NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 29 (2006) 185 – 202

Impact of acquisition protocols and processing streams on tissue segmentation of T1 weighted MR images

Kristi A. Clark,[a,*] Roger P. Woods,[a] David A. Rottenberg,[b] Arthur W. Toga,[c] and John C. Mazziotta[a,d]

[a] *Ahmanson-Lovelace Brain Mapping Center, Department of Neurology, David Geffen School of Medicine, University of California-Los Angeles, 660 Charles E. Young Drive South, Los Angeles, CA 90095, USA*
[b] *Departments of Radiology and Neurology, Minneapolis VA Medical Center, University of Minnesota, Minneapolis, MN 55417, USA*
[c] *Laboratory of Neuro Imaging, Department of Neurology, Division of Brain Mapping, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA*
[d] *Departments of Pharmacology and Radiological Sciences, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA*

Received 4 March 2005; revised 22 June 2005; accepted 11 July 2005
Available online 31 August 2005

The segmentation of T1-weighted images into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) is a fundamental processing step in neuroimaging, the results of which affect many other structural imaging analyses. Variability in the segmentation process can decrease the power of a study to detect anatomical differences, and minimizing such variability can lead to more robust results. This paper outlines a straightforward strategy that can be used (1) to select more optimal data acquisition and processing protocols and (2) to quantify the impact of such optimization. Using this approach with multiple scans of a single subject, we found that the choice of a segmentation algorithm had the largest impact on variability, while the choice of a pulse sequence had the second largest impact. The data indicate that the classification of GM is the most variable, and that the optimal protocol may differ across tissue types. Therefore, the intended use of segmentation data should play a role in optimization. Examples are provided to demonstrate that the minimization of variability is not sufficient for optimization; the overall accuracy of the approach must also be considered. Simple volumetric computations are included to illustrate the potential gain of optimization; these results show that volume estimates from optimal pathways were on average three times less variable than estimates from suboptimal pathways. Therefore, the simple strategy illustrated here can be applied to many studies to optimize tissue segmentation, which should lead to a net increase in the power of structural neuroimaging studies.

Keywords: Structural neuroimaging; MRI; Reliability; Validation; Segmentation; Tissue classification

Introduction

Longitudinal and multi-center structural neuroimaging studies have more power than smaller studies to conduct sophisticated studies of basic neuroanatomy and clinical disorders (Mazziotta et al., 1995; Giedd et al., 1996; Coffey et al., 2001; Courchesne et al., 2001; Hulshoff Pol et al., 2002; Salat et al., 2004). During the course of such studies, practical methodological issues often arise because these studies rely on the combination of data from different scanners or upgrades of the same scanner. While designing large-scale studies, investigators often need to decide on a pulse sequence and parameters, as well as the specific data processing protocol to use. Ideally these decisions would be made by acquiring and analyzing data from a “gold standard” that has the same complexities as a living human brain but with properties that are explicitly and thoroughly known. Although there is no such gold standard, several alternative strategies exist for the validation of a given data acquisition or analysis strategy: performance on a phantom, comparison to manual delineation, various data simulation techniques, or simple visual inspection. These methods are generally used to assess the accuracy of a technique; however, for a multi-center or longitudinal study, the reproducibility of a given result is equally important. The primary objective of the current study is to devise a strategy that can be used to evaluate the reproducibility of a data acquisition and/or analysis protocol.

The segmentation of a T1-weighted image into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) can be thought of as a signal detection problem, where the signal is the MR intensity of each tissue type. There are many sources of noise, such as partial voluming effects and noise due to the scanning environment. In a previous study, Holmes et al. showed that the process of averaging multiple scans of a single subject

[*] Corresponding author. Fax: +1 310 794 7406.
E-mail address: kaclark@ucla.edu (K.A. Clark).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2005.07.035

results in a higher signal-to-noise-ratio (SNR), as well as an increased contrast between gray matter and white matter (Holmes et al., 1998). In the current study, twenty images of one subject have been spatially aligned and averaged to create a "gold standard" image that has a high SNR. This gold standard brain can be used to quantify the reliability of an analysis technique, such as tissue segmentation, by comparing the final segmentation maps of any individual scan to that of the gold standard brain. The individual scans are all from the same subject and are all spatially aligned; therefore, any differences between an individual segmentation map and that of the gold standard can be attributed to the lower SNR and/or contrast that is present in the individual scans. Thus, if one technique reliably yields individual segmentation maps that are more similar to the gold standard segmentation map, then the results from that technique are more likely to be reproducible. In practice, investigators can collect multiple scans of one subject at the beginning of a large-scale study and use this strategy to choose an optimal data acquisition and/or data processing protocol guided by reliability. For example, if two pulse sequences are being considered for use in a large-scale study of GM volumes, this strategy can be used to determine if one pulse sequence yields GM volumes that are less sensitive to noise than a second pulse sequence. Similarly, this strategy can be used to measure how much the classification of any particular tissue type (e.g., GM) depends on any processing step (e.g., skull stripping).

While the focus of this study was reliability, it is important not to blindly choose the most reliable protocol without also considering the accuracy of that pathway. An analysis pathway can be reliable by consistently making mistakes, e.g., consistently classifying the thalamus as WM. At the very least, the most reliable pathways must be visually inspected for accuracy; ideally, the accuracy should be evaluated quantitatively.

Methods and materials

Experiment overview

A global overview of the data acquisition and analysis strategy is shown in Fig. 1. Twenty T1-weighted scans (10 MPRAGEs and 10 SPGRs) were collected from one subject and registered into a common space. A gold standard image was created by averaging the individual scans in order to achieve the best SNR (Figs. 1A and 2). Each image volume (individual images and the gold standard) was then processed in 3 steps: (1) noise reduction (e.g., correction for magnetic field inhomogeneities); (2) skull stripping; and (3) segmentation into GM, WM, and CSF. Several publicly available published algorithms were used for each processing step. Some of the software packages were also implemented using multiple values for one or more parameters (Fig. 1B). The segmentation map of each individual scan from each unique acquisition/analysis pathway was then compared to the segmentation map of the identically processed gold standard image. The comparison of an individual to the gold standard was done using a d-prime analysis: d-prime = z score (hit rate) − z score (false alarm rate), where the hit rate and false alarm rates are treated as probabilities and the z scores are computed from a standard normal distribution with unit variance. The hit rate is defined as: (hits) / (hits + misses); the false alarm rate is defined as: (false alarms) / (false alarms + correct rejections) (Fig. 3). The use of d-prime in this context derives from signal detection theory (Green and Sweets, 1966). One benefit of this type of analysis is that the robustness of the segmentation of an image can be reduced to three numbers: one each for GM, WM, and CSF. The resulting d-primes were then used in two ways: (1) to identify the optimal acquisition and analysis protocol tested (Figs. 4–6); and (2) to evaluate the relative impact of each acquisition/analysis step on the final segmentation map (Figs. 7 and 8). Finally, an example is provided to show how these differences in segmentation can affect volumetric analyses (Figs. 9 and 10). For this example, the following regions were identified and the gray matter volume was computed: cerebellum, frontal lobe, parietal lobe, temporal lobe, and occipital lobe. The dependence of the measurement error of the GM volume of each lobe on the GM d-prime was computed.

Data collection

One healthy normal volunteer (male, age 43) was scanned a total of 20 times during two separate sessions using a 1.5-T Siemens Sonata scanner in accordance with the rules and policies of the UCLA Institutional Review Board. Each session consisted of five scans each of MPRAGE (turboFLASH) and SPGR (gradient echo) pulse sequences in an alternating sequence. The data were collected in an interleaved fashion in order to minimize any time-varying confounds (e.g., movement, scanner drift). During the first session, the magnet was shimmed once at the beginning of the session; in the second session, the magnet was re-shimmed between each scan. The MPRAGE scans were collected sagittally with a 256 × 256 × 160 matrix, with 1.0 mm^3 resolution (TI/TE/TR/FA = 1100/4.38/1900/15°) (Fig. 2A). The SPGRs were also collected sagittally with a 256 × 256 × 160 matrix and 1.0 mm^3 resolution (TE/TR/FA = 9.2/22/30°) (Fig. 2B). The MPRAGEs were collected using 2 averages, yielding a total scan time of 16:16 min; the SPGRs were collected using 1 average, yielding a total scan time of 15:00 min. The averaging strategy of the MPRAGE was used in order to make the total scan time as similar as possible across the two different pulse sequences.

Alignment of scans

Each of the 20 scans was aligned to all of the other scans scan using a 6-parameter, rigid-body registration (AIR 5.2.5; Woods et al., 1998a). Any discrepancies in the pairwise registrations were reconciled to create one transformation matrix for each pair of images (Woods, 2003; Woods et al., 1998a); for example, the final transformation matrix for scan 1 to scan 2 used all of the information from registering scan 1 to scan 3, scan 3 to scan 2, etc. All of the transformation matrices were used to define a "common, average" space, into which all 20 of the images were resliced using a 3D scanline chirp-z interpolation model (Woods et al., 1998a). Then each image was intersected with a mask of the common coverage, so that each final image only contained data in voxels where every image contained data. Finally, all 20 images were averaged together to create a gold standard brain (Fig. 2C). The goal of creating the gold standard brain in this way was to avoid introducing a bias in the average brain towards either pulse sequence, and also to have one common gold standard that could be used for all of the images. Through visual inspection, it was verified that there were no non-linear distortions between the two pulse sequences; this was verified

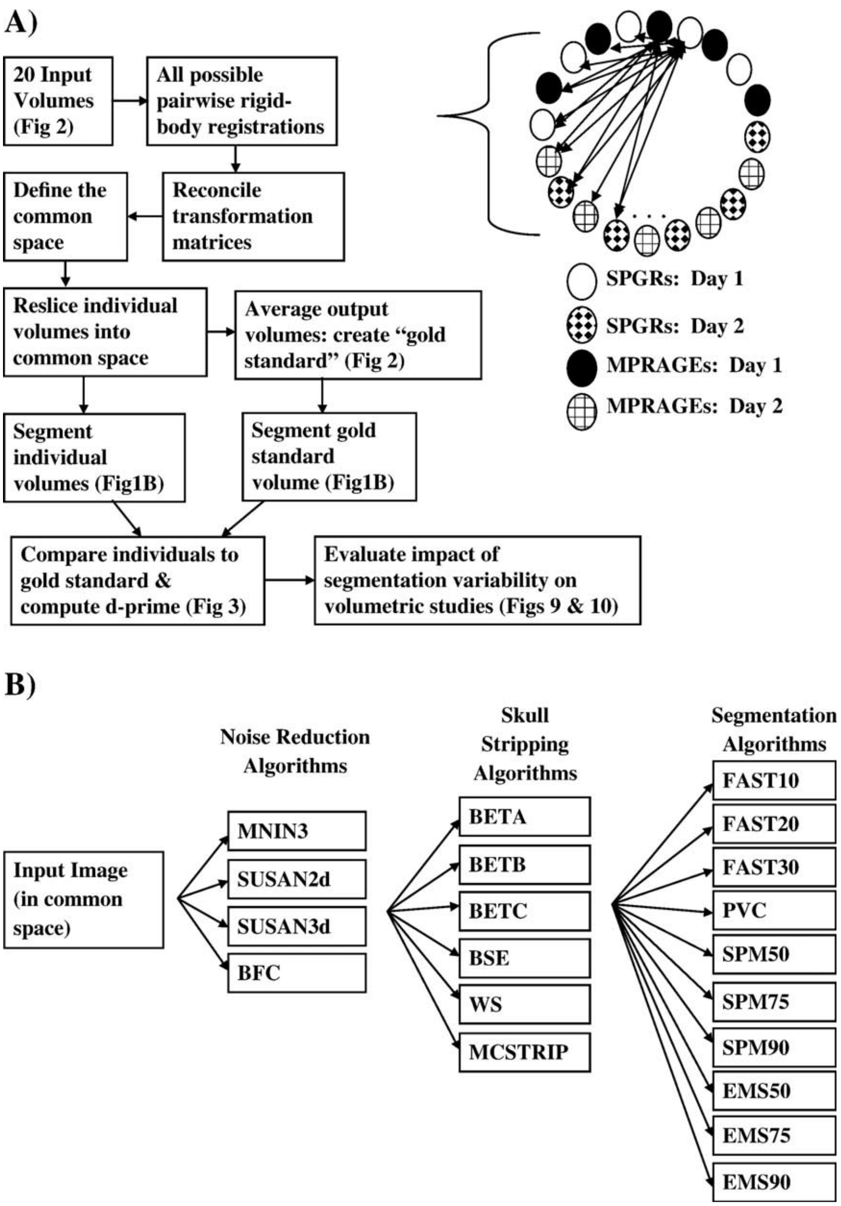
A)
20 Input Volumes (Fig 2)
All possible pairwise rigid-body registrations
Reconcile transformation matrices
Define the common space
Reslice individual volumes into common space
Average output volumes: create "gold standard" (Fig 2)
Segment individual volumes (Fig1B)
Segment gold standard volume (Fig1B)
Compare individuals to gold standard & compute d-prime (Fig 3)
Evaluate impact of segmentation variability on volumetric studies (Figs 9 & 10)
SPGRs: Day 1
SPGRs: Day 2
MPRAGEs: Day 1
MPRAGEs: Day 2
B)
Input Image (in common space)
Noise Reduction Algorithms
MNIN3
SUSAN2d
SUSAN3d
BFC
Skull Stripping Algorithms
BETA
BETB
BETC
BSE
WS
MCSTRIP
Segmentation Algorithms
FAST10
FAST20
FAST30
PVC
SPM50
SPM75
SPM90
EMS50
EMS75
EMS90

