Neuroinform (2007) 5:235–245
DOI 10.1007/s12021-007-9003-9

Feasibility of Multi-site Clinical Structural Neuroimaging Studies of Aging Using Legacy Data

Christine Fennema-Notestine · Anthony C. Gamst · Brian T. Quinn · Jenni Pacheco · Terry L. Jernigan · Leon Thal · Randy Buckner · Ron Killiany · Deborah Blacker · Anders M. Dale · Bruce Fischl · Brad Dickerson · Randy L. Gollub

Received: 26 September 2007 / Accepted: 5 October 2007 / Published online: 13 November 2007

Abstract The application of advances in biomedical computing to medical imaging research is enabling scientists to conduct quantitative clinical imaging studies using data collected across multiple sites to test new hypotheses on larger cohorts, increasing the power to detect subtle effects. Given that many research groups have valuable existing (legacy) data, one goal of the Morphometry Biomedical Informatics Research Network (BIRN) Testbed is to assess the feasibility of pooled analyses of legacy structural neuroimaging data in normal aging and Alzheimer's disease.

Brad Dickerson and Randy L. Gollub contributed equally.

C. Fennema-Notestine (✉) · T. L. Jernigan
Department of Psychiatry, University of California—San Diego,
9500 Gilman Drive, # 0841,
La Jolla, CA 92093-0841, USA
e-mail: Fennema@UCSD.edu

C. Fennema-Notestine · T. L. Jernigan · A. M. Dale
Department of Radiology, University of California—San Diego,
La Jolla, CA, USA

C. Fennema-Notestine · T. L. Jernigan
Veterans Affairs San Diego Healthcare System,
San Diego, CA, USA

A. C. Gamst
Department of Biostatistics, University of California—San Diego,
La Jolla, CA, USA

A. C. Gamst · L. Thal · A. M. Dale
Department of Neurosciences, University of California—San
Diego, La Jolla, CA, USA

B. T. Quinn · J. Pacheco · R. Buckner · B. Fischl · B. Dickerson ·
R. L. Gollub
Athinoula A. Martinos Center for Biomedical Imaging—MGH/
NMR Center, Charlestown, MA, USA

R. Buckner
Department of Psychology, Harvard University,
Cambridge, MA, USA

R. Buckner · B. Fischl
Department of Radiology, Harvard Medical School,
Boston, MA, USA

R. Killiany
Department of Anatomy and Neurobiology,
Boston University School of Medicine,
Boston, MA, USA

D. Blacker · R. L. Gollub
Department of Psychiatry, Massachusetts General Hospital &
Harvard Medical School, Boston, MA, USA

D. Blacker
Department of Epidemiology, Harvard School of Public Health,
Cambridge, MA, USA

B. Fischl
Computer Science & Artificial Intelligence Laboratory, MIT,
Cambridge, MA, USA

B. Dickerson
Department of Neurology, Massachusetts General Hospital &
Harvard Medical School, Boston, MA, USA

The present study examined whether such data could be meaningfully reanalyzed as a larger combined data set by using rigorous data curation, image analysis, and statistical modeling methods; in this case, to test the hypothesis that hippocampal volume decreases with age and to investigate findings of hippocampal asymmetry. This report describes our work with legacy T1-weighted magnetic resonance (MR) and demographic data related to normal aging that have been shared through the BIRN by three research sites. Results suggest that, in the present application, legacy MR data from multiple sites can be pooled to investigate questions of scientific interest. In particular, statistical analyses suggested that a mixed-effects model employing site as a random effect best fits the data, accounting for site-specific effects while taking advantage of expected comparability of age-related effects. In the combined sample from three sites, significant age-related decline of hippocampal volume and right-dominant hippocampal asymmetry were detected in healthy elderly controls. These expected findings support the feasibility of combining legacy data to investigate novel scientific questions.

Keywords MRI · Hippocampus · Asymmetry · Image processing · Statistical modeling

Introduction

The application of advances in biomedical computing to medical imaging research, the primary goal of the Biomedical Informatics Research Network (BIRN), is enabling scientists to conduct quantitative clinical imaging studies using data collected across multiple sites to test new hypotheses on larger cohorts, thus increasing the power to detect subtle effects. The pooling of valuable data can dramatically increase the sample size of populations of interest, potentially leading to an increase in sensitivity that may reveal scientific findings relevant to further our knowledge of normal development and aging, as well as neurodegenerative and other neuropsychiatric disorders. Given that many research groups have valuable existing (legacy) data, one goal of the Morphometry BIRN (mBIRN) Testbed has been to assess the feasibility of pooled analysis of legacy structural imaging data in normal aging and Alzheimer's Disease (AD). That is, although it is expected that pooled analysis of multi-site data from planned prospective studies will be successful given explicitly tailored acquisition and calibration methods (Czanner et al. 2006, Han et al. 2006), many research groups have valuable legacy data sets that have been collected using the best available methods for their site at any given point in time. The ability to pool such legacy data may significantly advance future work, for example, by allowing retrospective analyses of the relationship between newly acquired genetic data and existing brain structure information. The present study aims to test the hypothesis that such legacy collections of clinical and structural MRI data from different sites can be meaningfully reanalyzed as a larger combined data set by using rigorous data curation and image analysis methods.

Our approach to assess feasibility relies on the careful replication of known findings of biological interest in well-characterized samples to better understand how best to combine, process, and analyze the pooled data. Three primary sources of variability that influence the analysis of the pooled data include cohort differences across sites, image analysis methods employed for volumetric methods (Jack et al. 1995), and MR scan acquisition parameters (e.g., scan platform, field strength, voxel dimensions, sequence specification). This study aims to control for the first two sources of variability and examine the feasibility of pooling data from healthy elderly controls that vary primarily on MR scan acquisition parameters. The data were culled from three similar studies through the mBIRN by the University of California, San Diego (UCSD), Massachusetts General Hospital/Brigham and Women's Hospital (MGH/BWH), and Washington University (WashU). The resultant large sample was then analyzed with the same image processing methods, available in the FreeSurfer software (http://surfer.nmr.mgh.harvard.edu/; Dale and Sereno 1993, Dale et al. 1999, Fischl et al. 1999, 2002), to control for any differences related to image analysis. The use of this automated cerebral segmentation software considerably reduces the substantial amount of manual interaction required to process image data, which would be prohibitive for these larger cohort studies, and results have been shown to be similar to manually-defined regions of interest (e.g., Fischl et al. 2002). The third source of variability, related to MR scan acquisition, is an inherent difference across sites, particularly with respect to legacy data; even in prospective multi-site studies, differences in scanner hardware and software typically remain unavoidable (e.g., Mueller et al. 2005). Thus, this study may also provide direction for prospective acquisition of data from multiple sites that will require comparable data curation and analysis methods.

