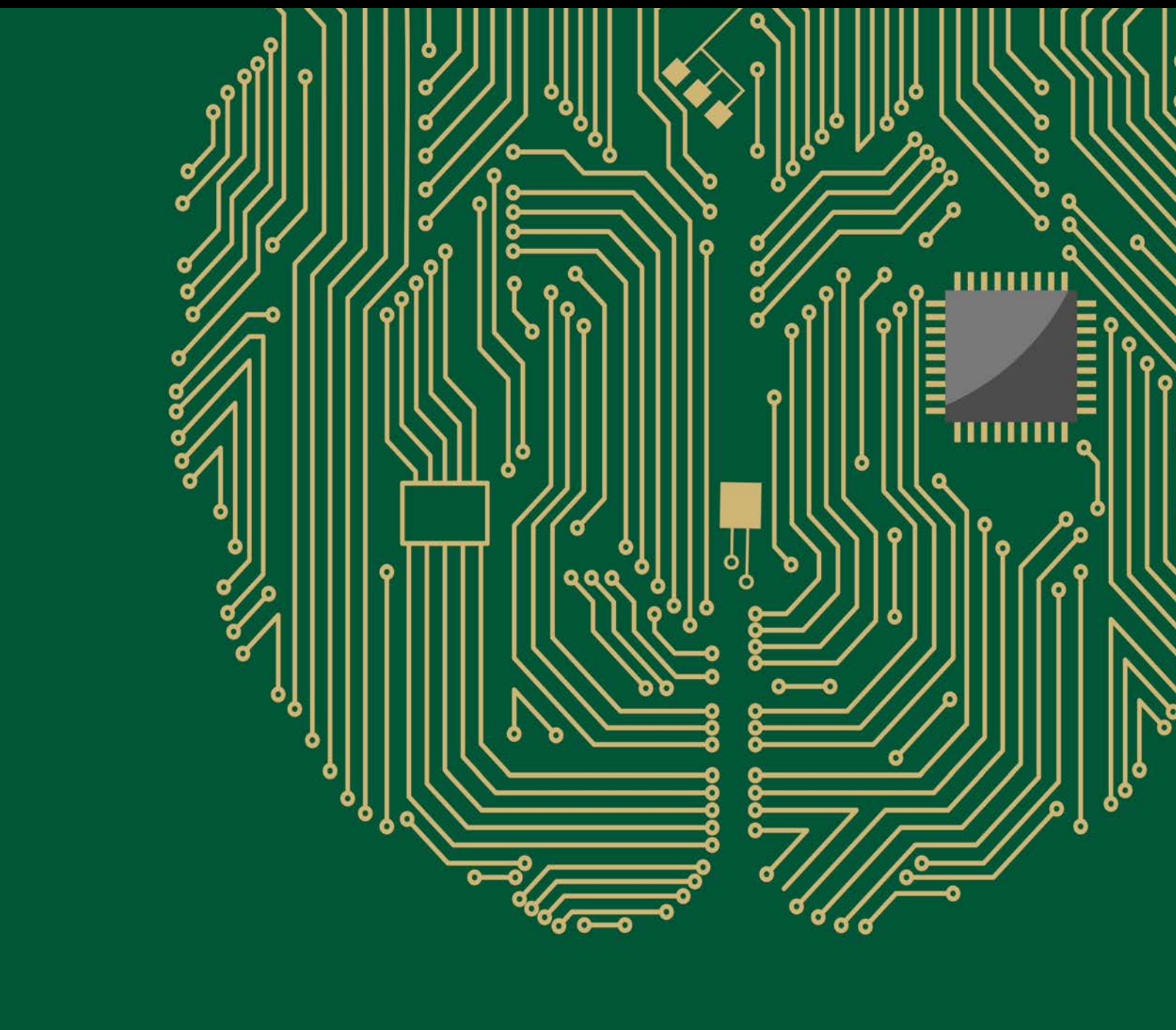
Computational Intelligence and Neuroscience

Simulation and Validation in Brain Image Analysis

Guest Editors: Jussi Tohka, Pierre Bellec, Christophe Grova, and Anthonin Reilhac



Simulation and Validation in Brain Image Analysis

Computational Intelligence and Neuroscience

Simulation and Validation in Brain Image Analysis

Guest Editors: Jussi Tohka, Pierre Bellec, Christophe Grova, and Anthonin Reilhac

Editorial Board

Contents

Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2016, Article ID 1041058, 2 pages
http://dx.doi.org/10.1155/2016/1041058

Editorial

Simulation and Validation in Brain Image Analysis

Jussi Tohka,[1,2] Pierre Bellec,[3,4] Christophe Grova,[5,6] and Anthonin Reilhac[7]

[1]*Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganes, Spain*
[2]*Instituto de Investigacion Sanitaria Gregorio Marãnon, Calle de Doctor Esquerdo 46, 28007 Madrid, Spain*
[3]*Unité de Neuroimagerie Fonctionnelle, Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal, Université de Montréal, Montreal, QC, Canada H3W 1W5*
[4]*Department of Computer Science and Operations Research, University of Montreal, Montreal, QC, Canada H3C 3J7*
[5]*Physics Department and PERFORM Centre, Concordia University, 7141 Sherbrooke Street West, Montreal, QC, Canada H4B 1R6*
[6]*Multimodal Functional Imaging Lab, Biomedical Engineering Department, McGill University, 3775 University Street, Montreal, QC, Canada*
[7]*CERMEP-Imagerie du Vivant, PET-MR Department, Boulevard Pinel, 69677 Bron, France*

Correspondence should be addressed to Jussi Tohka; jtohka@ing.uc3m.es

Received 6 June 2016; Accepted 6 June 2016

The development of novel data analysis methods in brain imaging has been extremely fast paced in recent years. Advanced analytical tools are seen today as instrumental to further our understanding of the brain in health and disease. Translation to practitioners has also been accelerating, with the release of free and open-source software implementations of new tools starting to become the norm. Novel methods in brain imaging can often be used to ask new and exciting questions, and we are concerned that this may have relegated their validation to a position of secondary importance. Consequently, we believe that brain imaging as a field needs to improve its standards for method validation and that validation in itself is a research area where new methodologies are very much needed. Better validation steps could dramatically improve brain imaging methods in the future, by exposing the strengths and limitations of competing methods under a variety of experimental conditions and image acquisition protocols. These progresses will lead to more efficient and reproducible science and help equip brain scientists with the arsenal of high-quality tools they need for big data analytics.

The major challenge faced by researchers when validating brain imaging methods is the lack of a ground truth measure against which the outcome of their methods are compared. Unlike fields such as machine learning in natural images, there are only few public benchmark brain imaging datasets that get widely included in validation studies within a specific subfield. This lack of benchmarks makes it very difficult to compare validation results published on different image analysis methods. Developing reference benchmarks is very challenging because the physiological (e.g., neuronal and metabolic) and physical (e.g., magnetic and electrical) processes underlying brain imaging are extremely complex. Therefore, incorporating experimental imaging data into method validation is highly challenging, perhaps besides valuable and often expensive reproducibility studies. This challenge can be sometimes overcome in specific applications, for example, in image segmentation, when an automated method is expected to replace the manual work by a human expert. Yet, generating manual segmentation requires countless hours of work. Relying on pure simulations, on the other hand, is also challenging because of the aforementioned complexity of the physiology and data generating process. Simulations relying on a simple linear mixture of ground-truth signals and white noise can only be seen as a little more than a sanity check. Sound validation using simulated images needs to encompass detailed biological models as well as the bias inherent to employed imaging technique(s).

This special issue tries to fill some of the gap that exists between the analysis method development and the method validation. We received 14 papers out of which 5 were selected

for the inclusion to the special issue. The accepted papers described validation techniques and their applications for multiple applications using different imaging techniques.

In "MRBrainS Challenge: Online Evaluation Framework for Brain Image Segmentation in 3T MRI Scans" A. M. Mendrik et al. describe an online platform to evaluate tissue segmentation in structural brain magnetic resonance imaging (MRI). They also report the evaluation results of MRBrainS13 competition for brain tissue segmentation of the aging brain. Manual segmentation is used as ground-truth to evaluate automatic segmentation algorithms.

In "Evaluation of Second-Level Inference in fMRI Analysis," S. P. Roels et al. compare different modeling and inference techniques for general linear model (GLM) analysis of group functional magnetic resonance imaging. Although the GLM has been ubiquitous in fMRI research for decades, distinct variants remain popular to this date, and the validity of their underlying assumptions is not yet well established. How these variants behave in terms of reproducibility of results, instead of sensitivity and specificity, is also incompletely understood. S. P. Roels et al. implemented mixed effect simulations as well as a bootstrap analysis on real data to shed light on these questions.

In "How Many Is Enough? Effect of Sample Size in Inter-Subject Correlation Analysis of fMRI" J. Pajula and J. Tohka evaluate intersubject correlation (ISC) based analysis of fMRI data with a large dataset consisting of 130 subjects. They utilize split-half resampling to disclose the reproducibility of the analysis results with different sample sizes and additionally compare the analysis results using a large 130-subject dataset to the analysis results with smaller sample sizes.

In "MEG Connectivity and Power Detections with Minimum Norm Estimates Require Different Regularization Parameters" A.-S. Hincapie et al. evaluate the tuning of regularization hyperparameter within the minimum norm source reconstruction context when applied to magnetoencephalography (MEG) data. They addressed the question using Monte Carlo simulations of MEG data, where they generated 21,600 configurations of pairs of coupled sources with varying sizes, signal-to-noise ratio (SNR), and coupling strengths, thus providing a detailed evaluation framework within a new application context, the analysis of resting state functional connectivity using MEG data.

In "BrainK for Structural Image Processing: Creating Electrical Models of the Human Head," K. Li et al. present BrainK, which is a set of automated procedures for characterizing the tissues of the human head from magnetic resonance images. The tissue segmentation and cortical surface extraction support the primary goal of modeling the propagation of electrical currents through head tissues. They compare the accuracies of BrainK's tissue segmentation and cortical surface extraction in relation to existing research tools (FreeSurfer, FSL, SPM, and BrainVisa). Their method is presenting and evaluating a very promising realistic head model, using finite element model, for electroencephalography (EEG) forward model. This is a necessary step to allow accurate source reconstruction (inverse problem) from high density EEG data.

Acknowledgments

We thank all the authors for submitting their papers to this special issue as well as the reviewers for providing their expertise and time to evaluate and improve the papers.

Jussi Tohka
Pierre Bellec
Christophe Grova
Anthonin Reilhac

Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2016, Article ID 1349851, 25 pages
http://dx.doi.org/10.1155/2016/1349851

Research Article

BrainK for Structural Image Processing: Creating Electrical Models of the Human Head

Kai Li,[1] Xenophon Papademetris,[1,2] and Don M. Tucker[1]

[1]*Electrical Geodesics, Inc., 500 E. 4th Avenue, Eugene, OR 97401, USA*
[2]*Yale University, New Haven, CT 06520, USA*

Correspondence should be addressed to Don M. Tucker; dtucker@egi.com

Received 29 May 2015; Revised 8 September 2015; Accepted 17 March 2016

Academic Editor: Christophe Grova

BrainK is a set of automated procedures for characterizing the tissues of the human head from MRI, CT, and photogrammetry images. The tissue segmentation and cortical surface extraction support the primary goal of modeling the propagation of electrical currents through head tissues with a finite difference model (FDM) or finite element model (FEM) created from the BrainK geometries. The electrical head model is necessary for accurate source localization of dense array electroencephalographic (dEEG) measures from head surface electrodes. It is also necessary for accurate targeting of cerebral structures with transcranial current injection from those surface electrodes. BrainK must achieve five major tasks: image segmentation, registration of the MRI, CT, and sensor photogrammetry images, cortical surface reconstruction, dipole tessellation of the cortical surface, and Talairach transformation. We describe the approach to each task, and we compare the accuracies for the key tasks of tissue segmentation and cortical surface extraction in relation to existing research tools (FreeSurfer, FSL, SPM, and BrainVisa). BrainK achieves good accuracy with minimal or no user intervention, it deals well with poor quality MR images and tissue abnormalities, and it provides improved computational efficiency over existing research packages.

1. Introduction

With the simple recording of the electroencephalogram (EEG), the brain's electrical activity can be measured with millisecond temporal resolution at the head surface. Dense array EEG (dEEG) systems now allow up to 256 channels to be applied quickly with full coverage of the head, assessing the fields from the basal as well as superior cortical surface [1, 2]. The cortex, with its laminar neural organization and with locally synchronous activity stemming from its columnar organization, is the primary generator of the far fields measured by head surface EEG [3]. Cortical sources can be modeled as point dipoles, and their contribution to surface activity can be reconstructed through electrical source analysis. The ambiguity of the inverse estimation in electrical source analysis can be minimized if the precise locations and orientations of the cortical sources are well specified. To a first approximation, the source dipoles can be assumed to be oriented perpendicular to the cortical surface, consistent with the orientation of the pyramidal neurons and cortical columns.

To model cortical sources with these properties, accurate cortical surface extraction is a key challenge. Furthermore, the volume conduction of the electrical potentials, from the cortex to the head surface, must be specified through characterizing the conductivity of each tissue compartment. The skull is the primary resistive medium in the head, and it must be modeled, preferably with bone density values from CT. Because the electrical boundary effects of the volume conduction are affected by discontinuities in current paths, for example, caused by holes in the skull (optical canals and foramen magnum), spherical shell or boundary element models provide only approximate electrical propagation from cortex to the surface, and more detailed (FDM or FEM) volumetric models are needed. Finally, the position of the electrodes must be specified accurately, for example, with geodesic photogrammetry [4].

Figure 1: BrainK GUI.

By describing the positions of electrodes and cortical surface targets, an accurate electrical head model also supports dense array approaches to transcranial current injection. The electrical head model can be validated with bounded Electrical Impedance Tomography (bEIT), in which current injection and potential recovery are analyzed, within the bounds of the tissue geometry from MRI, to test whether the electrical conductivity of head tissues estimated by the model is accurate in predicting the recovered potentials [5, 6]. With an accurate electrical head model, it is possible to modulate brain activity noninvasively, using dense array transcranial Direct Current Stimulation (tDCS), or transcranial Alternating Current Stimulation (tACS), in which patterns of source and sink electrodes are computationally optimized to target specific cerebral sites [7, 8].

In the present report, we review the methods implemented in BrainK. In addition, we report validation studies for accuracy and efficiency of both tissue segmentation and cortical surface extraction. Klauschen et al. [9] evaluated automatic tools for head tissue segmentation from MR images. These included FSL [10], FreeSurfer [11, 12], and SPM [13]. Although BrainK is designed as an automated tool, visualization and editing capability is provided to allow adaptation to unique image properties, such as the presence of lesions or tissue anomalies.

2. Overview of BrainK

Figure 1 shows the graphical user interface for BrainK, including the multiple steps of image processing. Although maximal accuracy for EEG source localization or transcranial neuromodulation requires the full complement of MRI, CT, and sensor photogrammetry, it is important to optimize the results possible with the imaging data available for a particular subject or patient. Figure 2 shows the architecture of BrainK that incorporates specific workflows designed to adapt to the available data for the person. Although it is possible to assume standard sensor positions for BrainK (when sensor placement is referenced to skull fiducials, as with the Geodesic Sensor Net), a more accurate EEG source localization workflow begins with sensor positions from photogrammetry, such as what is provided by the Geodesic Photogrammetry System (GPS). If no imaging data is available for the person, an Atlas head model, constructed from a database of MRIs and CTs for the appropriate ages [14], is then warped with nonlinear registration to fit the sensor positions (and thus the person's head shape), producing a *conformal Atlas head model*. If only a structural MRI (typically a volumetric T1) is available for the person, the sensor positions are registered to the MRI, a database skull (from CT) is registered to the MRI, and tissue segmentation and cortical surface extraction are conducted to create the *individual head model with Atlas skull*. The skull compartment includes either the original CT Hounsfield units for estimating bone density (and thus conductivity) voxel by voxel or a single skull segment, depending on the FDM computational model that will use BrainK's head model output. If both the CT and MRI are available for the person, the MRI is registered to the CT, which has more accurate dimensions than the MRI, to create the *individual head model with individual CT skull*.

For all workflows when the T1 MRI is available for the person, BrainK segmentation identifies the white matter (WM) and the gray matter (GM). It partitions these into two hemispheres and it differentiates cerebellum from cerebrum. In addition, an entire head mask and the two eyeballs are separated as well in the segmentation component. The eyeballs may be important for the electrical head model because of the large far fields generated by their cornea-retinal potentials (that must be separated from the brain signal in the EEG). For all workflow scenarios, a generic spherical sensor cloud is warped onto the head contour of the subject. For the scenarios of Atlas-to-MRI, MRI-to-CT, and CT-to-MRI, an additional GPS-to-head registration procedure is conducted to register the specific GPS sensor cloud, from one individual Geodesic Sensor Net application, onto the head contour. For the scenarios of Atlas-to-MRI-to-GPS and Atlas-to-GPS, the individual GPS data has already played a role in the skull registration and has been aligned with the head contour. In all scenarios, the skull registration is conducted, such that the resulting head segmentation includes the following tissue types: WM, GM, CSF, bone, flesh, and eyeball, in which the WM and the GM are further partitioned into two cerebral hemispheres and the cerebellum.

