◆ Human Brain Mapping 28:892–903 (2007) ◆

A Technique for the Deidentification of Structural Brain MR Images

Amanda Bischoff-Grethe,[1,2] I. Burak Ozyurt,[1] Evelina Busa,[3] Brian T. Quinn,[3] Christine Fennema-Notestine,[1,2] Camellia P. Clark,[1,2] Shaunna Morris,[1,2] Mark W. Bondi,[1,2] Terry L. Jernigan,[1,2] Anders M. Dale,[4] Gregory G. Brown,[1,2] and Bruce Fischl[3,5,6]*

[1]*Laboratory of Cognitive Imaging, Department of Psychiatry, University of California, San Diego, La Jolla, California*
[2]*Veterans Affairs San Diego Healthcare System, San Diego, California*
[3]*Athinoula A. Martinos Center—MGH/NMR Center, Charlestown, Massachusetts*
[4]*Department of Neurosciences, University of California, San Diego, La Jolla, California*
[5]*Department of Radiology, Harvard Medical School, Charlestown, Massachusetts*
[6]*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts*

Abstract: Due to the increasing need for subject privacy, the ability to deidentify structural MR images so that they do not provide full facial detail is desirable. A program was developed that uses models of nonbrain structures for removing potentially identifying facial features. When a novel image is presented, the optimal linear transform is computed for the input volume (Fischl et al. [2002]: Neuron 33:341–355; Fischl et al. [2004]: Neuroimage 23 (Suppl 1):S69–S84). A brain mask is constructed by forming the union of all voxels with nonzero probability of being brain and then morphologically dilated. All voxels outside the mask with a nonzero probability of being a facial feature are set to 0. The algorithm was applied to 342 datasets that included two different T1-weighted pulse sequences and four different diagnoses (depressed, Alzheimer's, and elderly and young control groups). Visual inspection showed none had brain tissue removed. In a detailed analysis of the impact of defacing on skull-stripping, 16 datasets were bias corrected with N3 (Sled et al. [1998]: IEEE Trans Med Imaging 17:87–97), defaced, and then skull-stripped using either a hybrid watershed algorithm (Ségonne et al. [2004]: Neuroimage 22:1060–1075, in FreeSurfer) or Brain Surface Extractor (Sandor and Leahy [1997]: IEEE Trans Med Imaging 16:41–54; Shattuck et al. [2001]: Neuroimage 13:856–876); defacing did not appreciably influence the outcome of skull-stripping. Results suggested that the automatic defacing algorithm is robust, efficiently removes nonbrain tissue, and does not unduly influence the outcome of

Contract grant sponsor: University of California; Contract grant sponsor: National Center for Research Resources at the National Institutes of Health (NIH); Contract grant number: U24 RR021382 and projects BIRN002 and BIRN004, M01RR00827, P41-RR14075, R01 RR16594-01A1; Contract grant sponsor: Mental Illness and Neuroscience Discovery (MIND) Institute; Contract grant sponsor: NIMH; Contract grant numbers: 5K08MH01642, R01MH42575; Contract grant sponsor: NIA; Contract grant numbers: R01 AG006849, AG12674, AG04085; Contract sponser: San Diego Alzheimer's Disease Research Center; Contract grant number: P50 AG05131; Contract grant sponsor: HIV Neurobehavioral Research Center; Contract grant number: MH45294; Contract grant sponsor: NIDA; Contract grant number: 5K01DA015499; Contract grant sponsor: Department of Veterans Affairs Medical Research Service; Contract grant numbers: Research Enhancement Award Program, VA Merit Review, and Mental Illness Research, Education, and Clinical Center grants.

*Correspondence to: Bruce Fischl, Ph.D., Athinoula A. Martinos Center—MGH/NMR Center, 149 Thirteenth Street, Rm. 2301, Charlestown, MA 02129. E-mail: fischl@nmr.mgh.harvard.edu

Received for publication 17 March 2006; Revised 5 June 2006; Accepted 22 June 2006

DOI: 10.1002/hbm.20312
Published online 12 February 2007 in Wiley InterScience (www.interscience.wiley.com).

the processing methods utilized; in some cases, skull-stripping was improved. Analyses support this algorithm as a viable method to allow data sharing with minimal data alteration within large-scale multisite projects. *Hum Brain Mapp 28:892–903, 2007.*

Key words: MRI; image processing; statistics; HIPAA; algorithms; humans

INTRODUCTION

To share human data in compliance with federal, state, and local regulations, including the recently enacted Health Insurance Portability and Accountability Act of 1996 (HIPAA, http://www.hhs.gov/ocr/hipaa/), it is crucial to have in place robust practices and procedures that protect the welfare of the individuals who participate in the research. These practices must include measures that ensure the privacy of the individual. For data to qualify as sharable under the "safe harbor" regulations, one of the HIPAA-defined identifiers that must be removed is "full face photographic images and any comparable images." With the increasing resolution of morphometric MR scans, it has become possible to reconstruct detailed images showing facial anatomy (Fig. 1a). Thus, in order to share unaltered MRI images, both sites are required to provide a waiver of consent. This becomes problematic in multisite projects, particularly, those with the goal of making data available to a larger research community.

The face recognition literature has suggested that internal facial features (i.e., eyes, nose, and mouth) are particularly relevant when recognizing a familiar individual [Bruce et al., 1999; Burton et al., 1999]. Therefore, automated techniques to obscure or remove an individual's facial features from structural MR images have become an important part of the data sharing process for large-scale, multisite projects such as the Biomedical Informatics Research Network (BIRN). In addition to reducing the ability to visually identify a subject, a method must be robust, removing only nonbrain tissue while leaving brain tissue intact (Fig. 1b and Fig. 2). It should be insensitive to pulse parameters, thereby working for a variety of 3D T1-weighted sequences. Finally, the outcome of such deidentification must not change the data in such a way as to have debilitating effects on later data processing and analysis.

Although numerous automated skull-stripping algorithms are available that might be considered for deidentification purposes, their performance may be influenced by a variety of factors, such as MR signal inhomogeneities, gradient performance, and extent of neurodegeneration in the subjects studied [Smith, 2002]. A detailed study of how such variables may influence the automated performance of four common skull-stripping techniques (Brain Surface Extractor (BSE), Brain Extraction Tool (BET), 3dIntracranial, and a Hybrid Watershed (HWA)) has shown that it is difficult to achieve satisfactory results for all datasets, especially, across different subject populations [Fennema-Notestine et al., 2006]. Even with some level of manual intervention, it is difficult to create "one-size fits all" parameters such that automated skull-stripping deidentifies the subject without loss of brain tissue; manual tuning for a particular scanner and/or pulse sequence may not produce consistent results across datasets collected under that protocol [Fennema-Notestine et al., 2006; Smith, 2002]. These manually optimized parameters may also be dependent upon area of interest, such that regions not currently under study (e.g., the superior parietal regions) may be sacrificed for better results within the regions to be studied (such as the anterior temporal lobe). This is highly significant for large multisite projects in which not every individual who might legitimately have access to images can be identified when consent is obtained. An example of this circumstance might occur when meeting government mandates to make research imaging-data public to the scientific community.

