evaluation of intersubject brain registration IEEE Trans Med Imag 22:1120–1130

Fennema-Notestine C, Ozyurt IB, Clark CP, Morris S, Bischoff-Grethe A, Bondi MW, Jernigan TL, Fischl B, Segonne F, Shattuck DW, Leahy RM, Rex DE, Toga AW, Zou KH, and Brown GG (2006) Quantitative evaluation of automated skullstripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction, and slice location Hum Brain Mapp 27 (2), 99–113 PMCID: PMC2408865

Yassa MA, Stark CE (2009) A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe NeuroImage 44(2):319-27

Collins DL, Zijdenbos AP, Kollokian V, Sled JG, Kabani NJ, Holmes CJ, Evans AC (1998) Design and construction of a realistic digital brain phantom IEEE Trans Med Imag 17(3):463–468

Aubert-Broche B, Griffin M, Pike GB, Evans AC, Collins DL (2006) Twenty new digital brain phantoms for creation of validation image data bases IEEE Trans Med Imag 25(11):1410–1416

Heimann T, van Ginneken B, Styner MA, Arzhaeva Y, Aurich V, Bauer C, Beck A, Becker C, Beichel R, Bekes G, Bello F, Binnig G, Bischof H, Bornik A, Cashman PM, Chi Y, Cordova A, Dawant BM, Fidrich M, Furst JD, Furukawa D, Grenacher L, Hornegger J, Kainmüller D, Kitney RI, Kobatake H, Lamecker H, Lange T, Lee J, Lennon B, Li R, Li S, Meinzer HP, Nemeth G, Raicu DS, Rau AM, van Rikxoort EM, Rousson M, Rusko L, Saddi KA, Schmidt G, Seghers D, Shimizu A, Slagmolen P, Sorantin E, Soza G, Susomboon R, Waite JM, Wimmer A, Wolf I. (2009) Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging. 28(8):1251-65.

Gerig G, Jomier M, Chakos M (2001) Valmet: a new validation tool for assessing and improving 3D object segmentation Med Image Comput Comput-Assist Interv: MICCAI 2001 LNCS2208:516–524

Shattuck DW, Mirza M, Adisetiyo V, Hojatkashani C, Salamon G, Narr KL, Poldrack RA, Bilder RM, Toga AW (2008) Construction of a 3D probabilistic atlas of human cortical structures NeuroImage 39:1064–80 PMCID: PMC2757616 URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2757616

Shattuck DW, Leahy RM (2002) BrainSuite: an automated cortical surface identification tool Med Image Anal 6(1):129–42

Tu Z, Narr K, Dollar P, Dinov I, Thompson P, Toga A (2008) Brain Anatomical Structure Segmentation by Hybrid Discriminative/Generative Models IEEE Tran on Medical Imaging 27(4):495-508

Woods RP, Grafton ST, Watson JDG, Sicotte NL, Mazziotta JC (1998) Automated image registration: II Intersubject validation of linear and nonlinear models J Comput Assist Tomogr 22:153–65

Collins, DL, Holmes, CJ, Peters TM, Evans AC (1995) Automatic 3-D model-based neuroanatomical segmentation Hum Brain Mapp 3(3):190–208

Ardekani B A, Guckemus S, Bachman A, Hoptman MJ, Wojtaszek M , Nierenberg J (2005) Quantitative comparison of algorithms for inter-subject registration of 3D volumetric brain MRI scans J Neurosci Methods 142:67–76

Vercauteren, T, Pennec, X, Perchant, A, Ayache, N, 2007 Non-parametric diffeomorphic image registration with the demons algorithm Med Image Comput Comput-Assist Interv: MICCAI 2007 4792, 319–326

Andersson J, Smith S, Jenkinson M (2008) FNIRT — FMRIB's non-linear image registration tool 13th Annual Meeting of the Organization for Human Brain Mapping Melbourne, Australia, Poster #496

Rueckert, D, Aljabar, P, Heckemann, RA, Hajnal, JV, Hammers, A (2006) Diffeomorphic registration using B-splines 9th International Conference on Medical Image Computing and Computer-Assisted Intervention: MICCAI 2006 LNCS4191:702–709

Chiang M-C, Dutton RA, Hayashi KM, Lopez OL, Aizenstein HJ, Toga AW, Becker JT, Thompson PM (2007) 3D pattern of brain atrophy in HIV/AIDS visualized using tensor-based morphometry NeuroImage 34:44–60

Hellier P, Barillot C, Mémin E, Pérez P (2001B) Hierarchical estimation of a dense deformation field for 3D robust registration IEEE Trans Med Imag 20:388–402

Christensen GE (1999) Consistent linear-elastic transformations for image matching Proceedings of Information Processing in Medical Imaging: IPMI 99 LNCA1613:224–237

Christensen G, Johnson H (2001) Consistent image registration IEEE Trans Med Imag 20:568–582

Johnson H, Christensen G (2002) Consistent landmark and intensity-based image registration IEEE Trans Med Imag 21:450–61

Avants BB, Epstein CL, Grossman M, Gee JC (2008) Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain Med Image Anal 12, 26–41

Friston KJ, Ashburner J, Poline JB, Frith CD, Heather JD, Frackowiak RSJ (1995) Spatial registration and normalization of images Hum Brain Map 2:165–189