3. Methods in BrainK

3.1. Image Segmentation. The MRI segmentation implements a cascade of automatic segmentation procedures in order to identify and separate the head tissues required for electrical head modeling, including the scalp, skull, cerebrospinal fluid (CSF), brain (gray and white matter), and the eyeballs. Finally, cortical surface extraction is performed to allow characterization of the normal of the cortical surface (allowing dipoles to be fit perpendicular to the cortex). The cascade mainly consists of two types of procedures: voxel classification on a region of interest (ROI) and the extraction of certain

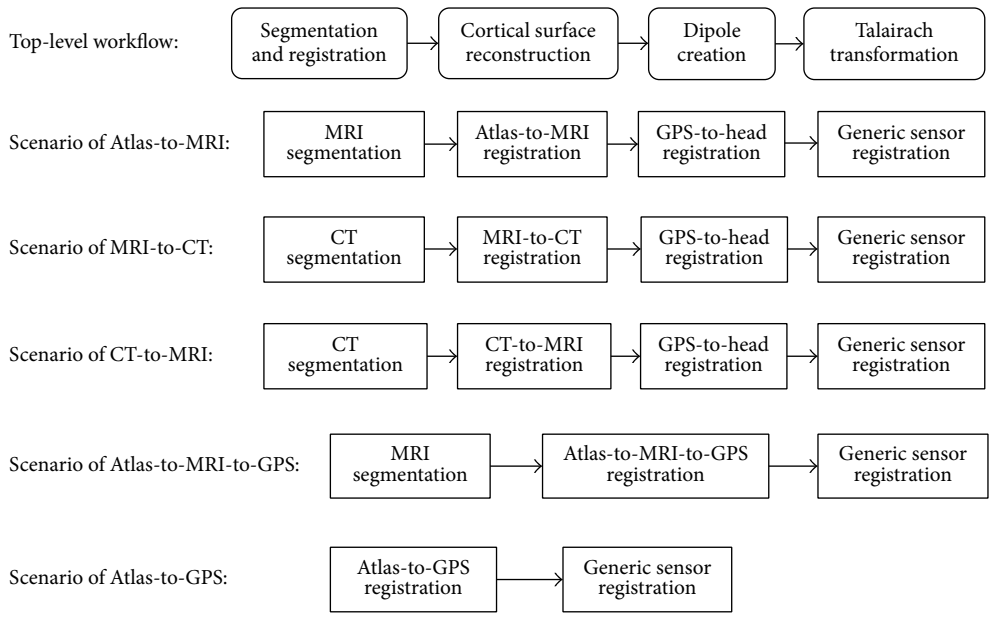


Figure 2: BrainK architecture.

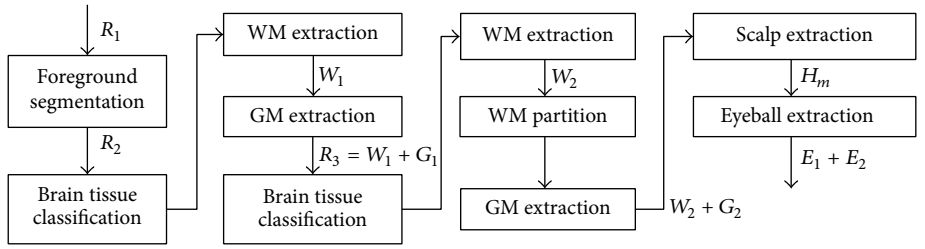


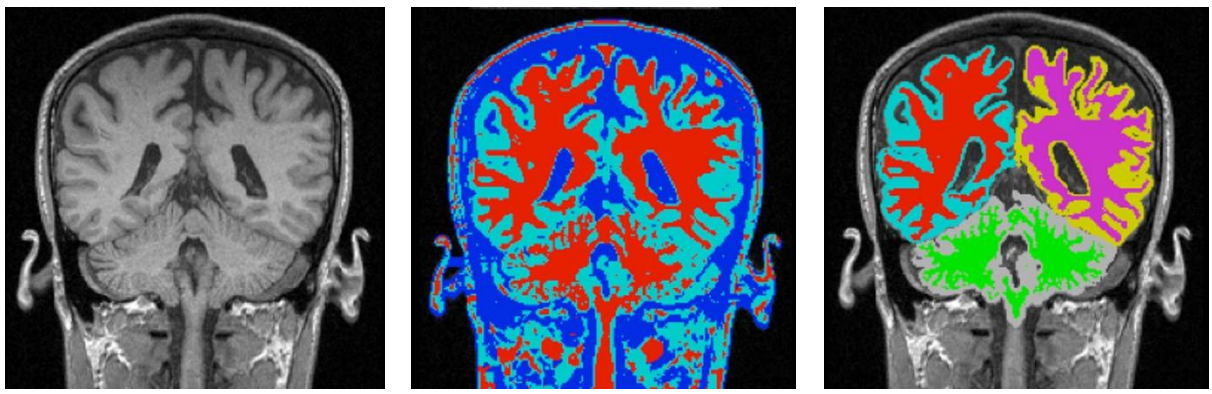
Figure 3: MRI segmentation workflow.

anatomical structures. The classification procedure aims at labeling all voxels in the ROI into different tissue types. The extraction procedure performs morphological operations on the voxel classification so that a certain anatomical structure, such as the WM, the GM, or the scalp, is separated from erroneous (false positive) segmentation results.

Given the MRI data, the segmentation cascade as shown in Figure 3 first takes the entire MR image space as the first ROI, R_1, and classifies all the voxels into two types: foreground and background. The following foreground extraction procedure does simple morphological operations to further improve the foreground segmentation resulting in the second ROI, R_2, in which all voxels are classified into three tissue types: WM, GM, and CSF, as shown in Figure 4(b). The major operation performed in this procedure is BrainK's novel relative thresholding (RT) technique (Section 3.1.1) [15]. Given the voxel classification on R_2, the WM extraction procedure then produces an initial WM volume W_1, which is used as the basis to identify the initial GM volume G_1 by the GM extraction procedure. Next, the union of G_1 and W_1 is taken as the third ROI, R_3, in which the WM/GM classification is refined with a different RT scheme. The new WM volume is then processed by the WM extraction procedure and a WM partition procedure, ending up with the second WM volume W_2.

This WM partition procedure not only separates the two cerebral hemispheres from each other but also makes an optimal cut between the cerebral WM and the cerebellar WM with the well-known maximum flow algorithm [16]. The result is specification of the following tissue types in W_2: two types of cerebral WM for the two hemispheres, respectively, and the cerebellar WM. W_2 then forms the basis for extracting the new GM volume G_2, which includes the two types of cerebral GM, respectively, for the two hemispheres and the cerebellar GM, as shown in Figure 4(c). The brain tissue segmentation (W_2 plus G_2) is then taken as a reference data for the scalp segmentation resulting in the head mask H_m. The scalp segmentation then conducts a binary thresholding procedure on R_1.

To achieve the above segmentation steps, BrainK implements a novel morphological image analysis (SMIA) technique (Section 3.1.3) and a cell complex based morphometric image analysis (CCMIA) method (Section 3.1.4). These play key roles in the WM extraction, topology correction, GM extraction, and scalp extraction. Together with the relative thresholding RT method, the SMIA and CCMIA steps form



(a) (b) (c)

Figure 4: MRI segmentation. (a) MR image. (b) RT result on the foreground. (c) MRI segmentation result.

the core techniques used in BrainK and will be described in the following subsections in detail. The eyeball extraction from the head mask is a special procedure using a priori knowledge about the eyeballs.

Although segmentation of the skull from the MRI is an important challenge, the characterization of bone properties is poor in the typical T1 MRI sequence compared to that provided by CT. The Hounsfield units provided by the CT image (measuring X-ray attenuation) are directly proportional to bone density and thus provide important information on the relative tissue properties of trabecular skull. For example, the CT Hounsfield units correlate $r = 0.83$ with water content in trabecular skull [17].

The CT image segmentation procedure first performs the simple yet robust thresholding [18] so that the entire image is classified into three voxel types: bone, flesh, and background. The original Hounsfield units are then retained as bone density and conductivity estimates. Morphology operations follow to smooth the initial segmentation. Finally, the maximum flow algorithm is applied to separate the brain volume from the rest of the soft tissue so that the cut between them is minimized. CT segmentation results in a voxel classification as well as a brain volume wrapped up by the cranium and eventually there are four tissue types in the segmentation: bone, flesh, brain, and background.

3.1.1. Relative Thresholding (RT). A well-known and prominent image artifact that challenges MR image segmentation is the intensity inhomogeneity (IIH) due to spatial distortions in the radio frequency (RF) gain in the RF coil [19]. The presence of IIH leads to a shading effect over the image and to significant overlap between histograms of different tissues. The result is that intensity based methods, such as thresholding and clustering, are typically unreliable in brain tissue segmentation. IIH correction can be performed prior to image segmentation, but the procedure may eliminate not only the image artifacts but also image signals that provide important information. When the correction is performed together with the segmentation, additional degrees of freedom are introduced into the problem formulation, making the optimization procedure more susceptible to local optima. In this paper, we present a new segmentation method referred to as relative thresholding (RT), which uses two global relative thresholds to compare with the local intensity contrasts, thereby bringing global and local information to segment the ROI into WM, GM, and CSF.

The nature of RT makes it robust against the influence of IIT without introducing an explicit IIH correction. Like traditional thresholding methods, RT enables exhaustive searching without possibilities of being trapped into local optima. It also incorporates various a priori knowledge such that it is robust to intersubject variability and to the convolution of cortical structures. Finally, RT conducts brain tissue segmentation without the need of a prior step of brain volume extraction.

In spite of intersubject variability, complexity of cortical structures, and variability on MR imaging sequences, we can make the following a priori structural, geometrical, and radiological observations: (1) Skull, CSF, GM, and WM form a layered structure from outside to inside; (2) the average intensities of skull, CSF, GM, and WM in local regions are in ascending order in T1-weighted MRI; (3) the cortex thickness is nearly uniform and is a very small value compared to the size of brains; in addition to the structural modeling described above, we also formulate an image model by incorporating a multiplicative low-frequency bias field b_i and an additive noise field ρ_i as follows:

$$y_i = b_i y_i' + \rho_i, \tag{1}$$

where y_i is the observed intensity at the ith voxel and y_i' is the unobservable true intensity without the influences of IIH and noise.

The structural modeling and the image modeling enable two procedures, both of which compare the intensity of a subject voxel x_i with that of a reference voxel x_j on its gradient path, which is a set of ordered voxels emanating from x_i and following the gradient direction at each voxel in the path. The first procedure, GM-WM differentiation, scans all voxels in the ROI, which are initialized as WM, and labels them as GM if there is such a reference voxel x_j that its distance from x_i is less than a distance threshold d_{gw} and the ratio between their

smoothed intensities z_i/z_j is less than a relative threshold t_{gw}; the second procedure, CSF-GM differentiation, scans each GM voxel x_i and relabels it as CSF if there is such a WM voxel x_k on its gradient path whose distance from x_i is less than another distance threshold d_{cw} ($d_{cw} > d_{gw}$) and the ratio between their smoothed intensities z_i/z_k is less than another relative threshold t_{cw} ($t_{cw} < t_{gw}$).

In order to find the optimal relative thresholds, BrainK exhaustively tries each pair among the Cartesian product of the candidate GM/WM relative threshold set and the candidate CSF/WM relative threshold set and chooses the pair that minimizes an objective function $h = w_w h_w + w_g h_g + w_c h_c$. In this equation, h_w is a metric measuring the homogeneity of the WM object and formulated as the sum of smoothed intensity differences of any pair of adjacent voxels labeled as WM. h_g and h_c are formulated in the same spirit for GM and CSF, respectively. w_w, w_g, and w_c are three weighting coefficients.

The second RT scheme takes the optimal relative thresholds and differs from the first scheme on the question of how to determine the reference intensity compared to the intensity of the subject voxel. The ROI R_3 is first reset and then an initial set W^0 of WM voxels are determined by thresholding R_3 with a traditional threshold that maximizes the objective function $f = f_e - f_r$, where f_e is the sum of intensity differences of all m pairs of adjacent voxels labeled as WM and non-WM and f_r is the sum of differences of the m pairs of adjacent voxels labeled as WM and with the greatest intensity differences. A minimal cardinality of W^0 is set to improve the robustness. The remaining region of R_3 is then segmented as a procedure of iteratively dilating W^0 by using the optimal GM/WM relative threshold to compare the subject voxel's intensity to a reference intensity computed by considering the WM voxels near the subject voxel.

3.1.2. GM and WM Extraction. The GM and WM extraction procedures are responsible for extracting the GM and the WM structure from the raw voxel classification in the ROI. We describe GM extraction first, assuming that WM extraction is already done. The GM extraction utilizes the following a priori knowledge: (1) the thickness of cortex is nearly uniform; (2) GM wraps around WM such that two tissues form a layered structure; (3) the average gray level of GM is lower than that of WM in any local region. Given the extracted WM volume W_2, the GM volume is first created by a gradient flow process, in which any voxel originally labeled as GM in the ROI is taken as the true GM if it can reach any WM voxels by following the gradient path emanating from itself within a given distance threshold. Each step in the gradient path is determined by looking at the gradient direction for each voxel. This initial GM volume is then further processed by traditional morphological operations to improve the segmentation, ending up with the GM volume G_2.

WM extraction is a more challenging process than GM extraction because it is responsible for extraction of WM itself, and it is also responsible for the brain volume extraction. The following a priori knowledge forms the basis of WM extraction: the WM is highly connected, but it has no topological defects. Topological defects or false positives present narrow bottlenecks. The extraction mainly consists of two steps: (1) WM localization, in which the center of one of the hemispheres is determined; (2) separation of WM from false positives. The localization of the WM center is essentially finding the WM voxel with the highest connectivity, which means that some quantitative measurement of connectivity should be taken for all WM voxels. Since topological defects, such as a tunnel inside the WM, can greatly influence the geometric measurement of the connectivity, the WM should be first topology-corrected before the connectivity measurement. Because the cortex surface (ribbon) must be defined in relation to the correct topology of the WM, WM topology correction is also essential for valid cortical surface reconstruction. BrainK introduces a skeletonization based morphological image analysis (SMIA) method and a cell complex based morphometric image analysis (CCMIA) framework. The following sections describe how these methods play central roles in WM extraction through achieving accurate connectivity measurement and topology correction.

3.1.3. Skeletonization Based Morphological Image Analysis (SMIA). BrainK's SMIA framework consists of a *surface skeletonization* procedure and a *curve skeletonization* procedure, both of which are based on its extensive topological point classification. Given a binary image, we say its foreground F has a *handle* whenever there is a closed path in F that cannot be deformed through connected deformations in F to a single point. A handle in F is referred to as a *tunnel* in its complement $\overline{F}$. A point in F is simple if it can be added to or removed from F without changing the topology of both F and $\overline{F}$. A *simple point*[20] is a central concept in digital topology and is the basis of our definition of what we call a *thick-simple point*. A point P is thick-simple with respect to F if it is simple with respect to F and the removal of P and any of its neighbors from F does not increase the number of tunnels and number of connected components in F. A point is *thin-simple* if it is simple but not thick-simple and can be classified as thick-surface and thick-curve points. Thick-simple points can also be further classified as surface-edge, curve-end, and other point types.

There are two steps for BrainK's surface skeletonization based on the above point classification: *thick-surface skeletonization* and *thin-surface skeletonization*. The former results in a discrete surface with thickness of at most two voxels, and the latter results in a final thin surface skeleton of one-voxel thickness. The thick-surface skeletonization iteratively removes boundary points of the object that are thick-simple at the moment of removal. The thin-surface skeletonization iteratively removes boundary points of the thick-surface skeleton that are thick-surface points, thick-surface-edge points, or thick curve-end points at the moment of removal. The curve skeletonization procedure iteratively removes the boundary points of the object that are thick-curve points or thick-simple, but not curve-end points, at the moment of removal. As a side product, BrainK's surface skeletonization procedure gives each skeleton point a depth metric as the

distance from the point to the boundary, whereas the curve skeleton procedure gives each skeleton point a wideness metric.

If the curve skeletonization procedure is conducted at a certain scale, then the points that become the thin-curve points can be checked if they are on a handle of the object. Removal of such handle points then corrects the topology defects by cutting the handle. BrainK supports a multiscale topology correction [21] method on the WM object so that the WM extraction is made more robust, and then the topology-corrected cortex can be generated based on the WM.