A more recent approach has suggested combining multiple automated skull-stripping methods within a single meta-algorithm to optimize results [Rex et al., 2004]. While this method showed improved results over individual algorithms, for optimal results it does require training for data sets with novel contrast or signal-to-noise characteristics. Further, should new algorithms be added, determination of the best overall algorithmic combination becomes intractable. Another concern is that many of these methods may remove certain elements, such as extracranial cerebrospinal fluid (CSF), which hold some importance in some fields of research. With recent advances in combining MRI with EEG/MEG, cranial features are important for identifying electrode placement with respect to a structural MRI. These features would be removed when a skull-stripping algorithm is applied. Skull-stripping methodology, then, may not be sufficiently reliable for large-scale, automated deidentification purposes.

In the current study, we introduce an automated "defacing" algorithm that removes only identifiable facial features from MR volumes, and we present the results of an investigation of the performance of this defacing algorithm on image sets that differed by age and diagnosis. First, image volumes were examined qualitatively for preservation of brain tissue after defacing. Second, to help quantify the outcome, we compared skull-stripped volumes, using Hybrid Watershed [Ségonne et al., 2004, in FreeSurfer] or Brain Surface Extractor [Sandor and Leahy, 1997; Shattuck et al., 2001], with both defaced and nondefaced datasets. These skull-stripping algorithms were selected based upon their performance in a previous analysis [Fennema-Notestine et al., 2006] as being fairly robust

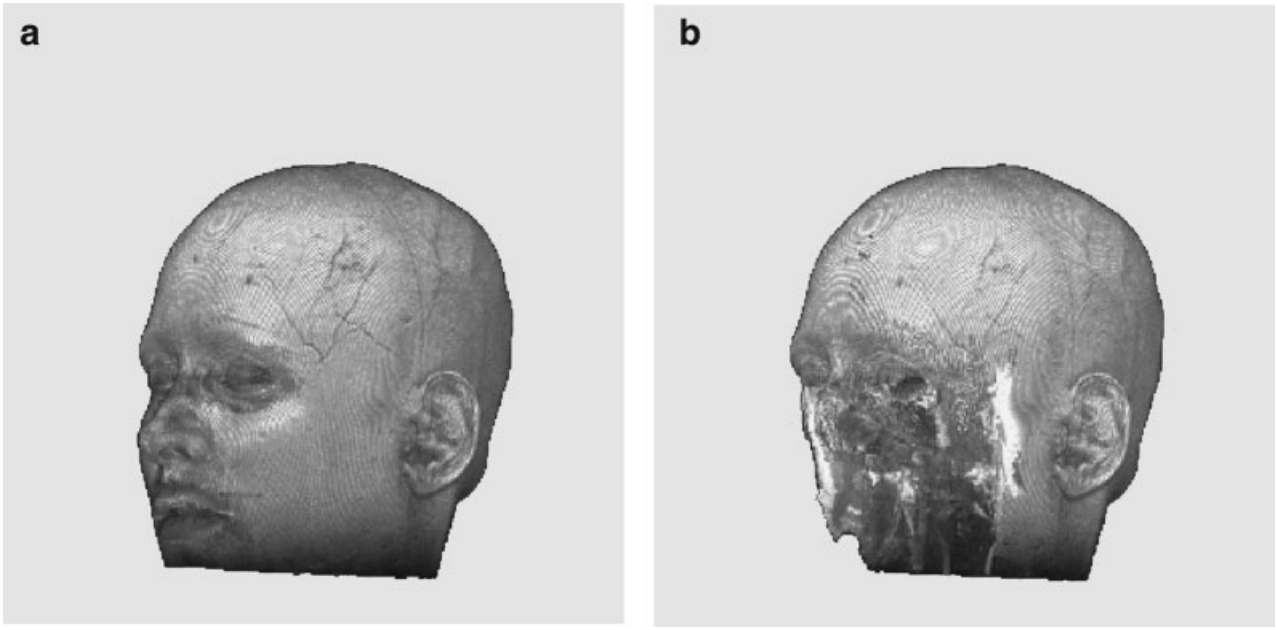

Figure 1.
An example of a 3-D reconstruction of a T1-weighted dataset (**a**) before and (**b**) after application of the defacing algorithm. The defacing algorithm removed identifying facial features while preserving brain tissue for future analyses.

across diagnoses. After visually inspecting whole brain volumes, we focused upon six slices in regions typically problematic in differentiating brain from nonbrain for more detailed quantitative analysis. These slices were compared to two manually-created gold standards to determine (1) similarity of results across methods; (2) sensitivity to classification of tissue as brain; and (3) ability to specify tissue as nonbrain. We hypothesized that defacing would successfully remove nonbrain tissue and not appreciably modify the performance of the skull-stripping algorithms employed.

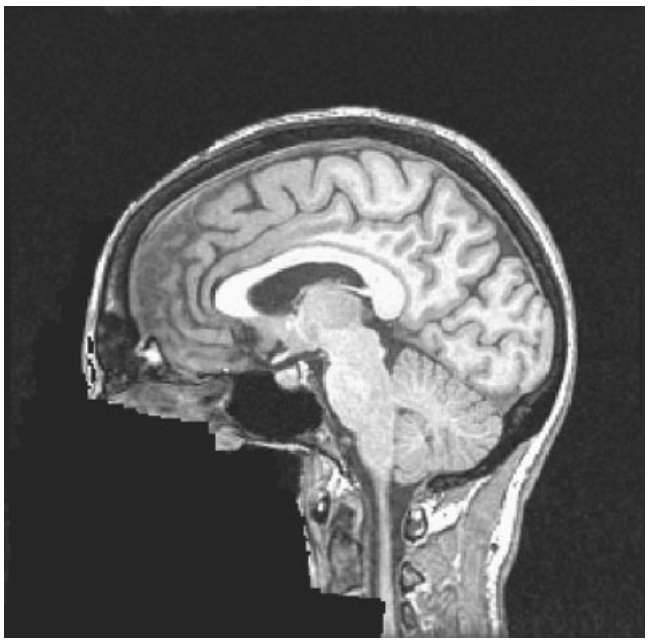

Figure 2.
A sagittal slice from a defaced dataset illustrating how nonbrain voxels in the face region are set to a fill value of zero.