The present aim is to determine the feasibility of detecting clinically meaningful neuroimaging findings in a combined sample of normal older individuals with legacy MR data from three mBIRN sites in preparation for investigations related to AD. We focus on hippocampal volume as our primary measure of interest given its role in normal aging, mild cognitive impairment (MCI), and AD. Hippocampal volume has been shown to decline with normal aging (Mu et al. 1999; Jernigan et al. 2001a; Allen

et al. 2005; Walhovd et al. 2005; van de Pol et al. 2006) and in AD (Jack et al. 2002; van de Pol et al. 2006), and this region has been an intense area of study in the search for an in vivo biomarker in individuals at risk for AD (Csernansky et al. 2005; Jack et al. 2005). Such a biomarker could be used in the diagnosis of AD and as a marker of effectiveness in therapeutic trials (Jack et al. 2003; Kantarci and Jack 2003).

Examination of hippocampal right/left asymmetry has also been of interest, particularly as it relates to neurodegenerative disorders, although the published results have been somewhat less consistent and have relied on a variety of methods (Jack et al. 1995; Pedraza et al. 2004; Raz et al. 2004). Several studies suggest that there is little or no asymmetry in healthy adults and that this does not change with normal aging (Mu et al. 1999; Raz et al. 2004). Other published findings, including a meta-analysis, support a right dominant hippocampal asymmetry in healthy adults (Pedraza et al. 2004). The existence of such an asymmetry is of even greater interest given reports of changes in asymmetry in individuals at elevated risk for and diagnosed with AD (Soininen et al. 1994; Soininen et al. 1995; Barnes et al. 2005) that may be related to concomitant cognitive changes (Finton et al. 2003).

In the current study, we explicitly tested the hypotheses that (1) hippocampal volume, as measured by our subcortical segmentation algorithm, shows the expected age-related volume decline; and (2) normal elderly controls demonstrate a consistent right dominant hippocampal asymmetry and the magnitude of this asymmetry may not change with normal aging. We also describe site effects and explore various statistical models that best fit the multi-site legacy data to help direct future work.

Methods

Legacy Data Cohort Legacy T1-weighted MR and demographic data from 133 healthy elderly control (HEC) participants were shared through the BIRN by University of California, San Diego (UCSD) (TL Jernigan; L Thal; D Salmon), Massachusetts General Hospital and Brigham and Women's Hospital (MGH/BWH) (M Albert; D Blacker; R Killiany), and Washington University (WashU) (R Buckner; J Morris). Data from the UCSD and WashU were collected as part of on-going Alzheimer's Disease Research Center (ADRC) studies; data from MGH/BWH were collected through an on-going study of prodromal Alzheimer's disease. Previous within-site published studies based on these samples include: UCSD (Jernigan et al. 2001a,b; Murphy et al. 2003; Jernigan and Fennema-Notestine 2004; Jernigan and Gamst 2005; Fennema-Notestine et al. 2006); MGH/BWH (Killiany et al. 2000, 2002); and WashU (Buckner et al. 2004, 2005, Fotenos et al. 2005, Head et al. 2005). Data from individuals over 60 years of age were included in this investigation. All HEC were considered neurologically and neuropsychologically normal at the time of scan as defined by project neurologists, psychiatrists, and clinical neuropsychologists. Individuals with a history of severe head injury, neurological illness (e.g., epilepsy), alcoholism, or psychiatric illness were excluded from the study.

Site cohorts were similar on education and Mini-Mental State Examination (MMSE) (Folstein et al. 1975) (all $t<1.5$, all $p>0.05$; Table 1). Site cohorts were similar on age range (Table 1), although MGH/BWH had a small but statistically significantly lower mean age in years relative to UCSD ($t(82)=2.5$, $p<0.05$) and WashU ($t(83)=2.5$, $p<0.05$). UCSD and WashU cohorts were not significantly different on age ($t(95)<1.0$, $p>0.05$). The MGH/BWH site had disproportionately fewer males relative to the other sites.

Pulse Sequence All MR data for this study were sagittal acquisition T1-weighted images collected during the mid- to late-1990s as follows:

UCSD: GE 1.5T Signa, gradient-echo (SPGR), TR=24 ms, TE=5 ms, flip angle=45°, FOV=24 cm, contiguous 1.2 mm sections, 256×192 matrix, NEX=2; single T1 acquisition.
MGH/BWH: GE 1.5T Signa, gradient-echo (SPGR), TR=35 ms, TE=5 ms, flip angle=45°, FOV=22 cm, contiguous 1.5 mm sections, 256×256 matrix, NEX=1; single T1 acquisition.

Table 1 Cohort demographics by site

Site	*N*	Age (mean, sd)	Gender (number, %)	Education (mean, sd)	MMSE (mean, sd)
MGH/BWH	36	71.9 (4.8) range 64–85	22F/14M 61%/39%	14.4 (2.3) range 10–19	29.3 (1.1) range 28–30
UCSD	48	74.6 (5.0) range 63–87	26F/22M 54%/46%	14.8 (3.3) range 6–20	29.4 (0.8) range 27–30
WashU	49	75.7 (7.9) range 62–89	24F/25M 49%/51%	14.5 (2.8) range 8–20	29.1 (1.1) range 26–30

Wash U: Siemens 1.5T Magneton, MP-RAGE, TR=9.7 ms, TE=4 ms, TI=20 ms, flip angle=10°, FOV=24 cm, contiguous 1.25 mm sections, 256×256 matrix, NEX=2; 4 T1 acquisitions were averaged.

Image Processing All T1 data were processed at MGH with a common image processing path that included bias correction, skull-stripping, registration to a common atlas space, and the application of an atlas-based FreeSurfer subcortical segmentation (Fig. 1) (Fischl et al. 2002). In addition, intracranial volume was estimated to assist in controlling for individual differences in head size (Buckner et al. 2004), as described below.

