Quantitative structural neuroimaging of mild traumatic brain injury in the Chronic Effects of Neurotrauma Consortium (CENC): Comparison of volumetric data within and across scanners

Elisabeth A. Wilde, Erin D. Bigler, Trevor Huff, Haonan Wang, Garrett M. Black, Zachary P. Christensen, Naomi Goodrich-Hunsaker, Jo Ann Petrie, CENC Neuroimaging Core Investigators including Tracy Abildskov, Brian A. Taylor, James R. Stone, Nicholas J. Tustison, Mary R. Newsome, Harvey S. Levin, Zili D. Chu, Gerald E. York & David F. Tate

To cite this article: Elisabeth A. Wilde, Erin D. Bigler, Trevor Huff, Haonan Wang, Garrett M. Black, Zachary P. Christensen, Naomi Goodrich-Hunsaker, Jo Ann Petrie, CENC Neuroimaging Core Investigators including Tracy Abildskov, Brian A. Taylor, James R. Stone, Nicholas J. Tustison, Mary R. Newsome, Harvey S. Levin, Zili D. Chu, Gerald E. York & David F. Tate (2016) Quantitative structural neuroimaging of mild traumatic brain injury in the Chronic Effects of Neurotrauma Consortium (CENC): Comparison of volumetric data within and across scanners, Brain Injury, 30:12, 1442-1451, DOI: 10.1080/02699052.2016.1219063

To link to this article: http://dx.doi.org/10.1080/02699052.2016.1219063

Published online: 11 Nov 2016.
Quantitative structural neuroimaging of mild traumatic brain injury in the Chronic Effects of Neurotrauma Consortium (CENC): Comparison of volumetric data within and across scanners

Elisabeth A. Wilde 1,2,3,4, Erin D. Bigler 5,6, Trevor Huff 6, Haonan Wang 6, Garrett M. Black 2, Zachary P. Christensen 6, Naomi Goodrich-Hunsaker 5, Jo Ann Petrie 5, CENC Neuroimaging Core Investigators including Tracy Abildskov 6, Brian A. Taylor 1,2,4, James R. Stone 7,8, Nicholas J. Tustison 7, Mary R. Newsome 1,2, Harvey S. Levin 1,2,3, Zili D. Chu 4, Gerald E. York 9, & David F. Tate 10

1Michael E. DeBakey Veterans Affairs Medical Center, Houston, TX, USA, 2Department of Physical Medicine and Rehabilitation, 3Department of Neurology, 4Department of Radiology, Baylor College of Medicine, Houston, TX, USA, 5Department of Psychology, 6Department of Neuroscience, Brigham Young University, Provo, UT, USA, 7Department of Radiology and Medical Imaging, 8Department of Neurological Surgery, University of Virginia, Charlottesville, VA, USA, 9Alaska Radiology Associates, Anchorage, AK, USA, and 10Missouri Institute of Mental Health, University of Missouri, St. Louis, MO, USA

Abstract

Background: An important component of the multicentre Chronic Effects of Neurotrauma Consortium (CENC) project is the development of improved quantitative magnetic resonance imaging (MRI) methods, including volumetric analysis. Although many studies routinely employ quality assurance (QA) procedures including MR and human phantoms to promote accuracy and monitor site differences, few studies perform rigorous direct comparisons of these data nor report findings that enable inference regarding site-to-site comparability. These gaps in evaluating cross-site differences are concerning, especially given the well-established hardware or software.

Methods: This study reports findings on (1) a series of studies utilizing two MR phantoms to interrogate machine-based variability using data collected on the same magnet, (2) a human phantom repeatedly imaged on the same scanner to investigate within-subject, within-site variability and (3) a human phantom imaged on three different scanners to examine within subject, between-site variability.

Results: Although variability is relatively minimal for the phantom scanned on the same magnet, significantly more variability is introduced in a human subject, particularly when regions are relatively small or multiple sites used.

Conclusion: Vigilance when combining data from different sites is suggested and that future efforts address these issues.

