Quantification of Inter-subject Variability in Human Brain: A Validation Framework for Probabilistic Maps

Amir M. Tahmasebi\textsuperscript{a}, Purang Abolmaesumi\textsuperscript{a,b}, Conor Wild\textsuperscript{c}, and Ingrid S. Johnsrude\textsuperscript{c,d}

\textsuperscript{a} School of Computing, Queen’s University, Kingston, ON, CANADA;  
\textsuperscript{b} Dept. of Electrical and Computer Engineering, Queen’s University, Kingston, ON, CANADA;  
\textsuperscript{c} Centre for Neuroscience Studies, Queen’s University, Kingston, ON, CANADA;  
\textsuperscript{d} Dept. of Psychology, Queen’s University, Kingston, ON, CANADA

ABSTRACT

Probabilistic maps are useful in functional neuroimaging research for anatomical labeling and for data analysis. The degree to which a probability map can accurately estimate the location of the structure of interest in a new individual depends on many factors, including the variability in the morphology of the structure of interest over subjects, the registration (normalization procedure and template) applied to align the brains among individuals and the registration used to map a new subject’s dataset to the frame of the probabilistic map. Here, we take Heschl’s gyrus (HG) as our structure of interest, and explore the impact of different registration methods on the accuracy with which a probabilistic map of HG can approximate HG in a new individual. We compare three registration procedures; high-dimensional (HAMMER); template-free B-spline-based groupwise; and segmentation-based (SPM5); to each other and to a previously published (affine) probabilistic map of HG.\textsuperscript{1} We quantitatively evaluate the accuracy of the resulting maps using evidence-based diagnostic measures within a leave-one-out cross-validation structure, to demonstrate that maps created using either HAMMER or SPM5 have relatively high sensitivity, specificity and positive predictive value, compared to a map created using the groupwise algorithm or compared to the published map.

Keywords: Magnetic resonance imaging, cerebral cortex, Heschl’s gyrus, inter-subject variability, image registration, probabilistic map, evidence-based diagnosis, evaluation framework

1. INTRODUCTION

Inter-subject morphological variability in the human brain has been a major challenge in conducting group analyses in cognitive neuroimaging studies. Due to lack of information on architectonic variability in human brain, functional coordinates are generally interpreted relative to a standard coordinate frame such as the Talairach\textsuperscript{2} or the AAL\textsuperscript{3} atlases. However, given inter-subject anatomical variability, localization of activation based on a single subject’s anatomy cannot be generalized over a population. To overcome this limitation, probability maps have been proposed.\textsuperscript{4} Probability maps allow statistical assessment of the location of a particular region in any image that has been transformed to the spatial frame of the map. Moreover, they provide a way to predict the position of a functional activation focus and provide a method for analyzing data in an anatomically informed way (i.e., region-of-interest (ROI)-based analysis).

The validity of any probability map is determined by its diagnostic utility - how well it can be used to estimate the location of a structure of interest in new individuals. This depends on many factors, including variability in the morphology of the structure of interest over subjects, the registration applied to align the brains among individuals constituting the map, including both the normalization procedure and the template (if any), and the registration used to map a new subject’s dataset to the space of the probabilistic map. It is to be expected that different registration techniques (which in practice can range from simple rigid-body transformation with six parameters\textsuperscript{5} to high-dimensional (high-d) deformable registrations with millions of parameters\textsuperscript{6}) would yield maps of different diagnostic utility, even when the same transformation method as was used to create the probability map is used to transform new subject data to the space of the probability map.

Here, we take human auditory cortex as our model structure. Heschl’s gyrus (HG) is found on the superior temporal plane in humans (Figure 1(a)) and is the approximate location of primary auditory cortex.\textsuperscript{7,8} The
morphology of HG is highly variable among individuals in terms of both geometry and topology, and it may appear as single, or with two or multiple folds\(^9\) (Figure 1(b)). Penhune \textit{et al.}\(^1\) introduced the first probabilistic atlas of HG. This probabilistic map was generated using labeled MRI data of 20 subjects, which were transformed to a standard stereotaxic space (MNI space\(^10\)) using affine registration. We use this popular published map (between 19 and 36 citations a year, according to the Thomson Corporation’s Web of Science), as a benchmark against which to evaluate other registration methods for map creation as well as for registration of new data.