To correct image bias, FreeSurfer employs the Non-parametric Non-uniform intensity Normalization method (N3) (Sled et al. 1998), which uses a locally adaptive bias correction algorithm to bring intensities into global alignment. This method was chosen for its applicability to un-skull-stripped image sets and for its excellent performance compared with other bias correction methods (Arnold et al. 2001). Removal of non-brain tissue, or skull-stripping, is performed by a hybrid watershed algorithm (HWA) (Segonne et al. 2004), a method that has been shown to be highly sensitive to retaining brain tissue (Fennema-Notestine et al. 2006). This HWA method is a hybrid of a watershed algorithm (Hahn and Peitgen 2000) and a deformable surface model (Dale et al. 1999) that was designed to be conservatively sensitive to the inclusion of brain tissue. In general, watershed algorithms segment images into connected components, using local optima of image intensity gradients. HWA uses a watershed algorithm that is solely based on image intensities; the algorithm, which operates under the assumption of the connectivity of white matter, segments the image into brain and non-brain

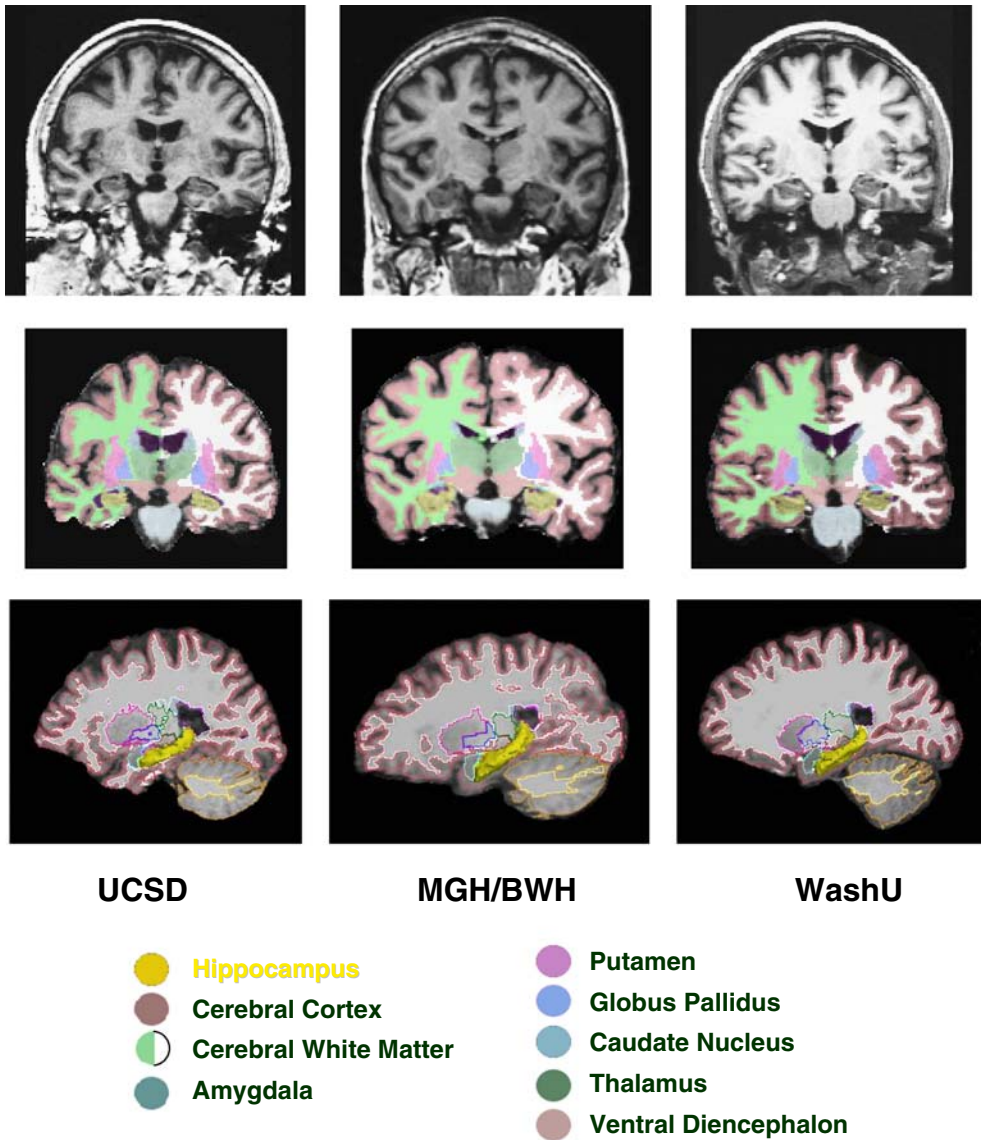

Fig. 1 Examples of automated results for a healthy elderly female control from each site. *Top row:* coronal section from original T1 volume; *Middle row*: coronal section from FreeSurfer subcortical segmentation (Fischl et al. 2002); *Bottom row*: sagittal section with 3D hippocampal model in 3DSlicer (http://www.slicer.org). Represented participants age ranged from 71 to 72 years; each had a perfect MMSE score of 30. Note the contrast differences between SPGR data (UCSD and MGH/BWH) and MPRAGE data (Wash U)

components. A deformable surface-model is then applied to locate the boundary of the brain in the image. Initial affine registration with Talairach space and high dimensional nonlinear volumetric alignment to the atlas were used (Fischl et al. 2004b); these processes were designed to be insensitive to pathology and maximize accuracy of final segmentation.

The FreeSurfer subcortical segmentation procedure assigns a neuroanatomical label to each voxel based on probabilistic information automatically estimated from an atlas. The template atlas employed herein was created with a separate cohort of 40 subjects from one of the sites studied, WashU, consisting of: young controls (n=10), middle-aged controls (n=10), older controls (n=10), and individuals with AD (n=10). Manual labeling was done by the Center for Morphometric Analysis (http://www.cma.mgh.harvard.edu/) and the atlas validated as in Fischl et al. (2002). The reliability of manual and automated segmentation is discussed in detail in Fischl et al. (2002). The labeling of each point in space is achieved by finding the segmentation for each dataset that maximizes the probability of input given the prior probabilities from the atlas. The probability of a class at each point is computed as the probability that the given class appeared at that location in the training set, modulated by the probability of the surrounding configuration of labels in the six cardinal directions, times the likelihood of getting the subject-specific measured intensity value from that class. An initial segmentation is generated by assigning each point to the class for which the probability is greatest, ignoring the probability of the neighborhood configuration. Given this segmentation, the neighborhood function is used to recompute the class probabilities. The data set is resegmented based on this new set of class probabilities. This is repeated until the segmentation does not change. This procedure has been shown to generate labels that are statistically indistinguishable from those of manual raters (Fischl et al. 2002). In this automated application, data were reviewed for gross technical errors only. On-going development work on the segmentation algorithm may further reduce differences related to acquisition sequence (Han and Fischl 2006).