3.1.4. Cell Complex Based Morphometric Image Analysis (CCMIA). CCMIA is motivated by the goal of representing the true connectivity of an object's skeleton considering the convolution of such objects as the cerebral WM, where neither the depth nor the wideness of the structure is good enough for this need. It is essentially a series of transformations on a space called the *cell complex* [22]. A cell complex is a topological space composed of points, segments, polygons, polyhedrons, and the generation to polytope in any dimension. Given a set of voxels X in the 3D binary image, we can construct a 3-dimensional cell complex, that is, *3-complex*, by creating a point for each voxel, an edge for every two connected points, a triangle for every three interconnected edges, and a tetrahedron for every four interconnected triangles. CCMIA proceeds as transforming the 3-complex to a 2-complex with points, segments, and polygons only and then the 2-complex to a 1-complex with points and segments only. The 3-2 transformation iteratively removes those boundary tetrahedrons by dropping one face f while a connectivity metric and a depth metric are accumulated on each f' of the remaining three faces for the removed tetrahedron as $\text{depth}(f') = \text{depth}(f) + d(f, f')$, while d represents the distance between the centers of the two faces. In a similar spirit, the 2-complex can be transformed to 1-complex and the remaining edges can be set with two accumulated metrics as $\text{wideness}(e') = \max(\text{widness}(e'), \text{wideness}(e) + d(e, e'))$ and $\text{connectivity}(e') = \text{connectivity}(e') + \text{connectivity}(e) + d(e, e')$.

After the raw WM object is topologically corrected, its surface skeleton is processed by CCMIA ending up with the 1-complex with points and edges only. The point incident to the edge of the greatest connectivity is identified as the WM center and a wideness threshold is used to break the bottleneck between the true WM and the false positives. Finally, the WM surface skeleton is dilated to restore the volumetric object.

3.2. Registration. Given the varying data available, flexible and accurate registration is an essential BrainK function. Whereas the MR image accurately differentiates soft tissues, the CT image accurately represents bone. When individual CT or even MRI is not available, BrainK provides registration of the sensor photogrammetry data to a digital head tissue Atlas of the appropriate age and gender. BrainK integrates both soft tissues and bones from multimodal images and from existing digital Atlases by various image registration techniques.

Image registration is a problem of finding optimal geometric transformations between images so that each point of one image can be mapped to a corresponding point of another image. There are generally rigid transformations and nonrigid ones. Rigid transformation involves only translation and rotation. Affine transformation is a typical nonrigid transformation and allows scaling and shearing. Another form of nonrigid transformation is an elastic transformation, allowing local deformation based on models from elasticity theory. BrainK implements rigid and affine transformations, as well as a landmark-based elastic transformation using thin-plate spline theory [23]. For different registration purposes, either one or more of them is used so that the simple transformation serves as initial state for the more sophisticated transformation.

Both the rigid transformation and the affine transformation can be represented by a simple 3×4 matrix. The key issue in applying such transformations is to optimize the similarity measure between the source and the target image, which in BrainK is represented in terms of landmarks. For the thin-plate spline transformation, the displacement field at any point is computed by interpolation from the vectors defined by a set of source landmarks and a set of target landmarks. It can be seen that landmark extraction plays a central role in all three transformations. BrainK is optimized for extracting landmarks based on the preceding image segmentation and on a priori anatomical knowledge. All registration procedures in BrainK undergo three steps: (1) landmarks extraction, (2) transformation coefficients determination, and (3) transformation execution. Note that three transformations all use landmarks but in different ways. The landmarks in rigid and affine transformation are used for constructing the similarity measurement, while there are source landmarks and target landmarks in thin-plate spine transformations and they are mapped to build the displacement field.

The thin-plate spline theory is based on an analogy to the approximate shape of thin metal plates deflected by normal forces at discrete points. Given n source landmarks $P_i = (x_i, y_i, z_i)$ in the source image and the corresponding target landmarks $P_i' = (x_i', y_i', z_i')$ in the target image, the thin-plate spline transformation maps any point $P = (x, y, z)$ in the source image to the point $P' = (x', y', z')$ in the target image as follows:

$$\begin{aligned} x' &= a_0 + a_1 x + a_2 y + a_3 z + \sum_{i=1}^{n} d_i r_i^2 \ln r_i^2, \\ y' &= b_0 + b_1 x + b_2 y + b_3 z + \sum_{i=1}^{n} e_i r_i^2 \ln r_i^2, \\ z' &= c_0 + c_1 x + c_2 y + c_3 z + \sum_{i=1}^{n} f_i r_i^2 \ln r_i^2, \end{aligned} \tag{2}$$

where $r_i^2 = (x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2$. The coefficients in the above equations together form a $(n + 4) \times 3$ matrix W, which can be obtained by solving the equation $LW = M$ that forces the matching of P_i and P_i', where M is formed by

organizing the coordinates of $P_i' = (x_i', y_i', z_i')$ and L is formed by composing the coordinates $P_i = (x_i, y_i, z_i)$ and the distance information $r_i^2 \ln r_i^2$.

When the individual CT as well as the MRI data is available for the person, the registration component in the MRI-to-CT scenario first segments the CT into the following tissue types: bone, brain, flesh, and background. The MRI segmentation is then aligned with the CT segmentation so that the brain tissues and the eyeballs recognized in the MRI data are properly transformed and put into the brain region defined by the CT data. The result of this transformation is that the entire head segmentation (WM, GM, CSF, bone, eyeball, and flesh) is now registered with the CT volume, where the CSF is redefined as the "brain" region recognized in the CT segmentation, but now subtracting the WM and GM that were inserted from the MRI segmentation.

When both MRI and CT are available, the user could also choose the CT-to-MRI registration so that the bone recognized in the CT is transformed and put into the MRI segmentation wrapping around the brain and the CSF is inserted between the brain and the cranium. However, the MRI-to-CT registration is preferred when the CT covers the entire head of the subject because MRI imaging introduces more geometry distortion than CT. However, when the CT only covers the cranium region and the MRI data includes the face and the jaw, the user could consider the CT-to-MRI registration in order to have adequate face and jaw data for the electrical head model.

When only the MRI data is available for the given subject, BrainK supports the skull Atlas-to-MRI registration, in which the bone tissue from an Atlas dataset is warped into the MRI space and the CSF is also taken as those regions between the cranium and the brain. Any region within the head mask that is not occupied by the GM, the WM, the CSF, the eyeballs, or the bone is set to *flesh*.

When only the MRI and the individual GPS data are available and if the MRI is badly warped (dimensionally distorted), the user could choose the MRI-to-GPS registration, so that the head region is warped to match the GPS sensor cloud, which has good metric properties of the actual head shape. The result is transformed segmentation data, which can be further processed by the Atlas-to-MRI registration described above.

Thus, the skull registration can be adapted to fit the available image data for the individual. When only the individual GPS data is available for a given subject, an Atlas dataset with all required tissue types can be warped to match the GPS sensor cloud with the so-called Atlas-to-GPS registration, producing an individual conformal Atlas that has the appropriate shape of the volume conductor for electrical head modeling.

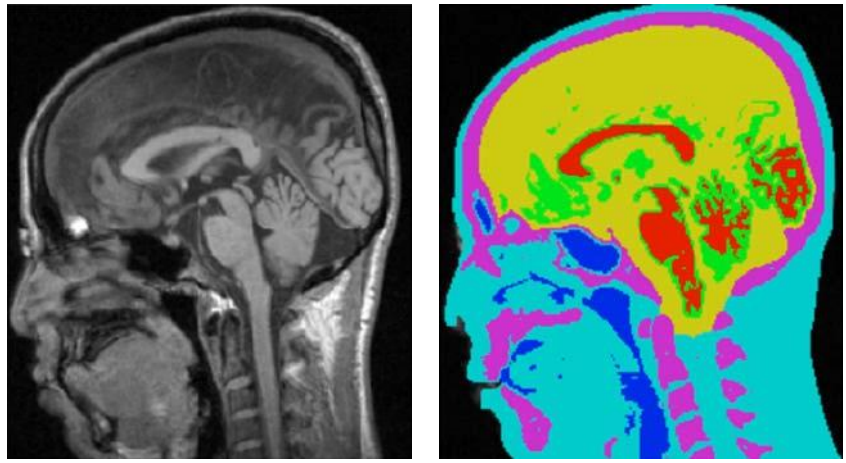
3.2.1. MRI/CT Registration. After MRI segmentation, we already have the MRI brain volume that is composed of WM and GM, but not CSF. The CT brain volume occupies spaces of WM, GM, and CSF. By applying morphological closing at a proper scale on the union of the MRI WM and GM, we obtain the MRI brain mask B_m, which has some inner CSF space filled and the brain contour smoothed but still misses some CSF voxels mainly wrapping around the superior contour of the brain. MRI segmentation also produces the MRI head mask H_m and the two eyeball masks E_1 and E_2. Correspondingly, we have the CT brain mask B_c and the CT head mask H_c. These data form the input set used to determine the transformation coefficients of the MRI/CT registration.

MRI/CT registration undergoes a rigid transformation followed by an affine transformation. Different sets of landmarks are extracted for the two transformations. The rigid transformation aims at only providing a good initial state for the affine transformation and its landmarks are simply those bordering voxels around B_m. The similarity measure to optimize the rigid transformation is the sum of squares of the distances from all landmarks to the contour of B_c. The optimal coefficients are found by multiscale exhaustive searching with acceptable computational efficiency and good global optimization performance.

The landmarks for the affine transformation are more sophisticated and consist of two parts: the inferior landmarks and the superior landmarks. The inferior landmarks are from the border voxels of the inferior part of B_m while the superior landmarks are from the superior contour of H_m. The inferior and the superior landmarks are partitioned so that they "visually" wrap up B_m but do not overlap. The motivation of using two sets of landmarks is to align MRI with CT in terms of both the brain matching and the scalp matching. However, the brain volume in the CT may contain significantly more CSF than that in the MRI at the superior part of the volume and we make the observation that their contours have reliable matching only at the inferior part of the brain. The landmark extraction thus essentially involves the partitioning of the cephalic space into the inferior part and the superior part. BrainK takes the following three landmarks as reference points to partition the cephalic space: the center of the brain mask B_m and the centers of the two eyeball masks E_1 and E_2. All three points can be automatically determined.

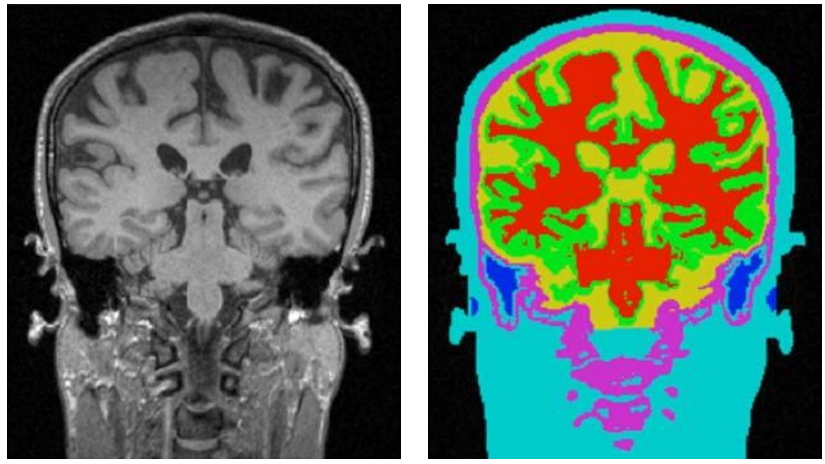
The similarity measure for the affine transformation is formulated as the sum of two terms. One is the sum of squares of the distances from all inferior landmarks to the contour of B_c and the other is the sum of squares of the distances from all superior landmarks to the contour of H_c. A gradient-descent optimization method is used to obtain the optimal affine coefficients. When all the optimal coefficients of both the rigid and the affine transformation are obtained, the MRI/CT registration can be performed from one direction while the other can be achieved by simply inverting the coefficient matrix. The result of MRI-to-CT registration is shown in Figure 5.

3.2.2. Skull Atlas-to-MRI Registration. Since the Atlas and the individual subject can have significant geometric differences, the Atlas-to-MRI registration undergoes an elastic thin-plate spline transformation following the rigid and the affine transformations. The rigid transformation provides a good initial state for the affine transformation, which in turn provides a good initial state for the elastic local transformation. The result of the Atlas-to-MRI registration is illustrated in Figure 6.



(a) (b)

Figure 5: MRI-to-CT registration. (a) Transformed MRI. (b) Fusion of MRI and CT segmentation.



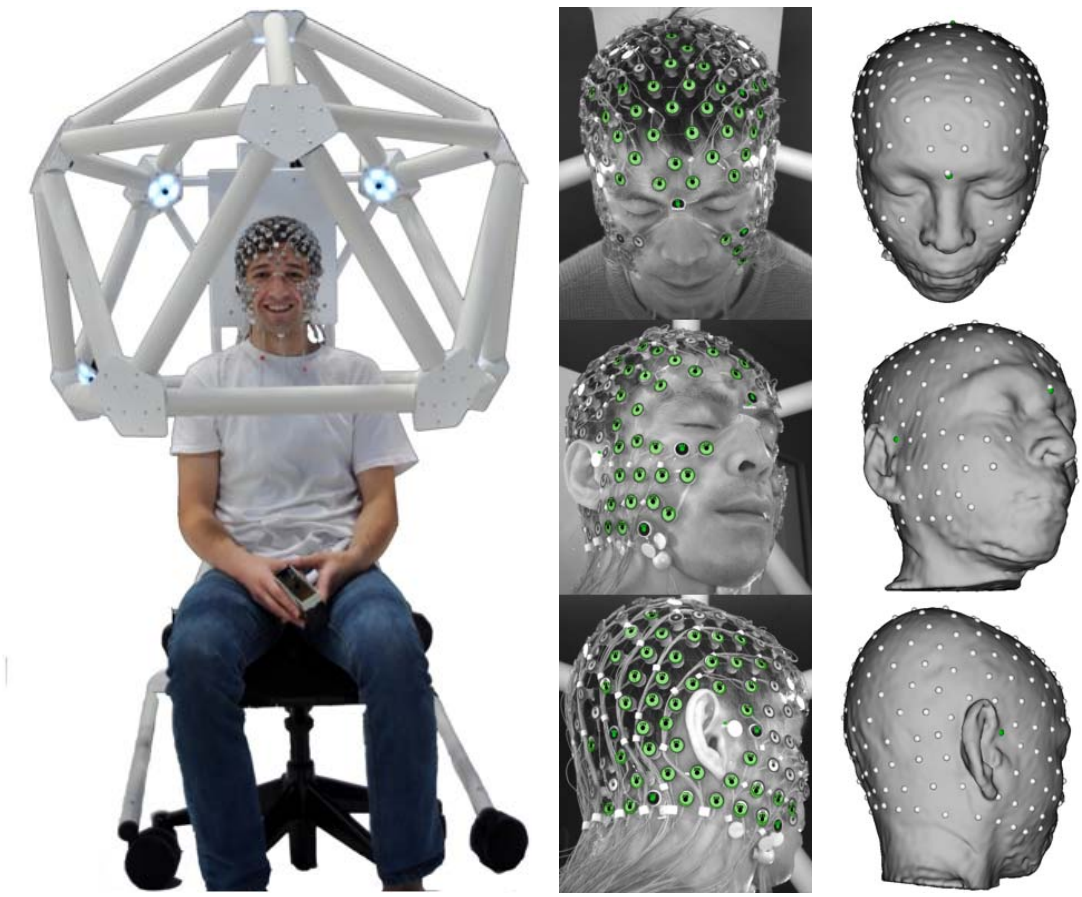
(a) (b)

Figure 6: Skull Atlas-to-MRI registration. (a) Transformed MRI. (b) Fusion of Atlas skull and MRI segmentation.

The rigid transformation and the affine transformation use the same set of landmarks to formulate the similarity measures. Let B_m be the MRI brain mask defined as in MRI/CT registration. Let B_a be the Atlas brain mask defined in the same way as B_m. The landmarks are then the set of border voxels of B_m and the similarity measure is the sum of squares of distances from all landmarks to the contour of B_a. The coefficients of the rigid transformation are found by multiscale exhaustive searching while those of the affine transformation are determined by gradient-descent searching.

The source landmarks used in the thin-plate spline transformation include several parts. The two primary parts are the inferior source landmarks and the superior source landmarks. Inferior source landmarks are distributed over the inferior part of the contour of the Atlas brain mask B_a. The superior source landmarks are placed over the superior part of the contour of the Atlas head mask H_a. Their extraction is very similar to the extraction of those used in the MRI/CT registration, but they are much more sparsely distributed to reduce the computational overhead.

The corresponding target landmarks in the subject head are determined automatically after the rigid and affine transformation are performed. For each superior source landmark, the corresponding superior target landmark is obtained as the intersection of the contour of the subject head mask H_m and the line between the superior source landmark and the subject brain center. We have two ways to obtain the inferior target landmarks according to their locations. The first method is the same as that obtaining the superior target landmarks and works for the posterior part of the inferior landmarks that are distributed around the cerebellum. For the second method that works for the anterior part of the inferior landmarks, we compute a field of distance from the contour of the brain mask B_m, and each inferior target landmark, which should be on the contour of B_m, is obtained by tracing from the corresponding source landmark following the gradient of the distance field.



(a) (b) (c)

Figure 7: (a) Geodesic Photogrammetry System (GPS). (b) Identification of EEG sensors in the GPS software. (c) Registration of EEG sensors with the MRI head surface in BrainK.