Fig. 1. Experimental Overview. (A) 20 inputs volumes were collected from one subject: 10 MPRAGEs and 10 SPGRs. All possible pairwise, rigid-body registrations of the 20 input volumes were computed and reconciled to define a common space into which each brain was resliced, and the outputs were averaged to create a "gold standard" image with high SNR. The individual volumes and the gold standard were all segmented in several ways (B), and the results were compared by computing a d-prime (Fig. 3). The d-primes were used in three ways: (1) to choose an optimal processing sequence (Figs. 4–6), (2) to evaluate the impact of each acquisition/analysis step on the final segmentation result (Fig. 8), and (3) to correlate variability in the segmentation process with variability in volumetric measurements (Figs. 9 and 10). (B) The data processing protocols are exemplified for one input image (the same protocols were used for each of the 10 MPRAGEs, the 10 SPGRs, and the gold standard volume). Each volume was processed with 3 noise reduction algorithms (1 parameter set for MNIN3, 2 parameter sets for SUSAN, 1 parameter set for BFC); 4 skull-stripping algorithms (3 parameter sets for BET, 1 each for BSE, WS, MCSTRIP); and 4 segmentation algorithms (3 parameter sets for FAST, EMS, SPM, and 1 for PVC). All of the possible combinations of algorithms were used (see text for details).

by creating averages of all of the SPGRs and all of the MPRAGEs separately and comparing those averages to the gold standard average of 20 images.

Noise reduction

Intensity variations due to hardware, such as radio frequency (RF) coil non-uniformities, add noise to structural images that can adversely affect the performance of downstream processing, such as skull stripping and segmentation (Simmons et al., 1994; Cohen et al., 2000; Jezzard, 2000). Three noise reduction algorithms were evaluated in this study. The non-parametric non-uniform intensity normalization algorithm (MNIN3; Sled et al., 1998) models the intensity non-uniformity as a smooth multiplicative field. The second algorithm examined was bias field correction (BFC; Shattuck et al., 2001), which also models the intensity non-uniformities as a multiplicative bias field but computes these models on local neighborhoods. It should be

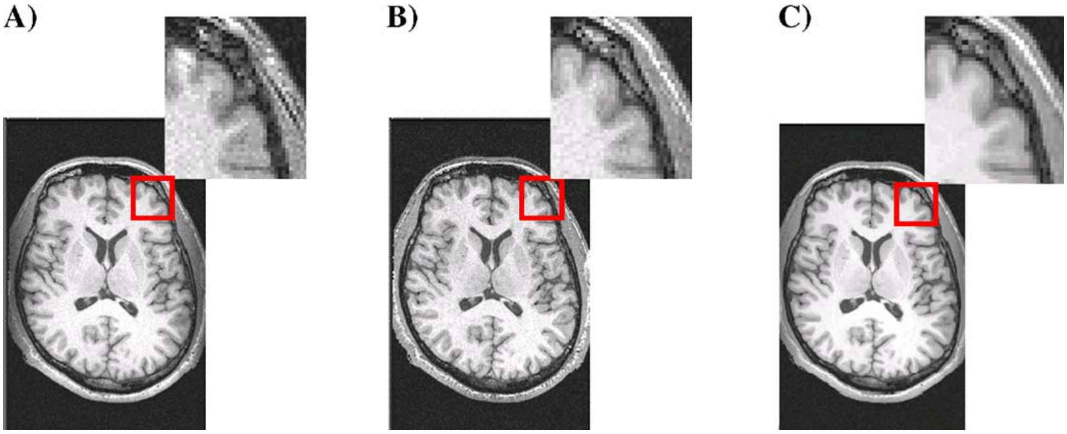

Fig. 2. Data Acquisition. (A) Sample MPRAGE image. (B) Sample SPGR image. (C) Gold Standard Image (average of 10 MPRAGEs and 10 SPGRs) with high SNR.

noted that BFC is designed to be used after skull stripping the data, an issue which is addressed in more detail in the Results section. A third algorithm, smallest univalue segment assimilating nucleus (SUSAN; Smith and Brady, 1997), uses a strategy of feature detection (edges and corners) to choose a local smoothing neighborhood to reduce the high-frequency noise present in the images.

MNIN3 (Sled et al., 1998) was applied using the stopping criterion of 0.0001 and a FWHM value of 0.04. BFC (Shattuck et al., 2001) was applied using the default parameters. The third algorithm, SUSAN (Smith and Brady, 1997), was used both in 2D mode and 3D mode (SUSAN2d and SUSAN3d) with a threshold value of 10.

Skull stripping

For almost all data analysis procedures, the skull must be removed from the image. Several methods exist for removing the skull; only automated methods were included in this study. The first algorithm examined was the brain extraction tool (BET; Smith, 2002), which uses a surface model approach that starts by finding the center of gravity and tessellating the surface using

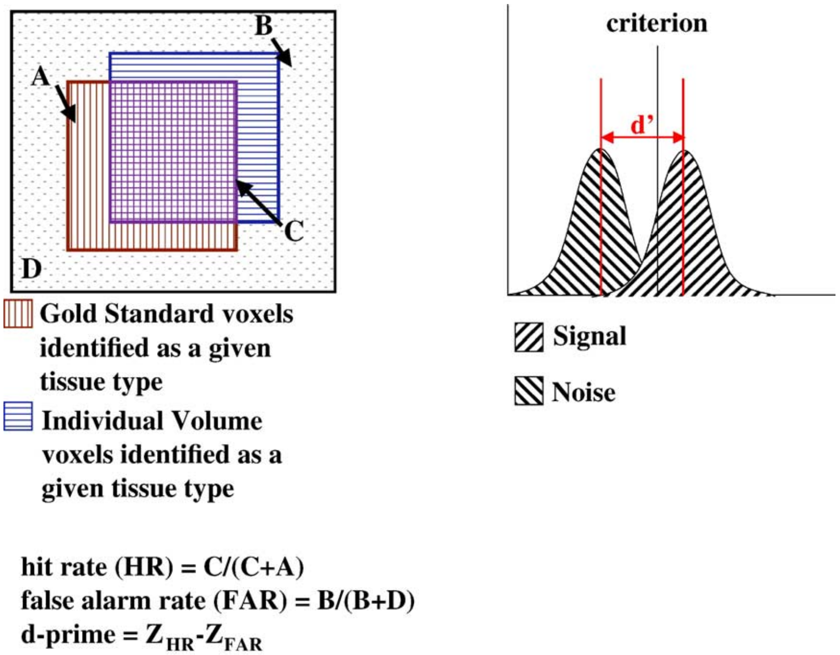
B
A
C
D
Gold Standard voxels identified as a given tissue type
Individual Volume voxels identified as a given tissue type
hit rate (HR) = C/(C+A)
false alarm rate (FAR) = B/(B+D)
d-prime = Z_{HR}-Z_{FAR}
criterion
d'
Signal
Noise

Fig. 3. Calculation of D-prime. The GM, WM, and CSF maps of each individual image are compared to the identically processed maps of the gold standard image. For each tissue type, every voxel in the image is labeled either A, B, C, or D where: A is **Miss** (i.e., classified GM in the gold standard, not-GM in the individual), B is **False alarm** (e.g., classified not-GM in the gold standard, GM in the individual), C is **Hit** (classified GM in both the gold standard and the individual), and D is **Correct Rejection** (e.g., classified as non-GM in both the gold standard and the individual). The same process is repeated for each tissue type. For every volume and every possible processing pathway, 3 d-primes are computed (one for each tissue type). The d-prime is defined as the difference between the *z* score of the false alarm rate from the *z* score of the hit rate. The hit rate is defined as: hits/(hits + misses), and the false alarm rate is defined as: false alarms/(false alarms + correct rejections).

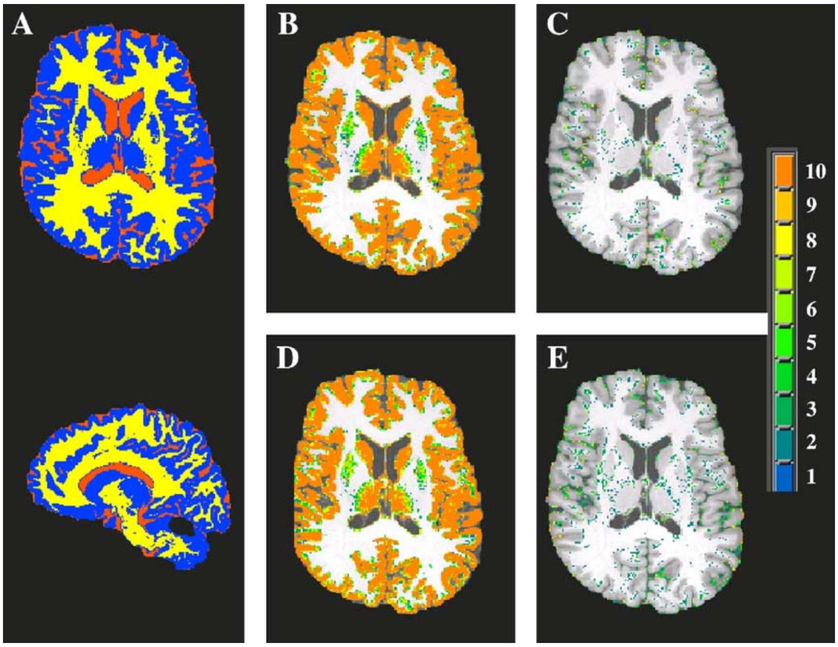

Fig. 4. Maximum d-prime for GM: comparison of MPRAGE and SPGR. (A) Tissue classification of the gold standard image segmented with the processing pathway associated with the maximum d-prime for GM: MNIN3, BSE, SPM50. Blue voxels were classified as GM, yellow voxels WM, and orange voxels CSF. For this pathway, the MPRAGE pulse sequence led a higher d-prime than the SPGR pulse sequence ($P < 0.001$). (B) Hits: MPRAGE. (C) False alarms: MPRAGE. (D) Hits: SPGR. (E) False alarms: SPGR. In panels B–E, the color of each voxel (see side scale) codes the number of images in which the voxel was labeled as hit (for panels B and D) or false alarm (for panels C and E).

connected triangles. A second approach, brain surface extraction (BSE; Sandor and Leahy, 1997; Shattuck et al., 2001), is primarily based on an edge detection approach combined with morphological procedures. The Minneapolis consensus strip (McStrip; Rehm et al., 2004) is a hybrid algorithm that involves warping to a template, intensity thresholding, and edge detection. The last algorithm tested uses a hybrid approach that combines watershed algorithms and deformable surface models (WS; Segonne et al., 2004).