We examined the influence of the probabilistic atlas on the quantification of hippocampal asymmetry. We sought to determine whether hippocampal asymmetry for a given individual would be a true reflection of their image data or if the priors introduced by the atlas would influence the final result. First, we examined the manually-created data used in the atlas to quantify the degree of asymmetry in the atlas; one-third of the 40 cases had larger left than right hippocampal volumes, two-thirds demonstrated a right-dominant asymmetry. The average left hippocampal volume was 3,521 mm^3 and right was 3,574 mm^3 in these 40 cases, reflecting a slight (~53 mm^3) right-dominance in the hippocampus. The range of asymmetry in these cases was −669 mm^3 (left-dominant) to 531 mm^3 (right-dominant). Subsequently, we selected two cases with strong right-dominant hippocampal asymmetry and applied our algorithm to the original volume and to the same volume with left-right reversed. The reversed volumes, then, represented manufactured strongly "left-dominant" asymmetry cases. We hypothesized that if the atlas was driving a right-dominant asymmetry, then the reversed data should also result in a right-dominant asymmetry, despite known, manufactured left-dominance. The original volume right-dominant asymmetry in these two cases was: Subject #1: L 2,627 mm^3, R 3,195 mm^3; Subject #2: L 3,137 mm^3, R 3,771 mm^3. The results from the reversed volumes demonstrated preservation of the manufactured left-dominant asymmetry; Subject #1REV: "L" 3,023 mm^3, "R" 2,688 mm^3; Subject #2REV: "L" 3,685 mm^3, "R" 3,218 mm^3. However, there was an attenuation of the asymmetry indicating that the atlas does in fact introduce a slight bias towards right hippocampal dominance. This work supports the use of the algorithm to provide realistic asymmetry results in our cohort; however, further investigation of this bias is warranted.

In addition, based on Buckner et al. (2004), we estimated total intracranial volume (eTIV) to compare across sites and to employ as a control for differences in head size where appropriate. A scaling factor was derived from a linear registration between each subject's data and the atlas within the FreeSurfer algorithm (Fischl et al. 2002). This scaling factor was used to calculate eTIV to control for individual differences in head size. This intracranial estimate was validated through comparisons with manual designations of ICV at each site through various methods. The WashU site validated the original eTIV method (Buckner et al. 2004); co-registered T1 and T2 weighted volumes were used to manually outline TIV on every tenth sagittal section for 147 individuals. These manual measurements were significantly correlated with the eTIV measures across a group of young and elderly normal controls, and AD participants (r=0.93). At the UCSD site, these eTIV measures were compared to intracranial volumes derived from tissue segmented PD-T2 FSE sequences of elderly controls (a subset of the cohort presented in this study) and AD participants (total n=56). In this case, the FSE intracranial volumes were well correlated with eTIV (r=0.87), with similar strength within the elderly controls (r=0.87, n=30) and AD participants (r=0.90, n=23).

Statistical Methods Combining cohorts from multiple sites will increase power due to the increase in sample size, although it may be important to appropriately model the likely increase in variability related to site-specific differences in the samples (e.g., age distribution) and MR scan acquisition parameters. Linear regression analyses were

employed to predict hippocampal volumes (left, right, and bilateral) or hippocampal asymmetry ratios (right/left hippocampal volume) with age. Findings for left, right, and bilateral hippocampal volumes were similar unless otherwise mentioned; statistics were reported to be inclusive of comparisons for each left, right, and bilateral volumes (e.g., all $t>2.0$, $p<0.05$). Since in the hippocampus variability decreased with increasing age, hippocampal volumes and eTIV were log-transformed to stabilize the variance in volume by age. Age was centered at its mean to eliminate the correlation between estimated slope and intercept in the fitted regression models. Regression models also considered log-transformed eTIV and proportionalized log-transformed hippocampal volumes when relevant for controlling for differences in head size. All comparisons were carried out at the conventional $\alpha=0.05$ level.

Within the linear regression analyses, we considered the strengths and weaknesses of various statistical models for the prediction of hippocampal variables from age with data combined from multiple sites. If there are no significant site effects, pooling the data without considering site as a factor would be preferable and would result in increased power based simply on increased sample size. Should significant site effects be demonstrated, combining the data will still increase power, however, the proper modeling of these differences is critical to optimizing the estimated average effects. We considered three different regression models and examined the best fit (likelihood ratio test) to these data:

Pooled Model Directly pools data across sites. This model assumes that site is not a relevant factor and does not model site-specific effects:

$$y_{ij} = a + bx_{ij} = e_{ij}$$

where i is the index for subjects, j is the index for sites, and a and b are the same coefficients for each site.

Fixed Effects Model Models data with site as a fixed effect. This model assumes that sites are completely different:

$$y_{ij} = a_j + b_j x_{ij} + e_{ij}$$

where i is the index for subjects, j is the index for sites, and there are six parameters, one (a, b) pair per site.

Mixed Effects Model Employs a mixed effects model with site as a random effect. Assumes that site-specific differences exist, although sites are comparable in some sense (e.g., the relationship between the age and hippocampus will be similar from site to site):

$$y_{ij} = a_j + b_j x_{ij} + e_{ij}$$

where i is the index for subjects, j is the index for site, and there are five parameters: (a_j, b_j); the (a_j, b_j) are assumed to be sampled independently from a Gaussian distribution with mean (a, b) (two parameters) and covariance matrix **S** (three parameters).

Results

Hippocampal segmentations (Fig. 1) were qualitatively reviewed primarily for major technical errors in image processing. Technical errors included failures in automated skull-stripping and failure in application of the segmented atlas. Skull-stripping failures were correctable with either manual editing to remove or replace tissue on a few sections or with adjustment of watershed parameters and re-running the data through the skull-stripping step. Minimal manual editing related to skull-stripping was required on occasion; this editing was performed by a single person. The automated segmentation failed on five cases due in large part to anatomical cases in which extreme ventricular size could likely not be mapped to the atlas; these cases

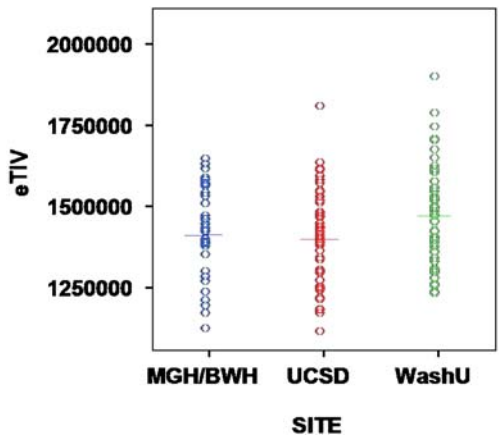

Fig. 2 Estimated total intracranial volume (eTIV; in mm^3) by site. Each *circle* represents an individual data point; *horizontal lines* represent mean volume