MATERIALS AND METHODS

MR Image Sets

Data collected using a common structural gradient-echo (SPGR) T1-weighted pulse sequence were examined. The datasets were collected on a GE 1.5T magnet located at the VA San Diego Healthcare System MRI Facility that was subject to regular hardware and software upgrades over time. Two large datasets were used for the purposes of qualitative visual inspection: 278 *Legacy* datasets were collected over 4 years in the mid to late 1990s (June 1994 to July 1998) using the following parameters: TR = 24 ms; TE = 5 ms; NEX = 2; flip angle = 45°; FOV = 24 cm; 1.2-mm contiguous sagittal sections. Sixty-four *Contemporary* datasets were collected over an 11-month period May 2002 to April 2003 using the following settings: TR = 20 ms; TE = 6 ms; NEX = 1; flip angle = 30°; FOV = 25 cm; 1.5-mm contiguous sagittal sections. A subset of these *Contemporary* data was employed in the qualitative skull-stripping assessment described in more detail later. The University

TABLE I. Diagnostic group information for datasets used in the statistical analyses

Diagnostic group	Age	Gender	MMSE
Young control	33.0 ± 15.1 (21–54)	2F/2M	N/A
Elderly control	74.5 ± 1.7 (72–76)	2F/2M	N/A
Unipolar depressed	40.8 ± 10.8 (21–54)	3F/1M	N/A
Alzheimer's disease	75.5 ± 1.7 (72–78)	1F/3M	23.2 ± 2.5 (22–27)

Values given are mean ± SD, and values inside parentheses indicate ranges. N/A: not available.

of California, San Diego, institutional review board approved all procedures, and written informed consent for image acquisition was obtained from all subjects.

Diagnostic Groups

Four populations were used throughout the analysis, consisting of depressed (DEP), Alzheimer's (AD), young control (YNC), and elderly control (ENC) groups. AD severity was measured with the Mini-Mental State Examination [Folstein et al., 1975]. Of the datasets used for qualitative visual inspection, the 278 *Legacy* datasets included 50 ENC, 92 AD, 96 YNC, and 40 DEP participants, and the 64 *Contemporary* datasets included 36 ENC, 4 AD, 5 YNC (3 subjects had 2 longitudinal sessions), and 8 DEP (all subjects had 2 longitudinal sessions) participants. From these, 16 *Contemporary* datasets (4 ENC, 4 AD, 4 YNC, 4 DEP) were selected for further quantitative statistical analysis. The YNC and DEP groups were similar on age and education, as were the ENC and AD groups (Table I) for this reduced dataset. *Legacy* datasets were not used in the statistical analysis due to their increased need for manual intervention during the skull-stripping process.

Defacing Algorithm

An algorithm was developed that uses models of nonbrain structures for removing facial features that may potentially allow the identification of a subject/patient from their MR scan. An atlas of face membership was created by manually labeling the facial features of 10 subjects. These facial features comprised the entire front of the head. To remove facial features from novel images, an optimal linear transform using both brain and nonbrain was computed for the input volume [Fischl et al., 2002]. Next, a brain mask was constructed by forming the union of all voxels whose prior probability of being any brain tissue was nonzero. This mask was then morphologically dilated n times ($n = 7$) to yield a binary volume, the nonzero values of which indicate the presence of brain tissue within nx millimeters. Here, x is the size of a voxel in millimeters, and the volumes were interpolated to ensure isotropic voxel dimensions. The number of dilations functions as a buffer and is related to the accuracy of the linear transform. It is essentially the distance from the brain in which one can be confident the linear transform can localize. The deidentification procedure involved finding all voxels that were outside the mask, but had a nonzero probability of being a facial feature, and setting them to zero. Voxels within x and $2x$ of the detected brain mask were removed if the Mahalanobis distance, using mean and covariances estimated from a manually labeled training set, to any brain tissue was low; that is, if the voxel intensity did not appear similar to brain tissue intensity. This was particularly useful for removing fatty tissue from the orbital areas, for example. As the face atlas was created from T1-weighted images, it therefore should be used only with T1-weighted datasets. The defacing algorithm took ~25 min per dataset to run on a Dell Precision Xeon 3.20 GHz with 2 GB RAM.

Automated Skull-Stripping Methods

To quantitatively assess whether the defacing algorithm removed brain tissue and/or influenced the performance of commonly used software, two different skull-stripping methods were applied to (1) 16 normalized, nondefaced datasets, and (2) the same 16 normalized datasets after defacing. The two skull-stripping methods employed included Brain Surface Extractor [Sandor and Leahy, 1997; Shattuck et al., 2001], a tool shown to have high specificity in finding the cortical surface, and a Hybrid Watershed algorithm [Ségonne et al., 2004], a relatively more sensitive tool that often results in a conservative strip that rarely removes any brain tissue [Fennema-Notestine et al., 2006]. For image normalization, nonparametric nonuniform intensity normalization [N3; Sled et al., 1998] was used; this locally adaptive bias correction algorithm was chosen for its applicability to raw, unstripped datasets and its performance relative to other methods [Arnold et al., 2001]. The two skull-stripping algorithms are briefly described as follows:

1. Hybrid Watershed Algorithm (v. 1.21). HWA [Fischl et al., 2002; Ségonne et al., 2004; in FreeSurfer, http://surfer.nmr.mgh.harvard.edu] is a hybrid of a watershed algorithm [Hahn and Peitgen, 2000]; it assumes white matter connectivity to determine a local optimum of the intensity gradient, and a deformable surface model [Dale et al., 1999], which is used to apply corrections when the connectivity assumption does not hold. Optionally, a statistical atlas can be used to verify and potentially correct the surface estimate. In the present study, the atlas-based option was not finalized for v. 1.21; therefore, for automated processing, the default parameters without the atlas option were utilized. On average, HWA required less than 8 min of processing time per dataset.
2. Brain Surface Extractor (v. 3.3). BSE [Sandor and Leahy, 1997; Shattuck et al., 2001; in BrainSuite, http://brainsuite.usc.edu/] uses anisotropic diffusion, edge detection, and morphologic erosion to segment the brain. Briefly, the algorithm first detects the boundary between brain and skull that is then filtered

with anisotropic diffusion to smooth small image gradients while retaining larger ones that correspond to strong edges in the image. Because noise in the image may lead to a result that does not separate the brain from the rest of the head, morphologic processing techniques are used to identify and refine the brain surface. The parameters employed were a sigma of 0.8 for the edge detection, with five iterations of the anisotropic filter at a diffusion constant of 5.0. BSE required about 15 s of processing time per dataset.

