Reproducibility and Reliability of Volumetric Measurements of Olfactory Eloquent Structures

David M. Yousem, MD1,2, Rena J. Geckle1,2, Richard L. Doty, PhD2, Warren B. Bilker, PhD3

Rationale and Objectives. The authors assessed the reliability and reproducibility of volumetric measurements of olfactory bulbs and tracts (OBTs) and temporal lobes (TLs) with magnetic resonance (MR) imaging.

Methods. Repetitive MR imaging of two phantoms and OBTs and TLs of patients was performed. Regions of interest were manually drawn around the relevant structures, and their volumes were measured at workstations. Intra- and interobserver variability for the two readers were measured. Intraclass and Pearson correlation coefficients and mean percentage differences were calculated.

Results. The measured phantom volumes were within 1.9%-12% of the true volumes with a variability of ≤5%. Intraclass and Pearson correlation coefficients were ≥0.919 for measurements by a single reader and ≥0.924 for measurements by different readers. The inter- and intraobserver variabilities were ≤4.2% for TL and 11.3%-14.6% for OBTs.

Conclusion. Volumetric measurements of the olfactory apparatus can be reliably and accurately reproduced from MR images.

Key Words. Brain, volume; brain, MR; magnetic resonance (MR); volume measurement; phantoms.

As part of a 5-year project to study the relationship between magnetic resonance (MR) imaging features of the olfactory system and clinical tests of olfaction, 62 subjects underwent volumetric measurements of the olfactory bulbs and tracts (OBTs) and temporal lobes (TLs). These measurements were performed by using contiguous 3-mm-thick two-dimensional T1-weighted images with a matrix of 256 × 256. The purpose of this study was to evaluate the accuracy and reproducibility of volumetric measurements of 3-mm-thick sections. An ISG Technologies Allegro Workstation (Toronto, Canada) was used to perform measurements. The precision of this technique and inter- and intraobserver variability were assessed in this regard.

MATERIALS AND METHODS
Phantom Studies

Phantom 1.—A phantom was constructed to simulate the shape of the OBT and the brain to determine the accuracy of our technique. This model...
A. B. c.

FIGURE 1. Representative MR images of the OBT and the TL. A, Coronal T1-weighted image through the olfactory bulb shows the typical irregular contour of the bulbs (arrows) outlined by cerebrospinal fluid above and the nasal vault below. B, The olfactory tracts can be seen posteriorly (arrows). Volumes of the OBTs were derived from these surface coil images (A, B). C, Volumes of the TLs were derived from T1-weighted coronal images through the brain by using a head coil.

Phantom 2.—After 2 years of experience and examination of more than 50 patients by using the volumetric sequences described later in the patient-studies section, another phantom was developed to assess the accuracy of the volumetric measurements. This phantom consisted of a loosely folded cabbage that had numerous air-filled interstices within it. This phantom was chosen to simulate the intricate gyral pattern of the brain and to provide a difficult interface that would simulate that of the TL or OBT and the cerebrospinal fluid. One of the individuals who had analyzed the first phantom and a third individual who had analyzed patient studies for the previous 2 years independently measured the volume of the cabbage on two separate occasions (D.M.Y., R.J.G.). The results were compared with true volumes determined with a water-displacement technique (the cabbage was cut into numerous sections to allow water into all the interstices). Because only one item was measured (two observers, two readings each), no intraclass correlation coefficients were assigned for this analysis.

Patient Studies

After a sagittal localizing image was obtained, 3-mm interleaved (no intersection gap) coronal T1-weighted images (500/15, two acquisitions, 256 × 256 matrix) through the OBTs were obtained with a 12-cm field of view by using a surface coil placed over the nasion (Fig 1A, 1B). These coronal T1-weighted images were used...
in the quantitative evaluation of the OBTs in 62 consecutive patients involved in our olfactory research. T2-weighted imaging was also performed through the OBT, but these images were not used for determining volumes.

An examination with a head coil included sagittal localizing imaging followed by 3-mm interleaved (no intersection gap) coronal T1-weighted imaging (600/11, one acquisition) with a 25-cm field of view and a 256 x 256 matrix. Then 5-mm interleaved axial T2-weighted imaging of the entire brain was performed with a fast spin-echo technique (4,000/80, one acquisition). The coronal T1-weighted images of the olfactory eloquent TLs were used for volumetric assessment of the TLs (Fig 1C). The TL was defined as that portion of the brain inferior to the sylvian fissure (not including the ventricular system) from the anterior edge of the middle cranial fossa to the posterior point at which the depth of the sylvian fissure was equal to that of adjacent sulci. This region of the brain has been shown to be involved in smell processing in studies of posttraumatic injury to the brain [1-3]. Informed consent was obtained from each patient who participated in this project.