Introduction

An important component of the multi-centre Chronic Effects of Neurotrauma Consortium (CENC) project is the development of improved methods to utilize information derived from neuroimaging studies, particularly magnetic resonance imaging (MRI). MRI is a robust neuroimaging method for the identification of neuropathological changes that occur over time following a traumatic brain injury (TBI [1]). However, brain imaging derived from MRI contains far more information than that which is typically conveyed in standard clinical reports. Routine radiology reports typically provide qualitative descriptions of macroscopic abnormalities to include the type of finding, description of finding and a differential diagnosis of the pathology consistent with the finding. Quantitative descriptions are typically confined to simple measures of lesion size or displacement of structures given the impracticality of performing extensive quantification of neurological tissues in current clinical workflows. This approach is well-designed to meet emergent clinical needs [2,3], but may leave potential quantitative indicators of diagnosis and/or prognosis unexamined [4]. The extraction of quantitative MRI data that are clinically relevant for the individual patient with TBI advances the field toward personalized, precision medicine and improved utilization of MRI-based findings [5]. However, for such quantitative scan information to be clinically useful, it must be derived with efficient methods that are reliable and reproducible and that
specifically relate to diagnostic decision-making, treatment and outcome.

Historically, the constraint to using quantitative MRI morphometric measures in radiological interpretation has been that operator-controlled measurement can be time-consuming and may invoke reliability concerns related to methodological differences and potential inter-observer variance. Fortunately, rapid advances in automated neuroimaging methods now allow ready quantification of every major brain structure and region of interest (ROI). Methods such as FreeSurfer (http://surfer.nmr.mgh.harvard.edu/) have facilitated automated brain image volumetric analysis and are widely used [6–8], including use in large consortia such as the Alzheimer’s Disease Neuroimaging Initiative (ADNI) and TBI-related consortia such as Transforming Research and Clinical Knowledge (TRACK-TBI-2). As one of the first publically-developed and open source image analysis software tools, FreeSurfer has become a standard approach for quantitative analysis of neuroimaging data and, given its wide use, will be a main focus of this paper.

As has been well documented at the group level in research reports, degenerative changes may be quantified following a clinically significant TBI [9–11] including a sub-set of patients with mild TBI [12–14]. Recent clinical recommendations advocate the use of MRI only in the presence of new, persistent or worsening symptoms in patients with sub-acute to chronic mTBI [15]. In active duty military, MRI is the modality of choice for mild TBI in the sub-acute to chronic stages [16]. Researchers have continued to explore applications for quantitative MRI to enhance diagnosis and prognosis. However, gaps remain in the ability to interpret quantitative MRI at the level of the individual patient. One of the primary obstacles in this pursuit has been the lack of available normative data, due, in part, to small samples of existing control groups, sub-optimally selected control groups and an inability to directly pool data collected across sites and scanner vendors. Recent efforts have promoted the aggregation and open-source availability of larger datasets, via multi-institutional consortia (e.g. TRACK-TBI-2, Concussion Assessment, Research and Education (CARE) Consortium, CENC) and publically accessible databases (e.g. Federal Interagency Traumatic Brain Injury Research (FITBIR) informatics system). For these efforts to be successful, careful monitoring of data quality, within and between sites is required, along with development of adequate strategies to address site differences. Additionally, quantified features must be reliably measured and replicated with a more complete understanding of factors that contribute to measurement error [17,18]. Without this knowledge, it is difficult to understand the relationship of the effect size of potential injury-induced changes to inherent measurement error. Addressing sources of error may be particularly important when the changes attributable to the disease are presumed to be subtle and occur within an organ that may be dynamically changing even in a healthy state. Additionally, it is well established that between 80–85% of TBIs occur on the mild end of the TBI spectrum [19]. Conventional MRI in mild uncomplicated TBI is defined by lack of imaging findings, with complicated mTBI often associated with minimal imaging abnormalities. Even group studies that find statistically meaningful differences using more advanced structural imaging modalities and analytic techniques conclude that such differences may be subtle [20], but clinically meaningful. As such, this remains an active area of investigation by several groups [21].

Successful utilization of MRI findings in mTBI requires attention to many technical parameters including differences in MRI equipment and platforms, evolving software differences, along with the within-subject variability inherent in any living organism. This is particularly relevant for Service Member and Veteran patients as they may be injured at one location (e.g. in theatre, on base or post), transferred to an appropriate hospital or clinic setting for evaluation and treatment and later transferred or re-assigned prior to re-evaluation. Since transfer between evaluation sites is common in active Service Members and since many of these Service Members later receive care via the Veterans Administration (VA) hospital system following separation from the military, the Veteran who has sustained a TBI could potentially undergo multiple MRI studies across widely different platforms. Indeed, challenges associated with data comparability collected across sites and across time remains an important obstacle that is highly relevant to the CENC Neuroimaging Core.