Manual Skull-Stripping

For more refined analyses, two anatomists manually stripped six sagittal slices from each of the 16 raw contemporary MR datasets to provide a criterion against which to judge the automated outcomes of skull stripping with/without prior defacing (Fig. 3). Note that whole brain volumes from which these slices were taken were visually inspected as described later. Both anatomists (CPC and SM) were experienced in neuroimaging, with training in both neuroscience and neuroanatomy. With the guidance of a trained neuroanatomist (CFN), the anatomists completed four sample datasets, not included in the present study, to formalize a set of criteria for skull-stripping. If the anatomists were unable to definitively classify tissue as brain or nonbrain, they were instructed to conservatively include the tissue. Six sagittal slices (three per hemisphere) that passed through regions that are difficult for both manual and automated skull-stripping methods were selected. These included slices passing through the anterior medial temporal, anterior inferior frontal, posterior cerebellar regions, and posterior occipital regions. These slices were chosen due to their difficulty in separating brain from nonbrain tissues on T1-weighted images. On average, it took 1 h for each anatomist to manually strip the six slices (about 10 min per slice) in a given dataset. The Jaccard similarity coefficient between these two neuroanatomists was 0.938 across all datasets; more detailed analyses of the anatomists' performance has been described elsewhere [Fennema-Notestine et al., 2006] and will therefore not be included in this report.

Data Processing

The 16 contemporary datasets selected for statistical analysis were processed in four ways: (1) the original images were normalized with N3, followed by skull-stripping with HWA or (2) BSE; (3) the original images were defaced, followed by image normalization with N3, and finally skull-stripping with HWA or (4) BSE. Therefore, for each initial dataset, there were four processed datasets for subsequent analysis.

Analytical Methods

All datasets were subjected to qualitative review. Defaced images were visually inspected to determine whether the defacing mask encroached upon brain tissue. Visualization was conducted using AFNI [Cox, 1996; http://afni.nimh.nih.gov/afni/] to examine each subject's structural image across all three planes by overlaying the mask onto the original anatomical image. After visual inspection to determine if there was a loss of brain tissue, the image was rendered into a three-dimensional image and further inspected to determine if facial features were adequately removed (Fig. 4).

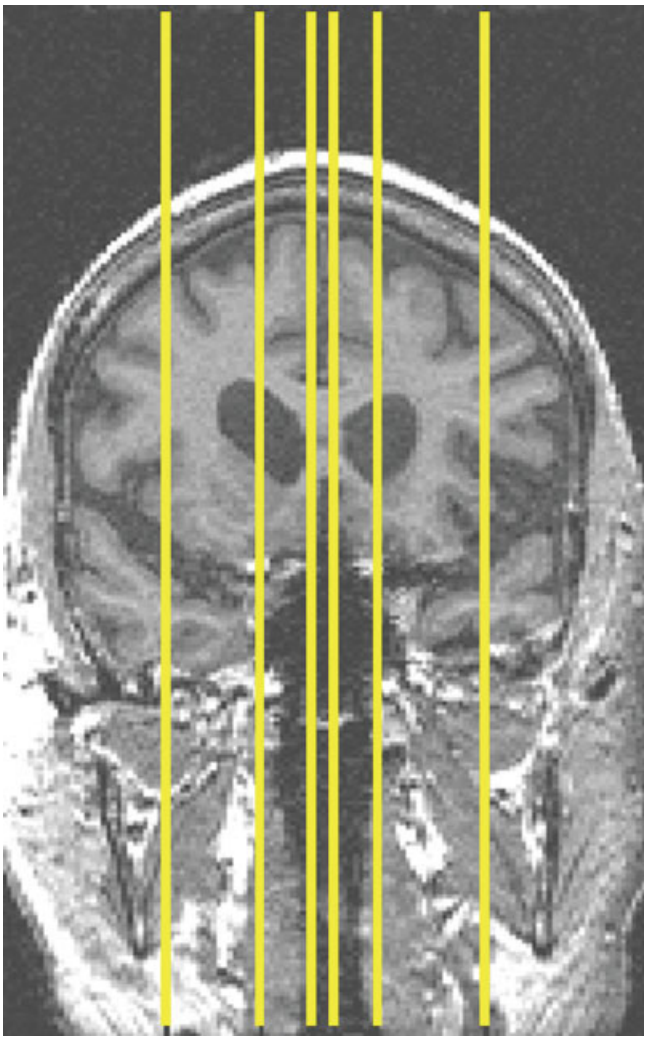

Figure 3.
Standard location of the manually stripped slices as demonstrated on a coronal image. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Analyses were conducted by using all sagittal slices in which each subject's representative slice contained brain tissue (total slices per dataset = 86). Within this analysis, which we will hereafter term the *Whole Brain* analysis, comparisons were made between skull-stripped and defaced + skull-stripped images, using either BSE or HWA as the skull-stripping method. This technique was selected