In the procedure of the registration, the Atlas skull's thickness is adjusted according to an estimated ratio of the subject skull's thickness to the Atlas skull's thickness. This is achieved by using another set of source landmarks and another set of target landmarks. For each superior source landmark, we determine its "mate landmark" over the inner side of the Atlas skull and on the line between the superior source landmark and the Atlas brain center. For each superior target landmark, there is also a "mate landmark" on the line between the superior target landmark and the subject brain center. For each pair of the superior source landmark S_i and its mate landmark S_i', we have a distance d_i^S; for each pair of the superior target landmark T_i and its mate landmark T_i', we have a distance d_i^T. A global ratio is then defined as d_i^T/d_i^S and its optimum is determined as the highest value so that the majority (such as 90%) of the superior target mate landmarks are off the mask B_m. For the few superior target mate landmarks that are within B_m according to the global ratio, the global value is decreased locally so that they are off the mask B_m.

For each posterior inferior source landmark, we determine its mate landmark over the outside of the Atlas skull and on the line between the source landmark and the Atlas brain center. For each posterior inferior target landmark, we also have a mate landmark on the line between the target landmark and the subject brain center while its location is determined according to the global thickness ratio described above. In order to further improve alignment performance, there are few other landmarks such as the centers of the eyeballs. We have a specific algorithm to automatically detect the subject's eyeballs, but the details are not described here for the limit of space. When all source landmarks and all target landmarks are determined, the transformation for any point in the space can be calculated as shown in (2).

3.2.3. Registrations to the EEG Sensor Positions. The EEG sensor positions are automatically localized in 3D with the Geodesic Photogrammetry System (GPS). GPS captures images from 11 cameras in fixed positions to allow identification of the sensors in the images and then computation of the 3D coordinates (Figure 7). The Atlas-to-MRI-to-GPS registration then consists of two stages, the MRI-to-GPS registration and then the Atlas-to-MRI registration, using the newly transformed MRI. Atlas-to-MRI registration has been described in the above subsection, and this subsection only deals with the MRI-to-GPS registration. There are three

steps. The first can be an automatic rigid transformation with other alternatives described below. The second is an affine transformation. The first two can bring the MRI to a good alignment state with the GPS data, but the GPS points may still be off the scalp. The third step then performs thin-plate spline transformation to deform the MRI so that each GPS point locates on the scalp. The landmarks used in the first two steps are the GPS points themselves, and the similarity measure is the sum of squares of the distances from the landmarks to the MRI scalp, that is, the contour of the head mask H_m. For the thin-plate spline transformation, the target landmarks are the GPS points while the source landmarks are those points on the contour of H_m so that each one is on the line connecting the brain center and the corresponding target landmark.

The first automatic step of the MRI-to-GPS registration can be replaced by a transformation with the user providing three fiducial landmarks on the MRI space: the nasion and the left and right periauricular points (where the jaw hinge meets the skull). BrainK is able to stretch the MRI so that the user-specified landmarks match the corresponding ones in the GPS space. This initial state can be directly fed to the thin-plate spline transformation or it may pass through an intermediate affine transformation. BrainK also allows the user to specify only the nasion landmark and does rigid transformation in the first step while keeping the corresponding nasion landmarks matched.

The Atlas-to-GPS registration is very similar to the MRI-to-GPS registration, except that the Atlas typically has greater geometric differences from the GPS than the subject's own MRI. The steps are similar to those for the MRI-to-GPS transformation just described. For the MRI-to-GPS registration, however, if they are to be used, the fiducial landmarks have to be manually marked on the MRI, whereas the fiducials are included in the digital Atlas dataset for the Atlas-to-GPS registration. In addition, one more fiducial landmark, the vertex, is used to accommodate the greater geometric difference between the Atlas and the subject.

The GPS-to-head registration can be seen as the inverse of the registration of MRI-to-GPS, but no thin-plate spline transformation is needed. The initial transformation, either automatic rigid transformation or the stretching in terms of fiducial landmarks matching, can put the GPS in a good initial state. If needed, an affine transformation can be performed to further improve the alignment. The final adjustment then moves each GPS point onto the scalp, if it is not on it yet, with an end result that is similar to finding the corresponding landmarks in the MRI-to-GPS registration.

When GPS is not available, a set of generic sensor positions may be used, such as those from the appropriate Geodesic Sensor Net (channel count and size) selected from the database. This registration has two steps. First, the generic spherical sensor cloud undergoes a translation transformation so that the center of the sensor point cloud overlaps with the center of the brain. Second, in a procedure that matches the physical conformation of the geodesic tension structure to the head, each sensor is mapped to a point on the scalp by the ray emanating from the brain center to the sensor's position.

3.3. Cortical Surface Reconstruction and Inflation. The well-known marching-cube isosurface algorithm [24] is performed on the cerebral cortex volume of each hemisphere to reconstruct the cortical surface mesh. In addition, BrainK supports cortical surface inflation by iteratively updating all points in the mesh, such that the new position of each point is the weighted sum of its old values with the weighted sum of the centroids of the triangles that are incident to the point.

3.4. Dipole Tessellation. The dipole tessellation operates in two modes. If the individual's MRI is available, the cortical surface is extracted and tessellated into patches, and a dipole is fit to each patch. If it is not available, then dipole *triples* (fitting x, y, z components of the unknown dipole orientation) are distributed evenly through the cortical gray matter.

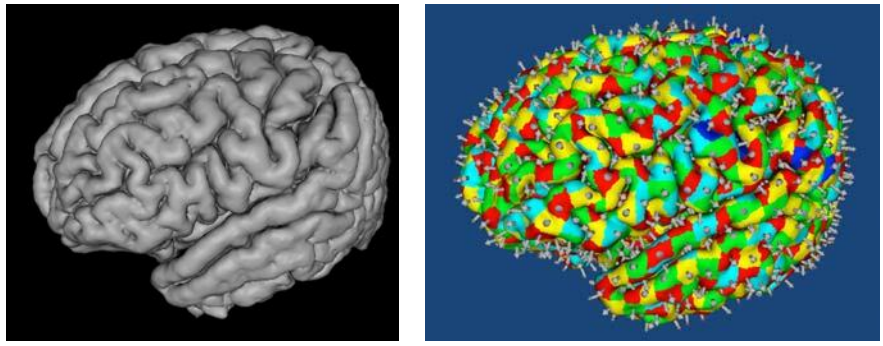
Whereas a trivial algorithm can distribute the triples, the even distribution of the oriented dipoles over the cortical surface is formulated as a sophisticated *graph partition* [25] problem. Given the cortical mesh from marching cubes, composed of a great number of interconnected tiny triangles, a graph is first constructed so that each *vertex* corresponds to a triangle and each *edge* incident to two vertices corresponds to the adjacency between the two corresponding triangles. The graph partition algorithm then divides the entire graph into a set of parts, with the constraints that (1) the mesh is divided into the same number of patches as the desired number of oriented dipoles, (2) all patches are very compact (the more like to a square or circle, the better), and (3) all patches have similar areas. The graph partition library Chaco [26] is used for the implementation.

Given the mesh partition performed as above, an equivalent oriented dipole is then defined for each patch, generated at the center of the patch, with its orientation defined by the mean of all surface normals of the triangles in the patch, weighted by their areas. This equivalent dipole thus reflects the orientation of the electrical source if the entire patch was synchronously active, thereby serving as a useful approximation for the resolution of cortical activity that is set by the graph partitioning decision. The cortical surface and the oriented dipoles are displayed in Figure 8.

3.5. Talairach Transformation. The Talairach transformation component brings all the data, the segmentation, the cortical surfaces, the sensors, and the dipoles, into the standard Talairach space [27, 28]. The transformation is determined with three points specified by the user: the anterior commissure (AC), the posterior commissure (PC), and one point forming the middle plane with AC and PC. The procedure maintains the topology correctness of the cortical surface that is guaranteed in either the MRI segmentation component or the Atlas dataset and then maintained through each registration operation.

4. Evaluation of BrainK's Tissue Segmentation and Cortical Surface Extraction

BrainK provides certain algorithms, such as skull bone density (X-ray CT) image registration and cortical surface dipole



(a) (b)

Figure 8: Cortical surface (a) and tessellation of the surface into ~1 cm^2 patches (about 1200 per hemisphere), each with an oriented equivalent dipole (b). Color is used here only to separate the cortical patches, such that every patch has a different color from its neighbors.

tessellation, which are specifically important to model the electrical properties of the human head. As described below in Section 8, efficient construction of accurate electrical volume conduction head models is important to improve the source analysis of human electrophysiological activity with dense array technologies. Accurate head tissue conductivities and geometries are also important to optimize the transcranial current delivery to stimulate the brain that is now possible with dense array technologies.

In addition to these somewhat unique capabilities, BrainK implements certain tissue segmentation and surface extraction algorithms that are useful in medical image processing generally and that can be evaluated in comparison with existing software that has become well known in the neuroimaging research community. Examples of the software examined here are FreeSurfer, SPM, FSL, and BrainVisa. To assure accuracy for medical applications, validation of BrainK's unique algorithms against well-known reference software not only is scientifically useful but also forms an integral part of medical quality verification and validation. In addition, medical use requires that the functionality is fast and intuitive, such that BrainK must meet standards of efficiency and ease of use.

Our current evaluation work pays particular attention to brain tissue segmentation due to the fact that the accuracy of brain tissue segmentation is critical for a variety of medical and neurological applications such as source localization and cortical dysplasia detection.

4.1. Comparative Packages. Below is a brief description of the four brain image analysis tools used for comparative performance evaluation. Although these tools provide varying functionality in neuroimage analysis, they all support automatic T1-weighted human brain MR image segmentation. Our comparative evaluation is focused on the image segmentation task.

4.1.1. FreeSurfer. FreeSurfer [29, 30] is a set of tools for reconstruction of cortical surfaces from structural MRI data and for the overlay of functional data onto the reconstructed surface. FreeSurfer is developed in the Nuclear Magnetic Resonance (NMR) Center, Massachusetts General Hospital. The cortical surface reconstruction pipeline in FreeSurfer mainly consists of three steps. First, a brain mask is extracted with alignment of the structure MR image to the Talairach Atlas and the bias field is corrected. Then, the brain volume is labeled as various cortical or subcortical structures in a procedure based on both a subject-independent probabilistic Atlas and subject-specific measured values. Finally, the cortical surfaces are constructed from the prior segmentation, which involves a topology correction procedure.

4.1.2. SPM. SPM (Statistical Parametric Mapping) is a statistical technique for testing hypotheses about functional imaging data [31]. SPM also refers to the software developed by the Wellcome Department of Imaging Neuroscience, University College London, to carry out such analysis. SPM features structural MRI segmentation as well as a series of functional neuroimage analysis. Structural MRI segmentation in SPM can be characterized as a circular procedure that involves alternating three processing steps [32]: a bias correction step that corrects the intensity inhomogeneity, a registration step that normalizes the image to standard tissue probability maps, and a segmentation step that classifies image voxels into different tissue types. As the segmentation results, SPM assigns each image voxel three probabilities with respect to three tissue types: CSF, GM, and WM. Our experiments were conducted with SPM5, which is old with respect to the latest SPM12, but their implementation is based on the same algorithm presented in [32], although the newer version makes use of additional tissue classes and multichannel segmentation and incorporates a more flexible image registration component, as stated in the release notes of SPM12.

4.1.3. FSL. FSL (the FMRIB Software Library) is a collection of functional and structural neuroimage analysis tools [33]. For structural segmentation, FSL applies the Brain Extraction Tool (BET) for segmenting brain from nonbrain regions in structural and functional data, and FAST (FMRIB's Automated Segmentation Tool) for bias field correction and brain segmentation into three tissue types: CSF, GM, and WM.

Structural MRI segmentation in FSL consists of two steps: using BET to extract the brain and using FAST to classify tissue types. BET performs skull stripping with a surface model [34]. The underlying method of FAST is based on an expectation-maximization algorithm combined with a hidden Markov random field (MRF) model [35]. Due to the regularization of the MRF model, FAST is supposed to be more robust to noise than standard finite mixture model based methods.

4.1.4. BrainVisa. BrainVisa [36, 37] is software developed at Service Hospitalier Frédéric Joliot (SHFJ) that encompasses an image processing factory and is distributed with a toolbox of building blocks dedicated to the segmentation of T1-weighted MR image. Structural MRI segmentation in BrainVisa consists of four main steps. First, the user prepares the data for segmentation by specifying several key landmark points including the anterior commissure (AC), the posterior commissure (PC), an interhemispheric point, and a left hemisphere point. A brain mask is then extracted including only white matter and gray matter integrating bias field correction [38] and histogram analysis [39]. This is followed by a hemisphere partition and removal of cerebellum with morphological image analysis [40]. Finally, cerebral gray matter and white matter are differentiated with histogram analysis [41].

5. Datasets

The evaluation was performed on three image datasets: a set of BrainWeb data with ground truth segmentation, a set of IBSR data with manually guided expert segmentation, and a set of real scans of subjects with either mild cognitive impairment or Alzheimer's disease.

5.1. BrainWeb. The BrainWeb dataset is a group of 8 realistic T1-weighted MR simulated images with ground truth segmentation provided by BrainWeb, a simulated brain database [42, 43]. All 8 MR images are simulated on a normal anatomical model. The resolution of the images is $1\,\text{mm}^3$. In the ground truth image, all voxels in the image are segmented into the following tissue types: Background, CSF, GM, WM, Fat, Muscle/Skin, Skin, Skull, Glial Matter, and Connective.

A variety of noise levels and levels of intensity inhomogeneity (i.e., intensity nonuniformity (INU)) are artificially introduced in the simulated images, as listed in Table 1. As stated in BrainWeb documentation [44], the "noise" in the simulated images has Rayleigh statistics in the background and Rician statistics in the signal regions. The "percentage noise" number represents the percent ratio of the standard deviation of the white Gaussian noise versus the signal for a reference tissue. The noise reference tissue used in our dataset is white matter. The meaning of the intensity inhomogeneity level is as follows. "For a 20% level, the multiplicative INU field has a range of values of 0.90–1.10 over the brain area. For other INU levels, the field is linearly scaled accordingly (e.g., to a range of 0.80–1.20 for a 40% level)." According to BrainWeb, the INU fields are realistic in that they are slowly varying fields of a complex shape and were estimated from real MRI scans.

Table 1: Noise levels and IIH levels of the BrainWeb datasets.

Dataset	1	2	3	4	5	6	7	8
Noise level	3%	3%	5%	5%	7%	7%	9%	9%
IIH level	20%	40%	20%	40%	20%	40%	20%	40%

5.2. IBSR. The IBSR dataset is a group of 18 T1-weighted real MR brain datasets. Their manually guided expert segmentation is included in the Internet Brain Segmentation Repository (IBSR) supported by the Center for Morphometric Analysis (CMA) at Massachusetts General Hospital [45]. The slice resolution of all datasets is 1.5 mm and the XY resolution varies from $1\,\text{mm}^2$ to $0.837\,\text{mm}^2$. The MR images have been "positionally normalized" into the Talairach orientation, but all five tools performed on this group of data assumed that the images were not normalized. The MR images were also processed by the CMA bias field correction routines, but it is not guaranteed that the intensity inhomogeneity is completely corrected. All five tools therefore treated the datasets as if no bias field correction had been performed.

Each MR image was manually segmented into 44 individual structures including 3rd ventricle, 4th ventricle, Brainstem, and left and right: accumbens area, amygdala, amygdala anterior, caudate, cerebellum cortex, cerebellum exterior, cerebellum white matter, cerebral cortex, cerebral exterior, cerebral white matter, hippocampus, Inf Lat vent, lateral ventricle, pallidum, putamen, thalamus proper, ventral DC, and vessel.

The 18 MR images represent various levels of image quality. To organize the evaluation by general image quality, we divided the images into two subgroups: the first 13 MR images with good quality and 5 more MR images with bad quality. Note that the ordering of the IBSR datasets is different from the original order. A map of the order we used to the original order is "1, 2, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 3, 4, 13." For example, when we refer to the 3rd dataset, it is actually the 5th in the original order.

5.3. Datasets Reflecting Neuropathology. In addition to the BrainWeb and the IBSR datasets, which were used for both quantitative and qualitative evaluation, we also tested five tools on an auxiliary group of 8 real MR images scanned from subjects with minor cognitive impairment or those with Alzheimer's disease. The resolution of these datasets is $1.139 \times 1.211 \times 1.211\,\text{mm}^3$. The source of these datasets is the Neurobiology Research Unit [46] in the University Hospital Rigshospitalet in Denmark. No ground truth or manual segmentation is provided for these datasets; the issue for the evaluation is primarily whether the abnormal brains can be segmented with apparently reasonable accuracy.

6. Quantitative Evaluation

In this section, we present a quantitative evaluation on the segmentation accuracy, robustness, and computational

efficiency of BrainK in comparison to other four packages. We use the widely used Dice metric [1, 4, 46, 47] as the measurement for segmentation accuracy. The standard deviation of the Dice metric provides a measure of segmentation robustness. Computational efficiency is measured as the required run time of each package.