We use three very different registration methods, including 1) a high-d deformable single-template pair-wise registration (HAMMER);\(^11\) 2) a template-free B-spline-based groupwise registration;\(^12\) and 3) the segmentation-based normalization of SPM5,\(^13\) to generate probability maps from labeled left and right HG of 20 individuals. We assess the three maps we create and compare them to the published map by evaluating their ability to ‘diagnose’ voxels that are actually Heschl’s gyrus in new subjects, whose image volumes have been transformed using the same registration method. We measure diagnostic utility using three measures from evidence-based medicine; namely sensitivity, specificity, and positive predictive value. This method could also be used to assess the diagnostic utility of maps when a different registration method is employed to bring data into the (approximate) space of the probability maps (as is often the case).

2. METHODS

2.1 Data Acquisition

T1-weighted anatomical images were acquired from 20 volunteer subjects (ages 18 – 26, right-handed) with voxel resolution of 1.0 mm\(^3\). MR imaging was performed on a 3.0 Siemens Trio MRI system located at Queen’s University. All subjects gave informed consent and the procedure was approved by the Queen’s University Health Sciences Research Ethics Board. MR data were stripped to remove skull and scalp using the Brain Extraction Tool (BET) of the FSL software (Oxford Centre for Functional MRI, Oxford University, UK) following the steps proposed by Brett\(^\ast\).

2.2 Heschl’s Gyrus Boundaries

Four raters labeled left and right Heschl’s gyrus volumes according to the criteria proposed by Penhune.\(^1\) MRIcron software\(^\dagger\) was used to display the images as well as to label and save the regions of interest. For cases with two or multiple Heschl’s gyri, only the most anterior one was painted. Final volumes of left and right HG were created by identifying voxels labeled as HG by at least three out of four raters.

\(^\ast\)http://imaging.mrc-cbu.cam.ac.uk/imaging/NormalizeSkullStripped.
\(^\dagger\)MRIcron: http://www.sph.sc.edu/comd/rorden/mricron/.
The inter-rater reliability measure or concordance is the degree of agreement among raters. In this study, we defined concordance as the average of the Dice’s similarity coefficients (Eqs. 1, 2), for all combinations of three out of four raters. In this study, concordance among the four raters was $74 \pm 6\%$ (mean±std).

$$
\begin{align*}
 d_i &= 2 \times \frac{|X \cap Y|}{|X| + |Y|} \\
 D &= \frac{1}{N} \sum_{i=1}^{N} d_i
\end{align*}
$$

2.3 Data Pre-processing

The structural MR volume data were affinely registered to a common reference frame (i.e., Colin27 or CJH27) using the SPM5 toolbox. The 12-parameter affine transformation guarantees the alignment of the volume centers and sizes among all the brains. The resulting transformation parameters were applied to the HG volumes of the corresponding subject.

2.4 Registration

After transforming all structural MR volumes to a standard stereotaxic space (by affinely registering them to the Colin27 template in MNI space), three different registration techniques were employed to warp the data:

A) HAMMER (Hierarchical Attribute Matching Mechanism for Elastic Registration)\textsuperscript{11} is an elastic registration technique, which utilizes an attribute vector for every voxel of the image. The attribute vector expresses the geometric features, which are calculated from the tissue maps to reflect underlying anatomy at different scales. Our application of the HAMMER algorithm proceeded in two steps: First, in order to generate the tissue map, the brain data is segmented into gray matter, white matter and cerebrospinal fluid using FMRIB’s Automated Segmentation Tool (FAST) of the FSL software package. Second, HAMMER registration is applied to warp the brain images to the Colin27 template.