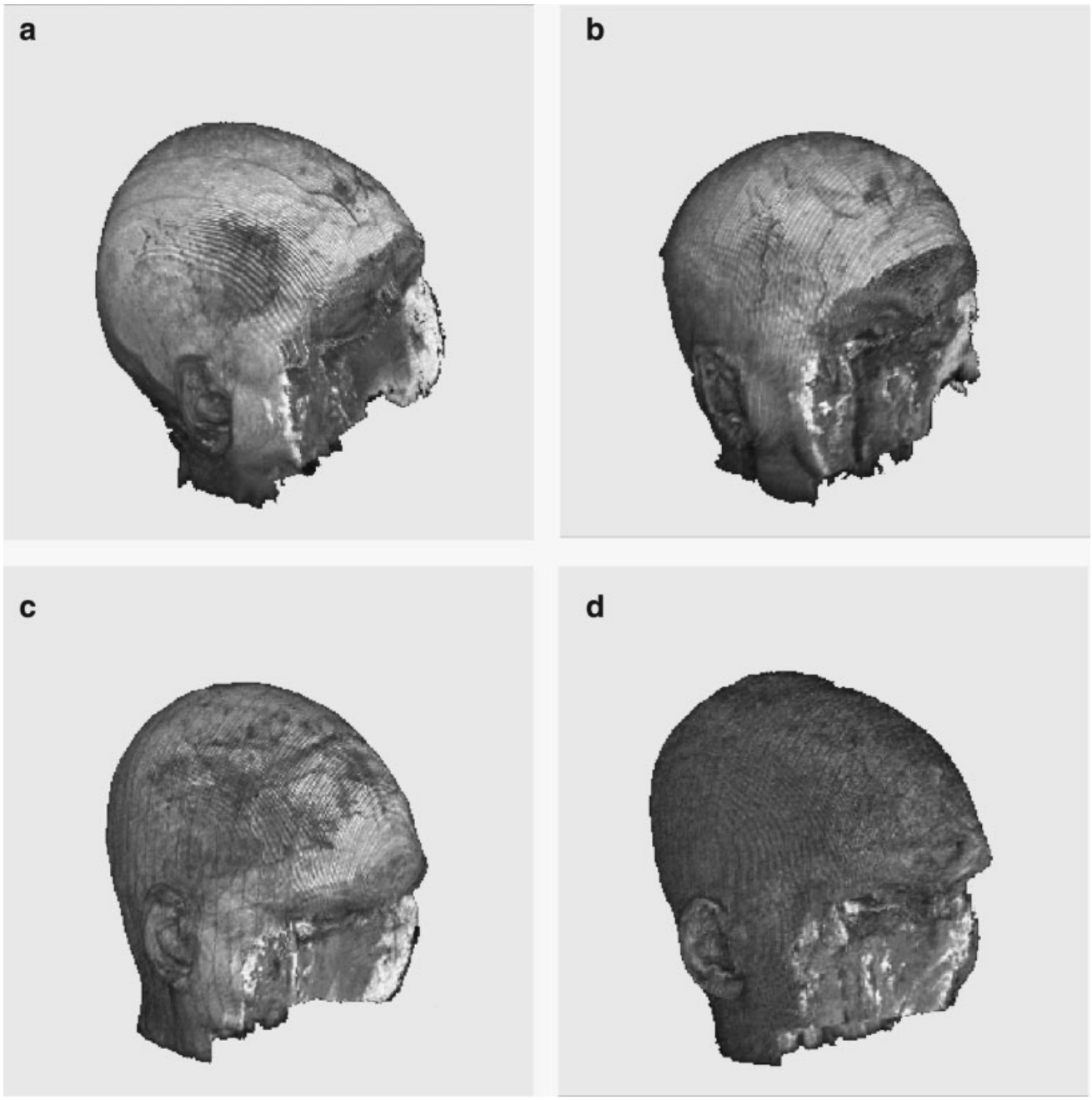

Figure 4.
Examples of successful application of the defacing algorithm. Top row: elderly control subject (left) and Alzheimer's patient (right); bottom row: young normal control subject (left) and unipolar depressed patient (right).

due to performance differences between the two automated methods. As previously stated, HWA tends to be highly sensitive to brain tissue and produces conservative skull-stripped images, whereas BSE is very specific and comes closer to the brain surface, although in some cases brain tissue may be removed [Fennema-Notestine et al., 2006]. Because the *Whole Brain* analysis relied solely upon automated methods, a second analysis was employed: Six sagittal slices, known to be problematic in skull-stripping, were selected from the HWA-stripped and the defaced + HWA-stripped datasets and compared statistically with manually stripped images by two trained anatomists (i.e., gold standard) to determine similarity across methods as well as ability to correctly classify voxels as brain or nonbrain. HWA was

selected due to its ability to generally conserve brain tissue across a number of patient populations [Fennema-Notestine et al., 2006]. Given that the six slices chosen were known to be difficult ones for determining brain vs. nonbrain, we felt it was prudent to use a conservative skull-stripping algorithm; this was consistent with the instructions to the trained anatomists, who were told to retain tissue if it were difficult to classify.

The four statistical methods chosen for both *Whole Brain* and *Six Slice* analyses were as follows:

1. Set-Difference. This technique examined the difference in the number of voxels left behind by skull-stripping only to the number of voxels that were removed by defacing only. The original image volume was skull-stripped, and the resulting mask was applied to the defaced image volume. The number of nonzero voxels was tabulated to determine how many voxels the defacing algorithm removed that the skull-stripping algorithm left behind in a stripped, nondefaced volume. This difference can occur if the defacing algorithm is overly sensitive to nonbrain tissue, or if it removes brain tissue from the original image.
2. Jaccard Similarity Index. The Jaccard measures the degree of correspondence, or overlap, for each image slice. It is formulated as

$$\mathrm{JSC}(A,B) = (A \cap B)/(A \cup B)$$

   where A, the criterion, and B are the images being compared [Jaccard, 1912]. For *Whole Brain*, the skull-stripped only datasets were the criterion, and the defaced + skull-stripped datasets were compared with them. Datasets stripped with HWA were treated separately from those stripped with BSE. In the *Six Slice* experiment, we used the manually stripped datasets as criterion, comparing the HWA and defaced + HWA datasets with them. A score of 1.0 indicates complete overlap or agreement, whereas a score of 0.0 indicates no overlap.
3. Hausdorff Distance Comparison. The Hausdorff examines the degree of mismatch between the contours of two image sets [Huttenlocher et al., 1993]. Given two finite point sets, $A = \{a_1,\ldots, a_m\}$ and $B = \{b_1,\ldots, b_n\}$, where A and B are sets of points along the contour of a skull-stripped brain slice, the Hausdorff distance is defined as

$$H(A,B) = \max(h(A,B), h(B,A))$$

   The directed Hausdorff distance from A to B is defined as

$$h(A,B) = \max_{a\in A} \min_{b\in B} \|a - b\|$$

4. Expectation-Maximization Algorithm. This algorithm calculates the maximum likelihood estimate of the underlying agreements among all methods [Warfield et al., 2002]. There are two main outcomes of this method:
   (a) Sensitivity: This metric determines the relative frequency of correct brain classification by one method relative to all methods.
   (b) Specificity: This determines the relative frequency of correct nonbrain classification by one method relative to other methods.

   The a priori probabilities for all voxels for each slice of each subject tested were set to 0.5; this indicates there was no initial knowledge about ground truth. Initial estimates for sensitivity and specificity were set to 0.9. The termination criterion for convergence set the root mean square error to <0.005.

The goal of using a variety of statistical outcomes was to demonstrate that the findings were robust, converged across methods, and were sufficiently generalizable. While it was expected that there would be some correlation amongst the different methods, inherently these metrics measure different aspects of the data.

Descriptive statistics (mean and standard error) were calculated for all analyses, both across and within each diagnostic group. This provided us with an initial impression of performance for all datasets, as well as whether a particular diagnosis produced unusual results with respect to other patient populations. These were followed by a series of mixed design analyses at the conventional α of 0.05 for statistical significance. Between-subjects effects were examined for diagnosis. Where appropriate, univariate within-subject repeated measures analysis of variance were examined for Slice (either the 86 slices that contain brain tissue within each subject or for Slices 1–6 as seen in Fig. 3), to potentially identify region-specific trends in defacing, and method (BSE for *Whole Brain* analysis only; HWA for both *Six Slice* and *Whole Brain* analyses; Gold Standard for *Six Slice* analysis only). All analyses used the Huynh-Feldt correction since sphericity could not be assumed. For all repeated measures reaching significance, partial η-squared (η^2) values were reported as an estimate of effect size. Because we did not wish to make assumptions regarding the *Six Slice* analysis distribution, a rank order analysis was employed.

RESULTS

Qualitative Review

To date, 342 T1-weighted datasets have been defaced and visually inspected; none of these defaced datasets have had brain tissue removed. Three-dimensional renderings of the defaced images indicated that identifying facial features (eyes, nose, mouth, chin) were removed from, on average, the nasion downward (Fig. 4).

