A Validation Framework for Probabilistic Maps using Heschl’s Gyrus as a Model

Amir M. Tahmasebi*a, Purang Abolmaesumi*b, Conor Wild*c, Ingrid S. Johnsrude*c,d

aSchool of Computing, Queen’s University, Kingston, ON, CANADA
bDepartment of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, CANADA
cCentre for Neuroscience Studies, Queen’s University, Kingston, ON, CANADA
dDepartment 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 a structure of interest in a new individual depends on many factors, including variability in the morphology of the structure of interest over subjects, the registration (normalization procedure and template) applied to align the brains among individuals for constructing a probability map, 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 assess and compare the goodness of fit of probability maps generated using five different registration techniques, as well as evaluating the goodness of fit of a previously published probabilistic map of HG generated using affine registration [25]. The five registration techniques are: a high-dimensional pairwise registration (HAMMER); three groupwise registration techniques (BSpline-based, implicit reference-based or IRG, and DARTEL) as well as a segmentation-based registration (unified segmentation of SPM5). The accuracy of the resulting maps in labeling HG was assessed using evidence-based diagnostic measures within a leave-one-out cross-validation framework. Our results demonstrated the outperformance of IRG and DARTEL compared to other registration techniques in terms of sensitivity, specificity and positive predictive value (PPV). All the techniques displayed relatively low sensitivity rates, despite high PPV, indicating that the generated probability maps provide accurate but conservative estimates of the location and extent of HG in

*Corresponding author. Address: Medical Image Analysis Laboratory, Cognitive Neuroscience of Communication and Hearing Laboratory, School of Computing, Queen’s University, Kingston, ON, CANADA. Tel: +1 (613) 533 2797, Fax: +1 (613) 533 6513.
Email address: tahmaseb@cs.queensu.ca (Amir M. Tahmasebi)
1. Introduction

Researchers rely on brain atlases to obtain crucial information about brain anatomical structure and function. Brain atlases are commonly used as aids in data processing, data visualization, and data interpretation. They provide an objective method for localizing activation foci, allowing researchers to make sense of their data and compare it with results from other studies. Paper-based atlases include neuroanatomical illustrations, photographs, and other imaging sources (see [34] for a review), but cannot be registered with imaging data with any precision. With advances in technology, many digital brain atlas frameworks have been proposed. These include the Visible Human Voxel-Man atlas [16], Montreal Neurological Institute’s (MNI’s) single subject atlas (AAL) [35], Harvard Brain Atlas [19] and Cerefy Neuroradiology Atlas (CNA) [24]. These atlases were constructed from very few samples - often a single subject - and cannot adequately capture the variability in morphology and extent of brain structures that exists across even a restricted population such as that of neurologically normal young individuals. To overcome the limitations of single-subject atlases, probabilistic atlases, developed by aggregating data over several individuals, have been proposed such as the ICBM452 [33] and MNI SPAM [11] (see [12, 31, 32] for more details on probabilistic atlases and more examples). Probability maps allow one to estimate the certainty of a region being a particular brain structure in any image (structural or functional) that has been transformed to the spatial frame of the map. They can also be used to analyze 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. The diagnostic power of a probability map depends on many factors including the anatomical variability of the structure; the registration applied to align the brains among individuals, including both the normalization procedure as well as the registration template (if any) applied to construct the atlas; and the registration used to map a new subject’s dataset to the space of the probabilistic map. Different registration techniques, which in practice can range from simple rigid-body transformation with a few parameters [7, 14, 37] to high-dimensional (high-d) deformable registrations with millions of parameters [6, 8, 13], will yield maps of different diagnostic utility even when the registration method used to construct the probability map is identical to the registration
Figure 1: (a) Three-dimensional visualization as well as three cross-sectional views of the transverse temporal gyrus of Heschl (HG); (b) transverse sections showing HG outlined in red in three individuals. These show the morphological (geometric and topological) variation of Heschl’s gyrus among individuals and between the two hemispheres of the same individual.
used to transform a new subject data to the coordinate frame of the probability map.

Here, we compare five very different registration methods: an implicit reference-based group (IRG) registration \cite{15}; an inverse-consistent diffeomorphic groupwise registration (DARTEL) \cite{2}; a template-free BSpline-based groupwise registration \cite{4}; a high-dimensional deformable pairwise registration (HAMMER) \cite{30}; and finally, the unified segmentation normalization \cite{3} of the SPM5 software package (Statistical Parametric Mapping: Wellcome Department of Cognitive Neurology, London, UK). The five selected registration techniques in this paper are considered as representatives of the two classes of registration algorithms: groupwise and pairwise methods. High-dimensional pairwise registration techniques such as HAMMER and the unified segmentation of SPM5 provide template-based alignment of the structure of interest among individuals. The groupwise registration techniques \textit{e.g.}, BSpline, IRG, and DARTEL avoid the anatomical bias introduced by choosing a specific template in typical pairwise registration. The effect of template bias in probability maps becomes even more dramatic when only a small group of subjects are used to make the map since the resulting maps give high probabilities where the anatomy is similar to the template and low probabilities elsewhere. In a groupwise registration, every brain in the study is given equal weight in determining the anatomy of the final map. In this paper, we begin to compare these two classes of registration methods.