BET (Smith, 2002) was applied with three different threshold gradients, which were optimized for each pulse sequence

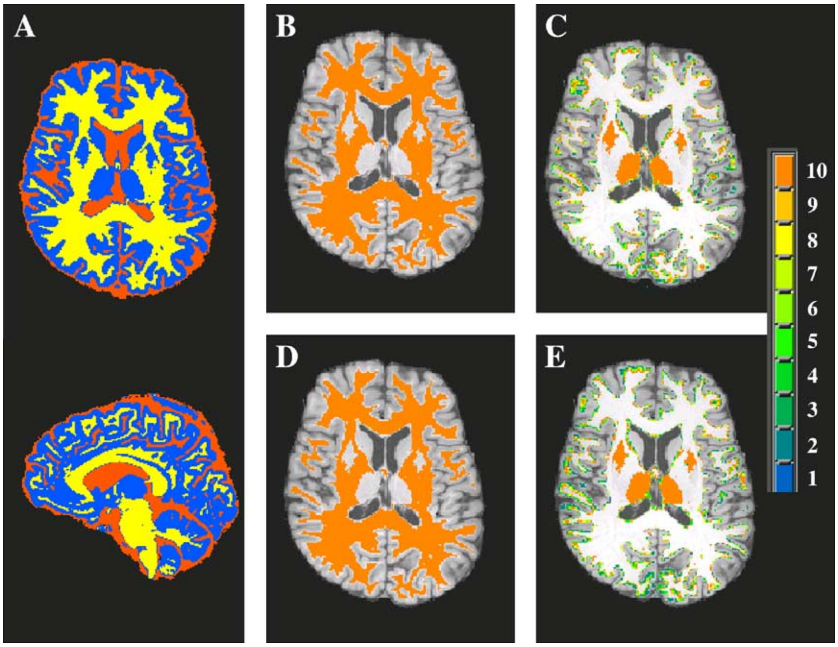

Fig. 5. Maximum d-prime for WM: comparison of MPRAGE and SPGR. (A) Tissue classification of the gold standard image segmented with the processing pathway associated with the maximum d-prime for WM: BFC, WS, FAST10. Blue voxels were classified as GM, yellow voxels WM, and orange voxels CSF. For this pathway, the MPRAGE pulse sequence led a higher d-prime than the SPGR pulse sequence ($P < 0.04$). (B) Hits: MPRAGE. (C) False alarms: MPRAGE. (D) Hits: SPGR. (E) False alarms: SPGR. In panels B–E, the color of each voxel (see side scale) codes the number of images in which the voxel was labeled as hit (for panels B and D) or false alarm (for panels C and E).

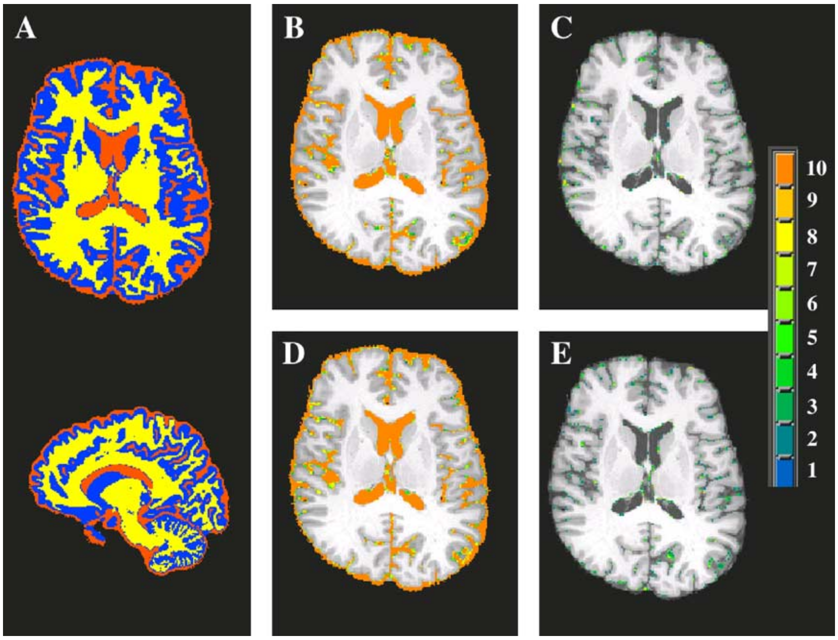

Fig. 6. Maximum d-prime for CSF: comparison of MPRAGE and SPGR. (A) Tissue classification of the gold standard image segmented with the processing pathway associated with the maximum d-prime for CSF: MNIN3, McStrip, FAST30. Blue voxels were classified as GM, yellow voxels WM, and orange voxels CSF. For this pathway, the MPRAGE pulse sequence led a higher d-prime than the SPGR pulse sequence ($P < 0.001$). (B) Hits: MPRAGE. (C) False alarms: MPRAGE. (D) Hits: SPGR. (E) False alarms: SPGR. In panels B–E, the color of each voxel (see side scale) codes the number of images in which the voxel was labeled as hit (for panels B and D) or false alarm (for panels C and E).

separately (SPGR: −0.03, −0.05, −0.07; and MPRAGE: −0.18, −0.2, −0.22; BETA, BETB, BETC, respectively). BSE (Sandor and Leahy, 1997; Shattuck et al., 2001) was applied with a value of 0.67 for the edge scale. The parameters for the McStrip meta-algorithm (Rehm et al., 2004) were optimized using the gold standard average of all 20 scans. The WS algorithm (Segonne et al., 2004) was applied to the SPGRs using a preflooding height of 41, and for the MPRAGEs the preflooding height was set to 15.

Interactions with skull stripping

It is likely that there are interactions between processing steps, e.g., the impact of a segmentation algorithm is affected by the choice of a skull-stripping algorithm. Therefore, a parallel set of analyses was performed to remove the scan-to-scan variability introduced by the skull-stripping algorithms. For these parallel analyses, each skull-stripping algorithm was run on the gold standard brain only, and then the resulting brain/background mask was applied to each individual scan. Thus, the d-primes from these parallel analyses do not reflect different amounts of non-brain tissue in the final images. It is possible that there are interactions between other steps in the processing stream (e.g., noise reduction interactions with segmentation algorithms), but there are no comparable methods for removing these effects. Therefore, in this paper, an analysis titled “skull strip individual scans” refers to analyses where each individual scan was skull stripped. Alternatively, “skull strip gold standard only” refers to analyses where only the gold standard brain was skull stripped and the brain/background mask was applied to each individual scan. Most of the interpretation of results, including the applications to volumetric studies, were done with the “skull strip gold standard only” data set.

Tissue segmentation

The final step of the tissue segmentation process is the implementation of the segmentation algorithm itself. In the absence of noise and partial volume effects, segmentation would be trivial, and a simple thresholding method could be used to identify CSF (lowest signal intensity), WM (highest signal intensity), and GM (intermediate). Because noise and partial volume effects are not trivial, simple thresholding strategies are not applicable. The four segmentation algorithms examined here each use different strategies to overcome the difficulties of partial voluming effects. Expectation maximization segmentation (EMS; Van Leemput et al., 2003) is based on a hidden Markov random field model and an expectation–maximization algorithm, which is initialized with a priori probability images of GM, WM, and CSF that have been provided by the Montreal Neurological Institute (Evans et al., 1993). FMRIB’s automatic segmentation tool (FAST; Zhang et al., 2001) is also based on a hidden Markov random field model and an expectation–maximization algorithm but does not use the a priori probability images. The segmentation algorithm used in the statistical parametric mapping package (SPM; Ashburner and Friston, 1997) uses a maximum likelihood “mixture model” algorithm that is initialized with a priori probability images of GM, WM, and CSF that have been provided by the Montreal Neurological Institute (Evans et al., 1993). The fourth algorithm, partial volume classifier (PVC; Shattuck et al., 2001), starts with a model that is based on the theoretical intensity values for both pure and mixed-tissue types as an initialization for a maximum a posteriori (MAP) classifier.

Two of the algorithms, EMS (Van Leemput et al., 2003) and SPM (Ashburner and Friston, 1997), were evaluated using the default parameters. Because the outputs of both of these algorithms are probabilistic values (e.g., 75% chance that a voxel is CSF),

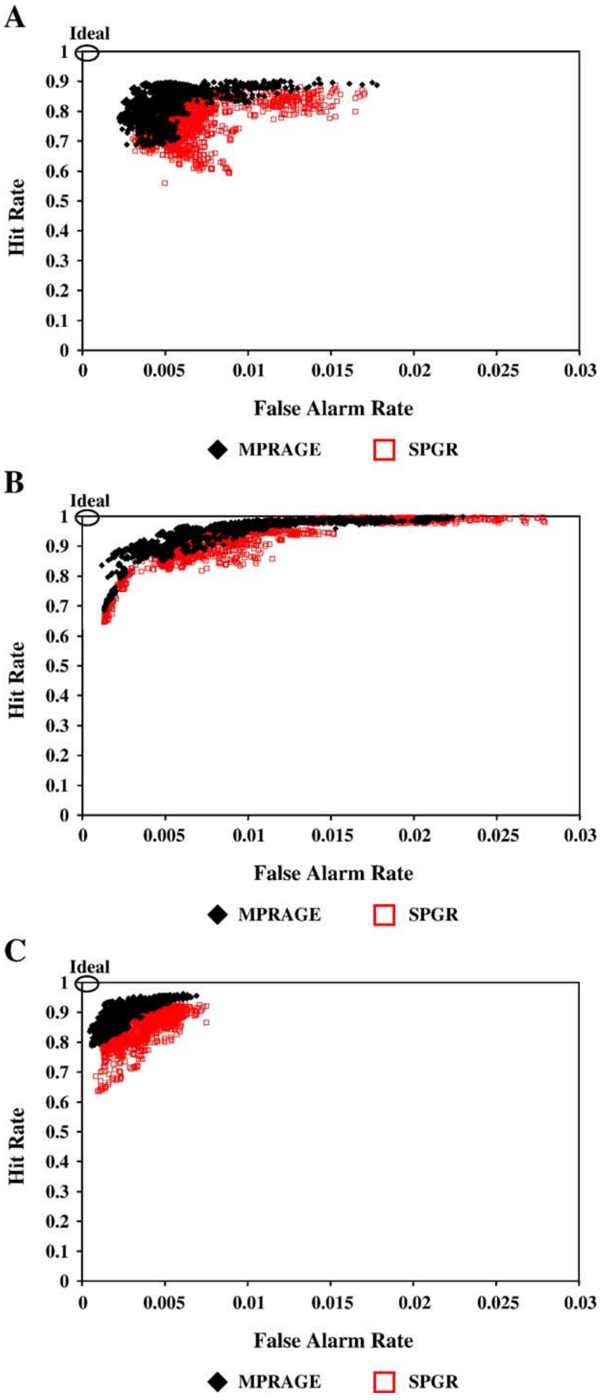

Fig. 7. Comparison of pulse sequence. Hit rate versus False Alarm Rate. Each point represents the agreement of one individual volume (processed with one unique processing sequence) with the identically processed gold standard, expressed as an ordered pair: (false alarm rate, hit rate). The ideal agreement between an individual volume and the gold standard volume would be in the top left-hand corner of the graph (0, 1), which represents 0% false alarms and 100% hits. All possible combinations of data processing protocols are represented. The red points indicate volumes that were acquired with the SPGR pulse sequence; black points represent volumes acquired with the MPRAGE pulse sequence. (A) GM; (B) WM; (C) CSF.

three thresholds were applied to the output probabilities (50%, 75%, and 90%); therefore, an SPM50 CSF map is a map that contains all voxels that were labeled CSF with >50% probability. Since both EMS and SPM involve an initial step of aligning to the MNI brain, this alignment was done once with the gold standard brain. Then the a priori maps were resliced into the space of the current data set; this was done so that all of the computations of tissue type counts were done in the same space for all algorithms and to avoid interpolation of the images. The third algorithm, FAST (Zhang et al., 2001), was applied using three different parameters for the number of iterations (10, 20, and 30; FAST10, FAST20, and FAST30, respectively). The fourth algorithm, PVC (Shattuck et al., 2001), was applied using the default parameters. While the current study focused on "hard" segmentation (each voxel is assigned only one tissue type label), future studies can be done to compare "hard" segmentation algorithms to "soft" (probabilistic labels) segmentation algorithms; a topic that is currently being investigated by Schaper et al. (2005).