For an individualized MRI quantitative approach to be effective, the above issues related to the viability of automated image analysis techniques and their reproducibility will be addressed in three separate studies investigating measurement variability within and between CENC MRI sites. The sections that follow will be topically organized around specific experiments designed to not only quantify variability in volumetric measurement, but also to attempt to elucidate specific sources of measurement variability. Each section includes a brief ‘Methods’ summary explaining the technical aspects of the experiment followed by a ‘Results’ section presenting the main findings. In this first manuscript, it is the authors’ intent to describe measurement results and highlight issues that may arise from simple volumetric measurements and analyses in multi-site data. Future manuscripts will highlight the ongoing effort to apply specific strategies designed to ameliorate variability within and between consortia sites and investigate replicability in other imaging modalities. While much of this paper focuses on technical issues, the clinical importance cannot be overstated as careful elucidation of sources of measurement error and solutions for ameliorating this error have broader implications for the entire field of TBI, including differential diagnosis, improving prognostic accuracy and informing treatment development and assessment.

Study 1. Machine variability (MR phantoms, same scanner)

Given the complexity of MRI scanner hardware, each scan session is likely to introduce subtle differences into the acquired data, potentially related to factors associated with fluctuations in applied gradients and magnetic fields. One method to examine signal variability is to scan and then re-scan an MRI phantom to measure signal intensity and geometry, both of which ultimately impact quantitative volumetric measurement of tissue since measurement is dependent, in part, upon intensity of the pixels in the image.
and accurate spatial rendering. Inanimate MRI phantoms do not have the limitations of ‘human phantoms’ (i.e. an individual who is scanned at each site for the purpose of eliminating sources of individual variability), given the absence of movement, tissue pulsation (i.e. cardiac pulsation) and dynamic physiological changes (e.g. hydration status, diurnal or circadian changes).

Periodic MR phantom object scanning is commonly performed as part of quality assurance (QA) procedures to monitor scanner uniformity and stability, particularly for studies that have a multi-site or longitudinal component. These measures are typically acquired at planned intervals that may be days to months in length and the results are compared against certain specifications to monitor performance. Although these procedures are implemented in the CENC, the aim was to first examine stability in repeated measures with higher frequency than those typically employed for routine QA purposes. As such, in the first experiment, two different phantoms were scanned on a single scanner. First, a manufacturer-supplied Siemens phantom was scanned multiple times within a short time window to examine signal intensity variability. Measures of signal intensity uniformity are intended to detect deficiencies in the scanner hardware (e.g. defective head coil or inadequate inductive decoupling of the body coil from the head coil, radio-frequency subsystem issue, magnetic field inhomogeneities or fluctuations in signal-to-noise-ratio). Second, the Alzheimer’s Disease Neuroimaging Initiative (ADNI) phantom (Magphan EMR051, The Phantom Laboratory, Salem, NY) was repeatedly scanned using the standard T1-weighted volumetric sequence to quantify the degree of geometric distortion (i.e. the degree to which dimensions in the images differ from known dimensions of the object) inherent in the MRI system. These measures are important in detecting other hardware issues such as a miscalibrated gradient or an abnormally high B₀ inhomogeneity.

Methods

Using the standard 1900 ml Siemens phantom plastic bottle (model #8624186 3.75 g NISO₂ X 6H₂O + 5g NACL) on a Siemens Tim Trio Scanner (Malvern, PA), the phantom was scanned repeatedly (7-times) using the following schedule: Scans 1–3 were run continuously without a pause between data acquisitions, Scan 4 was obtained after a 10-minute pause, Scan 5 was obtained after the MRI gantry table was moved out of the bore and then back into the same position without physically moving the phantom with respect to the table, Scan 6 was performed immediately after Scan 5 (no time gap) and Scan 7 was acquired following a 10-minute pause.

Additionally, several scanner-related variables were monitored and maintained constant during the scanning acquisition. The Center Frequency was set at 123.261 378 MHz and the Voltage at 353.568 V. Each of the coil elements were shown to be constant at the following values: HEP = 0.031, H4P = 0.030, HEA = 0.037, H2P = 0.036. After the scans were acquired, a post-scan analysis was done using OsiriX (www.osirix-viewer.com/). Using this software, a circular region of interest (ROI) of area 56.196 cm² (5714 pixels) was created. This ROI was then placed on the centre of each phantom scan, taking great care to place the ROI at the same position in each scan. Intensities within the ROI were subsequently recorded as a histogram with the mean, minimum and maximum values. The dependent variable was the mean voxel intensity within the defined space of the phantom.