We take human auditory cortex as our model structure. The transverse temporal gyrus of Heschl (HG) is found on the superior temporal plane in humans as shown in Figure 1(a) and is the approximate location of histologically and functionally defined primary auditory cortex in humans \cite{27}. The morphology of HG is highly variable among individuals in terms of both geometry and topology (Figure 1(b)). The size and shape of HG can vary: it sometimes appears as a single gyrus, or as two or multiple folds \cite{21}. When more than one transverse temporal gyrus is present, the primary auditory cortex is likely to be found on the anteriormost gyrus \cite{28}. In 1996, Penhune \textit{et al.} \cite{25} introduced the first probabilistic atlas of HG. This probabilistic map was generated using labeled MRI data of 20 subjects, which were transformed to a stereotaxic space using 12-parameter affine registration. We use this published map as a benchmark against which to evaluate the other registration methods.

As in the study by Penhune and colleagues \cite{25}, we used labeled MR images from 20 subjects to construct a probability map for Heschl’s gyrus. We evaluate and compare the quality of the constructed probability maps using a leave-one-out (LOO) cross validation: maps were created using labeled data from 19 datasets, and used to label the Heschl’s gyrus structure in the excluded dataset. The overlap between a probability map and the manually labeled Heschl’s gyrus volume can be measured in terms of hits (H), misses (M), false positives (FP) and true negatives (TN). “Hit” refers to those voxels in the manually labeled volume \textit{e.g.}, HG that are correctly identified as HG by the probability map; “Miss” refers to voxels in the manually labeled volume that are not identified by the probability map; FP refers to voxels that are incorrectly identified as HG,
and TN refers to voxels that are correctly identified as not being HG. These values can be combined to compute three measures of diagnostic utility taken from evidence-based medicine; namely sensitivity (Sn), specificity (Sp), and positive predictive value (PPV). Sensitivity is defined as the proportion of HG voxels (in a test subject) that are correctly identified by the map. Specificity measures the proportion of non-HG voxels that are correctly identified by the map. Finally, positive predictive value indexes the probability that a voxel, flagged as expressing HG according to the probability map, actually expresses HG. For each map, we use the same registration method to warp test subjects as was used to warp the brains constituting the map. However, these diagnostic utility measures could also be used to assess the goodness of fit of maps when a different registration method is employed to bring data into the (approximate) space of the probability maps.

Jaccard [18] and Dice [9] similarity coefficients, associated with sensitivity and specificity, have previously been used to evaluate the performance of segmentation methods [40, 26, 36]. The Dice coefficient (DC, defined in Eq. (1)) is the most common way to measure spatial overlap, and is a special case of the kappa statistic commonly used in reliability analysis, when there is a much larger number of background (negative) voxels than of target (positive) voxels [38]. Recently, Klein et al. [20] used Dice coefficient as a measure of mean overlap for comparing several nonlinear deformation algorithms applied to brain image registration including SPM5 unified segmentation and DARTEL. The DC can be calculated as:

\[ DC = \frac{2 \times H}{2 \times H + M + FP} \]

As can be seen from Eq. (1), the Dice coefficient does not distinguish between M and FP errors: the cost of making these two types of error is assumed to be equal. However, depending on the application, one type of error may be less desirable than the other. For example, when using a probability map as a region of interest for fMRI data analysis, a high FP rate would result in the averaging of signal from functionally different voxels, whereas misses would only increase the standard error of the estimated effect. Chang et al. [5] find the DC measure insufficient for analytical and systematic evaluation of probability maps, and show that Jaccard and Dice measures do not have sufficient power to discriminate small variations in segmented images. Although the DC does not tell the whole story, we consider it potentially useful and so include it here for comparison.

2. Materials and Methods

In this section, we will first describe the 20-volume dataset that we used and how we labeled the Heschl’s region in each volume. We will then describe how all the data were preprocessed, before detailing how the five different registration methods were applied, how probability maps were constructed, and finally how the maps were evaluated for goodness of fit.
2.1. Image Acquisition

T1-weighted anatomical images were acquired from 21 volunteer subjects (ages 18-26, 12 female, all right-handed) with a 3D MP-RAGE sequence (TR = 1769 ms, TE = 2.6 ms, flip angle $\alpha = 9^\circ$, voxel resolution of 1.0 mm$^3$). One brain dataset (female subject) was excluded from the analysis due to history of seizure. Therefore, 20 subjects’ data were included for this study. MR imaging was performed on the 3.0 Tesla Siemens Trio MRI scanner in the Centre for Neuroscience Studies MRI Facility 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) followed by manual editing, when required, to correct for inaccurate segmentation of brain from skull and scalp.

2.2. Heschl’s Gyrus Boundaries

Four students (undergraduate and graduate) at Queen’s University were given training by an expert in labeling HG. Four labelers independently painted left and right Heschl’s gyrus volumes according to the criteria proposed by Penhune et al. [25] and further described by Abdul-Kareem et al. [1]: HG is bounded anteriorly by the first transverse sulcus (TS) and posteriorly by the Heschl sulcus (HS) or sulcus intermedius (SI) when it extends at least half the length of the HG. The posteromedial boundary of Heschl is a line drawn from the medial end of the TS to the medial end of the HS. The anterolateral boundary is defined by the ending of the gyrus or in case of duplication by extending the lines defined by the TS and HS to the lateral border of the temporal plane. Table 1 shows the frequency of different left-right HG combinations based on Heschl duplication. The overall duplication frequency was 20% (4 out of 20) and 65% (13 out of 20) in left and right hemispheres, respectively. MRIcron software [23] was used to display the images as well as to label and save the regions of interest. HG was labeled separately in each hemisphere. Labelers viewed and painted the desired region in three orthogonal planes simultaneously to be able to make a more accurate definition of the borders.