6.1. Dice Metric. Let TP refer to the number of true positives, FP to false positives, and FN to false negatives. The Dice metric is then given by

$$\text{Dice metric} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}. \quad (3)$$

TP, FP, and FN are measured versus the manual segmentation of real MR datasets or the ground truth of simulated images. Note that when the segmentation is given as a probability between 0 and 1 for each image voxel for each tissue class (such as in the case of SPM), then TP, FP, and FN are calculated as the sum of the probabilities instead of discrete counting.

For quantitative evaluation using Dice metric, we must decide the tissue type on which the metric is measured. BrainVisa only performs cerebrum segmentation while FSL and SPM segment the entire brain into CSF, GM, and WM without extraction of the cerebrum. FreeSurfer also performs segmentation on the whole brain, but it segments the brain into a greater number of tissue types (including cerebral white matter and cerebral cortex). Currently, BrainK performs segmentation on the entire image, and it is able to differentiate cerebrum from cerebellum and one cerebral hemisphere from the other. In our quantitative evaluation, we must calibrate the segmentation of the five packages within a standard framework, so that common tissue types can be used for quantitative metric measurements.

For the BrainWeb datasets, we calibrated the segmentation of five tools to the segmentation of cerebral WM and cerebral GM and measured the Dice metrics with respect to these two tissue types. To enable this, we manually partitioned the ground truth whole brain (WM plus GM) at the brainstem to extract the cerebral WM and the cerebral GM. Cerebral WM and cerebral GM also need to be extracted for the SPM and FSL segmentation results. We use a procedure (described in the next paragraph) that almost "perfectly" partitions the segmentation results based on the ground truth partition. For BrainVisa and BrainK, no transformation in the calibration is required. For FreeSurfer, we just need to relabel all cerebral cortex voxels and all subcortical voxels excluding cerebral WM as cerebral gray matter.

Let TP-Cerebrum and TP-Cerebellum, respectively, denote the set of true positives of cerebrum and cerebellum in the segmentation of SPM and FSL. Let FP-Brain denote the entire false positives including those in cerebrum and cerebellum. The partition of the brain segmented by FSL and SPM is essentially the partition of FP-Brain into false positives in cerebrum and those in cerebellum, which is described as follows. For each voxel v in FP-Brain, if it has a shorter path in FP-Brain to TP-Cerebrum than any paths in FP-Brain from v to TP-Cerebellum, then v is taken as a false positive (of WM or GM) in cerebrum; otherwise, it is taken as a false positive (of GM or WM) in cerebellum.

For the IBSR data, we calibrated the segmentation of five tools to the segmentation of cerebral cortex and cerebral WM and measured the Dice metrics with respect to these two tissue types. These quantitative metrics give an evaluation on the accuracy of the cortical surface reconstruction which depend on segmentation of cerebral cortex and cerebral WM and are irrelevant to segmentation of subcortical gray matter tissues. Since FreeSurfer explicitly labels cerebral cortex and cerebral white matter, we do not need to do any transformation in the calibration. The calibration of FSL and SPM first conducts the brain partition to extract the cerebral WM and cerebral GM. Given the set of cerebral WM and cerebral GM segmented by FSL, SPM, BrainVisa, or BrainK, we measured the Dice metrics with respect to cerebral cortex and cerebral WM in the way described below.

6.2. Experiment Settings

6.2.1. FreeSurfer. We tested FreeSurfer on both the BrainWeb and the IBSR datasets in a fully automatic mode without any user intervention. An issue in collecting FreeSurfer segmentation results is the production of the cerebral cortex mask. There is a so-called "aseg" image and a "ribbon" image and both of these record voxels are labeled as cerebral cortex. The "ribbon" data is what FreeSurfer suggests to use, but it has more false negatives than the "aseg" data. In contrast, the "aseg" image is an intermediate result, and it has more false positives than the "ribbon" data. We applied a simple morphological closing operation on the union of the cortex ribbon and the subcortical structures so that certain true cerebral cortex voxels labeled in "aseg" but missed in "ribbon" are covered. This procedure apparently improved the performance of the cerebral cortex segmentation. The Dice metric for the "aseg" image was 0.7353, for the "ribbon" image was 0.8009, and for the "closed" image was 0.8212.

6.2.2. FSL. In our first batch of experiments with FSL, we let FSL automatically extract the brain and performed brain tissue classification on both the BrainWeb and the IBSR datasets. However, FSL generated poor results on the brain extraction and brain tissue classification on 6 of the IBSR datasets (dataset 5 to dataset 10). In our second batch of experiments, therefore, we used different parameters in FSL, obtained better brain masks for these datasets, and then repeated the brain tissue classification. The brain masks generated in the second batch of experiments were still not good enough. We therefore used the brain masks generated by FreeSurfer for the brain tissue classification in FSL. This delivered the best FSL brain segmentation performance on the 6 difficult IBSR datasets. The three batches of experiments on FSL show that the brain extraction algorithm of FSL is not robust on the IBSR datasets, but the brain tissue segmentation itself performed well given good brain masks. The performance of FSL on the 6 IBSR datasets with respect to the three batches of experiments is shown in Table 2.

Table 2: Dice metrics collected for FSL using different brain masks.

Brain masks	Tissue types	Dice metrics on 6 IBSR datasets					
		5	6	7	8	9	10
Default FSL Brain mask	Cerebral cortex	0.6591	0.6806	0.7268	0.6887	0.7713	0.6762
	Cerebral WM	0.8845	0.8928	0.8891	0.8335	0.9095	0.8792
Customized FSL Brain mask	Cerebral cortex	0.7608	0.7735	0.7772	0.7740	0.7854	0.7859
	Cerebral WM	0.8722	0.8747	0.8766	0.8711	0.8953	0.8767
FreeSurfer Brain mask	Cerebral cortex	0.7312	0.7559	0.7898	0.7587	0.8277	0.7471
	Cerebral WM	0.8862	0.8953	0.8912	0.8914	0.9146	0.9009

6.2.3. SPM. In our first batch of experimental tests with SPM, we used the default parameters and let SPM automatically perform brain tissue segmentation on the BrainWeb and the IBSR datasets. In the second batch of experiments, we changed the parameter "bias regularization" from the default "very light regularization" to "medium regularization" and reran SPM on the IBSR datasets. SPM is supposed to be used with greater bias regularization when it is known *a priori* that there is less intensity inhomogeneity in the image. Since the IBSR datasets were processed with bias field correction, the use of "medium regularization" rather than the default "very light regularization" improved the performance of SPM on almost all IBSR datasets. For the cerebral cortex, the mean Dice metric over the 18 ISBR datasets was 0.7609 for very light regularization and 0.7656 for medium regularization. The mean Dice metric for white matter was 0.8572 for very light regularization and 0.8707 for medium regularization. We used the best regularization for each dataset in our comparative evaluation.

6.2.4. BrainVisa. We tested BrainVisa on both the BrainWeb and the IBSR datasets automatically, with the exception that we manually specified landmark points including the AC point, the PC point, an interhemispheric point, and a left hemisphere point. BrainVisa produced an empty brain mask in the 9th IBSR dataset and was unable to generate brain masks for the 13th and the 18th datasets. In both cases, we set the Dice metrics to be 0.

6.2.5. BrainK. BrainK was tested on the BrainWeb and the IBSR datasets fully automatically.

6.3. Comparison Results

6.3.1. Segmentation Accuracy on the IBSR Datasets. We collected the Dice metrics with respect to cerebral cortex (Table 3) and cerebral WM (Table 4) using the five software packages on the IBSR datasets.

On average, BrainK performed best on cerebral cortex segmentation on all 18 IBSR datasets, good and bad, and on the 13 good datasets exclusively. In particular, BrainK's cerebral cortex performance is consistently better than the four other packages on the 13 good datasets except for the 4th dataset, where BrainK's performance is almost identical to the best, and the 12th dataset, where BrainK's performance is close to the best. The cerebral cortex segmentation performance of the five packages on the five bad datasets is similar, except that BrainVisa generated empty brain mask for the 18th dataset.

Table 3: Dice metrics of five tools with respect to cerebral cortex on the IBSR datasets.

IBSR datasets	Dice metrics with respect to cerebral cortex				
	BrainVisa	SPM5	FreeSurfer	FSL	BrainK
1	0.7461	0.7705	0.8039	0.7803	**0.8682**
2	0.7953	0.8048	0.8175	0.8121	**0.8619**
3	0.7674	0.8080	0.8362	0.8361	**0.8714**
4	0.7233	0.8356	**0.8641**	0.8028	0.8612
5	0.2875	0.4621	0.7794	0.7312	**0.8638**
6	0.6610	0.6609	0.8068	0.7559	**0.8441**
7	0.7108	0.7065	0.7888	0.7898	**0.8638**
8	0.6982	0.7670	0.7800	0.7587	**0.8790**
9	0	0.7595	0.8128	0.8277	**0.8700**
10	0.7707	0.6868	0.7729	0.7471	**0.8611**
11	**0.8688**	0.8396	0.8702	0.8833	0.8634
12	0.8596	0.8541	0.8458	0.8582	**0.8772**
13	0	0.8588	0.8647	0.8554	**0.8673**
14	0.8406	0.8426	0.8065	**0.8429**	0.8315
15	0.8441	0.8439	**0.8487**	0.8381	0.8457
16	0.8260	0.8263	0.8413	**0.8426**	0.8281
17	**0.8445**	0.8443	0.8076	0.8278	0.8379
18	0	**0.8539**	0.8337	0.8278	0.8379
Mean	0.6247	0.7656	0.8212	0.8121	**0.8557**
Mean on 13 good datasets	0.6068	0.7365	0.8187	0.8030	**0.8656**

To provide an estimate of the statistical significance of the comparison of BrainK with the other packages, we conducted paired comparison t-tests across these 18 datasets. BrainK was more accurate at segmenting cerebral cortex with the ISBR datasets than BrainVisa ($p < 0.02$), SPM ($p < 0.006$), FreeSurfer ($p < 0.0005$), and FSL ($p < 0.0009$). If we omitted the three datasets on which BrainVisa failed, the comparisons again showed BrainK to be more accurate than BrainVisa ($p < 0.002$), SPM ($p < 0.009$), FreeSurfer ($p < 0.001$), and FSL ($p < 0.002$).

Table 4: Dice metrics of five tools with respect to cerebral WM on the IBSR datasets.

IBSR datasets	Dice metrics with respect to cerebral WM				
	BrainVisa	SPM5	FreeSurfer	FSL	BrainK
1	0.8652	0.8927	0.7964	**0.8971**	0.8789
2	0.8899	0.8940	0.8208	**0.9160**	0.8926
3	0.8649	0.8936	0.8138	**0.9084**	0.8874
4	0.8596	0.9013	0.8489	**0.9168**	0.9070
5	0.4097	0.7312	**0.9240**	0.8862	0.9101
6	0.7970	0.7422	**0.9115**	0.8953	0.8954
7	0.8255	0.8734	**0.9147**	0.8912	0.9001
8	0.8110	0.8923	**0.9203**	0.8914	0.9170
9	0	0.9059	0.9179	0.9146	**0.9224**
10	0.8457	0.8744	**0.9069**	0.9009	0.8790
11	0.8975	0.8927	0.8711	**0.9142**	0.8969
12	0.8858	**0.9014**	0.8099	0.8988	0.9006
13	0	**0.8955**	0.8647	0.8673	0.8748
14	0.8613	**0.8678**	0.7824	0.8632	0.8622
15	0.8541	0.8880	**0.8746**	0.8743	0.8637
16	0.8792	**0.8948**	0.8479	0.8933	0.8713
17	**0.8619**	0.8551	0.8592	0.8585	0.8463
18	0	**0.8764**	0.8124	0.8323	0.8463
Mean	0.6893	0.8707	0.8610	**0.8900**	0.8862
Mean on 13 good datasets	0.6886	0.8685	0.8708	**0.8999**	0.8971

On average, FSL performed best on cerebral WM segmentation on the 18 IBSR datasets and on the 13 good datasets. However, BrainK's performance was very close to FSL's performance in this comparison. The performance of the five packages on the five bad datasets is similar, except that BrainVisa generated empty brain mask for the 18th dataset, and FreeSurfer gave particularly poor performance for the 14th dataset.

6.3.2. Segmentation Robustness on the IBSR Datasets. We calculated the standard deviations of the Dice metrics on the IBSR datasets. Together with the mean Dice metrics, these values indicate the segmentation robustness of the five packages on MR images scanned from different subjects. Greater mean Dice metric and lower standard deviation indicate greater robustness with respect to segmentation accuracy.

Two groups of standard deviations were calculated: one on the total 18 IBSR datasets and one on the 13 good IBSR datasets. BrainK demonstrated the lowest standard deviation of the software packages with respect to cerebral cortex on both the total 18 IBSR datasets and the 13 good IBSR datasets, as shown in Table 5. The lowest mean and standard deviation Dice metric with respect to cerebral cortex indicate that BrainK demonstrates the best accuracy and robustness with respect to cerebral cortex on the IBSR datasets. For cerebral WM, BrainK and FSL performed very similarly with respect to both the mean and the standard deviation of the Dice metric on both the total 18 IBSR datasets and the 13 good IBSR datasets. BrainK and FSL tied for the best accuracy and robustness with respect to cerebral WM on the IBSR datasets. Considering both cerebral cortex and cerebral WM, these results suggest that, in general, BrainK demonstrated the best overall performance on the IBSR datasets.

6.3.3. Segmentation Robustness with respect to Noise and IIH on the BrainWeb Datasets. As described in Section 3.1, the BrainWeb datasets vary in noise levels and intensity inhomogeneity (IIH) levels. The performance in Dice metrics of the five packages with respect to cerebral GM and cerebral WM is listed in Tables 6 and 7.

Among the five packages, FreeSurfer demonstrated the lowest performance variation over different noise levels. SPM and BrainVisa are similar in showing the highest performance variations over different noise levels. BrainK and FSL have medium performance variations over different noise levels, compared to the other three packages. Although FreeSurfer performed consistently over different noise levels, it also gave results with the lowest accuracy on average.

For each of the four noise levels, we also tested the packages on images with two different IIH levels. All five packages gave little variation over different IIH levels. The only exception is for BrainVisa with $N = 9\%$ and IIH = 40%. This is due to a poor brain mask.

It is worth noting that, in real MR scans, the intensity inhomogeneity may occur in various and unknown patterns. Such noise could occur together with other difficulties that may be not present in the simulated BrainWeb datasets. Therefore, we remark that our experiments with the BrainWeb dataset should not be taken to give a thorough and sufficient evaluation on the five packages with respect to the IIH robustness.

6.3.4. Computational Efficiency. The execution times of the fives packages tested on the IBSR datasets and the BrainWeb datasets are listed in Table 8. The experiments were all run on a single 2.8 Ghz Intel Xeon processor. Among the five packages, BrainVisa took the least amount of time, but it also produced the lowest segmentation accuracy and robustness on the IBSR datasets. FreeSurfer is well known for long execution times. However, these long execution times cover segmentation of more subcortical structures as well as reconstruction of cortical surfaces. It should be noted that most of the BrainK time was spent for topology correction,

Table 5: Standard deviation of Dice metrics of five tools on the IBSR datasets.

Sample groups		Standard deviations				
Datasets	Tissue type	BrainVisa	SPM5	FreeSurfer	FSL	BrainK
All IBSR datasets	Cerebral cortex	0.3155	0.1284	0.0308	0.0429	0.0192
	Cerebral WM	0.3351	0.0505	0.0473	0.0230	0.0225
13 good datasets	Cerebral cortex	0.3047	0.1414	0.0345	0.0476	0.0087
	Cerebral WM	0.3305	0.0593	0.0482	0.0143	0.0147

Table 6: Dice metrics of five tools with respect to cerebral GM on the BrainWeb datasets.

BrainWeb datasets		Dice metrics with respect to cerebral GM				
Noise level	IIH level	BrainVisa	SPM5	FreeSurfer	FSL	BrainK
3%	20%	**0.9292**	0.9173	0.8333	0.9242	0.9084
	40%	0.9247	0.9189	0.8342	**0.9268**	0.9086
5%	20%	**0.9197**	0.8989	0.8323	0.9193	0.8908
	40%	**0.9201**	0.8998	0.8323	0.9193	0.8858
7%	20%	0.8628	0.8673	0.8320	**0.9113**	0.8816
	40%	0.8740	0.8713	0.8312	**0.9127**	0.8827
9%	20%	0.8166	0.8255	0.8259	**0.8996**	0.8658
	40%	0.7836	0.8301	0.8264	**0.9019**	0.8678

Table 7: Dice metrics of five tools with respect to cerebral WM on the BrainWeb datasets.

BrainWeb datasets		Dice metrics with respect to cerebral WM				
Noise level	IIH level	BrainVisa	SPM5	FreeSurfer	FSL	BrainK
3%	20%	0.9550	0.9471	0.8849	**0.9672**	0.9558
	40%	0.9599	0.9533	0.8889	**0.9664**	0.9593
5%	20%	0.9552	0.9314	0.8824	**0.9567**	0.9494
	40%	0.9534	0.9315	0.8863	**0.9581**	0.9476
7%	20%	0.9325	0.8978	0.8779	**0.9448**	0.9382
	40%	0.9311	0.9008	0.8796	**0.9467**	0.9370
9%	20%	0.8926	0.8656	0.8757	**0.9332**	0.9296
	40%	0.8748	0.8701	0.8740	**0.9354**	0.9289

Table 8: Computation times of five tools on the IBSR and the BrainWeb datasets.