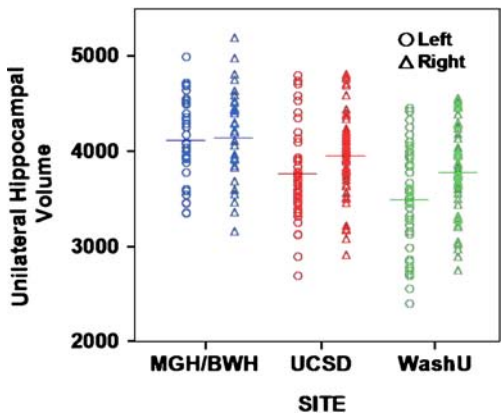

Fig. 3 Raw left and right hippocampal volume (in mm^3) by site. Hippocampal asymmetry would be represented by right divided by left volumes. Each *circle* or *triangle* represents an individual data point; *horizontal lines* represent mean volume

Table 2 Values for raw volumes (in mm^3) for left and right hippocampus (Hpc), hippocampal asymmetry (right/left hippocampal volume), and estimated total intracranial volume (eTIV) by site

Site	*n*	Left Hpc (mean, sd) range	Right Hpc (mean, sd) range	Hpc Asymmetry Ratio (R/L) (mean, sd) range	eTIV (mean, sd) range
MGH/BWH	36	4,113 (425)	4,144 (456)	1.01 (0.08)	1,411,060 (143,706)
		3,330–4,966	3,130–5,161	0.86–1.17	1,120,189–1,640,110
UCSD	48	3,763 (484)	3,959 (455)	1.06 (0.08)	1,399,341 (143,958)
		2,678–4,778	2,893–4,789	0.89–1.21	1,109,464–1,805,284
WashU	49	3,489 (546)	3,780 (471)	1.09 (0.09)	1,470,911 (153,180)
		2,386–4,436	2,727–4,533	0.92–1.34	1,225,932–1,894,744

were excluded. Subsequently, the segmented volumes were reviewed for gross errors related to hippocampal volume measures. Only one volume differed from the mean by greater than two standard deviations on the hippocampal measure, due to a technical error, and this case was excluded.

Modeling eTIV Site- and Age-related Effects Linear regression models supported significant site differences in eTIV, with WashU demonstrating larger eTIV relative to MGH/BWH and UCSD (t=2.17, p<0.05), whereas MGH/BWH and UCSD did not differ significantly (t<1.0; Table 2, Fig. 2). The two sites with SPGR data (UCSD and MGH/BWH) were more similar on mean eTIV than the MPRAGE data site (WashU). There was no evidence of an age-related effect on the eTIV measure at any site (all t< 1.0, p>0.05); in other words, this measure was relatively flat across the age range of the cohort studied. Including Site as a fixed (Fixed Effects Model) or random effect (Mixed Effects Model) did not fit the data better than the Pooled Model (likelihood ratio test = 1.8, p>0.05), suggesting that the heterogeneity in head-size across sites is insufficient to force the use of a mixed effects model. Hippocampal volumes were proportionalized to eTIV in subsequent analyses. Accounting for differences in head-size did not account for all of the observed site differences.

Modeling Hippocampal Volume across Normal Aging Linear regression models within each site individually revealed significant age-related decline in hippocampal volume at two of the three sites (MGH/BWH and WashU p<0.05, but not UCSD p>0.05). Possible explanations for the lack of a significant age-related effect at UCSD include statistical factors, such as sampling bias or inadequate power, or cohort-specific factors. After combining the data across sites, all models revealed significant age-related hippocampal volume decline as described below. Thus, the biologic effect of aging on hippocampal volume was reliably detected (regardless of which model was used) in the pooled sample despite the lack of significance at one of the sites.

Although in the Pooled Model (pooling data without site as a factor) age significantly predicted hippocampal volume decline (all t<−5.4, p<0.0001), the Fixed Effects Model demonstrated that there was significant site-to-site variability in hippocampal volume (Table 2, Fig. 3). WashU had the smallest hippocampal volumes (t<−3.0, p<0.05); UCSD and MGH/BWH did not differ significantly for right (t<1.0, p>0.05) and bilateral (t=1.6, p=0.11) hippocampal measures, although UCSD left hippocampal volumes were smaller on average relative to MGH/BWH (t> 1.6, p<0.05). In addition, sites differed for the effect of age on hippocampal volume (Figs. 4 and 5); within the UCSD sample, there was no significant age-related effect on hippocampal volume, whereas MGH/BWH and WashU both evidenced significantly smaller hippocampal volumes with age (p<0.05). MGH/BWH demonstrated the steepest decline with age relative to WashU and UCSD. Including site as a Fixed Effect fit the data better than the Pooled Model (likelihood ratio test, t=23.3, p<0.0001), supporting the inclusion of site-specific effects in the model.

While the sites were clearly different in some respects, we expected the relationship between the primary variables to be similar from site to site, and we would like to borrow strength from this assumption with a mixed effects model. That is, the effect of normal aging in humans should be similar across sites. The Mixed Effects Model takes this into account and revealed similar age-related hippocampal volume decline with age (all t≥2.5, p<0.05). This model provided a better fit relative to the Pooled Model (likelihood ratio test>13.0, p≤0.005). There were significant differences in the coefficients by site and forcing all coefficients to be the same introduced significant bias; site differences were large enough that the mean squared error of the Pooled analysis was larger than that for the Mixed Effects analysis.

For the hippocampus, heterogeneity between sites is important, and the Mixed Effects Model fits the data best and has a significant power advantage over modeling all sites separately (i.e., the Fixed Effects Model). When estimating the mean of such a multi-dimensional Gaussian distribution, the mean squared error of an estimator (in this

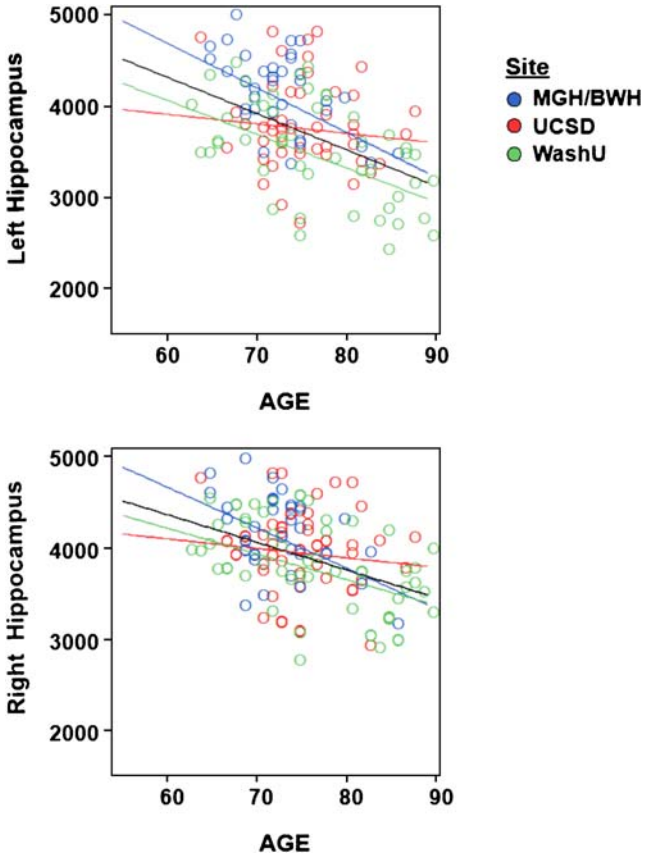

Fig. 4 Raw hippocampal volume (in mm^3) across normal aging for left (*top*) and right (*bottom*) hemisphere. Each *circle* represents a single subject; *color* represents site; the *black line* represents the results of linear regression including all sites. Linear regression within each site is represented by a line in the appropriate legend color