For measurement of geometric distortion, the ADNI phantom was also scanned on a Siemens Tim Trio Scanner. With the centre of the phantom at isocentre, a series of identical magnetization-prepared rapid acquisition gradient echo (MPRAGE) T1-weighted images with a resolution of 1 mm × 1 mm × 1.2 mm, TE/TR/TI = 3/2300/900, flip angle = 9°, 240 Hz/pixel bandwidth with a 256 × 256 matrix in the sagittal plane. As with the Siemens phantom, no shimming was performed between scans. Both 2D and 3D distortion corrections were performed as well as a zero filling interpolation to 0.5 mm × 0.5 mm in plane resolution for comparison. In addition, the phantom was then landmarked at 1 cm inferior to isocentre and re-shimmmed in the MR to test for repeatability.

Results

All seven trials with the Siemens chemical phantom demonstrated subtle differences in signal intensity, despite using the same scanner, phantom and pulse sequence, and there were no trials where replication yielded identical results. To simplify the presentation of findings, signal intensity histograms from the first four test trials are shown in Figure 1, where intensity histograms differ somewhat for each trial, despite using the same sequence and phantom, unchanged in any manner with regards to its physical and chemical makeup. Across all trials, the mean intensity was similar, with the percentage variance of 0.5%, which could be attributed to variability emanating from the machine and, possibly, observer variability in the placement of the ROIs where the measurements were obtained given the intensity gradient across the diameter of the phantom.

In the ADNI phantom results, the coefficient of variation between measurements (maximum distortion, mean distortion in the x, y, z directions) was less than 0.008, demonstrating excellent reproducibility. There was no statistical difference in the measurements when 3D distortion correction or interpolation was used. When 2D correction was utilized, the maximum distortion in the z-direction was 0.08 mm higher (outside the 95% confidence interval). Interestingly, re-shimming at 1 cm off isocentre caused an increase in the x and y distortions at 0.6 mm and 0.5 mm, respectively (p < 0.01). 2D correction reduced the y-direction distortion from 1.1 mm to 0.4 mm, while minimally affecting the x and z direction (max: 0.1 mm change in z). 3D distortion correction lowered the y-direction distortion to 0.4 mm, in addition to lowering the x-direction to 0.8 mm from 1.7 mm. This shows that correcting geometric distortion in 3D assists in more geometrically accurate images when the scanner characteristics change (i.e. shim, new imaging session, etc.). Measurement of geometric distortion is critical because it can detect changes in morphology of the image, which can lead to inaccurate volumetric measurement. Failure to monitor this distortion in multisite or longitudinal data will make it difficult to ascertain differences.
in morphology that may arise from the patient vs variability from the scanner.

**Study 2. Human phantom (within-subject variability in the same scanner)**

In addition to measurement with an MRI phantom, measurement using a living phantom is important due to the host of physical and physiological factors that could potentially influence the MRI signal in the same individual [18].

**Methods**

Mirroring, in part, the same procedure performed with the inanimate MR phantom and using the same scanner as that employed for Study 1, Study 2 involved a set of experiments in a single 23-year-old healthy male volunteer. As with the phantoms, the volunteer underwent a series of identical MPRAGE T1-weighted images (using sequences similar to CENC-designed acquisition parameters). The acquisition of these sequences is as follows on a Siemens Tim Trio Scanner (Malvern, PA): (1) localizer, (2) Scan 1 acquisition, (3) 2-minute pause without moving the subject, (4) Scan 2 acquisition, (5) moving the table out then back in the bore with the same location, (6) second localizer, (7) Scan 3 acquisition, (8) Scan 4 acquisition (without time delay between the Scan 3 and Scan 4 acquisitions), (9) 5-minute pause and (10) Scan 5 acquisition. Approximately 6 months prior to this same-day scanning, two additional MPRAGE sequences of the same individual were acquired, separated by 30 minutes (Scans 6 and 7).

The MR data from each acquisition was processed using FreeSurfer 5.3 and segmentations for several ROIs including the hippocampus, amygdala, thalamus, caudate, nucleus accumbens, putamen, pallidum, third ventricle, fourth ventricle, lateral ventricles, total white matter, total grey matter and intracranial volume (ICV) were derived for each scan.