<table>
<thead>
<tr>
<th>frequency of HG (L,R) in %</th>
<th>(1,1)</th>
<th>(1,2)</th>
<th>(2,1)</th>
<th>(2,2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>current study</td>
<td>30</td>
<td>50</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Penhune et al. [25]</td>
<td>70</td>
<td>20</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 1: Frequency of different right-left Heschl combinations. This table is adapted from Penhune et al. [25]. For comparison, numbers from Penhune’s study are also included in the table.

Figure 2 shows the union of HG regions, as independently painted by all four labelers, for brain volumes from four subjects. Labeled HG voxels are shown
with four colors in the map: red, yellow, green, and blue represent voxels that are labeled as HG by four, three, two and one out of four labelers, respectively. The percentage of voxels correctly identified by 2, 3, or 4 labelers out of the total volume of labeled voxels (i.e., the number of voxels labeled by 2, 3 and 4 labelers divided by the total number of voxels in the volume of union), averaged over the 20 subject volumes is given in Table 2. As can be observed from Figure 2, the significant difference between voxels labeled by only one person (in blue) and the other cases (i.e., two, three or four labelers) is at the anterolateral end of Heschl’s gyrus. Volumes of left and right HG in each individual were created by identifying voxels labeled as HG by at least three out of four people; this included approximately 56% of the volume of union as can be seen in Table 2. We were concerned that selecting for our ‘ground truth’ volume only those voxels labeled by all four observers would have been too conservative, since this would have failed to include a majority of voxels in the volume of union. To ensure that the ‘ground truth’ volume reflected all labelers equally, we computed the overlap between each labeler’s painted HG and the ground truth map for each subject, for each hemisphere. Repeated-measures ANOVA of overlap measures (labeler index: four levels; hemisphere: 2 levels) for 20 subject datasets revealed no significant difference, either among labelers or between hemispheres.

As can be seen in Figure 2, the middle part of Heschl’s gyrus was labeled with great consistency. However, the anterolateral and medial boundaries were harder to discern. For example, the anterolateral boundary, defined as the visible ending of HG, was often not a clear edge, but could appear as a gradual

![Figure 2: Axial cross-sections from the union of HG regions painted by all four labelers shown for four example subjects. Coordinates were kept in original subject space to avoid the interpolation and resampling artifacts due to spatial transformation. Four colors are used to distinguish among HG voxels that are painted by different numbers of labelers: red, yellow, green, and blue represent voxels that are labeled by four, three, two and one out of four labelers, respectively. The left hemisphere, represented by (L), is on the left.](image-url)
smoothing of the gyrus into the superior temporal plane. In these cases, there was little consistency among raters in the placement of the anterior boundary. Furthermore, a handful of subjects did not have a clearly defined single vs. double HG, but instead an ambiguous one-and-half HG (e.g., a heart shape); these cases presented a challenge for raters to reliably identify the medial end of Heschl’s sulcus.

2.3. Data Pre-processing

The structural MR images were first affinely registered (12-parameter) to a common reference frame (i.e., Colin27 or CJH27 [17] in MNI space) using SPM5 to align the volume centers and sizes of all the brains. The resulting transformations were applied to the HG volumes of the corresponding subject. Due to interpolation and resampling artifacts, the affinely registered HG volumes did not remain as binary images. To make them binary again, the volumes were thresholded at the value 10/255 (i.e., in these 8 bit images, voxels with values within the range [10, 255] were considered as HG and given the value 255, whereas the rest of the voxels, in the range [0,9] were considered background and given the value 0).

2.4. Inter-subject Registration

Affine-registered structural MR volumes were warped using five different registration techniques. The five registration methods were chosen from a wide range of available techniques including groupwise (A, B, C), and pairwise (D, E) methods:

A) The implicit reference-based group registration (IRG) [15] is a recently developed groupwise registration technique that jointly estimates the transformation...
from each image in the group to a “hidden” reference by optimizing the intensity difference of each pair of deformed images. The algorithm assumes a small deformation linear elastic model and uses the Fourier series to parameterize the deformation field.

B) DARTEL (Diffeomorphic Anatomical Registration Through Exponential Lie Algebra) has been proposed by Ashburner [2] as an alternative method of normalization in the SPM package. DARTEL is an algorithm for diffeomorphic image registration, which utilizes large deformations in an inverse consistent framework. DARTEL’s deformations are parameterized by a time-invariant velocity field. Similar to the unified segmentation method, DARTEL requires tissue classification of the brain images. Intensity averages of the gray and white matter images were generated to serve as an initial template for DARTEL registration. The template is iteratively updated after each step of the registration. DARTEL encodes the spatial transforms using approximately $6 \times 10^6$ parameters per subject (John Ashburner, communication on SPM Forum, March 24, 2009). DARTEL was primarily designed for voxel-based morphometry (VBM) studies, which is well-suited to its diffeomorphic nature.

C) The BSpline-based groupwise [4] registration uses the sum of voxel-wise entropies as a joint alignment criterion. For deformation, BSpline control points are overlaid uniformly over the image sets. The directions of movement of the BSpline 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.

D) HAMMER (Hierarchical Attribute Matching Mechanism for Elastic Registration) [30] 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 FM-RIB’s Automated Segmentation Tool (FAST) of the FSL software package. Second, HAMMER registration is applied to warp the brain images to the Colin27 template. HAMMER uses every voxel’s information in its hierarchical multi-resolution approach, and so the number of parameters for the deformation field is equal to the number of voxels within the volume.