Datasets	Computation times					
	BrainVisa	SPM5	FreeSurfer	FSL	BrainK	BrainK (topology correction)
IBSR	1.5 m	34 m	27.2 h	5 m	17 m	14 m
BrainWeb	1.6 m	20 m	24.5 h	9 m	21 m	18 m

which was not included in the execution times of the other three.

7. Qualitative Evaluation

In this section, we give a qualitative evaluation of the five packages based on the results of applying the packages to the IBSR datasets, the BrainWeb datasets, and the 8 pathological datasets with mild cognitive impairment or Alzheimer's disease. We first summarize and compare the segmentation functionalities of the five packages followed by the discussion of their automaticity. Finally, we present various segmentation abnormalities that we observed in the experiments.

7.1. Segmentation Functionalities. The following is a summarization and comparison on the main segmentation features of the five packages:

(i) *Bias Field Correction*. Freesurfer, SPM, FSL, and BrainVisa all integrate a bias field correction procedure, either prior to tissue classification or combined with the classification. BrainK, on the other hand, does not need an explicit bias field correction because the relative thresholding method is robust to bias field in arbitrary patterns.

(ii) *Brain Extraction*. FSL, FreeSurfer, and BrainVisa provide separate tools for brain extraction (i.e., skull stripping) prior to brain tissue classification, whereas SPM combines brain extraction with tissue classification. BrainK, on the other hand, performs brain extraction after tissue classification. Note that the brain mask generated by BrainVisa is meant to contain only GM and WM while the brain mask generated by FSL and FreeSurfer is meant to contain CSF as well as GM and WM.

(iii) *Tissue Classification*. FSL and SPM segment the brain volume into three tissue types: CSF, GM, and WM. BrainVisa extracts cerebral WM and cerebral GM. BrainK classifies the entire region of interest into CSF, GM, and WM, extracts the GM and WM, and separates cerebellum from cerebrum. BrainVisa and BrainK also provide cerebral hemisphere partition. FreeSurfer segments a whole brain into 37 individual structures including cerebral cortex, cerebral WM, a set of subcortical structures, brainstem, and cerebellar structures.

(iv) *Cortical Surface Reconstruction*. BrainVisa, FreeSurfer, and BrainK support cortical surface reconstruction, but FSL and SPM do not. A core requirement for cortical surface reconstruction is to insure that the topology of the cortical surface is correct.

(v) *Multichannel Segmentation*. It is worth noting that FSL and SPM12 provide multichannel segmentation such that T1- and T2-weighted images can be combined together for increased accuracy. However, T2-weighted images are often not available and when they are available, T2-weighted images often decrease the contrast between grey and white matter and hence lead to bad performance in practice.

7.2. Segmentation Automaticity and User Intervention. All five packages support highly automatic brain segmentation, with little or no user intervention. FreeSurfer allows the user to start the cortical surface reconstruction without any intervention. In case the segmentation is not satisfactory, FreeSurfer supports interactive tools to allow the user to modify the brain mask and to add control points to improve the intensity normalization of WM, a procedure that is extremely important for the performance of FreeSurfer's whole brain segmentation. FreeSurfer also supports interactive tools for editing the final results generated by the automatic processing.

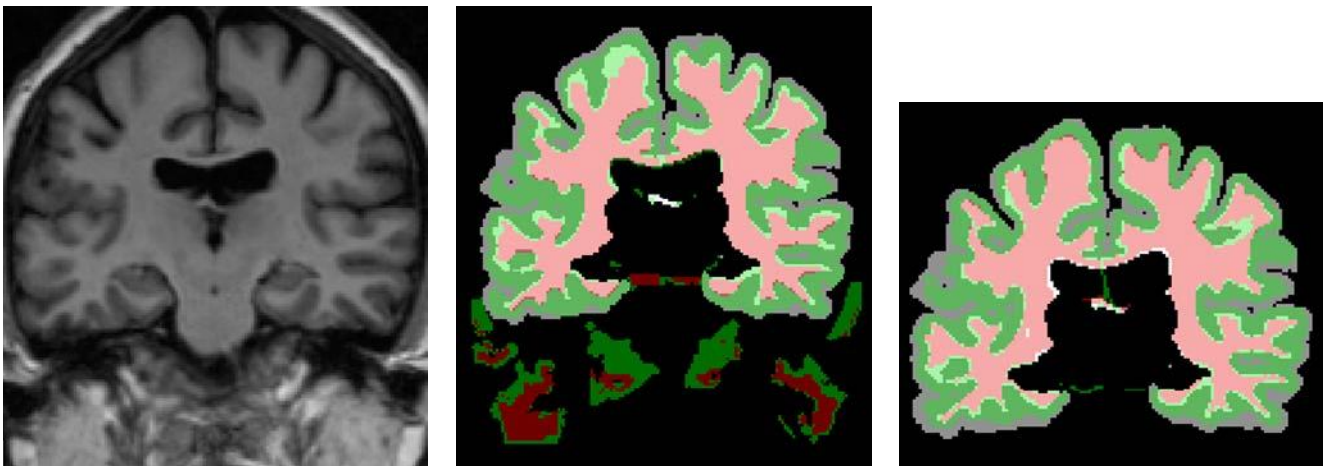
FSL also allows the user to start the segmentation without any intervention. In FSL, brain extraction and tissue classification are performed, respectively, by the BET and FAST algorithms. The BET performance substantially influences that of FAST. If the user is not satisfied with the brain extraction, FSL allows the user to select different parameters and rerun BET. However, our experiments with BET on the IBSR datasets and the pathological datasets show that BET cannot guarantee good brain extraction even with user intervention. FAST has custom options for the user to select whether to use the k-means segmentation or *a priori* probability maps for initial segmentation and to guide the k-means segmentation with manual intervention.

In SPM, brain segmentation can also be automatically started with the default parameters and SPM often generates good results. An important custom parameter of SPM is the one that controls the extent of bias field regularization. When any parameter is changed, the segmentation procedure has to be started over from scratch.

BrainVisa requires the user to prepare the data by first specifying several landmark points including the AC point, the PC point, an interhemispheric point, and a left hemisphere point. When the data is prepared by the user, BrainVisa automatically performs segmentation. BrainVisa supports interactive tools for the user to edit the segmentation results.

BrainK performs segmentation fully automatically, but the user also has the opportunity to tune the performance by changing key parameters such as the relative thresholds. Compared to the user intervention mechanisms in the other four packages, user intervention in BrainK is in the form of global parameter selection and has the following advantages. First, it is straightforward and requires little or no expertise to understand the meaning of the parameters and the criterion for selecting optimal ones. Second, it is very easy to operate by sliding a value bar. Third, it is very efficient and the user can obtain the effect of parameter selection in real time for such parameters as relative thresholds. Fourth, the parameters have global effect for segmentation and the user does not need to repeat similar operations for different local regions.

7.3. Segmentation Abnormalities. In examining the quality of the results, we first attempted to understand why BrainK consistently achieved better performance than the other packages



(a) MR image (b) BrainVisa result (c) FreeSurfer result

(d) FSL result (e) SPM result (f) BrainK result

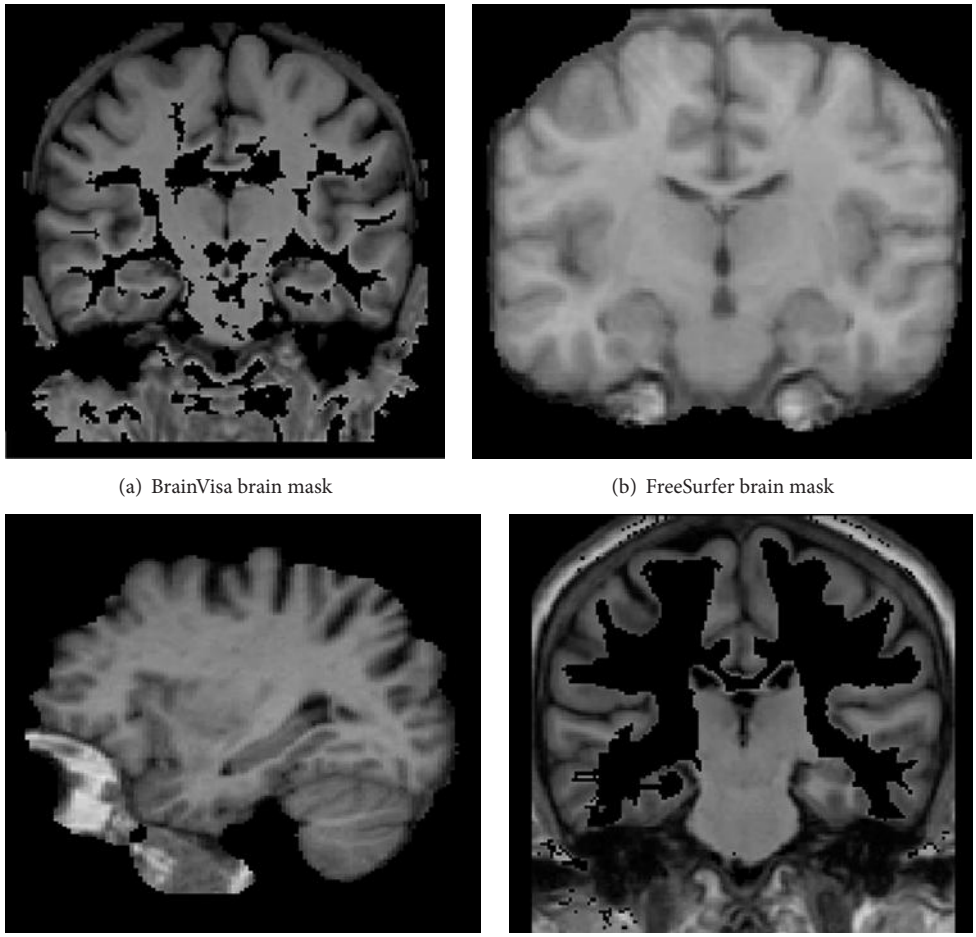
Figure 9: GM-shrinking phenomenon.

with respect to cerebral cortex segmentation on the 13 good IBSR datasets [48]. We found that there was a "shrinking" effect on the cerebral cortex segmentation for FreeSurfer, FSL, SPM, and BrainVisa, which gives rise to a significant number of false negatives. This problem did not occur or was much milder in BrainK. We think the underlying reason is that BrainK uses a new image modeling mechanism that can adapt to wider variations of GM intensities, whereas the statistical methods used in other packages were misled by these intensity variations, resulting in missing many GM voxels that had lower intensities.

The contrasting cortex segmentation of the five packages is shown for a representative IBSR dataset in Figure 9. Pink stands for correct WM, red for false WM negative, light green for correct GM, very light green for false WM negative and false GM negative, dark green for false GM positive, and gray false for GM negative. Note that, for SPM, light green represents correct GM segmentation, darker green represents false GM negative, and gray level represents false GM positive.

A common problem in FSL, BrainVisa, and SPM is the poor brain extraction. Some examples are shown in Figure 10. Poor generation of the brain mask was the primary factor in the very poor performance for SPM and BrainVisa on the IBSR datasets. Since we used the considerably better brain masks generated by FreeSurfer for the FSL analysis, the FSL analysis is not representative. FreeSurfer did not encounter poor skull stripping, but an inaccurate brain mask may still be generated, as shown in Figure 11(b), where some nonbrain voxels with high intensities are taken as brain tissues. BrainK, on the other hand, does not depend on a brain extraction preprocessing step and robustly generated clean cerebrum masks as the union of the cerebral white matter and the cerebral gray matter on all tested datasets.

We also found some other interesting abnormalities with FreeSurfer, as shown in Figure 11. For example, in Figure 11(c), FreeSurfer cut off a significant amount of cerebral WM and cortex at the top of the brain. In Figure 11(d), FreeSurfer was unable to segment the complete lateral ventricle of the subject with Alzheimer's disease. In Figure 12, FreeSurfer generated poor GM/WM segmentation even for a simulated image with excellent quality (noise level is 3% and IIH level is 20%) while BrainK and FSL generated excellent results. These abnormalities, we believe, are due to the overregularization of the *a priori* probability maps used in FreeSurfer.



(a) BrainVisa brain mask (b) FreeSurfer brain mask

(c) FSL brain mask (d) SPM GM mask

Figure 10: Poor brain masks.

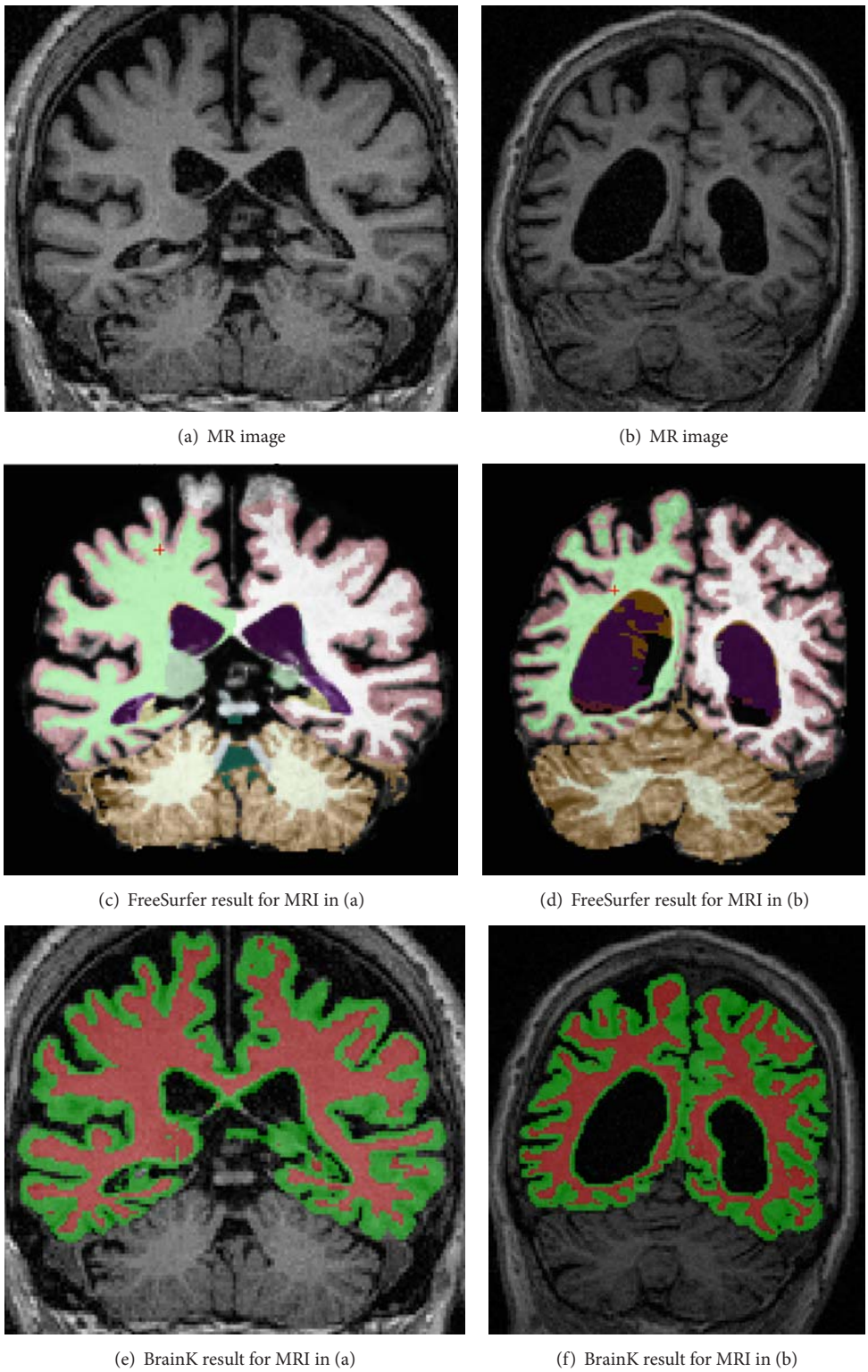
8. Applications

At this point, BrainK has built functional electrical head models for more than 200 persons. Critical evaluation of BrainK's performance for generating segmented volumes and cortical surfaces continues in the applications of the *Modal Image Pipeline*, EGI's commercial software that incorporates the BrainK algorithms and workflows. Validation tests of physical tissue boundaries are conducted in the routine workflow for each head model construction. A first step in the FDM electrical lead field generation [49, 50], EGI's head physics code, is to test whether there are physically realistic segmentation properties in the BrainK output required for accurate electrical modeling. For example, there must be at least one voxel of CSF between the outer gray matter of the cortex and the inner table of the skull.

There have been many applications of BrainK in research and development projects that have tested its capabilities and limitations. For example, although the skull Atlases are organized by age and gender in a database to minimize the warping required to fit an individual's MRI, tests have been conducted with skull CTs and MRIs of divergent shape (Figure 13), with good results.

