

The BIRN De-identification and Upload Pipeline (BIRNDUP)

BWH (SPL): S. Pieper, J. Sacks, Duke (NIRL): B. Boyd, UCI: D. Wei,

MGH (NMR): B. Fischl, UCSD (fMRI): A. Grethe, C. Fennema-Notestine, UCLA(LONI): D. Rex, A. Toga

Abstract:

Assembling and sharing a multi-site database of high-resolution structural MRI scans while protecting subject privacy to comply with HIPAA guidelines and local IRB requirements led the Morphometry BIRN to develop a new, customized software package: the BIRN De-identification and Upload Pipeline, or BIRNDUP. This package creates sharable output images from diverse input images and removes identifying subject information, including facial features. The key elements of BIRNDUP are as follows:

- 1) Data from either retrospective studies or newly acquired scans must be converted into a standardized image file format. Morphometry BIRN has chosen DICOM for its standard file format. The pipeline has been adapted to work with small variations in the standard files produced by the various scanner vendors or variations that arise due to conversion of retrospective data from other file formats to accommodate the image data obtained from the BIRN sites.
- 2) The images are sorted into a local directory structure that matches the organization and naming conventions of the BIRN archive. At this stage the user identifies the correct series for subsequent defacing and removes any bad series such as series with motion artifacts.
- 3) All subject identifiers such as Medical Record Number, Social Security Number etc. are replaced by a pseudo randomly generated unique identifier called the BIRN_ID
- 4) The scans are de-identified, first by removing any subject-identifying DICOM tags as specified by HIPAA (see 5. below). Next, the participating site does the Facial De-identification of the image by registering the scan to an atlas and removing image areas that correspond to facial regions. Since some series have less than optimal image contrast for facial de-identification, a mask generated by one series may be applied to the images in other series.
- 5) A human review, called the Go/NoGo step, is performed on each series to confirm that the facial de-identification has not removed any brain image data that may be required for BIRN analyses.
- 6) The defaced images are uploaded to the BIRN Storage Resource Broker (SRB) with the de-identified DICOM header fields as metadata. Special bulk SRB upload/download utilities are used to optimize upload time.

An evaluation version of the BIRNDUP software currently exists, including the steps outlined above. Planned improvements during the renewal period include the option of using different facial de-identification modules to accommodate a wider range of input scan types e.g. T2-weighted, and an inter-series registration step to accommodate patient motion when applying a mask from one series to another. There has been some discussion of promoting the BIRNDUP software as a solution for external groups or projects facing similar deidentification requirements but the ability to share the code are awaiting policy decisions from the BIRN Steering Committee.



Figure 1. Main interface to the BIRNDUP pipeline. The four panels correspond to the four main steps requiring user intervention in the deidentification and upload process.

1. Requirements for De-identification and Anonymization

A. Background

A specific aim of the Morphometry BIRN is to build a technically and legally sound basis to support data sharing. Protection of patient privacy has been an important goal from the beginning of the project. A great deal of effort and discussion has gone into determining the correct trade off between preserving as much important research data as possible from the original acquisitions while ensuring that at all of the participating institutions the process will meet the standards of the strictest interpretations of patient privacy. The challenges faced in this effort required a new set of software to be custom built by the collaborating groups. The BIRN De-identification and Upload Pipeline (BIRNDUP) is the result of a truly multi-institution effort to solve a unique set of requirements that are a consequence of the nature of the BIRN. In future we expect that other groups will face similar challenges and we hope that the BIRNDUP software can be widely shared to help address these needs.

B. Need for Facial De-identification

Due to recent changes in the guidelines for research on human subjects, the anonymization of participants has become of paramount concern. With the advent of neuroimaging as a popular research technique, subject de-

identification has gone beyond the removal of textual information (e.g., subject name) to the removal of facial features. The data produced by anatomical MR imaging methods can often be reconstructed to produce a 3-dimensional view of the participant's face and skull. Although these images lack relevant identifying features (such as hair, eyebrows, and skin pigmentation), or may be of poor quality, it is still possible to identify a subject, particularly if the subject is familiar (Bruce et al., 1999; Burton et al., 1999). This is particularly important if the subject is participating not as a normal control, but as a representative diagnosed with a particular illness (e.g., HIV, depression).