E) SPM5’s normalization [3], referred to as “unified segmentation”, includes a probabilistic framework, which integrates image registration, tissue classification, and intensity bias correction within the same generative model. The number of deformation parameters is on the order of $10^3$: the exact number
depends on the image field of view (FOV) (John Ashburner, communication on SPM forum, March 24, 2009). The unified segmentation technique requires tissue probability maps as the priors: here, the ICBM452 tissue probability maps were used.

The HG binary masks were warped using the deformation parameters generated by each registration technique. Again, due to interpolation and resampling artifacts, the warped HG volumes were thresholded to maintain them as binary masks. These masks were averaged across subjects to produce a 3D probability map in the standard coordinate frame (i.e., MNI space).

2.5. Map Construction and Comparison Framework

We employed a leave-one-out-based (LOO) cross-validation scheme to assess goodness of fit of the maps. Nineteen out of twenty HG images were averaged across subjects to produce a 3D probability map and this was used to estimate the location and extent of right and left HG in the excluded (i.e., test) subject. This process was repeated 20 times, excluding each subject in turn. The published map [25] was also used to evaluate the test cases. One should note that Penhune et al. used a different MRI scanner (1.5T Philips Gyroscan system vs. 3T Siemens Trio MRI scanner used in this study) and different realization of affine transformation (landmark-based affine transformation [10] vs. 12-parameter affine registration of SPM5) for probability map construction. Nevertheless, given that we registered our data to a space very similar, if not identical, to the one used by Penhune et al. (MNI452 vs. MNI305), we anticipate little difference in the end result of the two transformation methods.

We evaluate the quality of the probability maps in terms of their sensitivity (Sn), specificity (Sp) and positive predictive value (PPV) at four different thresholds, approximately 60%, 70%, 80% and 90% (these exact values were not achievable owing to the relatively small number of samples, N = 20). Those HG-labeled voxels from the excluded subject that were correctly identified by the probability map are hits, and those that were not identified in the excluded subject are misses; non-HG voxels in the excluded subject that were incorrectly classified as HG voxels by the generated probability map are FP voxels; and non-HG voxels that are correctly identified as such are TN. Figure 3 schematically depicts the H, M, FP, and TN regions.

Sensitivity, selectivity and positive predictive value are defined as following:

(1) Sensitivity:

\[ Sn = \frac{H}{H + M} \]  

(2) Specificity:

\[ Sp = \frac{TN}{TN + FP} \]  

(3) Positive predictive value:
Figure 3: Schematic depiction of hits, misses, false positive and true negative regions identified by comparing a probability map at a given threshold with the HG-labeled volume of a single subject in the right hemisphere. Note: these are not actual data.

The Dice coefficient can be formulated as a function of Sn and PPV:

$$DC = \frac{2 \times PPV \times Sn}{PPV + Sn}$$

(5)

In the comparison framework, left and right HG volumes were considered separately since previous work indicates increased inter-subject variability in the right hemisphere [21, 25]. Three-factor repeated-measures ANOVA (hemisphere: two levels; registration technique: six levels consisting of five registration techniques described in Section 2.4 as well as the probability map of Penhune et al. [25]; threshold: four levels) were conducted separately on Dice coefficients as well as each of the three diagnostic measures. Post-hoc comparisons (Sidak-corrected for multiple comparisons to achieve $\alpha = 0.05$) were used to identify significant differences when required.

When discriminability is not perfect, there is a tradeoff between hit and false positive rates (between Sn and 1-Sp). The poorer the discriminability, the higher the false positive rate is required to identify a majority of true positive voxels. One may be willing to risk false positive voxels in order to get as many hits as possible. Conversely, one may prefer to miss some voxels to in order to avoid mislabeling others. The tradeoff between true positive and false positive rates can be represented graphically. Such a plot, adapted from Signal
Detection Theory (SDT), is referred to as a Receiver Operating Characteristic (ROC) curve [41]. In an ROC curve, the true positive rate (Sn) is plotted as a function of the false positive rate (1-Sp) for a binary classifier system as its discrimination threshold is varied. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular probability threshold. A perfect classification (complete overlap between manual segmentation and the probability map) would be represented by an ROC curve passing through the upper left corner (100% sensitivity, 100% specificity). Therefore, the closer the ROC plot is to the upper left corner, the better the probability map was at discriminating HG voxels from non-HG voxels. Despite similarity in concept, one should note that Sn vs. (1-Sp) plots shown in this work cannot be referred to as ROC curves. This is due to the fact that some of the assumptions regarding the distribution of the noise and signal are not satisfied for the specific application considered in this study (i.e., HG voxels are considered as signal and non-HG voxels constituting the hemisphere volume are noise); In SDT, all noise trials are drawn from one distribution with regard to the chance of making a false positive error. Therefore, having more or less noise trials does not change the false positive rate. In other words, the “true” probability of getting a false positive is independent of the number of the noise trials, because all noise trials have an equal a-priori chance to lead to a false positive error. However, in this study, at least two distributions could be considered for the noise or non-HG voxels: (1) one distribution consisting of the voxels that are never identified as HG such as voxels in frontal lobe. Such voxels never contribute to false positive error; and (2) a second distribution consisting of the voxels that are located closer to HG, for instance in temporal lobe, and may mistakenly be labeled as HG voxels (leading to false positive error). Therefore, we avoid such confusion between ROC curves from SDT and the Sn vs. (1-Sp) plots given in this study by referring to the Sn vs. (1-Sp) plots as ROC-like curves. ROC-like curves were plotted for different probability threshold values for the left and right hemispheres, for all of the registration methods tried.