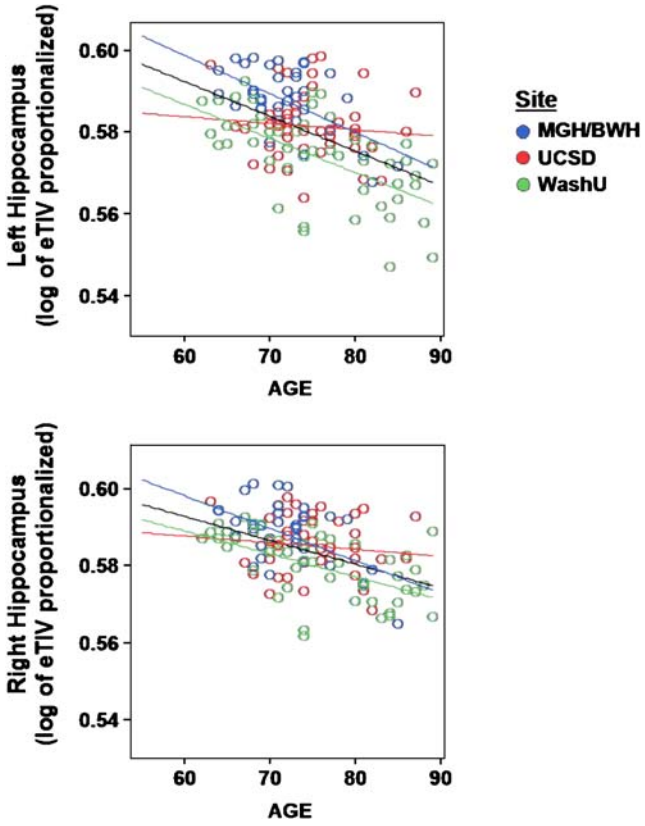

Fig. 5 Log transformation of hippocampal volume proportionalized by eTIV across normal aging for left (*top*) and right (*bottom*) hemisphere. Each *circle* represents a single subject; *color* represents site; the *black line* represents the results of linear regression including all sites. Linear regression within each site is represented by a line in the appropriate legend color

study, the least-squares estimates) that shrinks the sample means toward zero is uniformly smaller than the mean squared error of the sample mean itself (Stein 1981). The Mixed Effects Model has a shrinkage property similar to this and will essentially always outperform an all sites separately (or Fixed Effects) analysis.

Modeling Hippocampal Asymmetry Linear regression models within each site individually revealed significant right dominant hippocampal asymmetry (right greater than left; Table 2; Fig. 3) at WashU (t=7.3, p<0.001) and UCSD (t=4.9, p>0.001), but not at MGH/BWH (t<1.0, p>0.05); these findings were not different between genders (ps>0.05). After combining the data across sites, the Pooled Model revealed a significant hippocampal right dominant asymmetry overall (t=7.4, p<0.0001), and the asymmetry ratio appeared to change significantly with age (t=2.5, p<0.05). A direct interpretation of these Pooled Model results would suggest that the right dominant asymmetry increases in size with age. This Pooled Model, however, may lead to the wrong conclusion in this case, as suggested below with the models including site as a separate factor. As mentioned, the MGH/BWH cohort is slightly younger on average, and this age-related difference in sites should be accounted for in the interpretation of the findings. In the Fixed Effect Model, the right dominant asymmetry remained significant (t=5.1, p< 0.0001), whereas there was no significant change in asymmetry with age. It appears that the site-to-site variability in both age and asymmetry may have contributed to the Pooled Model findings of age-related changes in asymmetry. Modeling site effects may reduce the risk of spurious effects, although it cannot entirely eliminate them.

Employing instead the Mixed Effects Model, the asymmetry remained significant (t=2.4, p<0.05) and again there was no difference in the hippocampal asymmetry ratio across age (t=1.2, p>0.05). This supports the idea that the apparent age-related effects suggested by the Pooled Model may have been linked to site-specific differences in age. This Mixed Effects Model fit the data significantly better than the Pooled Model (t=8.8, p<0.05).

Discussion

This structural neuroimaging study demonstrates a viable analysis path for combining legacy MR data from multiple sites to investigate questions of scientific interest, preserving the ability to detect subtle effects despite variability in data acquisition methods and subject samples across sites. The investigation culled cohorts that were comparable with respect to demographic criteria and employed a single image processing path, which performed similarly across sites. The critical examination of statistical modeling approaches suggested that a mixed-effects model, employing site as a random effect, best fit the data and preserved power unnecessarily lost with fixed effects model. This Mixed Effects model accounted for site-specific effects while taking advantage of the expected comparability of age-related effects and similar image processing techniques across sites. The neuroimaging findings replicate previous findings of decline in hippocampal volume across the age range (Mu et al. 1999; Jernigan et al. 2001a; Allen et al. 2005; Walhovd et al. 2005; van de Pol et al. 2006) and provide support for right dominant hippocampal asymmetry in healthy elderly controls (Pedraza et al. 2004).

The use of the same morphometric analysis path for all sites' data was intended to reduce between-site variability by constraining gray matter segmentation and boundaries of the defined hippocampal region. Despite the application of this analysis path to relatively similar T1-weighted images, significant site differences in the magnitude of measured hippocampal volume and eTIV remained. This between-site variability for relatively comparable data is a major concern for combining MR data from multiple sites; for example, we expect that the average hippocampal volume and the effects of aging on hippocampal volume will be similar for people regardless of their city of residence. In the present study, the site differences are likely due to differences in data acquisition (e.g., vendor, pulse sequence, etc.) and some small cohort differences as well (e.g., slightly younger cohort for MGH/BWH). Our investigation suggests that a Mixed Effects statistical model best accounts for these differences between sites and provides a meaningful method for analyzing multi-site data. Recent prospective multi-site studies, such as the Alzheimer's Disease Neuroimaging Initiative (Mueller et al. 2005), will provide new avenues to further explore multi-site MRI analyses with carefully designed protocols that are more similar across vendors to reduce site-related effects.