As no standard metric of reliability has been established in the imaging literature related to this issue, many studies comparing agreement or similarity of volumetric data often report more than one metric [22]. As such, several metrics are reported in the current communication. Initially, the variability between scans was assessed by determining the absolute difference in volumes within a given ROI (minimum subtracted from maximum) in mm$^3$. Percentage difference along with the mean and standard deviation across all scans were also computed. In addition, several measures of overlap and overlap error were also calculated. This was accomplished by first creating an unbiased template [23] from the T1 MPRAGE sequence using the open-source Advanced Normalization Tools (ANTs) multivariate template construction pipeline. The scan label maps (aseg.mgz) from the FreeSurfer processing described above were then resliced to the template space using nearest neighbour interpolation by applying the transforms used to create the template. The segmentation labels were then examined for overlap similarity and dissimilarity. These steps resulted in the calculation of four commonly assessed measures [22] of overlap and two overlap error measures, which are described in Table I.

**Results**

Results reported in absolute and percentage difference are shown in Table II. Similar to the inanimate phantom, human phantom data also generated scan-to-scan variability in the same individual, with no scan identically replicating another. The greatest variability was observed in accumbens, amygdala, pallidum and third ventricle volumes and the least variability was observed in the larger overall measures such as total grey and white matter volumes as well as intracranial volume (ICV).

The overall average (all ROIs included) overlap and overlap error metric results are presented in Figure 2. As illustrated, total target and Dice coefficient measurements were both above 0.9, indicating good reproducibility when
structures were examined in aggregate. The Jaccard coefficient was above 0.8 and volume similarity was close to zero. False negative and positive overlap error was less than 0.1, again indicating similar error, regardless of whether the source or target images were considered. The range between the highest and lowest measures was minimal (< 0.02). Generally, the best overlap measures were observed between scans acquired sequentially and without any delay between the acquisitions. The two lowest overlap measures occurred between the scans acquired 5 minutes apart (data not shown).

Examination of individual ROI overlap and overlap error measures is presented in the radar plots in Figure 3. These plots show the general consistency of many measures across a broad range of ROIs that are commonly used in studies of TBI. However, similar to the percentage difference data presented above, there are clearly individual ROIs that have more variability in the way they are labelled and quantified including, but not limited to, the inferior lateral ventricles, pallidum, nucleus accumbens and third ventricle.

### Study 3. Within-subject but between-site variability

Between-site variability of MRI-derived quantitative measures has been well-studied with FreeSurfer [17], likely due, in part, to intensity differences that affect image segmentation and registration, differences in the application of the probability atlas and differences in head positioning. Given the results in Study 2 where a single scanner was utilized, it was anticipated that there would likely be even greater variability between measures derived from the same individual, but different scanners.

### Methods

As part of the initial site qualification and ongoing quality control efforts in place with the CENC, the same healthy 42 year-old female human phantom travelled to each scanner involved in the consortium to undergo imaging within the same year. For the sake of simplicity and ease of data presentation, only three sites (of those involved in the Observational study are presented. Platforms differed across the sites (3T

### Table I. Overlap and overlap error formulae and descriptions.

<table>
<thead>
<tr>
<th>Overlap measure</th>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target overlap</td>
<td>$TO = \frac{</td>
<td>S \cap T</td>
</tr>
<tr>
<td>Jaccard coefficient</td>
<td>$UO = \frac{</td>
<td>S \cup T</td>
</tr>
<tr>
<td>Dice coefficient</td>
<td>$MO = \frac{</td>
<td>S \cap T</td>
</tr>
<tr>
<td>Volume similarity</td>
<td>$VS = \frac{</td>
<td>S \cap T</td>
</tr>
<tr>
<td>False positive</td>
<td>$FP = \frac{</td>
<td>S \setminus T</td>
</tr>
<tr>
<td>False negative</td>
<td>$FN = \frac{</td>
<td>T \setminus S</td>
</tr>
</tbody>
</table>

* S. Source image; T. Target image; r. region.

### Table II. Means, standard deviations, absolute difference and percentage difference in FreeSurfer-derived measures of commonly used regions of interest.