3. Results

Figure 4 shows the average of 20 structural images warped using the five different registration techniques as well as the corresponding intensity histogram of the mean image. IRG appears to stand out from the other groupwise registration techniques, and HAMMER from the pairwise techniques, as they seem to provide relatively sharper images.

The five constructed probability maps and the published probability map proposed by Penhune et al. [25] are shown in Figure 5, overlaid on the Colin27 brain (in the MNI coordinate frame). The colour map depicts the probability value of a voxel belonging to HG, and a brighter colour corresponds to a probability value closer to 1.0. As can be observed from Figure 5, IRG, DARTEL, and HAMMER yield probability maps that are more concentrated over HG compared to the other techniques.
Figure 4: Three cross-sections of the mean warped brain volumes generated using five different registration techniques shown at ($x = 0$, $y = 0$, $z = 0$) in the MNI coordinate frame. The first three rows correspond to mean volumes generated using IRG, DARTEL, and BSpline-based groupwise registration methods for warping. The fourth row displays the Colin27 brain, which was used as the template for pairwise registration using HAMMER (row five). The last row corresponds to the average of brain volumes warped using unified segmentation technique of SPM5 and the ICBM452 tissue probability maps as the template. The intensity histogram of the mean volumes are also shown for each method of registration. Intensity values were normalized between 0 and 255.
Figure 5: Cross-sectional images of HG probability maps corresponding to five registration techniques (IRG, DARTEL, BSpline groupwise, HAMMER, and unified segmentation of SPM) as well as the published map [25] overlaid on Colin27 brain at \((x = -41, y = -20, z = +7)\) in the MNI coordinate frame. "actc" colormap was used to show each probability map as it provides a wider range of colors for better distinction. Each row displays a probability map constructed using a different registration technique. The sixth row depicts the previously published map proposed by Penhune et al. [25]. The colormap depicts the probability value of a voxel belonging to HG region (expressed as a proportion of individuals exhibiting HG at that voxel, out of the sample of 20).
Figure 6: Dice coefficient measure for six probability maps, IRG, DARTEL, HAMMER, unified segmentation of SPM, BSpline-based groupwise, and the published map [25] in a) Left hemisphere and b) Right hemisphere. Horizontal axes correspond to different probability maps (left) and probability threshold values (right). The vertical axis is the Dice coefficient value for every threshold value between 0.1 and 0.9, for different probability maps.

Figures 6(a) and 6(b) show the Dice coefficient at different probability threshold values for the six probability maps. The left and right hemispheres are shown separately. As can be observed from the figures, the overlap ratio (i.e., DC) increases with probability threshold value, but begins to decrease again at a threshold of approximately 40%, due to a decreased number of hits and increased misses. The decrement rate appears less for IRG- and DARTEL-based maps compared to the rest; these methods likely provide a better alignment of the individual labeled HG volumes across subjects, which results in higher hit rates and lower miss rates at higher thresholds. One should note that the point corresponding to the maximum DC value corresponds to the threshold at which the overall error (i.e., M + FP) is minimized while the hit rate is maximum. This threshold is given in the fourth column of Table 4 for different maps. For most of the maps a threshold around 40% yields minimum error and maximum hit rate.

Mean of DC measures for all probability maps (collapsed across threshold and hemisphere) are plotted in Figure 7(a). Results of a three-factor ANOVA on the Dice coefficient demonstrate a significant main effect of the registration method (statistics for significant effects are given in Table 4). Pairwise comparisons revealed the following order of performance: IRG & DARTEL > HAMMER > SPM5 > BSpline GW > published map.

The effect of hemisphere was also significant, such that the overlap ratio was greater in the left hemisphere. There was also a significant interaction between the hemisphere and the registration method: pairwise comparisons revealed that the effect of hemisphere was only significant for the published and BSpline-based groupwise maps (i.e., the poorest performing maps). As expected, there was a significant main effect of threshold cutoff: we observed a consistent decrease in DC with increasing threshold value above approximately 40% for all maps. The registration and threshold factors also interacted significantly, such that the
decrement rate above the 40% threshold was higher for the poorer performing maps. This result confirms our earlier observation (from Figures 6(a) and 6(b)) that most maps appeared to fall off in performance above a certain threshold, whereas IRG and DARTEL did not. There was also a significant three-way interaction among the three factors: the decline in DC measure was steeper for poorly performing maps with increasing the threshold.

Results of three-factor ANOVA on Sn, Sp, and PPV measures resembled those of the DC analysis (statistics for significant effects are given in Table 4). Figures 7(b), 7(c), and 7(d) show the mean of Sn, Sp, and PPV measures for all six probability maps, as well as the results of pairwise comparisons. We observed a significant main effect of registration method for all three measures: For Sn, IRG > DARTEL > HAMMER > SPM5 > published; For Sp, the main effect of registration method was only significant between IRG and DARTEL; For PPV measures, the only difference was that the SPM5 maps had greater PPV than the published maps. The main effect of hemisphere was only significant for PPV (Left > Right). There was a significant interaction between hemisphere and registration method for all three measures. Post-hoc tests revealed a significant difference between hemispheres (Left > Right) for the published map in all cases and also for the BSpline-based groupwise only for Sn and Sp measures (Left > Right). There was a main effect of threshold for all three measures, with consistently decreasing sensitivity and increasing PPV and specificity with increasing threshold value for all maps, as expected. There was also a significant interaction between the threshold and the registration method: similar to the DC results, the Sn, Sp and PPV changed more quickly as a function of increasing threshold (from the 40% threshold up) for the less accurate maps than for the more accurate ones. There was a significant three-way interaction among registration method, threshold, and hemisphere: the decrease in performance with increasing threshold for poorly performing maps was worse in the right hemisphere, consistent with the DC results.