At the beginning of the study, parameters were chosen for evaluation based on visual, subjective evaluation of performance on the gold standard image and one image from each pulse sequence. Multiple parameter sets were evaluated for an algorithm if the choice of the parameter was observed to have an effect on the output and if it was not visually obvious which parameter set was ideal. All of these algorithms (and parameters) were applied to every volume from both variations of the skull-stripping strategy. For the analysis of skull strip individual scans (each volume was skull stripped individually), there were a total of 170 pathways; each of the twenty original volumes (10 MPRAGEs and 10 SPGRs) were processed with all possible pathways. This led to a total of 3400 fully segmented images, of which 127 were rejected due to poor results. In the skull strip gold standard only (only the gold standard volume was skull stripped, and the results applied to the individual volumes) analysis, there were 190 pathways. This led to 3800 fully segmented images, of which 2 were rejected due to poor results. Poor results were identified by visual inspection.

Assessment

Each individually processed segmentation map was compared to the identically processed gold standard brain, and a d-prime was computed (Fig. 3). For example, in order to compute an individual GM d-prime, the GM mask of an individual scan is compared to the GM mask of the identically processed gold standard image. A voxel is considered a "hit" if the individual scan is labeled GM and the same voxel of the gold standard image is labeled GM. Similarly, a voxel is labeled a "miss" if the individual scan is *not* labeled GM, but the same voxel of the gold standard image is labeled GM. A voxel is labeled "false alarm" if the individual scan is labeled GM, but the gold standard is not; while a label of "correct rejection" is applied if neither the individual volume nor the gold standard label the voxel as GM. Thus, every voxel in the volume is labeled with one of these four terms. The hit rate is defined as the number of hits divided by the total number of voxels labeled GM in the gold standard image; similarly, the false alarm rate is the number of false alarms divided by the total voxels in the gold standard image that were not labeled as GM. Then a d-prime is computed for each segmented image by computing the difference between the *z* scores of the hit rate. For example, a hit rate of 50% and a false alarm rate of 50% is chance performance and the associated d-prime is 0; a hit rate of 80% and a false alarm rate of

5% corresponds to a d-prime of 2.5. The d-prime reflects the distance between the signal and noise means (Stanislaw and Todorov, 1999). Therefore, in the current context, if one pathway has a higher d-prime then that pathway is more robust to noise; in other words, the segmentation of an individual volume is more likely to correspond to the segmentation of the gold standard image (high SNR).

Many studies of reliability calculate an intraclass correlation coefficient, which is an appropriate measure of reliability if multiple individuals are being compared without a gold standard. For the current data set, the d-prime is more appropriate because the variability of the segmentation is measured as a deviation from the gold standard, rather than just agreement across individual segmentation maps. The d-prime is also a comprehensive metric of agreement because it is based on both the hit rate and the false alarm rate. Metrics that are solely based on percent overlap only show the hit rate while neglecting the false alarm rate. In an extreme case, an algorithm could label every voxel of one image as GM and show 100% hit rate, but this would not be an example of an optimal algorithm because the false alarm rate would also be

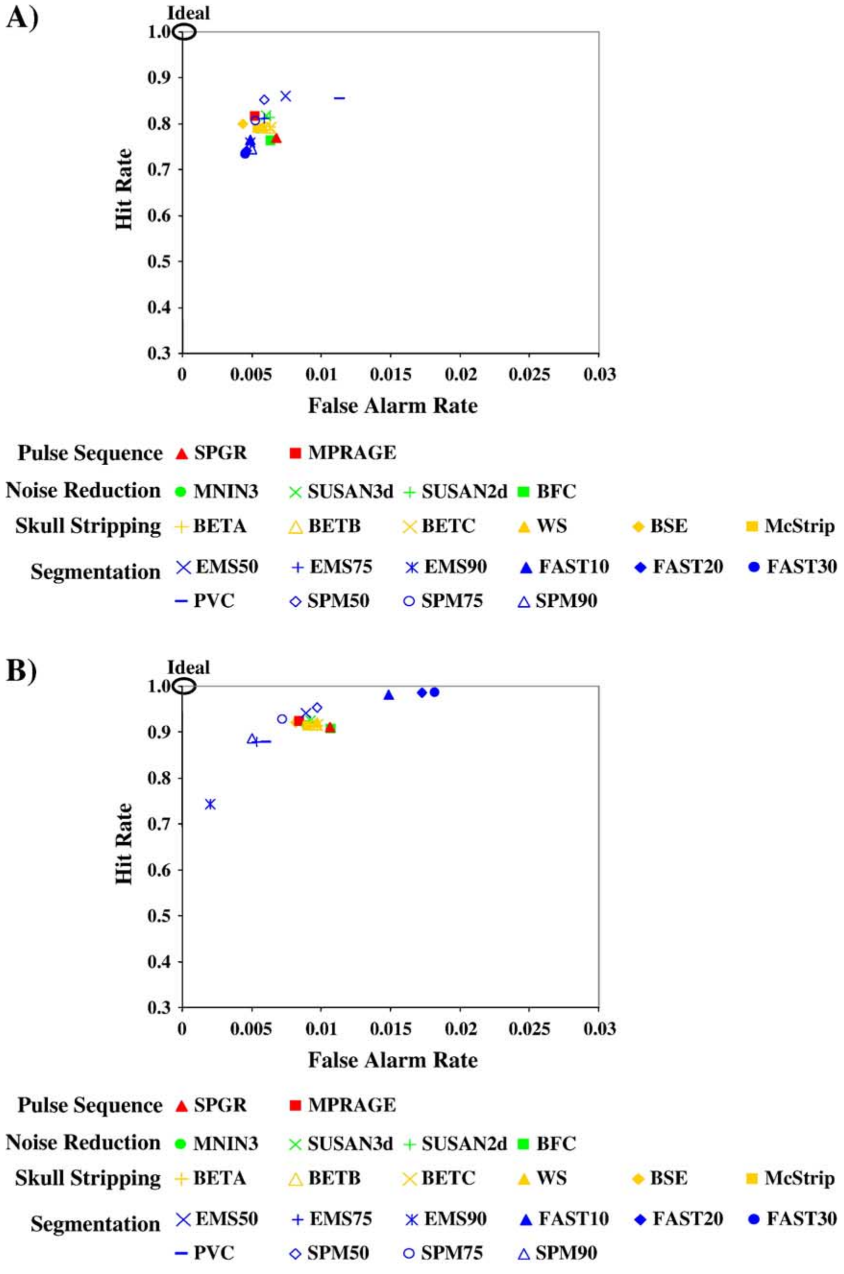

Fig. 8. Relative impact of data acquisition and data analysis. Graphs like those shown in Fig. 7 were reduced to one point that represents the average hit rate and average false alarm rate for each variable (e.g., the graph in Fig. 7A is reduced to the red square and red triangle in A). Red symbols are for the pulse sequence choices evaluated. Green symbols are for noise reduction algorithms. Gold symbols are for skull-stripping algorithms. Blue symbols are for segmentation algorithms. (A) GM; (B) WM; (C) CSF. These graphs indicate that the choice of segmentation algorithm has the most impact on segmentation reliability (blue points are the most scattered). Pulse sequence is the second most important, while noise reduction algorithms and skull-stripping algorithms are the least important (the points are clustered together).

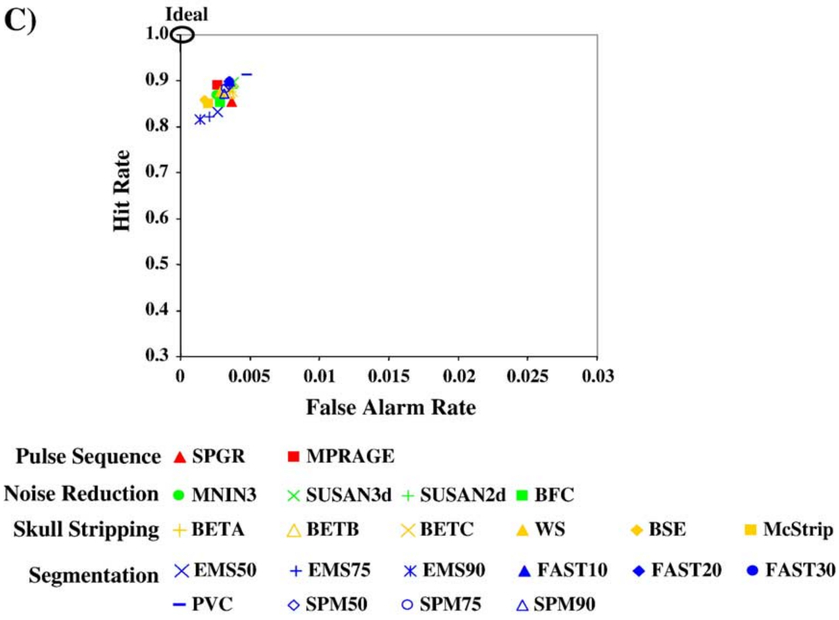

Fig. 8 (*continued*).

100%. The least informative metrics occasionally used in validation studies are those that only compute the volume correlations of a region, e.g., comparing the amount of GM in an individual volume to the amount of GM in the gold standard volume. This type of metric is misleading because volumes of regions can show a high degree of correlation while the spatial location of the regions can be very different. In an extreme case, two regions can be completely non-overlapping and still show an identical volume. Therefore, for many types of validation studies that compare individual data to that of a gold standard, a d-prime analysis is an appropriate metric to use.

Application to volumetric studies

The segmentation of an image into GM, WM, and CSF is a preprocessing step that is used in many volumetric analyses. It is currently unknown how variability in the segmentation process can affect volumetric analyses; therefore, a selected set of simple volumetrics were computed on the current data set.

A 3rd order non-linear warping algorithm (Woods et al., 1998b) was used to spatially align the single-subject Montreal Neurological Institute (MNI) brain (Evans et al., 1993; Collins et al., 1998) to the gold standard brain of the current study. The

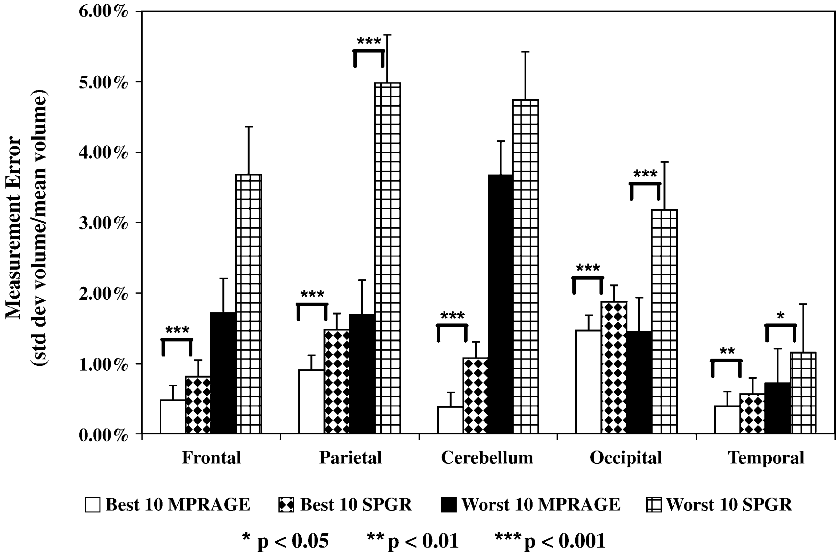

Fig. 9. Application to volumetric studies. The following bilateral regions were delineated on the volumes using an automated procedure (see Methods for details): frontal lobe, parietal lobe, cerebellum, occipital lobe, and temporal lobe. Within each pulse sequence, the GM volume of each region was calculated for volumes that were processed with the 10 “best” and 10 “worst” data analysis pathways (i.e., the 10 highest and 10 lowest GM d-primes, respectively). Outliers were defined as a volume error of 25% or more; this led to the removal of 10 out of 380 pathways. This graph shows that GM volume estimations (standard deviation of volumes estimated/mean volume) can be improved by an order of magnitude by using an optimized analysis pathway (high d-prime) compared to a poorly optimized pathway (low d-prime). Error bars show standard error.