For the 16 contemporary datasets used for subsequent quantitative analysis, visual inspection showed that defacing prior to automated skull-stripping with HWA tended

TABLE II. Mean and standard error for the Jaccard Similarity and Hausdorff Distance analyses (*Whole Brain* analysis)

Method	HWA vs. DEF + HWA	BSE vs. DEF + BSE
Jaccard similarity	0.965 (0.001)	0.967 (0.002)
Hausdorff distance	3.204 (0.075)	2.76 (0.245)

The skull-stripped only datasets were employed as criterion, with HWA as criterion for the DEF + HWA comparison, and BSE for the DEF + BSE comparison, respectively. BSE: Brain surface extractor; DEF: defaced; HWA: hybrid watershed.

to remove eyes and leave only a small amount of nonbrain tissue in the ventral anterior temporal lobes, whereas HWA without prior defacing tended to leave behind eyes in several datasets. For one dataset (DEP), visual inspection revealed that the image failed automated skull-stripping; HWA left behind a large amount of nonbrain tissue, including the face, on the ventral portion of the brain. It should be noted that when the image was defaced prior to skull-stripping, the nonbrain tissue present with skull-stripping only was removed. Thus, in this case, defacing significantly improved the subsequent performance of skull-stripping. However, due to the poor performance of automated HWA on this dataset, it was excluded from the statistical analyses so as not to skew the outcomes. In general, however, defacing the image prior to automated skull stripping with HWA did visually appear to improve the quality of skull-stripping (with respect to the removal of unwanted nonbrain tissue) in most datasets.

Visual inspection of automated skull-stripping with BSE (without prior defacing) tended to be more specific, but in some cases nonbrain tissue was left around the parietal region, and, in one case, some brain tissue was removed in the orbitofrontal cortex. Three left excessive amounts of nonbrain tissue. When the image was defaced prior to skull-stripping with BSE, some cases left extra tissue around the parietal region, and three left too much nonbrain tissue, typically retaining the skull (but not CSF) and tissues surrounding the spinal cord. In general, BSE (with or without prior defacing) removed a significant portion of brain tissue, although some voxels removed were close to the brain surface and were visually difficult to classify as brain or nonbrain. In two subjects (1 YNC, 1 AD), there was a tremendous disparity in the quality of the skull-stripping with vs. without prior defacing. The YNC outcome was poor with defacing followed by skull-stripping, but acceptable when only skull-stripped; defacing + skull-stripping retained most of the nonbrain tissue. Conversely, the AD outcome was improved when the image was defaced and skull-stripped; the nonbrain tissue left behind with skull-stripping only were removed when the data were both defaced and skull-stripped. These datasets were subsequently excluded from the statistical analyses.

Whole Brain Statistical Comparison

Here, the set difference comparison indicated that 2.612% (SD 2.433) of the voxels retained by skull stripping with HWA were removed by defacing. Visual examination indicated that the retained voxels were nonbrain tissue. These regions tended to be in the vicinity of the eyes. HWA uses intensity information during the normalization process, hence the removal of the eyes through defacing, which are quite bright on T1-weighted images, likely led to an improvement in intensity differentiation of brain from nonbrain. The more lateral slices tended to look the same with or without prior defacing.

The subsequent analyses using the Jaccard Similarity, Hausdorff Distance, and E-M methods were conducted by comparing the automated (HWA or BSE) stripped datasets with and without prior defacing. The mean descriptives for the Hausdorff Distance and Jaccard Similarity were not appreciably different (Table II). We found a significant main effect of slice for HWA for the Jaccard Similarity coefficient ($F(3.651, 40.158) = 4.886$, $P = 0.003$, partial $\eta^2 = 0.308$); all other analyses failed to reach significance. These slice effects were due to the combination of HWA with defacing performing slightly less conservatively than HWA alone. As previously noted, defacing prior to HWA tended to have no appreciable effect on the lateral slices. The slices around the eyes tended to be more conservatively stripped with HWA alone, such that more nonbrain tissue remained. None of the main effects or interactions reached significance for the Hausdorff Distance analyses.

The descriptive statistics for EM Sensitivity and Specificity for both HWA and BSE was at or near ceiling (Table III), indicating that for either case, defacing did not appreciably interfere with the abilities of HWA or BSE to differentiate brain from nonbrain tissue. Therefore, a repeated measures analysis was not pursued.

Six Slice Statistical Comparison

Using the set difference comparison, 2.538% (SD, 2.572) of the voxels retained by skull-stripping (with HWA) were removed by defacing. Visual examination showed these retained voxels were nonbrain tissue, and that defacing prior to skull-stripping tended to remove more nonbrain

TABLE III. Mean and standard error for the expectation–maximization algorithm for the sensitivity and specificity analyses (*Whole Brain* analysis)

Method	HWA	DEF + HWA	BSE	DEF + BSE
EM sensitivity	0.895 (0.008)	0.879 (0.008)	0.859 (0.009)	0.842 (0.009)
EM specificity	1.000 (0.000)	1.000 (0.000)	0.999 (0.000)	1.000 (0.000)

BSE: Brain surface extractor; DEF: defaced; HWA: hybrid watershed.

voxels in the ventral frontal areas than skull-stripping alone. Additionally, defacing prior to skull-stripping resulted in more nonbrain tissue removal in areas such as along the cerebellum ventrally and in superior frontal areas.

Descriptive analyses of the Jaccard Similarity and Hausdorff Distance methods (Table IV) suggest the results were similar both across methods (HWA with and without prior defacing) and across anatomists. However, defacing tended to improve performance, particularly in patient populations (Fig. 5). We compared the outcome of HWA with and without prior defacing with the gold standards (Anatomist 1 and Anatomist 2). A rank order analysis was used to determine if there were significant differences between methods (1) across all slices, and (2) on a slice by slice basis (Table V). For the Jaccard Similarity method, the across slices analysis revealed significance by anatomist (Anatomist 1: $P = 0.001$; Anatomist 2: $P = 0.009$), but only Slice 5 for Anatomist 2 ($P = 0.04$) was significant within slice. The Hausdorff Distance method likewise revealed across slice significance (Anatomist 1: $P = 0.007$; Anatomist 2: $P = 0.02$), but no within slice analysis reached significance (although it did approach significance for Slice 2 for both anatomists).

The descriptive analyses for EM sensitivity and specificity were at or near ceiling for the two methods (Table VI). Thus, defacing did not appreciably influence HWA in its ability to correctly classify tissue as brain or nonbrain. Due to the inherent difficulty in differentiating results with essentially no standard error, rank order analyses were not pursued.

TABLE IV. Mean and standard error for the Jaccard similarity and Hausdorff distance analyses (*Six Slice* analysis)

Method	Anatomist	HWA	DEF + HWA
Jaccard similarity	Anatomist 1	0.861 (0.010)	0.876 (0.010)
	Anatomist 2	0.871 (0.010)	0.887 (0.010)
Hausdorff distance	Anatomist 1	11.453 (0.128)	10.138 (0.113)
	Anatomist 2	11.383 (0.127)	9.952 (0.111)

DEF: defaced; HWA: hybrid watershed.