ROC-like curves corresponding to the left and the right hemispheres are shown in Figures 8(a) and 8(b), respectively. These plots suggest that DARTEL and IRG consistently outperformed the other methods, with the BSpline groupwise and published (affine) maps exhibiting the poorest accuracy.

Users need to threshold a probability map at a particular value in order to use it. This threshold can be very low (e.g., 1-5%, at which even voxels labeled as the structure of interest on only one constituent brain are considered) but are usually somewhat higher - typically 20% or more. Our analyses may help to inform about what threshold value may be appropriate for a particular application, depending on whether misses or false positives are equally undesirable, or whether one type is more undesirable than the other. For example, at low threshold values, the number of misses is lower than the number of false positives, whereas at high threshold values the number of misses will be higher than the number of false positives. We can identify the threshold at which the number of these two error types is the same. This ‘neutral point’ represents a threshold that is neither conservative nor liberal. In order to find the neutral point for all the maps, plots depicting Sn and PPV as a function of threshold
Figure 7: Mean and standard deviations for (a) Dice, (b) sensitivity, (c) specificity, and (d) positive predictive value measures of six different probability maps resulting from three-way ANOVA analysis, collapsed across threshold and hemisphere factors. (*) indicates a significant difference between the two corresponding methods.
Table 4: Statistics from three-way ANOVA analyses and post-hoc comparisons (Sidak-corrected for multiple comparisons to achieve $\alpha = 0.05$) for Dice coefficients and measures of sensitivity (Sn), specificity (Sp), and positive predictive value (PPV). Th. and Hem. stand for threshold and hemisphere, respectively. Only significant results are shown in the table. *$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$. 

<table>
<thead>
<tr>
<th>Main Effect</th>
<th>Dice</th>
<th>Sn</th>
<th>Sp</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registration</td>
<td>$F(5, 15) = 372.52^{***}$</td>
<td>$F(5, 15) = 237.11^{***}$</td>
<td>$F(5, 15) = 26.92^{***}$</td>
<td>$F(5, 15) = 24.33^{***}$</td>
</tr>
<tr>
<td>IRG vs. DARTEL</td>
<td>$t(19) = 4.40$</td>
<td>$t(19) = 7.00$</td>
<td>$t(19) = 5.67$</td>
<td>$-$</td>
</tr>
<tr>
<td>DARTEL vs. HMR</td>
<td>$t(19) = 4.38$</td>
<td>$t(19) = 6.30$</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>HMR vs. SPM5</td>
<td>$t(19) = 28.44$</td>
<td>$t(19) = 22.12$</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>SPM5 vs. published</td>
<td>$t(19) = 4.90$</td>
<td>$t(19) = 7.00$</td>
<td>$t(19) = 5.67$</td>
<td>$-$</td>
</tr>
<tr>
<td>Hemisphere</td>
<td>$F(1, 19) = 4.73^{*}$</td>
<td>$-$</td>
<td>$-$</td>
<td>$F(1, 19) = 17.83^{***}$</td>
</tr>
<tr>
<td>Left vs. Right</td>
<td>$t(19) = 2.25$</td>
<td>$-$</td>
<td>$-$</td>
<td>$t(19) = 4.25$</td>
</tr>
<tr>
<td>Threshold</td>
<td>$F(3, 17) = 176.39^{***}$</td>
<td>$F(3, 17) = 459.58^{***}$</td>
<td>$F(3, 17) = 136.17^{***}$</td>
<td>$F(3, 17) = 79.29^{***}$</td>
</tr>
<tr>
<td>Method x Th.</td>
<td>$F(5, 15) = 68.69^{***}$</td>
<td>$F(5, 15) = 162.51^{***}$</td>
<td>$F(5, 15) = 173.81^{***}$</td>
<td>$F(5, 15) = 2.24^{*}$</td>
</tr>
<tr>
<td>Method x Hem.</td>
<td>$F(5, 15) = 6.73^{**}$</td>
<td>$F(5, 15) = 4.17^{*}$</td>
<td>$F(5, 15) = 27.26^{***}$</td>
<td>$F(5, 15) = 15.05^{***}$</td>
</tr>
<tr>
<td>BSpline GW (L vs. R)</td>
<td>$t(19) = 2.35$</td>
<td>$t(19) = 4.36$</td>
<td>$t(19) = 3.33$</td>
<td>$-$</td>
</tr>
<tr>
<td>Published (L vs. R)</td>
<td>$t(19) = 7.21$</td>
<td>$t(19) = 6.33$</td>
<td>$t(19) = 10.33$</td>
<td>$t(19) = 7.94$</td>
</tr>
<tr>
<td>Method x Hem. x Th.</td>
<td>$F(5, 15) = 19.24^{**}$</td>
<td>$F(5, 15) = 39.77^{***}$</td>
<td>$F(5, 15) = 53.73^{***}$</td>
<td>$F(5, 15) = 5.85^{*}$</td>
</tr>
</tbody>
</table>
Figure 8: ROC-like curves (1 - specificity vs. sensitivity) across the probability threshold values (in 0.05 increments) for (a) left, and (b) right hemispheres. As can be observed from the figures, DARTEL and IRG consistently outperformed the other methods in both hemispheres.
were examined: see Figures 9(a) and 9(b). The threshold value corresponding to such a neutral point is given in Table 4 for all the maps. As can be observed from Figures 9(a) and 9(b) and Table 4, the optimal threshold value that would balance the false positive and miss errors varies among maps and hemispheres, ranging from 20% (i.e., published map) to 45% (i.e., IRG and DARTEL). Exact threshold values are not achievable due to the relatively small number of samples (i.e., N = 20).