The purpose of a defacing algorithm is to protect the identity of the participant. Face recognition studies have shown that these internal features (eyes, nose, mouth) are more important than external features (hair, ears, face outline) for recognizing familiar subjects, whereas both feature classes are of equal importance for identifying unfamiliar subjects (Clutterbuck and Johnston, 2002).

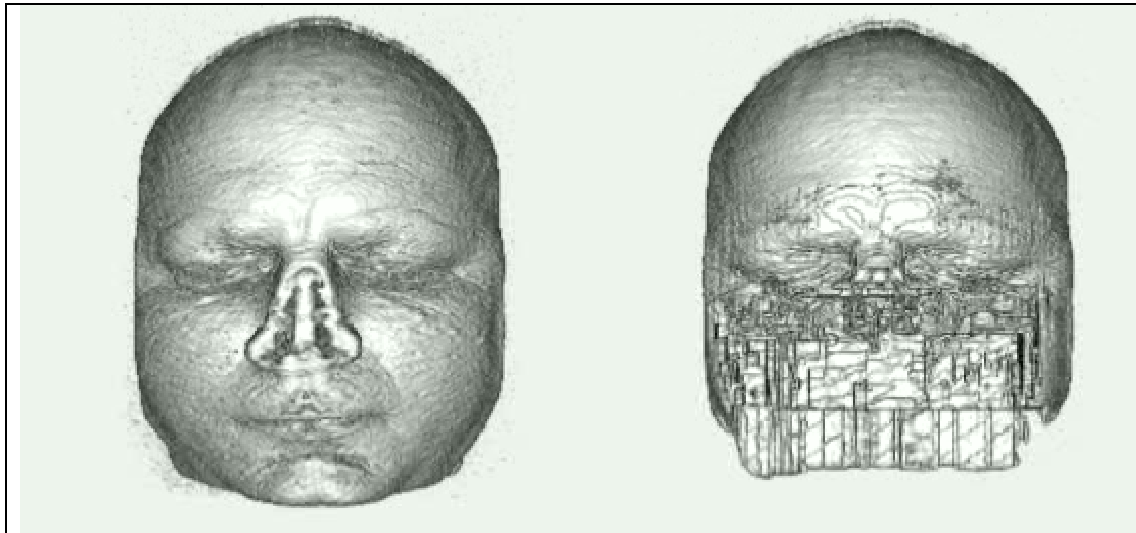


Figure 2. Illustration of volume rendered MR morphometry scan before and after application of the `mri_deface` software (as in this example, the front of the nose is typically cut off from a brain scan as are the ears; the remainder of the face on the left is arguably individually identifiable).

2. Steps for Using the BIRNDUP Tool

BIRNDUP is a software tool running on Linux that provides a single environment for anonymizing or de-identifying MR images and then uploading them to SRB. Its control logic is implemented in Tcl and it makes use of external programs implemented in Java, 'C' and Perl. BIRNDUP is built on 3D Slicer, which is particularly well adapted to obtaining highly refined visualizations of the results of defacing during the Go/NoGo step prior to upload.

A. Image Sorting

The GUI for sorting images is built upon 3D Slicer. A user can browse the directory tree and select the directory where DICOM images are located. The program then retrieves patient's name, patient ID, study ID, series number and flip angle from DICOM headers and presents image file names to the user in hierarchical order of patient, study and series. The user can toggle each series through "deface" for applying "mri_deface" to generate a mask, "mask" for applying "mri_mask" to remove facial features, "header de-id only" for uploading without defacing and "not uploading" for not uploading. The program only allows one series in a study to be marked as "deface".

After the user clicks a button 'OK', a popup screen shows up for the user to input a three-digit visit ID. All of the DICOM files are stored in local directories similar to SRB hierarchies. They are ready for the next step, de-identification.