Electrical source analysis with children requires age-appropriate electrical head models. Infants particularly require unique models because of their small head size and highly conductive skulls. BrainK has proven effective in adapting to the unique challenges of processing infant MRIs and CTs. Figure 14 shows results from a current project to create pediatric electrical head models directed by Sergei Turovets at EGI and by Linda Larsen-Prior and Fred Prior at University of Washington St. Louis. Figure 14(a) shows a normal infant's segmented MRI, with that infant's CT, registered to the segmented MRI, showing the thin skull and fontanel anterior to the vertex. Figure 14(b) shows the extracted cortical surface, and Figure 14(c) shows the dipole tessellated cortical surface, prepared for electrical source analysis.

The FDM electrical head models created by BrainK allow accurate estimation of the delivery of current, as well as its measurement, from the dense electrode array. With BrainK's anatomical results providing the geometry for the head physics code (Poisson equations for electromagnetic propagation in a finite difference model or FDM), lead fields are generated in the Modal Image Pipeline head physics code, describing the propagation from cortical sources to head



(a) MR image (b) MR image

(c) FreeSurfer result for MRI in (a) (d) FreeSurfer result for MRI in (b)

(e) BrainK result for MRI in (a) (f) BrainK result for MRI in (b)

Figure 11: FreeSurfer abnormalities on the pathological datasets compared to BrainK.

(a) BrainWeb MRI: N = 3% and IIH = 20%

(b) FreeSurfer result

(c) FSL result

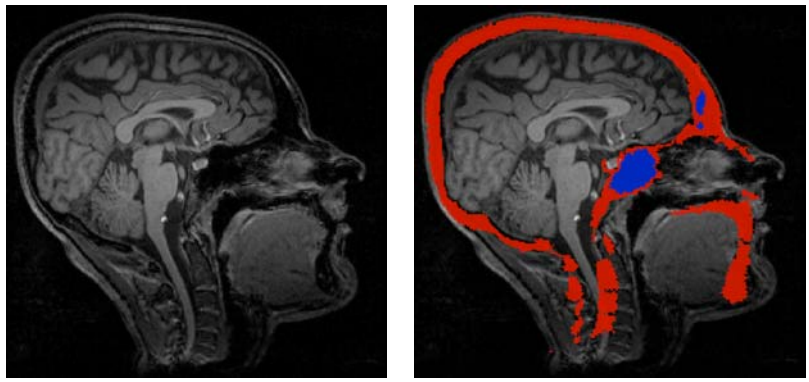
(d) BrainK result

Figure 12: BrainK segmentation on the BrainWeb MRI compared to FreeSurfer and FSL.

(a)

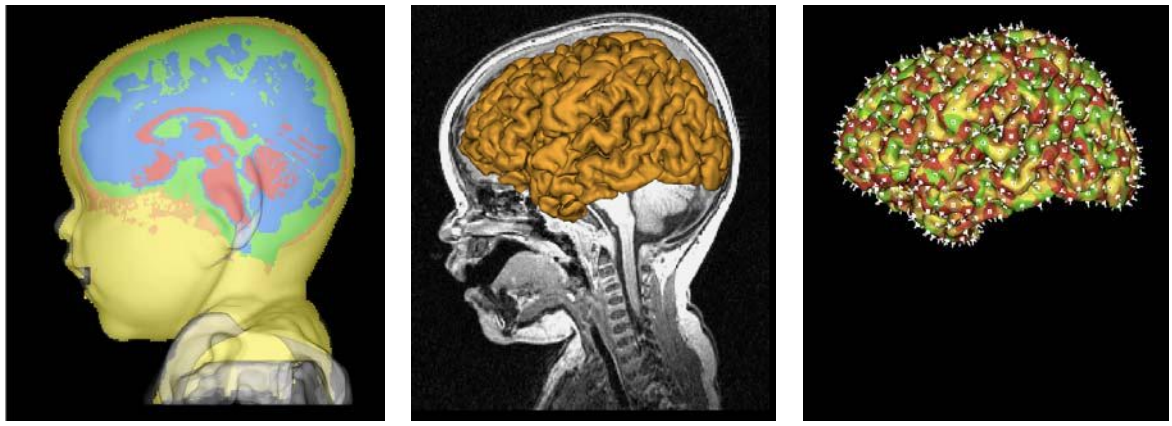
(b)

Figure 13: Continued.



(c) (d)

Figure 13: Skull Atlas-to-MRI registration with extensive warping of the skull to fit. (a) A volunteer in a 256 Net. (c) This person's MRI. (b) A database Atlas skull with a different shape. (d) Warp of the Atlas database skull to the volunteer's MRI.



(a) (b) (c)

Figure 14: Electrical head model for an infant. (a) Registration of skull from CT with segmented MRI. (b) Cortical surface extraction overlaid with MRI. (c) Dipole tessellation of cortical surface for electrical source analysis.

surface electrodes (for EEG analysis) and from head surface electrodes to cortical sources (for transcranial electrical neuromodulation). The results are visualized on BrainK's anatomy and registered electrode positions in the reciprocity engine developed by Erik Anderson of EGI (Figure 15). In this illustration, cortical sources were chosen in the approximate region of auditory cortex, with the equivalent dipole for each cortical patch visualized with a gold arrow (seen best in Figure 15(b)). The lead field computation with the head physics model then shows the resulting head surface fields in the color of both the scalp surface and the electrodes (red is positive; blue is negative). Given this generated surface field, the reciprocity theorem states that impressing current on those cortical sources can be accomplished by applying current in the same surface pattern (red for anodal and blue for cathodal). The lead field from the head physics code is again used to estimate the current density that is impressed on the cortex by electrical neuromodulation with this electrode pattern (Figures 15(c) and 15(d)). The accuracy of the electrical solutions is fully dependent on the detailed geometry provided by BrainK, including not only the accurate (<1 mm) registration of sensors to the head surface, but also the specification of the orientation of the cortical patches provided by BrainK's characterization of the cortical surface.

9. Conclusion

BrainK is a software package that implements certain tissue segmentation and surface extraction algorithms that are useful in scientific and medical image processing. It also applies certain algorithms, such as skull image registration and cortical surface dipole tessellation, which are specifically important to model the electrical properties of the human head. Accurate electrical head models are proving important for improving the source analysis of human electrophysiological activity with dense array technologies and for optimizing transcranial current delivery to modulate brain activity.

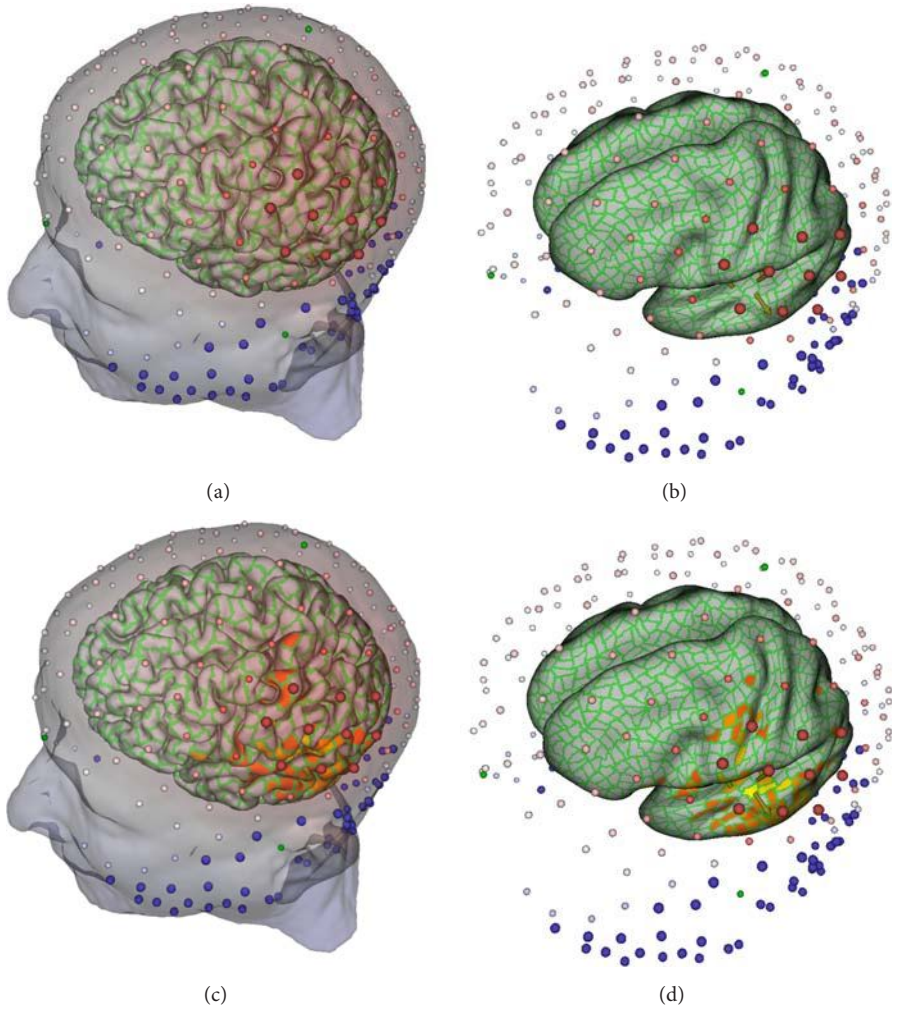


Figure 15: ((a), (b)) Using BrainK's anatomy in the reciprocity visualization environment software to show the electrical field at the head surface (positive potential at red electrodes and negative potential at blue electrodes) created by electrical activity in auditory cortex (dipoles shown as gold arrows seen best in the inflated cortex at (b)). ((c), (d)) Through reciprocity theory, the field pattern of the cortical sources can be used to estimate the optimal electrodes (sources: red, sinks: blue) to impress electrical current on the auditory cortex for neuromodulation. Here, the impressed cortical current density is imaged with a hot metal palette (gold, orange, and yellow), thresholded at 50% of maximum.

Competing Interests

Kai Li and Don M. Tucker are employees of EGI, a medical device company that is developing medical software from the methods described here. EGI and the University of Oregon hold a US patent (no. 8,478,011) on the relative thresholding method. Xenophon Papademetris is a paid consultant for EGI.

Acknowledgments

The first author would like to thank his Ph.D. advisors Allen Malony and Dr. Don Tucker for their long and continuing support in the procedure of the research work, Dr. Jasmine Song for her valuable help in revising the paper, and Dr. Phan Luu, Dr. Sergei Turovets, Dr. David Hammond, Chelsea Mattson, Colin Davey, Ryan Jayne, and Kyle Morgan for their constructive comments, advice, and requests provided in the process of using and testing the BrainK software. Xenophon Papademetris worked on this project as a paid consultant of Electrical Geodesics, Inc. The intellectual property of this software is protected by US Pat. no. 8,478,011.

References

[1] D. M. Tucker, "Spatial sampling of head electrical fields: the geodesic sensor net," *Electroencephalography and Clinical Neurophysiology*, vol. 87, no. 3, pp. 154–163, 1993.

[2] J. Song, D. M. Tucker, T. Gilbert et al., "Methods for examining electrophysiological coherence in epileptic networks," *Frontiers in Neurology*, vol. 4, article 55, 2013.

[3] D. L. Schomer and F. L. da Silva, Eds., *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, Lippincott Williams & Wilkins, 6th edition, 2011.

[4] G. S. Russell, K. J. Eriksen, P. Poolman, P. Luu, and D. M. Tucker, "Geodesic photogrammetry for localizing sensor positions in dense-array EEG," *Clinical Neurophysiology*, vol. 116, no. 5, pp. 1130–1140, 2005.

[5] S. Gonçalves, J. C. de Munck, J. P. Verbunt, R. M. Heethaar, and F. H. da Silva, "In vivo measurement of the brain and skull resistivities using an EIT-based method and the combined analysis of SEF/SEP data," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 9, pp. 1124–1128, 2003.

[6] Ü. Aydin, J. Vorwerk, P. Küpper et al., "Combining EEG and MEG for the reconstruction of epileptic activity using a calibrated realistic volume conductor model," *PLoS ONE*, vol. 9, no. 3, Article ID e93154, 2014.

[7] M. Dannhauer, R. MacLeod, S. Guler et al., "Comparison of patch electrodes and high density electrode arrays regarding volume conduction in tdcs using a realistic head model," in *Proceedings of the 19th Annual Meeting of the Organization for Human Brain Mapping (OHBM '13)*, Seattle, Wash, USA, June 2013.

[8] S. Guler, M. Dannhauer, R. MacLeod et al., "Dense electrode array current optimization for targeting and directionality of tDCS: studies in a realistic head model," in *Proceedings of the Organization for Human Brain Mapping Meeting (OHBM '13)*, Seattle, Wash, USA, June 2013.

[9] F. Klauschen, A. Goldman, V. Barra, A. Meyer-Lindenberg, and A. Lundervold, "Evaluation of automated brain MR image segmentation and volumetry methods," *Human Brain Mapping*, vol. 30, no. 4, pp. 1310–1327, 2009.

[10] S. M. Smith, M. Jenkinson, M. W. Woolrich et al., "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage*, vol. 23, pp. S208–S219, 2004.

[11] A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical surface-based analysis: I. Segmentation and surface reconstruction," *Neuroimage*, vol. 9, no. 2, pp. 179–194, 1999.

[12] B. Fischl, M. I. Sereno, and A. M. Dale, "Cortical surface-based analysis. II: inflation, flattening, and a surface-based coordinate system," *NeuroImage*, vol. 9, no. 2, pp. 195–207, 1999.

[13] J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage*, vol. 26, no. 3, pp. 839–851, 2005.

[14] S. Turovets, A. Salman, A. Malony, P. Poolman, C. Davey, and D. M. Tucker, "Anatomically constrained conductivity estimation of the human head tissues in vivo: computational procedure and preliminary experiments," in *Proceedings of the 7th Conference on Biomedical Applications of Electrical Impedance Tomography*, Seoul, South Korea, 2006.

[15] K. Li, A. D. Malony, and D. M. Tucker, "Automatic brain MR image segmentation by relative thresholding and morphological image analysis," in *Proceedings of the 1st International Conference on Computer Vision Theory and Applications (VISAPP '06)*, pp. 354–364, Setubal, Portugal, 2006.

[16] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press and McGraw-Hill, 2nd edition, 2001.

[17] J. Sierpowska, M. A. Hakulinen, J. Töyräs et al., "Interrelationships between electrical properties and microstructure of human trabecular bone," *Physics in Medicine and Biology*, vol. 51, no. 20, pp. 5289–5303, 2006.

[18] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[19] A. P. Zijdenbos and B. M. Dawant, "Brain segmentation and white matter lesion detection in MR images," *Critical Reviews in Biomedical Engineering*, vol. 22, no. 5-6, pp. 401–465, 1994.

[20] G. Bertrand, "Simple points, topological numbers and geodesic neighborhoods in cubic grids," *Pattern Recognition Letters*, vol. 15, no. 10, pp. 1003–1011, 1994.

[21] K. Li, A. D. Malony, and D. M. Tucker, "A multiscale morphological approach to topology correction of cortical surfaces," in *Medical Imaging and Augmented Reality: Third International Workshop, Shanghai, China, August 17–18, 2006 Proceedings*, vol. 4091 of *Lecture Notes in Computer Science*, pp. 52–59, Springer, Berlin, Germany, 2006.

[22] J.-D. Boissonnat and M. Yvinec, *Algorithmic Geometry*, Cambridge University Press, 1998.

[23] M. H. Davis, A. Khotanzad, D. P. Flamig, and S. E. Harms, "A physics-based coordinate transformation for 3-D image matching," *IEEE Transactions on Medical Imaging*, vol. 16, no. 3, pp. 317–328, 1997.

[24] B. K. Natarajan, "On generating topologically consistent isosurfaces from uniform samples," *The Visual Computer*, vol. 11, no. 1, pp. 52–62, 1994.

[25] B. L. Chamberlain, "Graph partitioning algorithms for distributing workloads of parallel computations," Tech. Rep. TR-98-10-03, Department of Computer Science & Engineering, University of Washington, 1998.

[26] B. Hendrickson and R. Leland, "The Chaco users guide: version 2.0," Tech. Rep. SAND 94-2692, Sandia National Laboratories, Albuquerque, NM, USA, 1994.

[27] J. Talairach and P. Tournoux, *Co-Planar Stereotaxic Atlas of the Human Brain*, Thieme, New York, NY, USA, 1988.

[28] J. L. Lancaster, M. G. Woldorff, L. M. Parsons et al., "Automated Talairach Atlas labels for functional brain mapping," *Human Brain Mapping*, vol. 10, no. 3, pp. 120–131, 2000.

[29] B. Fischl, D. Salat, E. Busa et al., "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.

[30] Freesurfer, https://surfer.nmr.mgh.harvard.edu.

[31] SPM, http://www.fil.ion.ucl.ac.uk/spm.

[32] J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage*, vol. 26, no. 3, pp. 839–851, 2005.

[33] FSL, http://www.fmrib.ox.ac.uk/fsl.

[34] S. M. Smith, "Fast robust automated brain extraction," *Human Brain Mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[35] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[36] Y. Cointepas, J.-F. Mangin, L. Garnero, J.-B. Poline, and H. Benali, "BrainVISA: software platform for visualization and analysis of multi-modality brain data," in *Proceedings of the 7th Human Brain Mapping Conference*, p. S98, Brighton, UK, 2001.