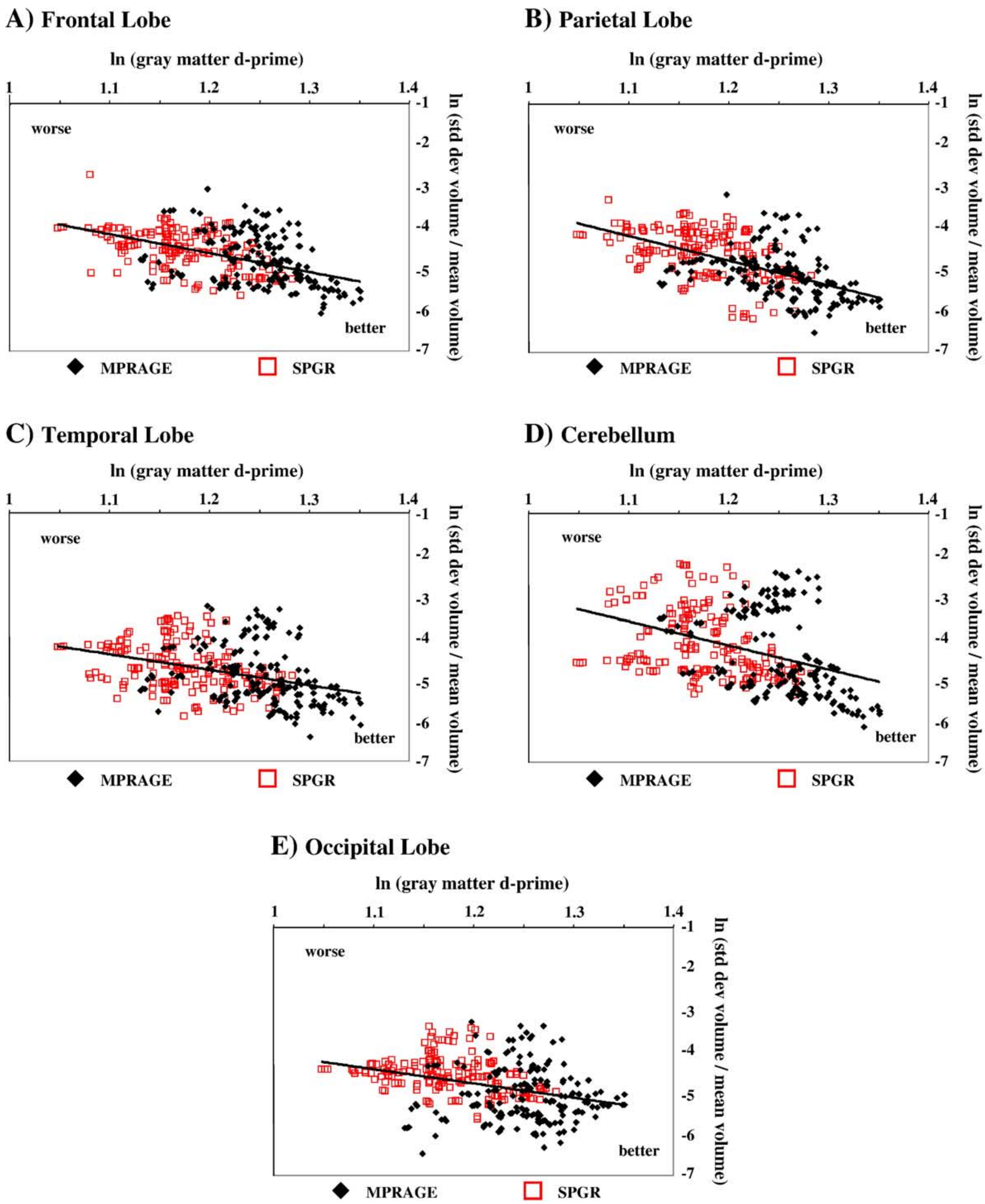

Fig. 10. Relationship between gray matter d-primes and gray matter volume measurement errors. For each region, the logarithm of the GM d-primes was plotted against the logarithm of the measurement error (standard deviation/mean volume) of the GM volume estimations. SPGR and MPRAGE points are displayed in separate colors, but the linear regressions were computed on the data as a whole. (A) Frontal lobe: ln (SD/mean volume) = −5.4 * ln (GM d-prime) + 2.2 ($r^2 = 0.24$, $F = 116$, $P < 0.001$). Parietal lobe: ln (SD/mean volume) = −6.0 * ln (GM d-prime) + 3.4 ($r^2 = 0.32$, $F = 173$, $P < 0.001$). Cerebellum: ln (SD/mean volume) = −6.8 * ln (GM d-prime) + 4.4 ($r^2 = 0.14$, $F = 61$, $P < 0.001$). Occipital lobe: ln (SD/mean volume) = −3.4 * ln (GM d-prime) + 0.4 ($r^2 = 0.12$, $F = 50$, $P < 0.001$). Temporal lobe: ln (SD/mean volume) = −3.8 * ln (GM d-prime) −0.29 ($r^2 = 0.13$, $F = 54$, $P < 0.001$). These plots indicate that there is a clear relationship between the GM d-prime and the measurement error in volumetrics. A larger GM d-prime indicates a smaller degree of variability in the estimation of GM volume. Therefore, by optimizing the segmentation pathway (maximizing the GM d-prime), the power and sensitivity of volumetric analyses can be maximized because the within subject variability of volume estimation can be minimized.

resulting transformation was then applied to the following regions of an atlas that were drawn manually on the MNI single-subject brain: cerebellum, frontal lobe, temporal lobe, parietal lobe, and occipital lobe (Tzourio-Mazoyer et al., 2002). This resulted in the delineation of the lobes on the gold standard brain; these delineations could then be transferred to the individual scans because all of the images were previously spatially aligned. While this process may not be as accurate as a manual delineation of the lobes, this is irrelevant for the current analysis because all of the scans are of the same subject and are spatially aligned. The GM volumes of each region were computed for all individual scans and the gold standard brain for every data acquisition/analysis pathway. The measurement errors of the GM volumes were computed for each pathway in order to characterize the relationship between the measurement error and the GM d-prime.

Results

Within each data set (both skull-stripping variations) and each tissue type (GM, WM, and CSF), descriptive statistics were computed, and the data were tested for normality using a one-sample Kolmogorov–Smirnov (KS) test. For a normal distribution, the mean and median are the same, the kurtosis has a value of 3, and the skewness a value of 0. The descriptive statistics and KS test results are summarized in Table 1. The results that follow are all reported using parametric statistics.

Optimal data acquisition and analysis

For every combination of pulse sequence, noise reduction, skull stripping, and segmentation protocol (e.g., MPRAGE-MNIN3-WS-EMS90), the average d-prime for each tissue type (GM, WM, and CSF) was computed (Table 2). Then the pathway associated with the maximum d-prime for each tissue type was identified (Figs. 4–6). The maximum d-primes indicate the optimal way to collect and analyze segmentation data, based on reproducibility. It should be noted that these results are necessarily limited to this particular scanner and these particular algorithms and parameter sets. It is probable that on a different scanner there would be a different optimal pulse sequence. It is also possible that a different algorithm or parameter choice of any of these processing algorithms could yield a more reliable segmentation of the current data set. Therefore, the results regarding the maximum d-primes are provided only as an illustration of how this method can be used to choose an optimal data acquisition and processing strategy.

The segmentation results of the data processing pathways that were associated with the maximum d-primes of each tissue type are shown in Figs. 4–6 (GM, WM, CSF, respectively). Each figure shows the segmentation of the gold standard volume (panel A), as well as the total number of hits and false alarms for the MPRAGE sequence (panels B and C) and the SPGR sequence (panels D and E). These figures are only shown for the analysis in which the gold standard volume was skull stripped and the result applied to individual volumes. While minimizing the variability of tissue segmentation across scans is important, it is also important to conduct studies that maximize the accuracy of tissue segmentation. The reason for this can easily be seen in Figs. 5C and E, the results of the pathway associated with the highest WM d-prime. The basal ganglia and thalamus of the gold standard image is labeled as GM, whereas all of the individual images labeled this area as WM (as evidenced by the high false alarm rate). Therefore, while optimizing data acquisition and analysis protocols, special attention should be paid to the accuracy of the results as well.

Table 1
Descriptive statistics of the d-primes

	Mean	Median	Skew	Kurtosis	KS test (D, p)
Skull strip individual scans					
GM	3.3750	3.3861	−0.0458	2.5535	(0.6, 0.22)
WM	3.9362	3.9262	0.1529	2.2787	(0.6, 0.23)
CSF	3.8146	3.8180	−0.2579	2.4843	(0.06, 0.19)
Skull strip gold standard only					
GM	3.3706	3.3802	−0.0431	2.4766	(0.06, 0.18)
WM	3.9674	3.9408	0.1420	2.1757	(0.06, 0.10)
CSF	3.9301	3.9308	−0.0833	2.1937	(0.07, 0.08)

The descriptive statistics and KS-test for normality results are summarized.

Table 2
Maximum d-primes

	Pulse sequence	Noise red.	Skull strip	Segment
Skull strip individual scans				
GM (3.8614)	MPRAGE	MNIN3	BSE	SPM50
WM (4.6517)	SPGR	BFC	BSE	FAST10
CSF (4.2378)	MPRAGE	MNIN3	BSE	PVC
Skull strip gold standard only				
GM (3.8614)	MPRAGE	MNIN3	BSE	SPM50
WM (4.7057)	MPRAGE	BFC	WS	FAST10
CSF (4.3806)	MPRAGE	MNIN3	McStrip	FAST30

The data acquisition and analysis choices that were associated with the maximum d-primes for each tissue type are summarized.

Notably, the GM d-primes are the lowest, indicating that the classification of gray matter is the most variable classification and is the least robust to noise. Also, the optimal strategy for GM is not the same as WM, which is not the same for CSF; therefore, the optimal strategy for segmentation will most likely vary depending upon the intended final use of the segmentation data. In order to further investigate this phenomenon, a post hoc analysis was conducted on the GM classifications of the deep brain structures (excluding the cortex). The maximum d-prime for GM of the deep brain structures was 4.5293 and was associated with the following pathway: MPRAGE-SUSAN2d-BETC-PVC, which is different from the pathway identified as optimal for global GM estimation. The MPRAGE still led to less variable GM classification than the SPGR (4.2184 > 4.1081, $P < 0.001$, Wilcoxon signed rank test). The GM classifications of deep brain structures showed that false alarm rates that were lower than any other tissue type (median = 0.02%) and highly variable hit rates (median = 74.68%, range: 54.44–92.98%). This indicates that the effect of noise on the classification of deep brain structures is that these structures tend to be misclassified as WM rather than the misclassification of nearby WM as GM. This example serves to highlight the importance of considering the final use of segmented data when implementing optimization procedures.

Failures

A few of the combinations of noise reduction algorithms and skull-stripping algorithms blatantly failed on these images (either large amounts of skull were kept or large amounts of cortex were removed). The following pairs were considered failures in the skull strip individual scans data set and were not included in the results: SUSAN2d-BSE, SUSAN3d-BSE, BFC-BSE, BFC-WS, SUSAN2d-McStrip, SUSAN3d-McStrip, and BFC-McStrip. For the skull strip gold standard only data set, there were fewer combinations of noise reduction algorithms and skull-stripping algorithms that failed because more of the combinations were successful on the gold standard volume than when applied to each individual brain. Therefore, the only pairs that were removed from this analysis were SUSAN2d-BSE, SUSAN3d-BSE, BFC-BSE, SUSAN2d-McStrip, and SUSAN3d-McStrip.

Impact of pulse sequence

Within each tissue type, paired t tests were computed between the mean d-primes of the MPRAGE volumes and the mean d-primes of the SPGR volumes, collapsing across data processing streams (Table 3). Paired t tests were used because each processing pathway (e.g., BFC-McStrip-PVC) was applied equally to both acquisition protocols. The t tests were computed separately for both variations of the skull-stripping application: (1) skull strip individual scans (top half of Table), and (2) skull strip gold standard only (bottom of table). In all 3 tissue types and both variations of the skull-stripping applications, the MPRAGE led to significantly higher d-primes than the SPGR (each test, $P < 0.001$, $df = 169$ for skull strip individual scans, $df = 189$ for skull strip gold standard only, paired). This indicates that, for this scanner, the MPRAGE sequence is more robust to noise than the SPGR; thus, the MPRAGE yields more reliable segmentation results.