B. Substitution of Unique Identifier (BIRN_ID) for original subject identifiers (Medical Record Number, Social Security Number etc.)

Associated with each subject's image and clinical/demographic/experimental data is a Unique Identifier (UID) called the BIRN_ID. This consists of a four-digit code corresponding to the institution at which the subject is registered (i.e. the institution from which the original upload is done), followed by a sequence of eight pseudo-randomly generated digits (for a technical description of the algorithm used see the last paragraph in this section).

The BIRN_IDs are created at upload time to replace any personal identifiers such as MRNs or SSNs, compliant with HIPAA guidelines for de-identification and anonymization of subject data (see Table 1). Duplicate eight digit sequences (which may occur very infrequently at a given site) are checked for and rejected, ensuring uniqueness of the BIRN_ID at the institution.

The BIRN ID generator code, which is implemented in Java, enables all of the above functionality. It is fully integrated with the BIRNDUP tool. For totally anonymized subject data, preserving continuity under successive downloads/uploads is not permitted under HIPAA guidelines. However, in certain restricted conditions, HIPAA guidelines do allow for the maintenance of a "link table" which enables preservation of continuity. The BIRN_ID generator provides for both modes of handling the BIRN_IDs as UIDs. Given a subject ID and a four-digit prefix, the BIRN_ID generator automatically creates a corresponding BIRN_ID. If appropriate, a link table can be maintained as well.

The algorithm for generating eight digit pseudo random sequences is called PRNG. The implementation (by the SUN provider) follows the IEEE P1363 standard, Appendix G.7. It compares the SHA-1 hash over a true random seed value concatenated with a 64-bit counter, which is incremented by one for each operation. Only 64 bits are used from the 160-bit SHA-1 output. SHA-1 is the Secure Hash Algorithm, as defined in Secure HASH Standard, NIST FIPS 180-1.

C. DICOM Header De-identification

The DICOM header de-identification procedure operates directly on the DICOM header information without modifying the pixel data content. A set of DICOM tags are identified and reviewed by Morphometry BIRN group members to ensure compliance with HIPAA restrictions and local IRB policy. The resulting translation is implemented within the "dcanon.tcl" program, which is part of the BIRNDUP package. This relies on widely available and accepted DICOM manipulation software provided by Dr. David Clunie (dicom3tools, <http://www.dclunie.com>). Testing has been performed to ensure that the resulting DICOM files are syntactically correct and thus should be compatible with any further processing that relies on DICOM formatted files.

Testing has also been performed to ensure that the process works with DICOM files from the scanners at all the participating Morphometry BIRN sites, since vendor differences in DICOM files are a well-documented phenomenon. Since the current sites include scanners by GE, Philips, Picker, and Siemens, we are optimistic that the tools are robust in the face of variations in file format. However each new scanner type that is added in future, and even software upgrades to existing scanners, must be tested for compatibility.

The "dcanon.tcl" program also manages the invocation of the appropriate defacing technique as selected by the user during the Image Sorting stage, the specific implementation of which is described below. The output image data from the defacing tool is then reassembled with the de-identified DICOM headers to form the output data files.

D. The Defacing Tool: *mri_deface*

To perform the facial de-identification process, a custom module was developed that uses models of non-brain structures for removing facial features, which may potentially allow the identification of a subject/patient from their MR scan.

In this work, manually labeling the facial features of 10 subjects, and using an optimal linear transform created an atlas of face membership. In order to remove facial features from novel images, the optimal linear transform is computed for the input volume as outlined in (Fischl et al., 2002). Next, a brain mask was constructed by summing the prior probabilities at each image location of all brain tissue. This mask was then morphologically dilated n times, to yield a mask that indicates the presence of brain tissue within $n \cdot v$ millimeters of each voxel. Here v is the size of a voxel in millimeters, and the volumes are interpolated to ensure isotropic voxel dimensions. The de-identification procedure then simply amounted to finding all voxels that were outside the mask, but had a non-zero probability of being a facial feature, and setting them to 0. Note that the face atlas was created from T1-weighted images, so it's better suited to operate on the same contrast data.