Volumetric analyses of the right and left OBT and TL were performed by two independent evaluators (D.M.Y., R.J.G.) on two separate occasions at least 1 week apart. Regions of interest were manually drawn around the OBTs and TLs at the workstation by using a cursor. The technique required windowing, removing irrelevant tissue (cerebrospinal fluid, fat) with thresholding techniques, and three-dimensional volumetric processing with a modified cubic spline interpolation between contiguous two-dimensional coronal sections. Inter- and intraobserver reliability were determined by using intraclass correlation coefficients to assess pairs of readings by individual readers and by different readers for right and left TL and right and left OBT volumes. The percentage difference in volumetric measurements between readers (interobserver difference) was computed by dividing the difference between the two readings by the most experienced evaluator’s reading and multiplying by 100. The percentage difference between readings for each individual observer was calculated by determining the difference between two volumetric measurements of a reader, dividing that difference by the first volumetric measurement performed, and multiplying by 100.

The data presented in this study represent our cumulative experience as of June 1996 with individuals who have normal (n = 18) and abnormal (n = 44) olfaction. The abnormal group included individuals with congenital, posttraumatic, or degenerative causes of olfactory loss. Data were available for 62 measurements of TL volumes and 55 measurements of OBT volumes by each of two observers.

RESULTS

Phantom Studies

Phantom 1.—For the first reader, the intraobserver difference in measured volume for four different sections was 4.8% for item 1 [(1.73 - 1.65)/1.65], 4.1% for item 2 [(2.51 - 2.41)/2.41], and 0.9% for item 3 [(1.018 - 1.009)/1.009] (Table 1). We believe these results show that repetitive imaging performed at different times and various section locations will produce volumetric results that are reproducible within a range of less than 5%.

Water displacement showed the volumes of items 1 and 2 to be 1.5 and 2.5 cm³, respectively. The mean values of the four MR series yielded a 12% error for item 1 and a 1.9% error for item 2. The measured volume of item 3 was 1.080 cm³, and the mean volume determined with the four MR series was 1.013 cm³; this volume yielded a mean error of 6.3%. Overall error for the three items therefore was 5.7%.

A second, less experienced reader performed a similar analysis of the four MR series of the phantom. For the second reader, the intraobserver difference between images based on four section locations was 8.9% for item 1, 6.5% for item 2, and 0.1% for item 3.

When the measured volumes were compared with the true volumes of 1.5 cm³ for item 1, 2.5 cm³ for item
TABLE 2: Phantom 2 Data

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Reader 1</th>
<th>Reader 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measured volume (cm³)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 1</td>
<td>1,280 (5.8)</td>
<td>1,096 (9.4)</td>
</tr>
<tr>
<td>Trial 2</td>
<td>1,291 (6.7)</td>
<td>1,202 (0.66)</td>
</tr>
<tr>
<td>Mean volume (cm³)</td>
<td>1,285.5 (6.2)</td>
<td>1,149 (5.0)</td>
</tr>
<tr>
<td>Intraobserver variation</td>
<td>0.85%</td>
<td>8.8%</td>
</tr>
<tr>
<td>Interobserver variation</td>
<td>10.6%</td>
<td>10.6%</td>
</tr>
</tbody>
</table>

Note.—The actual volume of phantom 2 was 1,210 cm³. Percentage differences from true volume are given in parentheses.

TABLE 3: Reproducibility of TL and OBT Volume Measurements

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Intraclass Correlation Coefficient*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reader 1, TL (self)</td>
<td>0.950 (0.932, 0.967)</td>
</tr>
<tr>
<td>Reader 1, OBT (self)</td>
<td>0.919 (0.873, 0.965)</td>
</tr>
<tr>
<td>Reader 2, TL (self)</td>
<td>0.950 (0.929, 0.971)</td>
</tr>
<tr>
<td>Reader 2, OBT (self)</td>
<td>0.925 (0.880, 0.971)</td>
</tr>
<tr>
<td>Reader 1-Reader 2, TL</td>
<td>0.924 (0.894, 0.955)</td>
</tr>
<tr>
<td>Reader 1-Reader 2, OBT</td>
<td>0.966 (0.950, 0.982)</td>
</tr>
</tbody>
</table>

*The 95% confidence intervals are given in parentheses.