B) The B-spline-based groupwise\textsuperscript{12} registration uses the sum of voxel-wise entropies as a joint alignment criterion. For deformation, B-spline control points are overlaid uniformly over the image sets. The directions of movement of the B-spline control points are calculated based on the gradient of the objective function with respect to the control point displacements. The control points are updated using gradient descents until the objective function converges to a local minimum. While minimizing the objective function, the sum of all the deformations is constrained to be zero in order to force the mean scale of the warped images to be the same as the original images.

C) SPM5’s normalization, referred to as unified segmentation,\textsuperscript{13} integrates image registration, tissue classification, and bias correction within the same generative model. This technique requires tissue probability maps as the priors and the ICBM452 tissue probability maps (SPM5 default) were used for the current evaluation.

Next, the warped HG volumes were thresholded to generate binary masks $M_{S_i}$, assigning value 1 to voxels $(p)$ corresponding to HG (i.e., $I(p) > 0$) and 0 to the background:

$$
\forall p \in S_i, M_{S_i}(p) = \left\{ \begin{array}{ll} 
1 & \text{if } S_i(p) \in \text{HG volume;} \\
0 & \text{else,} 
\end{array} \right.
$$

where $S_i$ refers to subject $i$’s volume data. The binary masks were summed and averaged across subjects to produce 3D probability maps for all three registration techniques. Figure 2 shows the average of the 20 warped images using the three techniques. The three constructed probability maps as well as the published probability map\textsuperscript{1} are shown in Figure 3.

Figure 2. Three cross-section views of the average brain volumes generated using different registration techniques shown at $(x = 0$, $y = 0$, $z = 0)$ planes: individual brain volumes were warped and averaged (left to right: sagittal, coronal, and axial views of the generated average brain; top to bottom: Colin27 template brain used for pairwise registration, HAMMER (HMR)-, SPM5-, and groupwise (GW)-based average brains.)
Figure 3. Cross-sectional images of Heschl’s gyrus probability maps shown at \((x = -41, y = -20, z = +7)\) planes. Each row in the figure depicts a probability map generated using different registration techniques (left to right: sagittal, coronal, and axial views; top to bottom: HAMMER (HMR), SPM5, groupwise, 12-parameter affine, and the probability map published by Penhune et al.). The colormap represents the probability of the selected voxel belonging to the HG region with white and black colors corresponding to the highest and the lowest probabilities, respectively.
2.5 Comparison Framework

To assess the diagnostic utility of the probability maps created using the three registration methods, a leave-one-out-based (LOO) cross-validation scheme was employed: 19 out of 20 HG images were used to generate a probability map and this was compared against the excluded (test) subject. This process was repeated 20 times to cover all possible combinations.

Besides these three probability maps, the published map was used to identify HG regions. This was accomplished by using 12 parameter affine normalization to the Colin27 template on the 20 datasets in our series. The same leave-one-out-based cross-validation scheme was employed as for the other three methods.

We evaluate the quality of the generated probability maps in terms of their prediction power at four different thresholds, approximately 60%, 70%, 80% and 90% (these exact values were not achievable in the new probability maps owing to the relatively small number of samples). Those HG-labeled voxels from the excluded subject that were correctly identified by the probability map are true positive (TP) voxels. HG voxels that were not identified in the excluded subject are false negatives (FN), and voxels that are incorrectly classified as HG voxels by the generated probability map are false positive (FP) voxels. Finally, the non-HG voxels that are correctly identified by the probability map are true negatives (TN). Figure 4 highlights the TP, FP, FN, and TN regions. These measures were combined as follows:

\[ Sn = \frac{TP}{TP + FN}; \quad Sp = \frac{TN}{TN + FP} ; \quad PPV = \frac{TP}{TP + FP} \]  \hspace{1cm} (4)

where \( Sn, Sp, PPV \) refer to Sensitivity, Specificity, and Positive Predictive Value, respectively. Left and right HG volumes were considered separately, since previous work indicates increased inter-subject variability in the right hemisphere. Three-factor repeated-measures MANOVAs (hemisphere: 2 levels; registration method: 4 levels; threshold: 4 levels) were conducted on each of the three diagnostic measures separately, and post-hoc comparisons (Sidak corrected for multiple comparisons to achieve \( \alpha = 0.05 \)) were used to identify significant differences when required. Moreover, receiver operating characteristic (ROC) curves (i.e., sensitivity vs. 1 - specificity) were plotted using mean values (Figures 5(a), 5(b)) for left and right hemispheres, separately.