It might also be useful to consider the thresholds at which sensitivity (i.e., hit rate) and PPV are 50%. At high thresholds where sensitivity drops below 50%, the map is generating more misses than hits. At low thresholds where PPV drops below 50%, the map is generating more false positives than hits. Both of these are undesirable. For each map, we can define a range of thresholds over which both values are higher than 50%. Figures 9(a) and 9(b) plot sensitivity and PPV as a function of threshold for the left and right hemispheres, respectively. The threshold values at which the sensitivity and PPV are at 50% for each type of map are also given in Table 5, and these values define what could be considered “an acceptable quality range” of thresholds for that map. As can be observed from Table 5, this range is from 5% to 85% for the IRG groupwise maps, 10-85% for the DARTEL maps, 10-75% for the HAMMER maps, 15-70% for the SPM5 maps, and almost nonexistent for the BSpline groupwise and published maps.

<table>
<thead>
<tr>
<th>Registration</th>
<th>FP = M</th>
<th>PPV = 50%</th>
<th>Sn = 50%</th>
<th>max(DC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L</td>
<td>R</td>
<td>L</td>
<td>R</td>
</tr>
<tr>
<td>IRG GW</td>
<td>45</td>
<td>45</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>DARTEL</td>
<td>40</td>
<td>45</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>BSpline GW</td>
<td>40</td>
<td>40</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>HMR</td>
<td>40</td>
<td>45</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>SPM5</td>
<td>40</td>
<td>40</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Published</td>
<td>35</td>
<td>20</td>
<td>65</td>
<td>65</td>
</tr>
</tbody>
</table>

Table 5: The first column gives the threshold values in each map corresponding to the neutral point where the miss and false positive rates balance. Second and third columns are the threshold values where the PPV and sensitivity rates reach 50%, respectively. The fourth column gives the threshold value where the Dice coefficient is maximum. (L) and (R) stand for left and right hemisphere.

4. Discussion

Results from the DC, Sn, Sp, and PPV measures, and the ROC-like curves all confirm that a more accurate registration technique yields a probability map with higher diagnostic power. In other words, the accuracy of the registration used in construction of a probability map directly influences the accuracy of the map in estimating the location and extent of the structure of interest. Any conclusion regarding the relative merits of groupwise and pairwise registration techniques cannot be generally made until more registration techniques
Figure 9: Plots of Sensitivity (red) and PPV (black) as a function of threshold for all map types for (a) left and (b) right hemispheres, separately. The intersection of Sn and PPV plots gives the threshold value in each map corresponding to the neutral point where the miss and false positive rates balance.
are examined; however, according to our preliminary results, some groupwise
techniques such as IRG and DARTEL outperform some high-dimensional pair-
wise registration techniques such as HAMMER and unified segmentation. These
groupwise techniques exhibited greater sensitivity than the pairwise methods,
which implies that they generate fewer misses. All the techniques displayed
relatively low sensitivity (hit rates), despite high PPV, implying that the gen-
erated probability maps are generally accurate but conservative in labeling HG
voxels. A conservative estimate is particularly appropriate for applications like
region-of-interest analyses in which signal is extracted from all voxels comprising
a region and these are averaged to estimate areal signal. If false positive errors
are very low, then signal from HG and non-HG voxels would not be averaged
together inappropriately.

According to the Dice measures reported in Figures 6(a) and 6(b), the IRG-
based map yields a better localization of HG compared to the HAMMER-based
map; however, from the Dice measure alone it is not clear whether the IRG map
yields more hits, fewer false positives, fewer misses or a combination of these.
On the other hand, by looking at the miss rate (1-Sn), one can easily conclude
that the HAMMER-based map results in more misses than the IRG-based map.

By separating misses and false positives, we were also able to calculate the
threshold at which equal numbers of these two types of error are committed (i.e.,
the ‘neutral point’). If false positives are worse, then the threshold should be
somewhat higher, if misses are worse, then the threshold should be lower. But
the threshold should probably also never range outside of the acceptable quality
range, defined as the range over which the number of hits is greater than the
number misses, and greater than the number of false positives. Outside of this
range, erroneously identified voxels exceed accurately identified voxels. Based
on the data presented in Table 5, we would suggest that a threshold set between
35% and 45% balances the two types of error, and (with a few exceptions) one
can comfortably vary the threshold from 20% below this neutral point to 30%
above. Another way of defining the optimal threshold, if the type of error does
not matter, would be to maximize hit rate and minimize the error (M + FP).
This measure is very similar to the Dice coefficient, which we observed to be
optimized around 40% - a threshold approximately around this value (between
30-50%) would minimize error rate and also maximize hit rate for most of the
maps.