TABLE 4: Reproducibility of TL and OBT Volume Measurements

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Pearson Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reader 1, TL (self)</td>
<td>0.952</td>
</tr>
<tr>
<td>Reader 1, OBT (self)</td>
<td>0.931</td>
</tr>
<tr>
<td>Reader 2, TL (self)</td>
<td>0.954</td>
</tr>
<tr>
<td>Reader 2, OBT (self)</td>
<td>0.945</td>
</tr>
<tr>
<td>Reader 1-Reader 2, TL</td>
<td>0.931</td>
</tr>
<tr>
<td>Reader 1-Reader 2, OBT</td>
<td>0.968</td>
</tr>
</tbody>
</table>

2, and 1,080 cm³ for item 3, the mean error for the three items was 6.5%.

The mean values determined by the second reader of 1.72 cm³ for item 1, 2.54 cm³ for item 2, and 1,013 cm³ for item 3 differed from those determined by the more experienced analyzer by 2.4%, 3.7%, and 0.1% (interobserver difference).

Phantom 2.—The actual volume of the cabbage was 1,210 cm³. The two readings by the first reader were within 1% of each other (intraobserver difference) and were within 7% of the true volume of the phantom (Table 2).

The intraobserver difference for the second reader was 8.8%. One of the readings was within 1% of the actual volume, whereas the second reading deviated by 9.4%, for a mean variation from the true volume of 5.0%. The interobserver difference was 10.6%. The mean value of the four readings of the two reviewers was 1,217 cm³, with less than 1% variation from the actual volume.

Patient Studies

Intraclass and Pearson correlation coefficients were used to assess the reproducibility of the quantitative measurements of the right and left OBTs and TLs for each reader. For the first reader the intraclass correlation coefficients ranged from 0.950 for the TL to 0.919 for the OBT (Table 3). For the second reader the correlation coefficients ranged from 0.950 for the TL to 0.925 for the OBT. Correlation coefficients and the kappa statistic have the same scale and are interpreted the same way: Values higher than 0.80 are regarded as almost perfect [4, 5]. Pearson correlation coefficients for the two readers ranged from 0.931 to 0.954 (Table 4). When the values were analyzed between readers, the intraclass correlation coefficients were 0.966 (OBT) and 0.924 (TL), while the Pearson correlation coefficients were 0.968 (OBT) and 0.931 (TL), which signified almost perfect reliability between readers.

The interobserver difference was 4.2% for the TL and 11.3% for the OBT. The degree of difference between the two readings of the same structure by the same individual (intraobserver difference) was an average of 3.3% for the TLs and 13.9% for the OBTs (Table 5).

The volumes of the OBTs ranged widely. Some patients had almost complete destruction of the OBTs after trauma, and control subjects had unaffected OBTs. The mean OBT volume for all patients was 104.4 mm³ with a standard deviation of 42.9 mm³. The mean TL volume was 70,201 mm³ with a standard deviation of 9,680 mm³.
DISCUSSION

Several volumetric MR studies have recently been published in the neuroradiology literature [6-14]. It is remarkable how few of these studies have involved the use of phantoms with known volumes to gauge the accuracy of the techniques employed. In many cases, the structure being measured (the hippocampus or TL in a living host) was not available for verification of measured volumes, and the researchers relied solely on (intraclass) correlation coefficients or similar tests to assess the veracity of their findings [9]. It is important to realize, however, that one can have excellent Pearson correlation coefficients but poor accuracy and wide variation among readings. The analogy would be three parallel lines spaced widely apart to represent two readers' volumes and the "truth"; the Pearson correlation coefficient would be high because there would be a uniform distance between the two readers' measurements (hence, parallel lines), but the accuracy could be poor if these lines were far from the parallel line of the truth. The intraclass correlation, which gauges the absolute agreement of measurements, would be low in this situation, and the percentage variation would be high. For example, if volumes measured by one reviewer were consistently 100 cm³ greater than those measured by a second reviewer for structures with true volumes of 10-30 cm³, the Pearson correlation coefficients would be excellent. However, the intraclass correlation coefficient and percentage variation would be terrible. For this reason, greater emphasis should be placed on the intraclass correlation coefficients (which were all 0.919 in our patient study data) when one cannot obtain true measures of volumetric accuracy. This idea is illustrated by the scenario in which two readers' lines (volumetric values) are perfectly superimposed but are still parallel to and far from the line of truth; Pearson and intraclass correlations might be perfect, but accuracy would still be low. The percentage variation would be small. Observers can agree on the wrong volume.