3. RESULTS AND DISCUSSION

ROC plots reveal that HAMMER and SPM5, but particularly HAMMER, consistently outperform the published map and the groupwise method. Results of three-factor MANOVAs demonstrate consistently decreasing sensitivity and increasing specificity with increasing threshold value, as expected. For sensitivity and PPV, we observed main effects of hemisphere (L larger values than R) and registration method (HAMMER and SPM5 yielding higher scores than the other two methods). The HAMMER method yielded probability maps with
higher sensitivity than did SPM5. The main effects of hemisphere and registration method were not significant for the specificity measure, although these factors interacted significantly, and both interacted with the threshold factor.

The template-free groupwise registration did not perform as well as HAMMER and SPM5. From this, we may conclude that the inclusion of a template (either single subject in the case of HAMMER, or multi-subject in the case of SPM5) significantly improves the alignment of labeled HG volumes. However, the B-spline-based method used in this paper should not be considered as the sole representative of the whole class of groupwise registrations. Perhaps, other groupwise techniques would provide a more accurate result.

Unsurprisingly, the affine-only normalization resulted in the map that yielded the poorest measures of diagnostic utility. The ROC curve for this map in the left hemisphere is close to the 45° line. This probably reflects greater anatomical variability in this hemisphere for Heschl’s gyrus as previously reported.\(^1\)

That HAMMER and SPM5 both yielded probabilistic maps of HG with greater diagnostic utility than the affine map is expected. What is more surprising is how similar they were to each other, given that HAMMER is a finer-grained nonlinear warping algorithm than SPM5’s smoothly nonlinear process, and that HAMMER registration was to a single-subject image, and not to a probabilistic template created from hundreds of brains (the ICBM452 tissue templates). It would be interesting to compare diagnostic measures that result when sample brains are warped using the same registration process as was used to create the probability map (as in this paper) and measures that result when sample brains are warped using a different registration process to bring the sample brains into the (approximate) space of the probability maps. In conventional practice, sample brain deformations and the probabilistic maps to which they are compared are often calculated using different registration algorithms, and rests on the assumption that the template spaces are comparable.

Shattuck \textit{et al.}\(^{16}\) have recently avoided this problem by creating three separate probabilistic atlases, one for each of the three common registration procedures. But as we demonstrate here, the question “how good is the fit?” can be answered quantitatively. Our method is key to exploring the assumption inherent in Shattuck \textit{et al.’s} approach that registration method does matter,\(^{16}\) and may yield evidence that researchers need not process their data using exactly the registration technique used to create the probabilistic maps they plan to utilize in their analysis.

Figure 5. Comparison of ROC curves for (a) left, and (b) right hemispheres, corresponding to probability maps generated using different registration techniques (HAMMER, SPM5, groupwise, and the probability map published by Penhune \textit{et al.}\(^1\)). HMR and GW refer to HAMMER and groupwise registration, respectively.
4. CONCLUSIONS AND FUTURE WORK

In this work, we explored the impact of registration accuracy on the precision and usefulness of a probability map. We observed that maps created using either HAMMER or SPM5 have high sensitivity, specificity and positive predictive value, compared to maps created using the B-spline-based groupwise algorithm or compared to the published maps. The proposed quantitative framework may be adapted to evaluate and quantify the validity of probability maps created using other registration techniques.

An interesting potential extension of this work is to evaluate the accuracy of probability maps in identifying functional regions. For example, to the extent that the probability map of HG accurately captures primary auditory cortex, primary auditory cortical activity should be observed in voxels that fall within this map, and the same diagnostic validity measures can be applied to quantify the ‘goodness of fit’ between functionally activated regions and probability maps.

REFERENCES