While the current combined sample size is not dramatically larger than other published work, the present study nevertheless provides support for combining data from multiple sites. The analyses confirm previous reports of age-related changes in hippocampal volume over a narrower age range than many of these previous studies (e.g., Jernigan et al. 2001a; Allen et al. 2005; Walhovd et al. 2005). It is easier, in general, to demonstrate a significant age effect over a wide age range, but for most studies it is much more difficult to estimate accurately from the data what the age effect is within a narrower age-range, such as in individuals between 60 and 89 as presented here. Studies like this, drawing from multiple cohorts, could permit one to define more clearly the precise shape of the age curves, which are likely not truly linear (e.g., Jernigan et al. 2001a; Allen et al. 2005, Jernigan and Gamst 2005; Walhovd et al. 2005), and if the cohorts were sufficiently similarly constructed, to assess the generality of these shape attributes. Within-subject longitudinal studies, particularly with respect to potential change in asymmetry, would be the gold standard, although such studies spanning many years are rare. Nevertheless, the approach applied herein could be modified to apply to pooled longitudinal studies to address just this question.

The present work focused on the hippocampal volumetric measure for data collected at the same field strength. The applicability of this approach to other regions and to the combination of data from different field strengths should be assessed. For example, the combination of SPGR and MPRAGE data from other brain regions, such as those that lie on edges near cerebrospinal fluid (e.g., the caudate nucleus), may present additional challenges. Furthermore, with the increase of collaborative neuroimaging efforts worldwide, through projects such as the BIRN (http://www.nbirn.net) and the ADNI (Mueller et al. 2005), work has focused on refinements in the segmentation algorithm that may further decrease site-specific variance due to differences in acquisition sequence (Fischl et al. 2004a), and on paradigms to prospectively collect data implementing multi-site calibration tools (Jovicich et al. 2005). Improvements will include more extensive investigations of potential asymmetry bias in atlas applications, untangling methods performance from true asymmetry, and extending such studies to include more extensive data on laterality. Future work will assess the generalizability of this analytic approach to the related AD cohorts, which will undoubtedly be more variable across sites.

Acknowledgments This research was supported by a grant (#U24 RR021382) to the Morphometry Biomedical Informatics Research Network (BIRN, http://www.nbirn.net), that is funded by the National Center for Research Resources at the National Institutes of Health, U.S.A. Additional support was provided by: the University of California, San Diego, Department of Medicine; San Diego ADRC NIA P50 AG05131; Washington University, St. Louis ADRC NIA P50 AG05681; NIA R01 AG12674, R01 AG06849, PO1 AG04953, and P01 AG03991; a Research Enhancement Award Program and VA Merit Review grant from the Department of Veterans Affairs Medical Research Service; Howard Hughes Medical Institute; NCRR R01 RR16594-01A1, M01 RR00827, P41 RR14075, and P41 RR13642; Mental Illness and Neuroscience Discovery (MIND) Institute; NINDS

R01 NS052585-01; NIH Roadmap for Medical Research U54 EB005149; and NIBIB R01 EB002010. Anders M. Dale is a founder and holds equity in CorTechs Labs, Inc, and also serves on the Scientific Advisory Board. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

Work based on these MRI data have been published separately for studies performed within each site locally, including: UCSD (Jernigan et al. 2001a, b; Murphy et al. 2003; Jernigan and Fennema-Notestine 2004; Jernigan and Gamst 2005; Fennema-Notestine et al. 2006); MGH/BWH (Killiany et al. 2000, Killiany et al. 2002); and WashU (Buckner et al. 2004, 2005, Fotenos et al. 2005, Head et al. 2005). Preliminary findings related to the combined data analysis were presented at the Society for Neuroscience 2005 meeting (Fennema-Notestine et al. 2005); work related to the combined data has not been published elsewhere.

References

Allen, J. S., Bruss, J., Brown, C. K., & Damasio, H. (2005). Normal neuroanatomical variation due to age: The major lobes and a parcellation of the temporal region. *Neurobiology Aging, 26*(9), 1245–1260 (discussion 1279–1282).

Arnold, J. B., Liow, J. S., Schaper, K. A., Stern, J. J., Sled, J. G., Shattuck, D. W., et al. (2001). Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects. *NeuroImage, 13*(5), 931–943.

Barnes, J., Scahill, R. I., Schott, J. M., Frost, C., Rossor, M. N., & Fox, N. C. (2005). Does Alzheimer's disease affect hippocampal asymmetry? Evidence from a cross-sectional and longitudinal volumetric MRI study. *Dementia and Geriatric Cognitive Disorders, 19*(5–6), 338–344.

Buckner, R. L., Head, D., Parker, J., Fotenos, A. F., Marcus, D., Morris, J. C., et al. (2004). A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: Reliability and validation against manual measurement of total intracranial volume. *Neuroimage*, *23*(2), 724–738.

Buckner, R. L., Snyder, A. Z., Shannon, B. J., LaRossa, G., Sachs, R., Fotenos, A. F., et al. (2005). Molecular, structural, and functional characterization of Alzheimer's disease: Evidence for a relationship between default activity, amyloid, and memory. *Journal of Neuroscience, 25*(34), 7709–7717.

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

Czanner, S., Han, X., Pacheco, J., Wallace, S., Busa, E., van der Kouwe, A., et al. (2006). *Test–retest reliability assessment for longitudinal MRI studies: Effects of MRI system upgrade on morphometric analysis of structural MRI data*. 12th Annual Organization for Human Brain Mapping Meeting, Florence, Italy.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage, 9*(2), 179–194.

Dale, A. M., & Sereno, M. I. (1993). Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: A linear approach. *Journal of Cognitive Neuroscience, 5*, 162–176.

Fennema-Notestine, C., Gollub, R., Fischl, B., Quinn, B., Pacheco, J., Gamst, A., et al. (2005). Feasibility of multi-site clinical structural neuroimaging studies of legacy data: Aging and Alzheimer's disease. *Society for Neuroscience* (Abstract).

Fennema-Notestine, C., Ozyurt, I. B., Clark, C. P., Morris, S., Bischoff-Grethe, A., Bondi, M. W., et al. (2006). Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: Effects of diagnosis, bias correction, and slice location. *Human Brain Mapping*, *27*(2), 99–113.

Finton, M. J., Lucas, J. A., Rippeth, J. D., Bohac, D. L., Smith, G. E., Ivnik, R. J., et al. (2003). Cognitive asymmetries associated with apolipoprotein E genotype in patients with Alzheimer's disease. *Journal of the International Neuropsychological Society, 9*(5), 751–759.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron, 33*(3), 341–355.