Statistical Analysis: In a preliminary analysis to quantify the effects of defacing on a volume, 22 of the 278 datasets described in section 2.E below were bias corrected with N3 (Sled, 1998) and skull-stripped using a hybrid watershed algorithm (Dale, 1999). Our previous work suggested that this technique might be the most conservative skull stripping procedure for these pulse sequences and patient populations (see the Skull Stripping section in the appendix to Project 2: Skullstripping progress for renewal, C. Fennema-Notestine, 2004). Within this subject population, the skull stripping tended to be conservative, with any remaining non-brain tissue predominately classified as cerebrospinal fluid. A set difference was calculated using the skull-stripped and defaced volumes to determine the percentage of voxels removed via defacing that were included in the stripped volume: $0.03 \pm 0.07\%$ of the voxels included in the skull-stripped image volume were removed by the defacing algorithm. The variability is most likely due to the hybrid watershed algorithm retaining non-brain tissue that the defacing algorithm removed. These results suggest that the automatic defacing algorithm is robust and efficiently removes non-brain tissue that would have similarly been removed via skull stripping.

To date, a total of 278 retrospective datasets (consisting of middle-aged and elderly controls, Alzheimer's patients, memory impaired subjects, and depressed subjects) have been processed and inspected at the UCSD BIRN site using the BIRNDUP software. Of these, 192 datasets (elderly controls, Alzheimer's, and memory impaired) have been uploaded to the BIRN database. Additionally, we have processed 64 contemporary (current BIRN protocol) datasets (elderly controls, middle aged controls, Alzheimer's, and depressed). Because the BIRN protocol collects two structural FLASH/SPGR scans, one with a flip angle of 30 degrees and another with a flip angle of 5 degrees (referred to as flip 30 and flip 5, respectively), the flip 30 (T1-weighted) was processed with *mri_deface*, and the flip 5 (proton density weighted) was defaced by applying the mask created during the defacing of the flip 30.

The later stages of the pipeline have been adapted to work with small variations in the standard files produced by the various scanner vendors or variations that arise due to conversion of retrospective data from other file formats: for example, UCI's Picker scanner needs to use the 45 degree flip angle rather than the 30 degree for *mri_deface*. Also the Philips scanner at JHU has slightly different DICOM header fields that needed to be handled. Retrospective data converted to DICOM also typically has missing or at least different header fields. At this point the only outstanding issue that needs to be resolved is retrospective Siemens format images from Washington University that need to be made compatible with *mri_deface*.

E. Go/NoGo (Visual Inspection) Module

The Go/NoGo step provides an organized method for visually inspecting the results of the defacing process, and verifying that each series is acceptable for upload to the BIRN SRB. The Go/NoGo script locates all of the series in a given directory and allows the user to view the MPEG files associated with each one. User inspection involves verifying that facial features have been sufficiently removed from the images without removing any

brain. In the step prior to the Go/NoGo step, four MPEG files are created for each series (sagittal view, coronal view, axial view, and a 3D rendering). The Go/NoGo graphical interface allows the user to open/close, play/pause, or step through each of the MPEG files. If closer inspection of the series is required, the user can also click a button that will launch 3D Slicer and load the series for visualization. Upon completing the review of a series, the user chooses to either approve or defer the series. After all of the series have been reviewed, the Go/NoGo script creates two text files, one for the approved series and the other for the deferred series. Each file contains the directory path to its series. The approved series file is then edited by the user and serves as the input to the next step in the de-identification pipeline where the series are uploaded to the SRB.