We therefore believe that it is critical to provide interobserver and intraobserver percentage variation not only for measured volumes but also for phantom data if in vivo standards of reference cannot be obtained. Multiple readings by multiple observers should be performed to establish the reproducibility of the methods before single-observer readings can be justified [9]. We also believe that providing data for phantoms with known volumes and a configuration analogous to that of the measured structures is also very helpful in assessing the reliability of a technique. Blatter et al [9] have called for the development of sophisticated brain phantoms to assess accuracy, calibrate instruments, and conduct routine quality control. Until such phantoms become available, use of artificial objects remains a reasonable alternative.

The techniques we currently use to measure volume have resulted in excellent rates of accuracy with our phantoms. Our intraclass correlation coefficients and variation among readings are similar to those reported for assessing accuracy and reliability in measurements of the amygdala [6], the hippocampal formation [6, 11], the basal ganglia [13, 14], the prostate [15], and the TLs [10]. The generalization that the smaller the structure measured with MR, the greater the degree of variability applies to our measurements of OBTS, for which mean intra- and interobserver variation ranged from 11.3% to 14.6%; TL values, however, ranged from 3.1% to 4.2%. Similar trends have been noted when measurements of subsegments of the TLs have been compared with full TL volumes [6] and brain volumes [9].

Kohn et al [16] used phantoms constructed from agarose gel and graphite powder that had volumes of 4.5-32.5 cm³. For three measurements by three reviewers at three separate times, the mean values were 3.3, 3.5, and 3.4 cm³ for the smallest phantom and 33.5, 33.8, and 33.7 cm³ for the largest phantom [16]. When our method of assessing mean percentage error is applied to these measurements, they deviate from the true volume (4.5 cm³ and 32.5 cm³) by as much as 27% [(4.5 - 3.3)/4.5] for the smallest phantom but just 4% [(33.8 - 32.5)/32.5] for the largest phantom. Clearly the smaller the lesion, the greater the potential error when 1 voxel is or is not erroneously included [16, 17]. When one is dealing with volumes as low as our mean OBT volume of 104.4 mm³, the potential for inaccuracy is large.

As to the issue of three-dimensional gradient-echo volumetric sequences versus spin-echo sequences for volumetric analyses, it is intuitive that the thinner the section profile, the more likely one is to achieve accuracy in measured volumes, provided that contiguous imaging is performed. Despite this axiom, Kohn et al [16] found no greater accuracy in measuring phantom volumes when they decreased their section thicknesses from 5 to 3 mm. When measuring the OBT in particular, the main reason for choosing spin-echo imaging is the susceptibility artifact that arises from the underly-
ing aerated sphenoid and ethmoid air cells and the bone. Globe-motion artifact is also problematic. Both of these artifacts are greater with three-dimensional Fourier transform gradient-echo pulse sequences than they are with two-dimensional spin-echo sequences. In the TLs, the presence of artifacts from underlying aerated temporal bones and the need to cover a greater imaging range (precluding the use of submillimeter three-dimensional Fourier transform gradient-echo sequences without markedly lengthening image times) can be an issue. The susceptibility artifacts can be addressed most readily by using greater angles of nutation and shorter echo times and by applying spoiler gradients to the T1-weighted three-dimensional Fourier transform gradient-echo sequences.

Ideally we would like to perform cadaver studies to verify our accuracy in measuring the volumes of the OBTs and TLs. The OBTs could be severed at the junction with the brain, and a water-displacement technique could be used to measure the actual volume. Unfortunately, the paucity of cadaveric brains available for research purposes, the variable presence of intracranial air in decapitated cadavers, and the difficulty of removing the brain from the skull while keeping the OBTs intact have precluded this type of verification procedure.

Nonetheless, the data provided here suggest that 3-mm contiguous two-dimensional spin-echo T1-weighted images of the OBTs and TLs provide data that can be analyzed at a workstation to provide reproducible and reliable measurements of volume.

REFERENCES