On average, 3560 voxels out of a total volume of 1,040,000 voxels were labeled
as HG (in each hemisphere) in the 180 datasets (20 brains × 6 registration
algorithms) considered here. Although the precise voxels that were labeled in
each dataset depend on that individual’s morphology and on the registration
method used, there are many thousands of voxels that were never labeled as
HG. In consequence, the numbers of true negatives were artificially high and
consequently, specificity values were very high. We could modify the specificity
values somewhat by considering a different population of voxels: instead of
considering the whole brain, we could have considered only voxels that were
labeled as HG in at least one dataset, using at least one registration algorithm.
Such a volume would have dramatically reduced the overall number of non-HG
voxels, decreasing the number of true negatives and thus reducing the sensitivity values.

Examination of the average structural images (Figure 4) reveals two Heschl’s gyri (i.e., a double fold) in the right hemisphere of the IRG- and DARTEL-based average brain images, but only a single gyrus in the HAMMER-based average image. This could be due to the fact that thirteen (out of twenty) subjects had a double HG in the right hemisphere versus only four subjects with a double on the left. Since the Colin27 brain (which was used as the template image for HAMMER registration) has only a single HG in the right hemisphere, the pairwise registration enforced the specific topology of the template brain (i.e., a single gyrus) on every subject. On the other hand, the groupwise registration techniques use ‘hidden’ templates that are constructed using the group data and are updated iteratively with every stage of the registration, and are an optimal match to the group’s anatomy. Generally speaking, morphological similarities among individuals comprising a group will be better preserved with groupwise registration.

Examination of ROC-like curves (Figures 8(a) and 8(b)) highlights an interesting hemispheric asymmetry in the brain; specifically, the curves for groupwise based maps in the right hemisphere, particularly for the BSpline-based groupwise map, are closer to the 45° line, indicating poorer discriminability of HG from non-HG voxels. As mentioned previously, groupwise techniques better reflect the group’s dominant morphology as there is no fixed template to enforce a specific anatomical bias during the registration. Therefore, this result from observing ROC-like plots may reflect greater anatomical variability in the right hemisphere for Heschl’s gyrus (as previously reported [25]), which would result in less overlap among registered HGs from different subjects. Furthermore, ANOVA results for the Dice and PPV measures revealed a significant main effect of hemisphere, with greater values in the left hemisphere. This is also consistent with greater anatomical variability of HG in the right hemisphere [25].

In this paper, we compared diagnostic measures that result when sample brains are warped using the same registration method as was used to create the probability map. In the future, it may be interesting to compare measures that result when sample brains are warped using a registration technique different to that used to generate the probability map. In conventional practice, sample brain deformations and the probabilistic maps to which they are compared are often calculated using different registration algorithms. For instance, as can be observed from Figures 9(a) and 9(b), the published map yields low Sn and PPV values when compared against the affinely registered data. However, Penhune’s map has been successfully utilized in different studies to extract the HG region. This could be due to the fact that nonrigid registration techniques (such as SPM5’s normalization) have been used to transform subject datasets to the space of the probability map and therefore, a better alignment was achieved compared to the affine registration in our study. However, the accuracy of this method rests on the (possibly incorrect) assumption that the template spaces are comparable. Shattuck et al. [29] recently suggested that registration method does matter, and to avoid this problem have proposed three separate probabilis-
tic atlases; one for each of the three most common registration procedures. As we demonstrate here, the question of “how good is the fit?” can be answered quantitatively. The overlap measures and validation framework proposed in this work could be the key to exploring the assumption inherent in Shattuck et al.’s approach, and yield evidence that researchers need not process their data using exactly the registration technique used to create the probabilistic maps.

5. Conclusions

We assess how well six different probability maps of an anatomically defined region of interest, generated using different registration techniques, could estimate the location of this region in a new individual. For this purpose, we selected a variety of registration algorithms from different classes (i.e., pairwise and groupwise) to construct probability maps for Heschl’s gyrus, the anatomical landmark for primary auditory cortex in the human brain. Furthermore, we demonstrated how three measures taken from evidence-based medicine - namely sensitivity, specificity, and positive predictive value - can offer greater insight into the diagnostic value of a probabilistic map than typical overlap measures such as the Dice coefficient.

In general, the more accurate registration techniques yield probability maps of higher diagnostic power. Among the selected registration methods, DARTEL and IRG groupwise techniques outperformed the pairwise registration methods such as HAMMER and SPM5, resulting in probability maps that could estimate the location and extent of the structure of interest most accurately. However, generalization of our conclusions to the whole class of groupwise versus pairwise registration methods requires further evaluation that includes more registration techniques from both classes. We also assess the probability threshold that would minimize overall error while maximizing the number of hits (Dice measure), as well as defining an “acceptable quality range” of thresholds between the thresholds at which the maps generate more errors than hits and true negatives. These may serve as useful guidelines when users of probability maps are considering the threshold value(s) at which to apply them. The proposed quantitative framework may be adapted to assess the goodness of fit of probability maps created for other structures of interest and using other registration techniques, as long as some ‘ground truth’ labeled volume of the structure of interest can be developed, as is the case, for example, with probabilistic maps created from cytoarchitectonically parcellated volumes [22, 39].

6. Acknowledgements

The authors would like to thank Drs. J. Ashburner, G.E. Christensen, D. Shen, X. Geng, and X. Wu for their help in utilizing some of the registration techniques. This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through Discovery grants to ISJ and PA.
and a studentship award to AMT, by the Ontario government through a scholarship (OGS) to AMT and an Early Researcher Award to ISJ. ISJ is supported by the Canada Research Chairs Program.

References