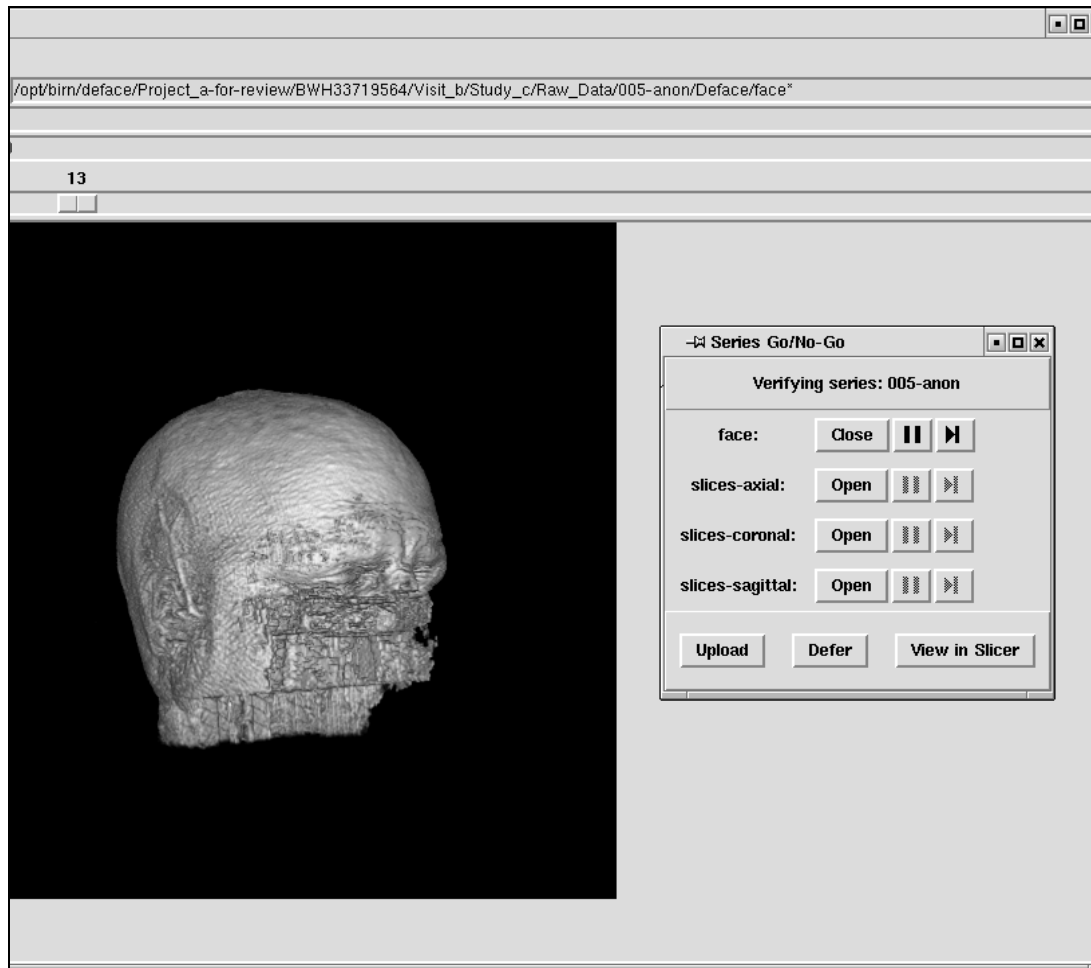


Figure 3. Review screen for accepting or rejection the defacing of the current series. The image rotates as an MPEG loop while being reviewed by the user.

F. Upload Script

The upload script uploads structural MR images and their corresponding meta-data into SRB in a batch fashion. The upload script, written in Perl, comes with utility libraries, which can be used for DICOM image manipulation, more convenient SRB access, etc. The upload script rewrites the DICOM headers of the MR images to add or update DICOM fields so that each DICOM file has the mandatory header fields as agreed by the Morphometric BIRN SRB/Data Integration group. The script creates and applies BIRN specific UUIDs such as SeriesInstanceUID to the DICOM header of each image file. Based on user selection the script can do in-place DICOM header anonymization if the header anonymization has not applied to the images beforehand. Also the script can rename the DICOM files before image upload. After the DICOM headers are fixed, the

header validation step is invoked. After the DICOM series headers are validated, the script creates a container per subject series, creates the collection hierarchy in the SRB, and uploads the image data. The next step is upload of the metadata extracted or computed from the DICOM header in combination with the user-supplied metadata to the corresponding collection levels in the uploaded subject series.