The graphs in Fig. 7 contain one point from every segmentation map that was generated through all of the possible combinations of data acquisition and analysis procedures. Each point is represented as an ordered pair of (false alarm rate, hit rate). The d-prime is computed by subtracting the z score of the false alarm rate from the z score of the hit rate; therefore, the d-prime is a summary of the information shown in Fig. 7. These graphs clearly demonstrate that the MPRAGE performs better than the SPGR because the black points (MPRAGE) are overall closer to the ideal of 0% false alarms and 100% hits, which is the point (0,1) in the top left-hand corner of the graph. All 3 tissue types are shown: GM (Fig. 7A), WM (Fig. 7B), and CSF (Fig. 7C). These figures are only shown for the analysis in which the gold standard volume was skull stripped and the result applied to individual volumes.

Table 3
Effects of data acquisition

	GM (mean d-primes)	WM (mean d-primes)	CSF (mean d-primes)
Skull strip individual scans			
MPRAGE	3.5072	4.0120	3.9630
SPGR	3.2429	3.8603	3.6662
	$t = 51.30$, $P < 0.001$	$t = 16.03$, $P < 0.001$	$t = 56.21$, $P < 0.001$
Skull strip gold standard only			
MPRAGE	3.4962	4.0456	4.0736
SPGR	3.2451	3.8891	3.7866
	$t = 48.52$, $P < 0.001$	$t = 17.50$, $P < 0.001$	$t = 78.66$, $P < 0.001$

The mean d-primes for GM, WM, and CSF are summarized for both the MPRAGE and SPGR pulse sequences. For all 3 tissue types, the MPRAGE yields a significantly less variable segmentation.

Table 4
Effects of noise reduction

	GM (mean d-primes)	WM (mean d-primes)	CSF (mean d-primes)
Skull strip individual scans			
MNIN3	3.3704	3.9692	3.8436
BFC	3.2328	3.7628	3.7631
SUSAN2d	3.4192	3.9896	3.7891
SUSAN3d	3.4444	3.9631	3.8353
Skull strip gold standard only			
MNIN3	3.3845	3.9870	3.9750
BFC	3.2435	3.8974	3.8590
SUSAN2d	3.4256	4.0037	3.9138
SUSAN3d	3.4538	3.9889	3.9680

The mean d-primes for GM, WM, and CSF are summarized for all of the noise reduction algorithms: MNIN3, BFC, SUSAN2, and SUSAN3.

Impact of noise reduction

Paired t tests were computed between each pair of noise reduction algorithm, separately for each tissue type. The t tests were all corrected for multiple comparisons with a Bonferroni correction, by dividing the alpha by the number of tests, which was 6. Multiple t tests were conducted rather than an ANOVA because there were missing data values due to the noise reduction skull-stripping algorithm failures. The mean d-primes for each tissue type and each algorithm are shown in Table 4. The t tests were computed separately for both variations of the skull-stripping application: (1) skull strip individual scans (top half of Table), and (2) skull strip gold standard only (bottom of table).

For the skull strip individual scans data, all of the GM d-primes were significantly different from each other ($P < 0.001$, corrected). In the WM and CSF d-primes, all of the differences were significant ($P < 0.001$, corrected) except for MNIN3 versus SUSAN3d (ns). For the skull strip gold standard only data, all of the GM d-primes were significantly different from each other ($P < 0.001$, corrected). In the WM d-primes, most differences were significant ($P < 0.003$, corrected); SUSAN3d did not differ from MNIN3 (ns) and was only marginally different from SUSAN2d ($P < 0.02$, corrected). All of the CSF d-primes were significantly different from each other ($P < 0.001$, corrected), except the MNIN3 did not significantly differ from either SUSAN algorithm (ns). In most cases, the SUSAN algorithm performed better than the others; however, the mean differences in the d-primes across noise reduction algorithms were rather small. It should also be noted that the MNIN3 algorithm was the only one that did not have any failures with the skull-stripping algorithms.

It should be noted that in the current data analysis tree (Fig. 1B), BFC and BSE are not applied in the order recommended by the authors of the methods. BFC in particular is based on a model that assumes the data have already been skull stripped; therefore, the failures with BFC are not unexpected. Nonetheless, BFC was part of the optimal pathway for WM segmentation. Attempts to systematically reverse the order of the analysis for companion purposes were largely unsuccessful. The only skull-stripping algorithm that correctly skull stripped the raw data was BET, so a fully automated processing stream based on BFC of skull stripped data was limited to the sequence BET-BFC. Of course it is possible that the other skull-stripping algorithms would perform more robustly in other data sets. Average d-primes were computed for each of the noise reduction algorithms when applied after BETA, BETB, and BETC. In all three tissue types, the d-primes associated with MNIN3 went down, while the d-primes for BFC went up. In the GM and the CSF, the values for SUSAN2d and SUSAN3d went up, while in the WM they went down. In all three tissues types the order of d-primes from largest to smallest was SUSAN3d > SUSAN2d > BFC > MNIN3. Future studies can be done to further investigate the effect of the order of the analysis pathway.

Impact of skull stripping

Matched paired *t* tests were also conducted between the mean d-primes of each pair of skull-stripping algorithms, within each tissue type (Table 5). The *t* tests were corrected for multiple comparisons with a Bonferroni correction by dividing the alpha by 15 (number of *t* tests). The d-primes associated with the skull strip individual scans data reflect two types of variability: that due to differences between algorithms or parameter choices of a given algorithm, and that due to scan-to-scan variability in the implementation of any skull-stripping algorithm. The d-primes associated with the skull strip gold standard only data only reflect the variability due to algorithms and parameters because only the gold standard brain was actually skull stripped with an algorithm; all of the individual scans were skull stripped by applying the mask from the gold standard brain.

For the skull strip individual scans data, all of the pairs were significantly different (P < 0.003, corrected), except BETA-McStrip (P < 0.04, corrected) and BETA-WS, BETC-WS (ns). In the WM, all of the pairs were significant. In the CSF, all of the pairs were significant (P < 0.001, corrected) except BSE was only marginally different from BETB (P < 0.02, corrected) and BETA (P < 0.007, corrected). For the skull strip gold standard only data, all of the differences in the GM d-primes were significant (P < 0.001, corrected), except none of the BET algorithms differed from each other (ns). In the WM, most of the differences were significant (P < 0.005, corrected), except the difference between BETA and BETB was only marginally significant (P < 0.04, corrected) and there was no significant difference between McStrip and WS (ns). In the CSF, McStrip and WS differed significantly from the BET algorithms (P < 0.001, corrected) and BSE was marginally different from the BET algorithms (P < 0.03, corrected). BSE was associated with the highest d-prime for all three tissues types; however, it only worked in images that were first corrected with MNIN3.

Many of the *t* tests of the skull-stripping algorithms were not significant, and the differences in d-primes across algorithms were rather small. This indicates that the choice of a skull-stripping algorithm has a relatively minor impact on segmentation variability. Furthermore, many of the differences across algorithms were reduced by removing the scan-to-scan variability of the skull-stripping algorithm (as was the case for the skull strip gold standard only data). This means that of the variability in tissue segmentation that can be attributed to skull-stripping algorithms, most of it is due to the scan-to-scan variability of the skull-stripping algorithms themselves rather than variability across algorithms. Therefore, in terms of segmentation variability, there is more to be gained by reducing the scan-to-scan variability of any given algorithm rather than by choosing an optimal algorithm. It should be noted that this study did not address the issue of accuracy of skull-stripping algorithm. While the evidence of the current study suggests that the choice of a skull-stripping algorithm has little impact on tissue segmentation variability, it is likely that the accuracy and/or reliability of skull stripping can affect other processing steps, e.g., cross-registration.

Table 5
Effects of skull stripping

	GM (mean d-primes)	WM (mean d-primes)	CSF (mean d-primes)
Skull strip individual scans			
BETA	3.3637	3.8935	3.8778
BETB	3.3722	3.9170	3.8896
BETC	3.3528	3.8705	3.8581
BSE	3.5006	4.0671	4.0379
McStrip	3.3794	4.0167	3.7082
WS	3.3801	4.0356	3.5335
Skull strip gold standard only			
BETA	3.3518	3.9375	3.9027
BETB	3.3545	3.9500	3.9052
BETC	3.3462	3.9165	3.9023
BSE	3.4999	4.0649	4.0376
McStrip	3.3962	3.9961	3.9614
WS	3.3850	4.0266	3.9677

The mean d-primes for GM, WM, and CSF are summarized for all of the skull-stripping algorithms: BETA, BETB, BETC, BSE, McStrip, and WS.

Impact of segmentation algorithm

The mean d-primes associated with each segmentation algorithm for each tissue type are summarized in Table 6. Within the skull strip individual scans data set, an ANOVA was computed for each tissue type: GM (F(9,297) = 79, P < 0.001); WM (F(9,297) = 202, P < 0.001); CSF (F(9,297) = 9; P < 0.001). Within the skull strip gold standards only data set, an ANOVA was computed for each tissue type: GM (F(9,333) = 122, P < 0.001); WM (F(9,333) = 280, P < 0.001); CSF (F(9,333) = 76, P < 0.001). ANOVA analyses were conducted because there were no missing data values, and post hoc pairwise comparisons are not reported because the number of *t* tests would be 45 for each skull-stripping variation. The differences among the mean d-primes were larger than any other factor; therefore, the choice of a segmentation algorithm has the largest impact on the variability of tissue segmentation. These effects are largest in the WM, as indicated by the large range of d-prime values in the WM compared to the other tissue types.

Relative impact of acquisition and analysis protocols

The process of optimizing every step of data acquisition and analysis protocols can be time consuming and labor intensive; therefore, it is more beneficial to optimize protocols that have a large effect on the final variability rather than a minimal effect. The graphs in Fig. 8 show the relative impact of each acquisition/analysis protocol on the final segmentation variability for each tissue type, as computed on the skull strip gold standard only data set. These graphs were generated by computing an average hit rate and false alarm rate for each pulse sequence and algorithm. For example, in Fig. 8A (GM), the MPRAGE pulse sequence has a point at (0.005, 0.816) and the SPGR pulse sequence has a point at (0.007, 0.769), indicating that when the data were averaged across

Table 6
Effects of segmentation

	GM (mean d-primes)	WM (mean d-primes)	CSF (mean d-primes)
Skull strip individual scans			
EMS50	3.5286	3.9537	3.7168
EMS75	3.4174	3.7495	3.7720
EMS90	3.3119	3.5594	3.8897
FAST10	3.3409	4.2348	3.8339
FAST20	3.2829	4.2516	3.8389
FAST30	3.2531	4.2430	3.8376
SPM50	3.5823	4.0062	3.8244
SPM75	3.4406	3.8925	3.8153
SPM90	3.2676	3.7784	3.7931
PVC	3.3248	3.6926	3.8245
Skull strip gold standard only			
EMS50	3.5348	3.9430	3.7644
EMS75	3.4143	3.7345	3.8052
EMS90	3.2993	3.5482	3.9032
FAST10	3.3277	4.3166	4.0000
FAST20	3.2657	4.3374	4.0005
FAST30	3.2391	4.3286	3.9969
SPM50	3.5832	4.0387	3.9704
SPM75	3.4362	3.9145	3.9522
SPM90	3.2570	3.7926	3.9216
PVC	3.3491	3.7195	3.9865

The mean d-primes for GM, WM, and CSF are summarized for all of the segmentation algorithms: EMS50, EMS75, EMS90, FAST10, FAST20, FAST30, PVC, SPM50, SPM75 and SPM90.

all data processing choices, the MPRAGE yielded an average false alarm rate of 0.5% and a hit rate of 81.6%, while the SPGR yielded an average false alarm rate of 0.7% and a hit rate of 76.9%. Therefore, by selecting the MPRAGE pulse sequence of this scanner instead of the SPGR sequence, the false alarm rate is decreased while the hit rate is increased. These graphs indicate that the choice of segmentation algorithm has the most impact on segmentation reliability (blue points are the most scattered); the choice of pulse sequence has the second greatest impact. The choices of noise reduction and skull-stripping algorithms have the lowest effect, as evidenced by the fact that the points are clustered. Also, it is clear from these graphs that the classification of WM is the most sensitive to acquisition and analysis protocols, while the CSF is the least sensitive. Therefore, studies that rely heavily on the reliable segmentation of WM have the most to gain from optimization.