<table>
<thead>
<tr>
<th>Volume</th>
<th>Mean ($mm^3$)</th>
<th>SD ($mm^3$)</th>
<th>Difference ($mm^3$)</th>
<th>% Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left Hippocampus</td>
<td>5 209.71</td>
<td>76.54</td>
<td>216.50</td>
<td>4.24%</td>
</tr>
<tr>
<td>Right Hippocampus</td>
<td>5 186.64</td>
<td>92.46</td>
<td>267.60</td>
<td>5.35%</td>
</tr>
<tr>
<td>Left Amygdala</td>
<td>1 989.09</td>
<td>61.00</td>
<td>151.80</td>
<td>7.87%</td>
</tr>
<tr>
<td>Right Amygdala</td>
<td>2 008.16</td>
<td>92.32</td>
<td>271.10</td>
<td>14.33%</td>
</tr>
<tr>
<td>Left Caudate</td>
<td>4 048.67</td>
<td>61.11</td>
<td>146.70</td>
<td>3.70%</td>
</tr>
<tr>
<td>Right Caudate</td>
<td>4 216.49</td>
<td>72.27</td>
<td>210.80</td>
<td>5.16%</td>
</tr>
<tr>
<td>Left Putamen</td>
<td>6 390.03</td>
<td>89.89</td>
<td>261.20</td>
<td>4.21%</td>
</tr>
<tr>
<td>Right Putamen</td>
<td>5 811.67</td>
<td>89.89</td>
<td>261.20</td>
<td>4.21%</td>
</tr>
<tr>
<td>Left Thalamus</td>
<td>9 075.29</td>
<td>222.46</td>
<td>649.10</td>
<td>7.41%</td>
</tr>
<tr>
<td>Right Thalamus</td>
<td>8 106.53</td>
<td>99.79</td>
<td>290.30</td>
<td>3.65%</td>
</tr>
<tr>
<td>Left Pallidum</td>
<td>1 675.87</td>
<td>89.46</td>
<td>275.70</td>
<td>18.04%</td>
</tr>
<tr>
<td>Right Pallidum</td>
<td>1 722.29</td>
<td>59.97</td>
<td>147.90</td>
<td>9.14%</td>
</tr>
<tr>
<td>Left Accumbens</td>
<td>689.21</td>
<td>58.29</td>
<td>162.00</td>
<td>27.36%</td>
</tr>
<tr>
<td>Right Accumbens</td>
<td>775.19</td>
<td>37.17</td>
<td>299.50</td>
<td>7.80%</td>
</tr>
<tr>
<td>Corpus Callosum</td>
<td>4 016.17</td>
<td>76.18</td>
<td>218.60</td>
<td>9.67%</td>
</tr>
<tr>
<td>Third Ventricle</td>
<td>881.59</td>
<td>87.38</td>
<td>213.07</td>
<td>2.80%</td>
</tr>
<tr>
<td>Fourth Ventricle</td>
<td>2 324.20</td>
<td>76.18</td>
<td>218.60</td>
<td>9.67%</td>
</tr>
<tr>
<td>White Matter</td>
<td>516 209.07</td>
<td>4234.98</td>
<td>11 894.60</td>
<td>0.68%</td>
</tr>
<tr>
<td>Total Grey Matter</td>
<td>769 642.87</td>
<td>8731.38</td>
<td>21 183.07</td>
<td>2.80%</td>
</tr>
</tbody>
</table>

The data shown here was generated using the same scanner and a single subject. ICV, intracranial volume.
Siemens Tim Trio at the Michael E. DeBakey Veterans Administration Medical Center (MEDVAMC) in Houston, 3T Siemens Verio at Brooke Army Medical Center (BAMC)/San Antonio Military Medical Center (SAMMC) and 3T Philips Ingenia at Hunter Holmes VA Medical Center/Virginia Commonwealth University (VCU)). The human phantom was scanned at each site using established volume acquisition sequences derived from the Alzheimer’s Disease Neuroimaging Initiative (ADNI-2) consortia, developed to be as comparable as possible across platforms. The phantom individual was scanned twice on one scanner for measurement of intra-site comparability. The MR data from each acquisition site was processed using FreeSurfer 5.3 and volumetric comparisons were made between the sites using the same methods and metrics as described above. For ease in data presentation, only a select number of more global (i.e. ICV, total grey matter and total cortical white matter) and as well as specific regions (i.e. right amygdala, left accumbens, left hippocampus) are graphically presented, although all ROI volumes as described in Study 3 were obtained.

Results
As with the results obtained in Study 2, Study 3 also demonstrated a failure to *identically* replicate ROI volumes when the same subject was scanned across sites and coefficients of variability again reflect substantive differences in certain regions. For ease in data presentation, only a select number of more global (i.e. ICV, total grey matter and total cortical white matter) and as well as specific regions (i.e. right amygdala, left accumbens, left hippocampus) are graphically presented, although all ROI volumes as described in Study 3 were obtained. Figure 4 depicts site variability for these volumes as well as for smaller sub-cortical structures such as the amygdala, hippocampus and nucleus accumbens. For the more global volumetric measures, the differences in data obtained across scanners is relatively minimal. However, for smaller structures, the variability in measurement is increased.