DISCUSSION

During recent years, there has been an increase in the number of large-scale projects that entail sharing of data across sites. NIH has initiated a data sharing policy, which requires researchers with NIH-funded grants above a certain monetary threshold to make their final research data available to other investigators. These data include human subject data acquired for basic or clinical research. With the recent enactment of HIPAA, researchers in the neuroimaging field have the added complication of removing identifying facial features from morphometric scans, in order to make the images unlike a facial photograph, without the removal or distortion of brain tissue. Most university institutional review boards require HIPAA compliance; therefore, in order to share data it must be deidentified as described by the Privacy Rule. These rules include the omission of "full facial photographic images and any comparable images," unless informed consent is obtained from the subject to share facial images. One solution has been to apply skull-stripping to the data, as is suggested by the fMRI Data Center, a neuroimaging data repository at Dartmouth College (http://www.fmridc.org). However, our experience has shown that automated skull-stripping algorithms are far from perfect and might remove brain tissue due to a variety of issues, including the subject population and scanner performance during data acquisition [Fennema-Notestine et al., 2006; Smith, 2002]. Human intervention is often required to minimize brain tissue loss, a time consuming process that is untenable when working with large datasets. Additionally, the variation in the performance of different automated skull-stripping algorithms further brings into question whether potentially vital information may be retained with one algorithm but removed by another. Therefore, as part of the BIRN initiative, we explored a possible solution to automate the deidentification of morphometric T1-weighted images that would not remove brain tissue or extracranial CSF. The defacing algorithm has been approved by the Institutional Review Boards within the BIRN consortium as sufficient for deidentification of anatomical MRI images, thus allowing for the sharing of neuroimaging data across research sites associated with the project. Our algorithm protects against casual identification of subjects. While skull-stripping takes the anonymization one step further than defacing, it may not be useful under all conditions. The loss of cranial features interferes with research combining MRI and EEG/MEG, and the technique may remove certain tissues and fluids, such as extracranial CSF, that are of interest for some fields of research.

The defacing algorithm employed herein has been conclusively shown to remove identifying facial features without disturbing brain tissue, and provides a reliable method that can be applied automatically with little human intervention required to review the outcome. The algorithm is very robust; our visual inspection of 342 datasets (some of them of poor quality) failed to find datasets in which brain tissue was removed. While the processing time is greater than that of the more widely used skull-stripping algorithms [25 min, compared with 15 s to 8 min as reported by Fennema-Notestine et al., 2006], our experience has been that it often takes far longer to skull-strip images due to manual tuning of the parameters. The algorithm can handle a variety of data formats (DICOM, AFNI, ANALYZE, etc.), and optional parameters allow users to, for example, adjust the defacing radius (i.e., distance from the brain that is stripped), as well as the intensity values of the removed voxels.

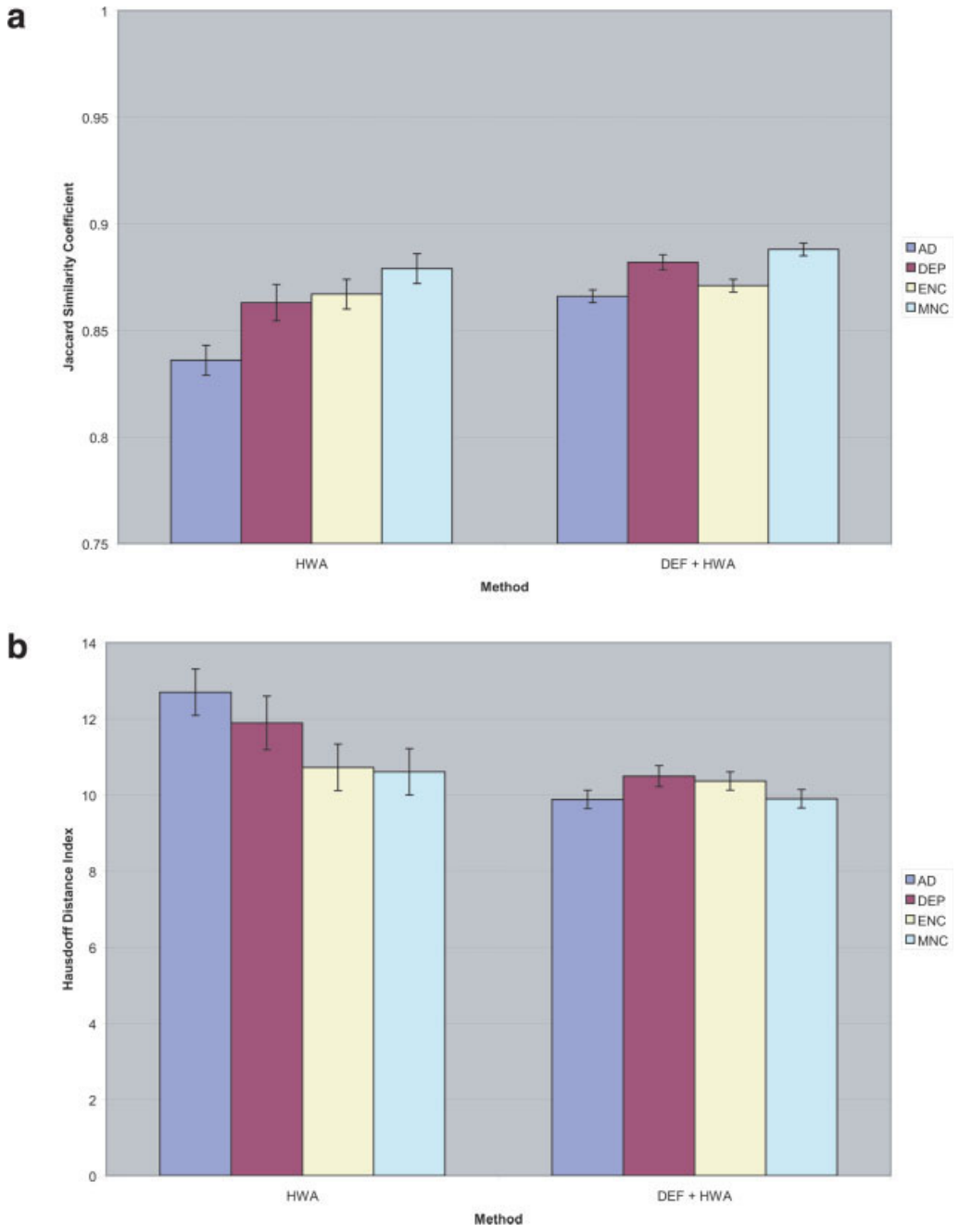

Figure 5.