Application to volumetric studies

The previous results indicated that different data acquisition and analysis choices can impact the variability in tissue segmentation; however, these results do not indicate how variability in tissue segmentation can impact volumetric measurements. Therefore, the relationship between the GM d-prime and GM volumetrics was explored. For every fully segmented image within each acquisition/processing pathway (e.g., MPRAGE-MNIN3-BETC-EMS75 pathway consisted of 10 images), GM volumes for each of the following regions were calculated: frontal lobe, parietal lobe, cerebellum, occipital lobe, and temporal lobe. The ratio of the standard deviation to the mean GM volume was computed for each pathway and each region. This analysis was only conducted on the skull strip gold standard only data set. Outliers were defined as those pathways that showed a ratio greater than 25%; this resulted in the removal of 10 out of 380 pathways. From the remaining pathways, the 10 “best” (highest GM d-prime) and 10 “worst” (lowest GM d-prime) analysis pathways were chosen for each pulse sequence. For the MPRAGEs, 9 out of 10 of the best pathways used the SPM50 segmentation (the 10th used SPM75), while the worst 10 pathways used a mix of the FAST algorithms and EMS50. For the SPGRs, 7 out of 10 of the best pathways used the EMS50 segmentation (the other 3 used SPM50), while the worst 10 pathways used a mix of the FAST algorithms and EMS50. For both pulse sequences, 8 out of 10 of the best analysis pathways used a parameter set of the SUSAN algorithm (e.g., SUSAN2d or SUSAN3d) for the noise reduction. Most of the worst pathways used BFC for a noise reduction algorithm, consistent with the developer’s recommendation that BFC should not be used with data not previously skull stripped.

In 4 out of 5 of the regions and both pulse sequences, the best 10 analysis pathways yielded a more reliable estimate of GM volume than the worst 10 analysis pathways, as evidenced by a lower ratio of standard deviation to mean volume ($P < 0.03$; Fig. 9); the difference in the occipital lobe was not significant. When the data were averaged across pulse sequences and analysis pathways, the mean volume estimation error (ratio of standard deviation to mean volume) for the best pathways was 0.94%, while the mean for the worst pathways was 2.70%. This means that on average, optimal pathways yield measurements that are three times less variable than suboptimal pathways. The mean volume estimation error across all regions was 0.72% for the 10 best MPRAGE pathways and 1.16% for the 10 best SPGR pathways. In the occipital lobe, optimization made the least difference; the best pathways yielded an estimation error of 1.67% and the worst pathways 2.30%. In the cerebellum, optimization made the most difference; the best pathways yielded an estimation error of 0.73% and the worst pathways 4.21%. This means that the potential benefit of using an analysis pathway that has a high GM d-prime versus a low GM d-prime is on average three-fold and can be as much as six-fold.

The relationship between GM d-primes and GM volume estimations was examined, and the results are shown in Fig. 10. In the figure, the logarithm of the GM d-primes are plotted on the x-axis, while the logarithm of the ratio of the standard deviation to the mean volume estimates are plotted on the y-axis. Each point represents an analysis pathway; red points represent the SPGR pulse sequence and black points the MPRAGE sequence. The linear regressions were computed on the data as a whole (not distinguishing between MPRAGE and SPGR) and are summarized here. Frontal lobe: ln (SD/mean volume) = −5.4 * ln (GM d-prime) +2.2 ($r^2 = 0.24$, $F = 116$, $P < 0.001$). Parietal lobe: ln (SD/mean volume) = −6.0 * ln (GM d-prime) + 3.4 ($r^2 = 0.32$, $F = 173$, $P < 0.001$). Cerebellum: ln (SD/mean volume) = −6.8 * ln (GM d-prime) + 4.4 ($r^2 = 0.14$, $F = 61$, $P < 0.001$). Occipital lobe: ln (SD/mean volume) = −3.4 * ln (GM d-prime) + 0.4 ($r^2 = 0.12$, $F = 50$, $P < 0.001$). Temporal lobe: ln (SD/mean volume) = −3.8 * ln (GM d-prime) −0.29 ($r^2 = 0.13$, $F = 54$, $P < 0.001$). These regressions indicate that there is a clear association between GM d-primes and volume estimates of GM in the major lobes of the brain. Specifically, as the GM d-prime increases, the variability in volume estimation decreases. Therefore, by optimizing the segmentation process, through the selection of a maximum d-prime, the variability in volume estimation will be minimized.

Discussion

Clinical and basic neuroanatomical structural neuroimaging studies usually quantify one or more metrics of brain structure, including total brain volume, lateral ventricle shape, frontal lobe gray matter volume, caudate volume, and many others. Most of these metrics depend upon the accurate and reliable tissue segmentation of a structural scan, i.e., the labeling of gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). Recent reviews in several disorders have shown convergent evidence of structural abnormalities (Lingford-Hughes et al., 2003; Haldane and Frangou, 2004); however, studies in other disorders show more inconsistent results (Sheline et al., 2002; Palmen and van Engeland, 2004). While many inconsistencies may be due to inhomogeneities in clinical populations and high inter-subject anatomical variability, it is also probable that data acquisition and analysis methods introduce variability into the estimation of the metrics. The role of data acquisition/analysis optimization is to increase the power of any given study by reducing the variability and/or increasing the accuracy of the measurements. The current study examined a method for optimizing data acquisition and analysis protocols with the objective of reducing the variability of tissue segmentation that is introduced through measurement error. By minimizing variability due to measurement error, true inter-subject anatomical variability can be studied more effectively. However, it is not enough to simply reduce the variability in the measurements; future studies that examine methods for optimizing the accuracy of tissue segmentation will also be important. Even though many structural studies depend on tissue segmentation, there are also several other procedures that lead to the final structural metric (e.g., registration, manual delineation). Future studies that reduce the variability and increase the accuracy of these procedures will also lead to more consistent results.

In optimization studies, two main points are usually considered: the implementation of the optimization strategy and its overall impact. In structural neuroimaging, the implementation of optimization strategies can be difficult because of the lack of a true gold standard. Some optimization strategies may be time consuming and/or labor intensive, and few studies have examined how different acquisition and analysis choices can impact the accuracy and/or reliability of tissue segmentation and volumetrics. In a recent study, Li and Mirowitz (2004) used a composite phantom as a gold standard in order to evaluate the effects of data acquisition choices (pulse sequence and parameters) on the overall quality of the image. They found that the excitation flip angle (FL), echo time (TE), and repetition time (TR) play critical roles in determining the quality of the image, as can be seen by intensity inhomogeneities and ghosting artifacts (Li and Mirowitz, 2004). It is not clear, however, how these differences in image quality translate into differences in the accuracy and/or reliability of downstream processes; therefore, the impact of optimizing these pulse sequence parameters has not been established. Arnold et al. (2001) evaluated the performance of six different noise reduction algorithms on both real and simulated data. They were primarily interested in the accuracy of the algorithms and found that the two locally adaptive methods that they studied (MNIN3 and BFC) were superior to the four non-adaptive methods (Arnold et al., 2001). From this study it is not clear how much the choice of an inhomogeneity algorithm can impact the calculation of structural metrics. In the current study, it was found that the choice of a noise reduction algorithm has very little effect on tissue segmentation variability; however, the application of some noise reduction algorithms prior to skull stripping can cause the skull-stripping algorithms to fail.

Lee et al. (2003) used data that are publicly available from two repositories (http://www.bic.mni.mcgill.ca/brainweb/ and http://www.cma.mgh.harvard.edu/ibsr/) to compare the accuracy of automated and semi-automated skull-stripping methods. They found that while the semi-automated methods were more accurate, they were also more time consuming and oftentimes more variable (Lee et al., 2003). It is unclear how these differences in variability and accuracy affect tissue segmentation, cross-registration, and/or volumetric calculations; therefore, the question of whether or not it is worth the extra time to use semi-automated methods remains unanswered. Boesen et al. (2004) compared several automatic skull-stripping algorithms for accuracy, by comparing to manual delineation, and for reproducibility, by comparing performance across repeat scans of a single subject. These results found that McStrip (Rehm et al., 2004) outperformed the other algorithms studied both in terms of accuracy and reproducibility. In the current study, the skull-stripping algorithms were compared in terms of their impact on tissue segmentation variability; BSE was found to lead to less variability than the other skull-stripping algorithms. However, in the current study, the overall impact of skull-stripping algorithms on tissue segmentation variability was small. Schnack et al. (2004) recently computed intraclass coefficients (ICCs) for several volumetrics on six subjects on six different scanners in order to quantify the reproducibility of volumetrics across different hardware. The aim of this study was to evaluate to what extent data could be combined across hardware, and also to calibrate analysis procedures from each site to increase the reliability (Schnack et al., 2004). This study described an optimization process that is relatively straightforward to implement; the ICC is an interpretable metric of variability, and these data explicitly show that optimization, or calibration, is important because there is a significant impact of calibration on the ICCs of volumetrics. Each of these studies examined how optimizing one aspect of either data acquisition or data analysis could influence the accuracy and/or reliability of structural neuroimaging. In the current study, several aspects were examined at once to examine how each of these factors can impact the reliability of tissue segmentation and the resulting computed volumetrics. However, it should be noted that the current study only examined reliability; future studies can be done to assess the impact of these factors on the accuracy of tissue segmentation.

The primary goal of the current study was to outline a general strategy that can be used to optimize tissue segmentation reliability. This strategy has three main steps: (1) collecting multiple T1-weighted images on a single subject; (2) creating an average of all of the scans to be used as a “gold standard”; and (3) comparing the segmentation maps of the individual volumes to that of the gold standard by computing a d-prime. While in the current study this strategy was implemented to evaluate several sources of variability at once, the same strategy could easily be adapted to evaluate only one or two sources of variability. For example, in order to choose among several pulse sequences, multiple images can be collected on one subject with each pulse sequence. The gold standard can be created by averaging the multiple scans. Then all of the individual volumes and the gold standard can be segmented with the standard analysis packages used in the laboratory. D-primes can be computed for each pulse sequence, and the pulse sequence that has the largest d-prime is the one that is optimal for that particular scanner in terms of minimizing variability. A similar approach can be used if only one pulse sequence is available, but the investigator wishes to choose among several tissue segmentation algorithms. However, caution must be used in this approach; one cannot just blindly choose the maximum d-prime without also considering the overall accuracy of the approach. For example, see Fig. 5, which shows the segmentation of images for the pathway that had the maximum d-prime for WM classification. Although this pathway yielded results that were less variable than other pathways, the resulting classifications were clearly inaccurate in that the basal ganglia and thalamus were consistently labeled WM, not GM. The optimal pathway is one that both minimizes variability while also achieving acceptable accuracy.

In the current study, two pulse sequences, three noise reduction algorithms (four total parameter sets), four skull-stripping algorithms (six total parameter sets), and four segmentation algorithms (ten total parameter sets) were evaluated simultaneously. The main effects of each acquisition/processing step were each examined (Tables 3–6, and Figs. 7 and 8). Post hoc statistics showed that, for this particular scanner, the MPRAGE pulse sequence performed better than the SPGR pulse sequence. For most of the *t* tests, the SUSAN algorithm performed better than other noise reduction algorithms; although the MNIN3 noise reduction algorithm was the only one that had no failures with skull-stripping programs. The differences among skull-stripping algorithms were small, and many of the tests were non-significant. The impact of segmentation algorithms was the largest; the methods that started with an a priori estimate (e.g., SPM50 and EMS50) performed the best on GM, and the FAST algorithms performed the best on WM. Generally, choosing a different parameter for a given algorithm led to less variability than choosing a different algorithm (e.g., BETA and BETB were more similar than BETA and WS). Also, there is clear evidence that the optimal strategy for measuring GM is different from the optimal strategy for measuring WM. Furthermore, the optimal strategy as determined by global GM measurements may differ from the optimal strategy for deep brain GM measurements. Therefore, the intended use of the segmentation data should play a role in the optimization process. Overall, the GM d-primes were the lowest, indicating that the classification of GM is the most variable; while the WM d-primes showed the most variability, indicating that the classification of WM is most affected by optimization.