As before, this study also calculated the Dice, Jaccard and total target coefficients as well as volume similarity, false negative and false positive overall overlap and overlap error.
measures, which are presented in Figure 5. In addition to the average overlap and overlap error, the specific site comparison metrics are also presented. The average measures appear to indicate adequate similarity between measures across the sites, although these measures are consistently less than the within-subject, within-scanner results from study 2, as expected. The total target and the Dice coefficient were all above 0.8, while the Jaccard coefficient was greater than 0.75. The volume similarity was close to 0.0 and was between –0.02 and 0.02. The false positive and negative values were consistently between 0.10 and 0.14, with the false positive value being consistently lower than the false negative value. Although not tested statistically, the best overlap metrics were observed between the scans acquired at the same site (Scanner 1 – Time 1, Scanner 1 – Time 2). The variability in the values obtained between the different sites did not appear to have a discernible pattern that would indicate that one scanner performed worse than any other.

Direct comparison of Study 2 and Study 3 (see Table III) metrics using an independent t-test (unequal variance assumed) demonstrated significant differences for all the measures, but volume similarity ($t$-value = –1.39, $p$-value = 0.20). The effect size for each comparison is also provided and noted to be very robust across the different measures.

Examination of individual ROIs overlap and overlap error for this experiment is presented in the radar plots in Figure 6. Again, these plots show general consistency between most measures across the range of ROIs. However, similar to the data presented in Study 2, there are individual ROIs that have more variability in the way they are labelled and quantified, as observed in the various overlap and overlap error measures. Similar excessive variability was noted for the inferior lateral ventricles, nucleus accumbens and third ventricle. Differences in smaller structures such as the basal ganglia are evident in direct comparison of the segmentation maps overlaid on T1-weighted images, as seen in Figure 7.

Studies 1–3 demonstrate measurement variability that can complicate multi-site and longitudinal imaging data. The ~0.5% variability in voxel intensity from the MR scanner is likely something that will represent a constant, but minimal, source of variance. Likewise, measures for larger and more global areas are generally consistent across these three scanners, where similar imaging parameters and rigorous quality assurance procedures were in use across sites (e.g. all scanners routinely fall within specification at < 2 mm in 3D accuracy using the ADNI phantom and protocol). However, some of the structures in the within-subject (Study 2) and between-scanner variability (Study 3) are substantial.
varying between less than 5% for some ROIs to over 10% for others.

**Discussion**

The results contained in this report suggest that MRI volumetric measurement performed in a manner consistent with many studies in clinical research may be subject to some degree of measurement error, despite the fact that morphometric measurement is one of the quantitative MRI techniques generally considered to be relatively robust to site differences. The results of the MRI phantom data presented suggest that the scanner itself does introduce some degree of measurement difference, albeit relatively negligible when a stable object is imaged. The amount of variability detected in these studies is unlikely to exceed differences that are considered clinically meaningful, even in populations where the effects are anticipated to be rather subtle, provided that the scanners are undergoing rigorous and consistent QA to monitor adequate functioning.

However, additional variability is introduced in imaging living organisms. The second study, that involved measurement of the same individual on the same scanner, indicated variability of a greater magnitude, particularly in the measurement of smaller structures. This variability was present despite the relative stability and accuracy of the MR system based on the results of phantom testing in Study 1. Such differences were present even under conditions where the subject was scanned continuously (without lapses in time) and without any movement of the gantry or re-positioning. Additional testing is underway to examine the potential impact of other sources of biological difference (e.g., hydration, time of day, hormonal influences, etc.) that may also contribute to additional variability within an individual subject. Clearly, it is also likely that some measurement error is also introduced during post-processing analysis, and this also remains an area of additional exploration.

**Table III. Means, standard deviations and statistical tests for the within and across sites human phantom data.**

<table>
<thead>
<tr>
<th></th>
<th>Across Site Mean (SD) (n=6)</th>
<th>Within Site Mean (SD) (n=10)</th>
<th>t-value (p-value)</th>
<th>Effect Size (Cohen’s d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>0.877 (0.007)</td>
<td>0.906 (0.004)</td>
<td>-8.81 ( &lt;0.001)</td>
<td>4.87</td>
</tr>
<tr>
<td>Jaccard</td>
<td>0.785 (0.008)</td>
<td>0.829 (0.004)</td>
<td>-12.73 ( &lt;0.001)</td>
<td>7.10</td>
</tr>
<tr>
<td>Dice</td>
<td>0.880 (0.005)</td>
<td>0.906 (0.002)</td>
<td>-12.48 ( &lt;0.001)</td>
<td>6.99</td>
</tr>
<tr>
<td>Volume Similarity</td>
<td>-0.007 (0.008)</td>
<td>-0.001 (0.007)</td>
<td>-1.39 (0.20)</td>
<td>0.73</td>
</tr>
<tr>
<td>False Negative</td>
<td>0.123 (0.007)</td>
<td>0.094 (0.004)</td>
<td>8.81 ( &lt;0.001)</td>
<td>4.87</td>
</tr>
<tr>
<td>False Positive</td>
<td>0.117 (0.004)</td>
<td>0.093 (0.003)</td>
<td>11.63 ( &lt;0.001)</td>
<td>6.17</td>
</tr>
</tbody>
</table>

SD, standard deviation.