Since some information such as clinical assessments for the subjects are not available from the DICOM headers, the user needs to supply a metadata input file, which has 15 comma-separated entries per image series. The upload script is configured through a single configuration file where all the site-specific parameters and user options are specified. The source code, dependencies and detailed documentation on metadata input file and configuration file can be obtained from the BIRN CVS server by running: “cvs checkout upload2”.

3. The John Doe Algorithm

A. General description

An alternative facial de-identification method is also under development at UCLA (LONI), (Shattuck et al., 2003) to address possible problems arising from replacing facial features with black pixels rather than the grayscale data that is normally present. Some post-processing programs rely on overall properties of the image in order to function properly. The most important current use of this method is to preserve the critical brain regions that shouldn't be changed, while modifying the facial feature, which must be changed to protect privacy. The John Doe algorithm warps the existing grayscale information rather than removing it. Thus, instead of removing the skull and scalp, regions in the image data are distorted to alter the face of the subject. By altering only the outer layers of the scalp, the method does not change the neuroanatomy in the volume.

More generally, additional constraints can be placed on the process to preserve specific regions of the image. By providing John Doe with a mask identifying the brain tissue, it can be ensured that the brain voxels will be unaffected by the method. A second constraint allows restriction of the deformations to the face and chin, which leaves the upper region of the scalp unchanged, thereby reducing the effect of the method on EEG source localization results

B. The algorithm:

John Doe processes an MRI volume using a sequence of threshold, mathematical morphology, edge detection, and spatial filtering. Firstly, John Doe computes a scalp mask using a threshold followed by mathematical morphological processing. The scalp mask is distorted to alter facial features. John Doe then computes a deformation that maps voxels in the outer layers of the new mask to positions in the original scalp mask. These values are then mapped to the new MRI volume, producing an anonymization effect. John Doe is typically applied using a brain mask computed using BSE (Shattuck et al., 2001) This mask is dilated and used to define regions that are guaranteed to remain unchanged in the anonymized data. John Doe optionally registers the brain to an atlas; from this registration, John Doe computes a mask that protects the upper half of the scalp. By doing so, effects on EEG source localization are reduced.

C. Validation:

Over 250 T1 MRI volumes from the ICBM database were processed using the John Doe procedure. Skull stripping was applied using BSE before and after anonymization. In all cases, the skull stripping masks were exactly identical to the initial masks. Since none of the data inside these regions were changed, the stripped brains were also identical before and after anonymization. In research methods for which skull stripped brains are used, no subsequent results will be affected by the use of the anonymization procedure. In order to estimate the impact of the anonymization procedure on subsequent EEG source reconstruction, forward models were calculated from tissue classifications of the original and anonymized MRI for a single subject. Even after the anonymization procedure, the RAP-music result (Ermer et al., 2001) still provides better source localization

than simpler methods that use spherical head models. Currently human studies are being conducted to quantify John Doe's ability to obscure subject entity.

4. FLIRT – A Tool for Refined Masking

Intra-series registration is listed as one of the goals for the facial de-identification task in the main body of the Project 2 text, but the near term integration is planned for the renewal period. A set of tests has been performed to confirm that registration is in fact needed and a viable strategy has been developed. In particular, UCSD has done testing of a tool called FLIRT, described below, with promising results. FLIRT and/or some of the other registration tools will be chosen during the implementation phase of the renewal period (first year).

The current BIRNDUP implementation relies on the mask from the optimal contrast series being applied to the other series in the study. To handle situations in which the subject has moved between the series, the following registration methodology is being investigated and is under consideration for integration into the BIRNDUP software.

FMRIB's Linear Registration Tool, part of a suite of neuroimaging tools known collectively as FSL (FMRIB Software Library, <http://www.fmrib.ox.ac.uk/fsl/>) is an automated, robust algorithm for the linear (and affine) registration of intra- and inter-modal brain images (Jenkinson and Smith, 2001; Jenkinson et al., 2002). FLIRT relies upon image intensity at each voxel rather than landmarks in order to perform its registration. Therefore, a cost function is used to rate how good the alignment is. This cost function can be as simple as calculating the mean absolute difference between voxel intensities (useful for intramodal registration) to more complicated methods to account for differences in mapping intensities across image modalities. Several transformations are calculated for a given alignment, and the global minimum of the cost function is sought to produce the best possible transformation.