Mean (standard error bars) for (**a**) Jaccard Similarity Coefficient (JSC) and (**b**) Hausdorff Distance for Diagnosis by Method relative to the manually stripped slices for Anatomist 1 (*Six Slice* analysis). DEF: defaced; HWA: Hybrid Watershed. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

While our primary goal was to determine that the defacing algorithm did not remove brain tissue, it is worthwhile noting that defacing did not have a detrimental effect on subsequent data processing. Overall, defacing prior to automated skull-stripping did not interfere with the chosen skull-stripping techniques. In some cases, defacing prior to skull stripping improved the quality of automated skull-stripping, such that more nonbrain tissue was removed. In one case, defacing prior to skull-stripping achieved poor results; this is not a limitation of defacing

TABLE V. The *P*-values from a rank order analyses by slice for the Jaccard similarity and Hausdorff distance methods (*Six Slice* analysis)

Method	Anatomist	Across Slices	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Slice 6
Jaccard similarity	Anatomist 1	0.010*	0.724	0.120	0.373	0.152	0.221	0.206
	Anatomist 2	0.009**	0.494	0.178	0.178	0.178	0.040*	0.120
Hausdorff distance	Anatomist 1	0.007**	0.129	0.051	0.494	0.319	0.182	0.450
	Anatomist 2	0.020*	0.455	0.080	0.184	0.130	0.252	0.176

**$P < 0.01$; *$P < 0.05$.

per se, but does clearly suggest one be mindful of the skull-stripping methodologies applied following defacing. Defacing likely influenced BSE's edge-detection algorithm; selecting a different set of parameters may have improved the outcome. Because the purpose of this experiment was to determine if defacing removed brain tissue by using automated skull-stripping as a metric for analysis, manual intervention to improve results was not pursued.

It should also be made clear that the two skull-stripping algorithms used, HWA and BSE, use different methodologies to remove skull and nonbrain tissue, and hence have different outcomes whether or not defacing was applied before automated skull-stripping. These differences would influence the whole-brain analyses; HWA showed a main effect of slice when comparing automated skull stripping with and without prior defacing. Because HWA tends to be conservative to the point of leaving nonbrain tissue, including CSF, behind, whereas defacing operates primarily on the slices in which prominent facial features are present (e.g., eyes vs. cheek), the effect of slice is no doubt related to the voxels that defacing removed which HWA might retain. This was supported by the set-difference comparison. The voxels retained by skull-stripping with HWA that were removed by defacing were generally located in the regions surrounding the eye. These differences may be reduced had we chosen to manually select parameters that would give the best skull-stripping performance; however, our goal was not to review the merits of skull-stripping algorithms, nor examine their capabilities with and without human intervention.

One limitation of the proposed algorithm is that it can only be applied to T1-weighted datasets since the face atlas was constructed with T1-weighted images. However, if a T1-weighted image is acquired in addition to other image types (e.g., proton density or T2-weighted images), the mask generated during the defacing process of the T1-weighted image may be used to deface these other co-registered image types as well. Our preliminary exploration with defacing non-T1-weighted data has shown that a minimal amount of effort on the part of the researcher is required, and that visual inspection of these non-T1-weighted images indicated that brain tissue was untouched. An improvement to this algorithm would be the creation of T2-weighted and proton density atlases to enable it to function on differently weighted acquisitions.

Overall, we determined that the defacing algorithm does an effective job of removing facial features without sacrificing brain tissue. The results of defacing do not interfere with subsequent data processing, and in fact in some cases appears to make subsequent skull stripping more robust. The algorithm is fully automated and can be scripted to process large quantities of data, making it easy to deidentify data for subsequent sharing in multisite projects.

TABLE VI. Mean and standard error of the sensitivity and specificity from the expectation–maximization (EM) analysis for each method (*Six Slice* analysis)

Method	Anatomist 1	Anatomist 2	HWA	DEF + HWA
EM sensitivity	0.883 (0.010)	0.893 (0.010)	0.999 (0.011)	0.998 (0.011)
EM specificity	1.000 (0.011)	1.000 (0.011)	0.992 (0.011)	0.998 (0.011)

DEF: defaced; HWA: hybrid watershed.

ACKNOWLEDGMENTS

Principal Investigator 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. The authors thank John Olichney for his comments.

REFERENCES

Arnold JB, Liow JS, Schaper KA, Stern JJ, Sled JG, Shattuck DW, Worth AJ, Cohen MS, Leahy RM, Mazziotta JC, Rottenberg DA (2001): Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects. Neuroimage 13: 931–943.

Bruce V, Henderson Z, Greenwood K, Hancock PJB, Burton AM, Miller P (1999): Verification of face identities from images captured on video. J Exp Psychol Appl 5:339–360.

Burton AM, Wilson S, Cowan M, Bruce V (1999): Face recognition in poor-quality video: Evidence from security surveillance. Psychol Sci 10:243–248.

Cox RW (1996): AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162–173.

Dale AM, Fischl B, Sereno MI (1999): Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9:179–194.

Fennema-Notestine C, Ozyurt IB, Brown GG, Clark CP, Morris S, Bischoff-Grethe A, Bondi MW, Jernigan TL, Fischl B, Segonne F, et al. (2006): Quantitative evaluation of automated-skull-stripping methods applied to contemporary and legacy images: Effects of diagnosis, bias correction, and slice location. Hum Brain Mapp 27:99–113.

Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, et al. (2002): Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. Neuron 33:341–355.

Fischl B, Salat DH, van der Kouwe AJW, Makris N, Segonne F, Quinn BT, Dale AM (2004): Sequence-independent segmentation of magnetic resonance images. Neuroimage 23 (Suppl 1):S69–S84.

Folstein M, Folstein S, McHugh P (1975): Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12:189–198.

Hahn HK, Peitgen H-O (2000): The skull stripping problem in MRI solved by a single 3D watershed transform. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention. Lecture Notes in Computer Science. New York: Springer, Berlin Heidelberg, Vol. 1935. pp 134–143.

Huttenlocher DP, Klanderman GA, Rucklidge WJ (1993): Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Mach Intell 15:850–863.

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

Rex DE, Shattuck DW, Woods RP, Narr KL, Luders E, Rehm K, Stolzner SE, Rottenberg DA, Toga AW (2004): A meta-algorithm for brain extraction in MRI. Neuroimage 23:625–637.

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

Ségonne F, Dale AM, Busa E, Glessner M, Salat D, Kahn HK, Fischl B (2004): A hybrid approach to the skull stripping problem in MRI. Neuroimage 22:1060–1075.

Shattuck DW, Sandor-Leahy SR, Schaper KA, Rottenberg DA, Leahy RM (2001): Magnetic resonance image tissue classification using a partial volume model. Neuroimage 13:856–876.

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

Smith SM (2002): Fast robust automated brain extraction. Hum Brain Mapp 17:143–155.

Warfield SK, Zou KH, Wells WM (2002): Validation of Image Segmentation and Expert Quality with an Expectation-Maximization Algorithm. Heidelberg, Germany: Springer-Verlag. pp 298–306.