For Cohen’s d, 0.5–0.79 is considered a medium effect size and > 0.8 is considered a large effect size.

**Figure 6.** Radar plots illustrating the overlap and overlap error values for individual ROIs in the FreeSurfer processed imaging data for the between-site phantom data. Illustrates that most ROIs have ‘good’ overlap, although several demonstrate reduced similarity. Similar ROIs appear to have diminished similarity when comparing the within-site and between-site phantom data.
Finally, the use of different scanners has long been known to produce some variability in measurement and the third study comparing the quantitative results of the same individual scanned on different magnets again increased the observed variability of measures, particularly for smaller sub-cortical regions of interest. Differences in these measures are likely multifactorial and include the same sources of error described above as well as additional error introduced by vendor-specific differences. Repeat measurement over time of the human phantom yielded more similar results than measurement obtained across different sites, suggesting that it is these differences (in hardware and software) that account for the most significant source of measurement variability.

Clearly there are notable advantages to multi-site and longitudinal investigation and the presentation of data related to variability in measurement derived across scanners is not intended to imply that such studies should not be undertaken or cannot produce meaningful results. Rather, the data presented here are intended to increase awareness regarding the need for vigilance in how data are used, the need for scrutiny in understanding the limitations of their accuracy and to highlight the need to evaluate and develop solutions to better address these issues. The data suggest that global measures may be less vulnerable to differences in measurement across scanner vendor, model and software. However, the use of more detailed sub-cortical structures may require additional attention to these and other sources of variability, and corrective strategies to address these differences should be carefully considered beyond the implementation of rigorous and consistent quality assurance efforts and acquisition parameter harmonization that are in use in many studies.

Identification of variability in these data have led to additional efforts by CENC scientists to apply and examine methods that can minimize sources of variability. The basic steps of FreeSurfer processing including registration, intensity normalization, segmentation and labelling likely contribute to some of the measurement error noted. Each of these areas of imaging post-processing are very active fields of research and several advances have been made that hold promise for improving results by reducing variability in measurement error. The CENC imaging core is actively examining several of these areas and data will be presented in future manuscripts that outline the results of these efforts.

Despite limitations, the great advantage of quantitative neuroimaging approaches as outlined herein is that they have the potential to bring objectivity to the assessment of TBI, especially mTBI. If these issues can be satisfactorily addressed, volumetric measurement could presumably be used not only for group-level studies, but also for enhanced understanding of individual patients, particularly to monitor change over time. Given the likelihood for continued conflict that puts Military personnel in harm’s way, increasing the risk for mTBI and PTSD, the importance of having reliable markers of injury cannot be overstated (see Cook and Hawley [24]).

Figure 7. FreeSurfer generated segmentation results on the same individual scanned on three different scanners. Note the subtle variation in volume estimations of the basal ganglia that lead to differences in quantitation.
References


9. Bigler ED, Abildskov TJ, Wilde EA, McCauley SR, Li X, Merkley TL, Fearing MA, Newsome MR, Scheibel PhD, and Xiaqiu Li, PhD, as well as CENC site neuroradiologists and technologists including John L. Ritter, MD, Jorge de Villasante, MD, Michael Lennon, MD, Rajan Agarwal, MD, Robert Cadrain, Marice L. Brown and William Wollfacke. We also wish to thank CENC leadership (Drs. David X. Cifu, Ramon Diaz-Arrastia, Rick Williams and Col. Sidney R. Hinds) for their support and the CENC Government Steering Committee for their oversight.

Declaration of interest
This project was jointly supported by the Department of Defense (W81XWH-13-2-0095), and the US Department of Veterans Affairs (I01 RX 00174). The authors report no conflicts of interest. The views expressed here are those of the authors and do not necessarily reflect the official position of the Department of Defense, The Department of Veterans Affairs, nor the US Government.

CENC volumetric data within and across scanners 1451