Ashburner J, Friston KJ (2005) Unified segmentation NeuroImage 26(3):839–851

Ashburner J (2007) A fast diffeomorphic image registration algorithm NeuroImage 38:95–113

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens T EJ, Johansen-Berg H, Bannister PR, Luca MD, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, Stefano ND, Brady JM, Matthews PM, (2004) Advances in functional and structural MR image analysis and implementation as FSL NeuroImage 23:S208–19

Klein A, Ghosh SS, Avants B, Yeo BT, Fischl B, Ardekani B, Gee JC, Mann JJ, Parsey RV (2010) Evaluation of volume-based and surface-based brain image registration methods Neuroimage 51(1):214-20. PMCID: PMC2862732 [Available on 2011/5/15]

Fischl B, Sereno MI, Dale AM (1999) Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system NeuroImage 9(2):195-207

Yeo BT, Sabuncu MR, Vercauteren T, Ayache N, Fischl B, Golland P (2010) Spherical demons: fast diffeomorphic landmark-free surface registration. IEEE Trans Med Imaging 29(3):650-68.

Shattuck DW, Prasad G, Mirza M, Narr KL, Toga AW (2009) Online Resource for Validation of Brain Segmentation Algorithms NeuroImage 45(2):431-9 PMCID: PMC2757629

URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2757629

Smith SM (2002) Fast robust automated brain extraction Hum Brain Map 17(3):143–55

Ségonne F, Dale AM, Busa E, Glessner M, Salat D, Hahn HK, Fischl B (2004) A hybrid approach to the skull stripping problem in MRI NeuroImage 22(3):1060–75

Shattuck DW, Sandor-Leahy SR, Schaper KA, Rottenberg DA, Leahy RM (2001) Magnetic resonance image tissue classification using a partial volume model NeuroImage 13:856–76

Crum WR, Camara O, Rueckert D, Bhatia KK, Jenkinson M, Hill DLG (2005) Generalised overlap measures for assessment of pairwise and groupwise image registration and segmentation Medical Image Computing and Computer-Assisted Intervention: MICCAI 2005 LNCS3749:99–106

Jaccard P (1912) The distribution of flora in the alpine zone New Phytol 11:37–50

Dice LR (1945) Measures of the amount of ecologic association between species Ecology 26:297-302

Dubuisson M-P, Jain A (1994) A modified hausdorff distance for object matching In: Pattern recognition, Vol 1, Conference A: Proceedings of the 12th IAPR international conference on Computer Vision & Image Processing 1:566–8

Warfield SK, Zou KH, Wells WM (2004) Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation IEEE Trans Med Imaging 23(7):903-21. PMCID: PMC1283110

Hill DL, Batchelor PG, Holden M, Hawkes DJ (2001) Medical image registration Phys Med Biol 46(3):R1-45

Crespo-Facorro B, Kim J, Andreasen N C, Spinks R, O'Leary D S, Bockholt H J, Harris G & Magnotta V A (2000) Cerebral cortex: a topographic segmentation method using magnetic resonance imaging Psychiatry Research 100(2):97-126

Kim JJ, Crespo-Facorro B, Andreasen NC, O'Leary DS, Zhang B, Harris G, Magnotta VA (2000) An MRI-based parcellation method for the temporal lobe NeuroImage11(4), 271-88.

Caviness V S, Meyer J, Makris N, Kennedy D (1996) MRI-based topographic parcellation of human neocortex: an anatomically specified method with estimate of reliability Journal of Cognitive Neuroscience 8(6):566-587

Desikan RS, Segonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS & Killiany, RJ (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest NeuroImage 31(3): 968-980.

Carper R Courchesne E (2000). Inverse correlation between frontal lobe and cerebellum sizes in children with autism. Brain, 123:836-844.

Carper R, Courchesne E (2005). Localized enlargement of the frontal cortex in early autism. Biological Psychiatry 57(2):126-133.

Carper, R., Moses, P. Tigue, Z. D. & Courchesne, E. (2002). Cerebral lobes in autism: early hyperplasia and abnormal age effects. NeuroImage 16(4):1038-1051.

Tourville J, Guenther F (2003) A cortical and cerebellar parcellation system for speech studies Boston University Technical Reports CAS/CNS-03-022

McCarley R, Shenton M (2008). Structural MRI laboratory manual: MR image acquisition and image processing tools and neuroanatomical regions of interest, Harvard University.

Buchanan R, Francis A, Arango C, Miller K, Lefkowitz D, McMahon R, Barta P, Pearlson, G (2004) Morphometric assessment of the heteromodal association cortex in schizophrenia American Journal of Psychiatry 161:322-331.

Zhou S. Suzuki M. Takahashi T. Hagino H. Kawasaki Y. Matsui M. S. H & Kurachi M. (2007) Parietal lobe volume deficits in schizophrenia spectrum disorders Schizophr Research 89(1-3) 35-48.

Ono M, Kubik S, Abernathy, C D (1990) Atlas of the cerebral sulci Stuttgart: Georg Thieme Verlag.

Duvernoy, H. M. (1999). The human brain: Surface, blood supply, and three dimensional anatomy. New York: Springer-Verlag Wien.

Mai J, Paxinos G, Voss T (2008) Atlas of the human brain New York: Elsevier.

Klein A, Mensh B, Ghosh S, Tourville J, Hirsch J (2005) Mindboggle: automated brain labeling with multiple atlases BMC Med Imaging 5(7)