All segmentation algorithms face the same problems of partial voluming effects and noise in the data. The average cortical GM thickness of the human brain is 1.5–4 mm, whereas the typical resolution of a T1-weighted image is 1.0–1.5 mm. In some anatomical regions (e.g., thalamus and cerebellum) the partial voluming effects are even worse, leading to even less contrast between tissue types, particularly between GM and WM, making these regions particularly difficult to segment accurately. One way to address this issue is to use anatomical knowledge by initializing the segmentation process with a priori probability images of GM, WM, and CSF. One potential drawback of this approach is that the initialization step depends on co-registering an individual image with the template of the a priori images. This will necessarily lead to a higher degree of accuracy for subjects whose anatomy is more similar to the template anatomy. It is beyond the scope of this study to evaluate the degree to which the benefit of anatomical knowledge offsets potential biases due to co-registration; however, two algorithms that use a priori information were included in this study—EMS and SPM. Both of these algorithms yielded the most reliable estimates of GM classification. Notably, the threshold of 50% was more reliable than higher thresholds (75% and 90%).

Many structural neuroimaging studies use *t* tests to compare volumetrics between two groups (e.g., patients and controls). The power of a *t* test depends on the effect size and the sample size, where the effect size is defined as the ratio between the difference in means and the pooled standard deviation (Cohen, 1977). If the data analysis method introduces variability into the volumetric estimates, then this will increase the standard deviation of values within each group, effectively diminishing the effect size. Therefore, by optimizing the data acquisition and analysis procedures that lead to volume estimations, the variability due to volume estimation can be reduced, thereby increasing the power of the study to detect anatomical differences. The degree to which optimization will increase the power is an empirical question that will most likely vary across studies. However, regardless of the study, the power will almost always be increased by the optimization of data acquisition and data analysis procedures. The d-prime is a fairly simple calculation that can be implemented in any study. All that is required is to collect multiple scans of a single subject, varying the factor(s) of interest. At the stage of study design, d-prime computations can be used to choose optimal data acquisition parameters. After the data has been collected, d-prime computations can still be used to choose optimal data analysis parameters. In some cases, optimizing tissue segmentation can yield volumetric measurements that are six times less variable. The results of this study indicate that the greatest gains can potentially be made by optimizing the tissue segmentation algorithms themselves, and the second most efficient optimization is on the data acquisition protocol.

Conclusion

This study examined a method for comparing the reliability of several acquisition/analysis pathways that lead to tissue segmentation. This work has implications for several types of structural neuroimaging studies, which can benefit from such optimization. Several conclusions can be drawn from this study; in particular, the intended use of segmentation data should play a role in the optimization process. For example, the optimal pathway for GM differs from the optimal pathway for WM. Also, it is not sufficient to simply blindly choose the most reliable pathway without also considering the accuracy. Ideally both accuracy and reliability should be optimized simultaneously. Several results from this study indicated that the biggest improvements in the reliability can be made by optimizing the segmentation algorithms themselves, secondarily through optimization of acquisition protocols. The results also showed that the classification of GM is the most variable, while the classification of WM is the most sensitive to acquisition/analysis choices. Finally, it was demonstrated that the optimization of the acquisition/analysis protocol on the basis of tissue segmentation reliability directly leads to a less variable estimation of GM lobe volumes.

Acknowledgments

Support for this work was provided by a grant from the Human Brain Project (P20-MHDA52176), the National Institute of Biomedical Imaging and Bioengineering, National Institute of

Mental Health, National Institute for Drug Abuse, National Cancer Institute, and the National Institute for Neurologic Disease and Stroke. This work was supported by a training grant #R01 MH071940 from the National Institute of Mental Health and the National Institute of Neurological Disorders and Stroke. Additional support was provided by a resource grant from the National Center for Research Resources (P41 RR013642). Additional support was provided by a grant from the National Institute of Health (P20-EB02013). For generous support, the authors also wish to thank the Brain Mapping Medical Research Organization, Brain Mapping Support Foundation, Pierson-Lovelace Foundation, The Ahmanson Foundation, Tamkin Foundation, Jennifer Jones-Simon Foundation, Capital Group Companies Charitable Foundation, Robson Family, William M. and Linda R. Dietel Philanthropic Fund at the Northern Piedmont Community Foundation, Northstar Fund, and the National Center for Research Resources grants RR12169, RR13642 and RR08655.

References

Arnold, J.B., Liow, J.S., Schaper, K.A., Stern, J.J., Sled, J.G., Shattuck, D.W., Worth, A.J., Cohen, M.S., Leahy, R.M., Mazziotta, J.C., Rottenberg, D.A., 2001. Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects. Neuroimage 13 (5), 931–943.

Ashburner, J., Friston, K., 1997. Multimodal image coregistration and partitioning—A unified framework. Neuroimage 6 (3), 209–217.

Boesen, K., Rehm, K., Schaper, K., Stoltzner, S., Woods, R., Luders, E., Rottenberg, D., 2004. Quantitative comparison of four brain extraction algorithms. Neuroimage 22 (3), 1255–1261.

Coffey, C.E., Ratcliff, G., Saxton, J.A., Bryan, R.N., Fried, L.P., Lucke, J.F., 2001. Cognitive correlates of human brain aging: a quantitative magnetic resonance imaging investigation. J. Neuropsychiatry Clin. Neurosci. 13 (4), 471–485.

Cohen, J., 1977. Statistical Power Analysis for the Behavioral Sciences, Rev. ed. Academic Press, New York.

Cohen, M.S., DuBois, R.M., Zeineh, M.M., 2000. Rapid and effective correction of RF inhomogeneity for high field magnetic resonance imaging. Hum. Brain Mapp. 10 (4), 204–211.

Collins, D.L., Zijdenbos, A.P., Kollokian, V., Sled, J.G., Kabani, N.J., Holmes, C.J., Evans, A.C., 1998. Design and construction of a realistic digital brain phantom. IEEE Trans. Med. Imaging 17 (3), 463–468 (Jun.).

Courchesne, E., Karns, C.M., Davis, H.R., Ziccardi, R., Carper, R.A., Tigue, Z.D., Chisum, H.J., Moses, P., Pierce, K., Lord, C., Lincoln, A.J., Pizzo, S., Schreibman, L., Haas, R.H., Akshoomoff, N.A., Courchesne, R.Y., 2001. Unusual brain growth patterns in early life in patients with autistic disorder: an MRI study. Neurology 57 (2), 245–254.

Evans, A.C., Collins, D.L., Mills, S.R., Brown, E.D., Kelly, R.L., Peters, T.M., 1993. 3D statistical neuroanatomical models from 305 MRI volumes. Proceedings of the IEEE Nuclear Science Symposium and Medical Imaging Conference, pp. 1813–1817.

Giedd, J.N., Snell, J.W., Lange, N., Rajapakse, J.C., Casey, B.J., Kozuch, P.L., Vaituzis, A.C., Vauss, Y.C., Hamburger, S.D., Kaysen, D., Rapoport, J.L., 1996. Quantitative magnetic resonance imaging of human brain development: ages 4–18. Cereb. Cortex 6 (4), 551–560.

Green, D.M., Sweets, J.A., 1966. Signal Detection Theory and Psychophysics. Wiley, New York.

Haldane, M., Frangou, S., 2004. New insights help define the pathophysiology of bipolar affective disorder: neuroimaging and neuropathology findings. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 28 (6), 943–960.

Holmes, C.J., Hoge, R., Collins, L., Woods, R., Toga, A.W., Evans, A.C., 1998. Enhancement of MR images using registration for signal averaging. J. Comput. Assist. Tomogr. 22 (2), 324–333.

Hulshoff Pol, H.E., Schnack, H.G., Bertens, M.G., van Haren, N.E., van der Tweel, I., Staal, W.G., Baaré, W.F., Kahn, R.S., 2002. Volume changes in gray matter in patients with schizophrenia. Am. J. Psychiatry 159 (2), 244–250.

Jezzard, P., 2000. Physical basis of spatial distortions in magnetic resonance images. In: Bankman, I.N. (Ed.), Handbook of Medical Imaging: Processing and Analysis. Academic Press, San Diego, pp. 425–438.

Lee, J.M., Yoon, U., Nam, S.H., Kim, J.H., Kim, I.Y., Kim, S.I., 2003. Evaluation of automated and semi-automated skull-stripping algorithms using similarity index and segmentation error. Comput. Biol. Med. 33 (6), 495–4507.

Li, T., Mirowitz, S.A., 2004. Fast multi-planar gradient echo MR imaging: impact of variation in pulse sequence parameters on image quality and artifacts. Magn. Reson. Imaging 22 (6), 807–814.

Lingford-Hughes, A.R., Davies, S.J., McIver, S., Williams, T.M., Daglish, M.R., Nutt, D.J., 2003. Addiction. Br. Med. Bull. 65, 209–222.

Mazziotta, J.C., Toga, A.W., Evans, A., Fox, P., Lancaster, J., 1995. A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM). Neuroimage 2 (2), 89–101.

Palmen, S.J., van Engeland, H., 2004. Review on structural neuroimaging findings in autism. J. Neural. Transm. 111 (7), 903–929.

Rehm, K., Schaper, K., Anderson, J., Woods, R., Stoltzner, S., Rottenberg, D., 2004. Putting our heads together: a consensus approach to brain/non-brain segmentation in T1-weighted MR volumes. Neuroimage 22 (3), 1262–1270.

Salat, D.H., Buckner, R.L., Snyder, A.Z., Greve, D.N., Desikan, R.S., Busa, E., Morris, J.C., Dale, A.M., Fischl, B., 2004. Thinning of the cerebral cortex in aging. Cereb. Cortex 14 (7), 721–730.

Sandor, S., Leahy, R., 1997. Surface-based labeling of cortical anatomy using a deformable atlas. IEEE Trans. Med. Imag. 16 (1), 41–54.

Schaper, K., Jarvis, T., Boesen, K., Rottenberg, D., 2005. Evaluation of brain grey-white ratios using automated tissue segmentation packages. Human Brain Mapping Conference, Toronto.

Schnack, H.G., van Haren, N.E., Hulshoff Pol, H.E., Picchioni, M., Weisbrod, M., Sauer, H., Cannon, T., Huttunen, M., Murray, R., Kahn, R.S., 2004. Reliability of brain volumes from multicenter MRI acquisition: a calibration study. Hum. Brain Mapp. 22 (4), 312–320.

Segonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. Neuroimage 22 (3), 1060–1075.

Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., 2001. Magnetic resonance image tissue classification using a partial volume model. Neuroimage 13 (5), 856–876.

Sheline, Y.I., Mittler, B.L., Mintun, M.A., 2002. The hippocampus and depression. Eur. Psychiatr. 17 (Suppl. 3), 300–305.

Simmons, A., Tofts, P.S., Barker, G.J., Arridge, S.R., 1994. Sources of intensity nonuniformity in spin echo images at 1.5 T. Magn. Reson. Med. 32 (1), 121–128.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imagi. 17 (1), 87–97.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155.

Smith, S.M., Brady, J.M., 1997. SUSAN—A new approach to low level image processing. Int. J. Comput. Vis. 23 (1), 45–78.

Stanislaw, H., Todorov, N., 1999. Calculation of signal detection theory measures. Behav. Res. Methods Instrum. Comput. 31 (1), 137–149.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15 (1), 273–289.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 2003. A unifying framework for partial volume segmentation of brain MR images. IEEE Trans. Med. Imag. 22 (1), 105–119.

Woods, R.P., 2003. Characterizing volume and surface deformations in an atlas framework: theory, applications, and implementation. Neuroimage 18 (3), 769–788.

Woods, R.P., Grafton, S.T., Holmes, C.J., Cherry, S.R., Mazziotta, J.C., 1998a. Automated image registration: I. General methods and intrasubject, intramodality validation. J. Comput. Assist. Tomogr. 22 (1), 139–152.

Woods, R.P., Grafton, S.T., Watson, J.D., Sicotte, N.L., Mazziotta, J.C., 1998b. Automated image registration: II. Intersubject validation of linear and nonlinear models. J. Comput. Assist Tomogr. 22 (1), 153–165.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation–maximization algorithm. IEEE Trans. Med. Imag. 20 (1), 45–57.