Fischl, B., Salat, D. H., van der Kouwe, A. J., Makris, N., Segonne, F., Quinn, B. T., et al. (2004a). Sequence-independent segmentation of magnetic resonance images. *NeuroImage, 23*(Suppl 1), S69–S84.

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage, 9*(2), 195–207.

Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D. H., et al. (2004b). Automatically parcellating the human cerebral cortex. *Cerebral Cortex, 14*(1), 11–22.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*(3), 189–198.

Fotenos, A. F., Snyder, A. Z., Girton, L. E., Morris, J. C., & Buckner, R. L. (2005). Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. *Neurology, 64*(6), 1032–1039.

Hahn, H. K., & Peitgen, H.-O. (2000). The skull stripping problem in MRI solved by a single 3D watershed transform. *Proc. MICCAI, LNCS 1935*, 134–143.

Han, X., & Fischl, B. (2006). *Intensity renormalization for improved brain MR image segmentation across scanner platforms*. 12th Annual Organization for Human Brain Mapping Meeting, Florence, Italy.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., et al. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage, 32*(1), 180–194.

Head, D., Snyder, A. Z., Girton, L. E., Morris, J. C., & Buckner, R. L. (2005). Frontal-hippocampal double dissociation between normal aging and Alzheimer's disease. *Cerebral Cortex, 15*(6), 732–739.

Jack, C. R. Jr., Dickson, D. W., Parisi, J. E., Xu, Y. C., Cha, R. H., O'Brien, P. C., et al. (2002). Antemortem MRI findings correlate with hippocampal neuropathology in typical aging and dementia. *Neurology, 58*(5), 750–757.

Jack, C. R. Jr., Shiung, M. M., Weigand, S. D., O'Brien, P. C., Gunter, J. L., Boeve, B. F., et al. (2005). Brain atrophy rates predict subsequent clinical conversion in normal elderly and amnestic MCI. *Neurology, 65*(8), 1227–1231.

Jack, C. R. Jr., Slomkowski, M., Gracon, S., Hoover, T. M., Felmlee, J. P., Stewart, K., et al. (2003). MRI as a biomarker of disease progression in a therapeutic trial of milameline for AD. *Neurology, 60*(2), 253–260.

Jack, C. R. Jr., Theodore, W. H., Cook, M., & McCarthy, G. (1995). MRI-based hippocampal volumetrics: Data acquisition, normal ranges, and optimal protocol. *Magnetic Resonance Imaging, 13*(8), 1057–1064.

Jernigan, T. L., Archibald, S. L., Fennema-Notestine, C., Gamst, A. C., Stout, J. C., Bonner, J., et al. (2001a). Effects of age on

tissues and regions of the cerebrum and cerebellum. *Neurobiology Aging, 22*(4), 581–594.

Jernigan, T. L., & Fennema-Notestine, C. (2004). White matter mapping is needed. *Neurobiology Aging, 25*(1), 37–39.

Jernigan, T. L., & Gamst, A. C. (2005). Changes in volume with age–consistency and interpretation of observed effects. *Neurobiology Aging, 26*(9), 1271–1274 (discussion 1275–1278).

Jernigan, T. L., Ostergaard, A. L., & Fennema-Notestine, C. (2001b). Mesial temporal, diencephalic, and striatal contributions to deficits in single word reading, word priming, and recognition memory. *Journal of the International Neuropsychological Society, 7*(1), 63–78.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., et al. (2005). Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *NeuroImage, 30*(2), 436–443.

Kantarci, K., & Jack, C. R. Jr. (2003). Neuroimaging in Alzheimer disease: an evidence-based review. *Neuroimaging Clinics of North America, 13*(2), 197–209.

Killiany, R. J., Gomez-Isla, T., Moss, M., Kikinis, R., Sandor, T., Jolesz, F., et al. (2000). Use of structural magnetic resonance imaging to predict who will get Alzheimer's disease. *Annals of Neurology, 47*(4), 430–439.

Killiany, R. J., Hyman, B. T., Gomez-Isla, T., Moss, M. B., Kikinis, R., Jolesz, F., et al. (2002). MRI measures of entorhinal cortex vs hippocampus in preclinical AD. *Neurology, 58*(8), 1188–1196.

Mu, Q., Xie, J., Wen, Z., Weng, Y., & Shuyun, Z. (1999). A quantitative MR study of the hippocampal formation, the amygdala, and the temporal horn of the lateral ventricle in healthy subjects 40 to 90 years of age. *American Journal of Neuroradiology, 20*(2), 207–211.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., et al. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America, 15*(4), 869–877, xi–xii.

Murphy, C., Jernigan, T. L., & Fennema-Notestine, C. (2003). Left hippocampal volume loss in Alzheimer's disease is reflected in performance on odor identification: a structural MRI study. *Journal of the International Neuropsychological Society, 9*(3), 459–471.

Pedraza, O., Bowers, D., & Gilmore, R. (2004). Asymmetry of the hippocampus and amygdala in MRI volumetric measurements of normal adults. *Journal of the International Neuropsychological Society, 10*(5), 664–678.

Raz, N., Gunning-Dixon, F., Head, D., Rodrigue, K. M., Williamson, A., & Acker, J. D. (2004). Aging, sexual dimorphism, and hemispheric asymmetry of the cerebral cortex: Replicability of regional differences in volume. *Neurobiology Aging, 25*(3), 377–396.

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

Sled, J. G., Zijdenbos, A. P., & Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging, 17*(1), 87–97.

Soininen, H., Partanen, K., Pitkanen, A., Hallikainen, M., Hanninen, T., Helisalmi, S., et al. (1995). Decreased hippocampal volume asymmetry on MRIs in nondemented elderly subjects carrying the apolipoprotein E epsilon 4 allele. *Neurology, 45*(2), 391–392.

Soininen, H. S., Partanen, K., Pitkanen, A., Vainio, P., Hanninen, T., Hallikainen, M., et al. (1994). Volumetric MRI analysis of the amygdala and the hippocampus in subjects with age-associated memory impairment: Correlation to visual and verbal memory. *Neurology, 44*(9), 1660–1668.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics, 9*(6), 1135–1151.

van de Pol, L. A., Hensel, A., Barkhof, F., Gertz, H. J., Scheltens, P., & van der Flier, W. M. (2006). Hippocampal atrophy in Alzheimer disease: Age matters. *Neurology*, *66*(2), 236–238.

Walhovd, K. B., Fjell, A. M., Reinvang, I., Lundervold, A., Dale, A. M., Eilertsen, D. E., et al. (2005). Effects of age on volumes of cortex, white matter and subcortical structures. *Neurobiology Aging*, *26*(9), 1261–1270 (discussion 1275–1268).