A shell script has been written, "coreg_deface_flip5", which uses FLIRT and several other tools to remove facial features from the flip 5 anatomical image. This script first converts the flip 30, the flip 5, and the mask generated by "mri_deface" to ANALYZE format (the image format required by FLIRT). FLIRT then calculates the transformation matrix (using a mutual information cost function) needed to transform the mask to the flip 5 images' orientation. The mask is binarized, transformed, and then rebinarized, as the transformation process interpolates voxels along the edges. If these partial-volume voxels were rounded to an integer, the mask would "shrink," potentially leaving identifiable tissue (e.g., eyes) behind. The transformed mask is then applied to the flip 5 structural image.

For visual inspection of the static images, AFNI (Analysis of Functional NeuroImages, by Bob Cox) is used. The ANALYZE images are converted to AFNI format. Because AFNI often calculates the origin of an image on-the-fly, the defaced flip 5 images may not have the same origin as the unedited flip 5 images, despite the fact that the voxel size and extent are retained. Therefore, after conversion of the defaced flip 5 image (and the flip 5 mask) to AFNI, the origin is set to be the same as the unedited flip 5. This ensures consistency when using AFNI's graphic viewer tools for visualization purposes.

To date 65 subjects (37 elderly controls, 4 Alzheimer's patients, 8 young normal controls, 16 depressed) have undergone coregistration using this tool. Upon completion, these images underwent visual inspection to ensure 1) that no brain tissue was removed, and 2) that facial features had been adequately removed. Thus far, only 1 (an Alzheimer's subject) exhibited brain tissue loss on the flip 5 image. This is remedied by shrinking the mask before applying it to the flip 5 image. All datasets had an acceptable amount of facial features removed. All determinations were made by visual inspection.

Table 1 Required fields for de-identification under the Safe Harbor guidelines of HIPAA:

1. Names
2. Postal address information other than town or city, state and zip code (geographic subdivisions smaller than a state, except for initial three digits of zip code unless this abbreviation contains less than 20,000 people)
3. Dates directly related to an individual (visit dates, birth dates)
4. Phone numbers
5. Fax numbers
6. Email addresses
7. Social Security numbers
8. Medical record numbers
9. Health Plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet protocol (IP) address numbers
16. Biometric identifiers (including finger and voice prints)
17. Identifiable photographic images
18. Other unique identifiers

References

- Bruce V, Henderson Z, Greenwood K, Hancock PJB, Burton AM, Miller P (1999) Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied* 5:339-360.
- Burton AM, Wilson S, Cowan M, Bruce V (1999) Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science* 10:243-248.
- Clutterbuck R, Johnston RA (2002) Exploring levels of face familiarity by using an indirect face-matching measure. *Perception* 31:985-994.
- Jenkinson M, Smith S (2001) A global optimisation method for robust affine registration of brain images. *Med Image Anal* 5:143-156.
- Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17:825-841.
- Shattuck DW, Rex DE, Darvas F, Leahy RM, Toga AW. JohnDoe: Anonymizing MRI data for the protection of research subject confidentiality. 9th Annual Meeting of the Organization for Human Brain Mapping, New York, New York, *Neuroimage Abs*, 2003.
- Shattuck DW, Sandor-Leahy SR, Schaper KA, Rottenberg DA, Leahy RM. Magnetic resonance image tissue classification using a partial volume model. *Neuroimage* 13(5):856-76, 2001.
- Ermer JJ, Mosher JC, Baillet S, Leahy RM. Rapidly recomputable EEG forward models for realistic head shapes. *Phys Med Biol* 46(4):1265-81, 2001.
- Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*. 2002 Jan 31;33(3):341-55. *Brain Morphometry BIRN* 25