[37] BrainVisa, http://brainvisa.info.

[38] J.-F. Mangin, "Entropy minimization for automatic correction of intensity nonuniformity," in *Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA '00)*, pp. 162–169, Hilton Head Island, SC, USA, June 2000.

[39] J.-F. Mangin, O. Coulon, and V. Frouin, "Robust brain segmentation using histogram scale-space analysis and mathematical morphology," in *Proceedings of the Medical Image Computing*

and Computer-Assisted Interventation (MICCAI '98), W. M. Wells, A. Colchester, and S. Delp, Eds., pp. 1230–1241, MIT, Boston, Mass, USA, 1998.

[40] J.-F. Mangin, J. Régis, and V. Frouin, "Shape bottlenecks and conservative flow systems," in *Proceedings of the IEEE/SIAM Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA '96)*, pp. 319–328, San Francisco, Calif, USA, 1996.

[41] J.-F. Mangin, O. Coulon, and V. Frouin, "Robust brain segmentation using histogram scale-space analysis and mathematical morphology," in *Medical Image Computing and Computer-Assisted Interventation—MICCAI '98*, W. M. Wells, A. Colchester, and S. Delp, Eds., vol. 1496 of *Lecture Notes in Computer Science*, pp. 1230–1241, 1998.

[42] C. Cocosco, V. Kollokian, R.-S. Kwan, and A. Evans, "BrainWeb: online interface to a 3D MRI simulated brain database," *NeuroImage*, vol. 5, no. 4, p. S425, 1997.

[43] BrainWeb, http://www.bic.mni.mcgill.ca/brainweb.

[44] CMA: The Center for Morphometric Analysis at Massachusetts General Hospital, http://www.cma.mgh.harvard.edu/ibsr/.

[45] NRU: The Neurobiology Research Unit, http://nru.dk.

[46] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy, "Magnetic resonance image tissue classification using a partial volume model," *NeuroImage*, vol. 13, no. 5, pp. 856–876, 2001.

[47] K. van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classification of MR images of the brain," *IEEE Transactions on Medical Imaging*, vol. 18, no. 10, pp. 897–908, 1999.

[48] K. Li, *Neuroanatomical segmentation in MRI exploiting a priori knowledge [Ph.D. thesis]*, University of Oregon, 2007.

[49] A. Salman, S. Turovets, A. Malony, J. Eriksen, and D. M. Tucker, "Computational modeling of human head conductivity," in *Proceedings of the 5th International Conference on Computational Science (ICCS '05)*, Atlanta, Ga, USA, May 2005.

[50] A. Salman, S. Turovets, A. Malony et al., "Noninvasive conductivity extraction for high-resolution eeg source localization," *Advances in Clinical Neuroscience and Rehabilitation*, vol. 26, no. 1, pp. 27–28, 2005.

Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2016, Article ID 3979547, 11 pages
http://dx.doi.org/10.1155/2016/3979547

Research Article

MEG Connectivity and Power Detections with Minimum Norm Estimates Require Different Regularization Parameters

Ana-Sofía Hincapié,[1,2,3,4] Jan Kujala,[4,5] Jérémie Mattout,[4] Sebastien Daligault,[6] Claude Delpuech,[6] Domingo Mery,[2] Diego Cosmelli,[3] and Karim Jerbi[1,4]

[1]*Psychology Department, University of Montreal, Montreal, QC, Canada H2V 2S9*
[2]*Department of Computer Science, Pontificia Universidad Católica de Chile, 7820436 Santiago de Chile, Chile*
[3]*School of Psychology and Interdisciplinary Center for Neurosciences, Pontificia Universidad Católica de Chile, 7820436 Santiago de Chile, Chile*
[4]*Lyon Neuroscience Research Center, DyCog Team, Inserm U1028, CNRS UMR5292, 69675 Bron Cedex, France*
[5]*Department of Neuroscience and Biomedical Engineering, Aalto University, 02150 Espoo, Finland*
[6]*MEG Center, CERMEP, 69675 Bron Cedex, France*

Correspondence should be addressed to Karim Jerbi; karim.jerbi@umontreal.ca

Received 15 October 2015; Revised 19 January 2016; Accepted 14 February 2016

Academic Editor: Jussi Tohka

Minimum Norm Estimation (MNE) is an inverse solution method widely used to reconstruct the source time series that underlie magnetoencephalography (MEG) data. MNE addresses the ill-posed nature of MEG source estimation through regularization (e.g., Tikhonov regularization). Selecting the best regularization parameter is a critical step. Generally, once set, it is common practice to keep the same coefficient throughout a study. However, it is yet to be known whether the optimal lambda for spectral power analysis of MEG source data coincides with the optimal regularization for source-level oscillatory coupling analysis. We addressed this question via extensive Monte-Carlo simulations of MEG data, where we generated 21,600 configurations of pairs of coupled sources with varying sizes, signal-to-noise ratio (SNR), and coupling strengths. Then, we searched for the Tikhonov regularization coefficients (lambda) that maximize detection performance for (a) power and (b) coherence. For coherence, the optimal lambda was two orders of magnitude smaller than the best lambda for power. Moreover, we found that the spatial extent of the interacting sources and SNR, but not the extent of coupling, were the main parameters affecting the best choice for lambda. Our findings suggest using less regularization when measuring oscillatory coupling compared to power estimation.

1. Introduction

Healthy brain function is largely mediated by coordinated interactions between neural assemblies in different cortical and subcortical structures. As a result, the study of neuronal processes requires techniques that can reliably measure the spatiotemporal dynamics of large-scale networks. To this end, it is important to assess the modulations of local activations as well as long-range coupling between brain areas. Over recent years, the quantification of neuronal interactions in a given behavioral task or brain state has been the focus of a large body of research, and various techniques are currently used to detect and probe the role of functional connectivity (e.g., [1, 2]). In particular, a widely used technique for the detection of large-scale interactions among neural assemblies is magnetoencephalography (MEG) [3–5]. This is primarily due to its high temporal resolution, which is in the same order of magnitude as the neuronal processes themselves (milliseconds). Unlike electroencephalography (EEG), MEG measures magnetic fields that are less affected by the skull and brain tissue and provides whole-head coverage, which is mandatory for the assessment of large-scale brain networks. The use of MEG has advanced our comprehension of the mechanisms underlying functional brain connectivity [6] involved in sensory, motor, and higher-order cognitive tasks [3, 7–11] higher-order cognitive tasks resting state [6, 12–18].

The ability to measure brain interactions at the source level, rather than between sensor channels, is an important prerequisite if we want to make useful inferences about the anatomical-functional properties of the network. To this end, source reconstruction techniques are applied in order to estimate the spatiotemporal activity of the cortical generators underlying the recorded sensor-level MEG data. Solving this relationship is an ill-posed inverse problem for which numerous methods have been developed [4]. The differences between the various techniques mainly stem from the assumptions they make about the properties of the neural sources and from the way they incorporate various forms of a priori information, if any is available. Conceptually, solving the MEG inverse problem boils down to solving a system of equations that is underdetermined (i.e., no unique solution) since we generally have far more sources (thousands) than measurements (hundreds). Identifying a solution to such a problem can be achieved by imposing constraints on the sources. A typical constraint is to minimize the source power. Among these methods, the Minimum Norm Estimate (MNE) relies on minimizing the L2-norm and is one of the most widely used techniques [4, 7, 8, 18–37]. By contrast, estimates obtained by minimizing the L1-norm are referred to as Minimum-Current Estimates (MCE) [34, 38]. While the L2-norm assumes a Gaussian a priori current distribution, the L1-norm assumes a Laplacian distribution [39].

In principle, MNE looks for a distribution of sources with the minimum (L2-norm) current that can give the best account of the measured data. As the problem is ill-posed, MNE generally uses a regularization procedure that sets the balance between fitting the measured data (minimizing the residual) and minimizing the contributions of noise [22, 40, 41]. The Tikhonov or Wiener regularization [42, 43] and SVD truncation are among the most widely used regularization procedures. Regularization may be considered a necessary evil: it is required to stabilize the solution of the inverse problem, yet too much regularization leads to overly smooth solutions (spatial smearing). Since we do not have a precise model of brain activity and noise, the choice of the optimal amount of regularization is a nontrivial step for which no magical recipe exists [44]. The relationship between optimal regularization and the patterns of underlying generators is still poorly understood. In particular, the effect of regularization on the detection accuracy of Minimum Norm Estimates of source power and interareal source coupling has not been thoroughly investigated. This raises the question of whether one should use the same regularization coefficient for MNE-based power and connectivity analyses, as is generally done, or would one benefit from optimizing the regularization coefficients separately for each analysis? Furthermore, how does the optimal regularization coefficient, in such configurations, depend on sensor-level SNR, source size, or coupling strength?

With these questions in mind, we performed extensive Monte-Carlo simulations, creating over 20,000 pairs of coupled oscillatory time series in MEG source spaces, and computed the resulting surface MEG recordings. Then, we estimated source power and coherence using the MNE framework with variable degrees of Tikhonov regularization. Moreover, we used an approach based on an area under the curve (AUC) to identify the optimal regularization coefficient (lambda) in the case of power and coherence analyses. We found a systematic difference between the optimal lambdas in each analysis: for source-level coherence analysis the optimal lambda was two orders of magnitude smaller than the best lambda for power detection. Lastly, our findings are broken down as a function of SNR, cortical patch size, and corticocortical coupling strength.

2. Methods and Materials

In this section, we first present the MEG inverse problem formulation and the MNE framework. Then, we describe the simulation, reconstruction, and performance assessment procedures.

2.1. The MEG Inverse Problem. The inverse problem looks for an estimation of the active sources $\mathbf{S}$ ($3n_{\text{sources}} \times$ time points) that generates the measurements $\mathbf{M}$ ($n_{\text{channels}} \times$ time points) recorded at the sensors. The dimension $3n$ of the columns of $\mathbf{S}$ accounts for the three components of the source n in the x, y, and z directions. According to anatomical observations, the main generators of MEG are located in the grey matter and their orientation is perpendicular to the cortical sheet [45]. Here, we used a constrained orientation approach (perpendicular to the cortical surface), and hence the dimensions of $\mathbf{S}$ are reduced to $n_{\text{sources}} \times$ time points. Assuming a linear relationship between the measurements and the active sources, the problem is modeled as

$$\mathbf{M} = \mathbf{LS} + \mathbf{N}, \tag{1}$$

where $\mathbf{L}$ is the lead field matrix ($n_{\text{channels}} \times n_{\text{sources}}$) and $\mathbf{N}$ is additive noise applied at the MEG channels ($n_{\text{sources}} \times$ time points). The lead field matrix describes how each source contributes to the measurements at each sensor, given a specific head conductivity model and a source space. As the number of sources is usually much higher than the number of sensors, the lead field matrix is highly underdetermined and thus not invertible. The estimation of the activity of the sources requires the definition of an inverse operator $\mathbf{W}$:

$$\widehat{\mathbf{S}} = \mathbf{W}^T\mathbf{M}, \tag{2}$$

where $\widehat{\mathbf{S}}$ represents the estimated sources ($n_{\text{sources}} \times$ time points) and the superscript T denotes matrix transpose.

2.2. Minimum Norm Estimate (MNE). As the MEG inverse problem is ill-posed, a regularization scheme is needed [22], and one of the most common options is the Tikhonov regularization [42, 43]. MNE calculates an inverse operator $\mathbf{W}$ by searching for a distribution of sources $\widehat{\mathbf{S}}$ with minimum currents (L2-norm) that produces an estimation of the measurements ($\mathbf{L}\widehat{\mathbf{S}}$) most consistent with the measured data ($\mathbf{M}$). The solution is a trade-off between the norm of the estimated regularized sources current $\lambda^2\|\widehat{\mathbf{S}}\|^2$ and the norm of the quality of the fit they provide to the measurements $\|\mathbf{M} - \mathbf{L}\widehat{\mathbf{S}}\|^2$. Assuming the noise $\mathbf{N}$ and the sources strength

S to be normally distributed with zero mean and covariance matrix **Q** and **R**, respectively, a general form of the MNE inverse solution can thus be given as [35, 46, 47]

$$\widehat{\mathbf{S}} = \underset{\mathbf{S}}{\operatorname{argmin}} \left\{ \left\| \mathbf{Q}^{-1/2} (\mathbf{M} - \mathbf{LS}) \right\|^2 + \lambda^2 \left\| \mathbf{R}^{-1/2} \mathbf{S} \right\|^2 \right\}, \quad (3)$$

where λ is the Tikhonov regularization parameter. Thus, the inverse operator **W** is defined as

$$\mathbf{W} = \mathbf{R}\mathbf{L}^T \left(\mathbf{L}\mathbf{R}\mathbf{L}^T + \lambda^2 \mathbf{Q} \right)^{-1}, \quad (4)$$

where the superscript −1 denotes matrix inverse.

The Minimum Norm Estimate embodies the assumption of independently and identically distributed (IID) sources, which corresponds to an identity matrix **R** in the above formula. Alternatively, **R** can incorporate more informed (spatial) assumptions, yielding a so-called weighted Minimum Norm Estimate [28, 35, 44]. Here, we use the general and classical minimum norm solution. The noise covariance matrix **Q** was computed from the actual noise which was added to the sensors in each simulation.

2.3. Simulations. We simulated 117s of oscillatory activity in pairs of cortical sources with different degrees of coupling in the alpha band (9–14 Hz). We computed the resulting sensor-level data through forward modeling based on a 275-channel CTF MEG system configuration. Our sources consisted of current dipoles placed at the vertices of a tessellated MNI-Colin 27 cortical surface, which was segmented and tessellated using FreeSurfer [48] and downsampled to 15028 vertices. Different strengths of coupling were obtained by forcing the time series of the second source to have a certain level of coherence with the first source time series. The magnetic fields at the sensors were calculated using a single sphere head model and constraining the orientations of the sources to be normal to the cortical surface. The simulated data were generated using a combination of custom MATLAB code, in addition to functions from Brainstorm [49] and FieldTrip [50] toolboxes. The source time courses were simulated by first setting the base frequency of the oscillator (e.g., 12 Hz) and inducing a small jitter to its instantaneous frequency across time points. The frequency modulated time courses were then generated using an exponential function. This procedure causes fluctuations in the phase relationship between two oscillators with the same base frequency, allowing us to achieve coherence levels below 1. The frequency jittering was performed randomly in a loop until the desired coherence level (e.g., 0.4) between the time courses of the two oscillators was reached.

We randomly selected two locations (seeds) for each simulated pair of sources (600 pairs). Next, for each pair, we varied three additional parameters in the simulations: the spatial extent of the sources, the strength of the coupling between the two sources, and the signal-to-noise ratio (SNR) at the sensors. These parameters are described in more detail below.

(i) Patches and Point-Like Sources. We simulated point-like sources (i.e., 1 dipole) and cortical patches (with surface areas: 2, 4, or 8 cm^2). The activity of a cortical patch was simulated by placing identical time series at the vertices that make up the patch. As described above, the vertices were obtained via tessellation of the MNI brain (Colin 27).

(ii) Coupling Strength. When generating the time series for each of the two sources of a pair of simulated generators, we defined the coupling by setting the alpha-band coherence to either 0.1, 0.2, or 0.4. We simulated time series that were 7000 samples long (600 Hz sampling frequency). Note here that by actually simulating true coherence at the source level, we circumvent debates about spurious coupling that arises from field spread. As such, we assess here the ability to recover truly coherent cortical activity.

(iii) Signal-to-Noise Ratio (SNR). White noise was added to the sensor signals in order to achieve three levels of SNR (0 dB, −20 dB, and −40 dB), calculated as the ratio of the Frobenius norm of signal and noise amplitudes at the sensors.

In summary, we randomly chose 600 different pairs of cortical location configurations, for which we varied source size (4 levels: point-like, 2, 4, or 8 cm^2), coupling strength (3 levels: 0.1, 0.2, or 0.4), and SNR (3 levels: 0 dB, −20 dB, and −40 dB). This yielded a total of 21,600 sets of simulated MEG sensor-level data. Next, we evaluated the effect of varying the Tikhonov regularization parameter λ on detection performance of MNE, independently, for power and for coherence mapping.

2.4. Power and Coherence Reconstructions. The previous section described the cortical power and coherence configurations that were generated in the MEG data simulation step. Now, in order to assess our ability to reconstruct these simulated (i.e., known) source configurations, we first reconstruct the source time series by applying MNE to the simulated sensor data, and then we compute spectral power and coherence from these estimated time series.

Most MEG studies that use MNE select a single lambda value once and for all to be used throughout the study; the same regularization coefficient is hence used for both power and coupling estimations (assuming both are performed within the study). This is not the case here, precisely because our objective is to test whether optimal lambda values differ between power and coherence mapping.

Therefore, in order to find the regularization coefficient providing the most accurate reconstruction of (a) power and (b) interareal coherence across all 21,600 simulated configurations, we used multiple values for λ ranging from $1e-11$ to $1e-5$ (7 values). In addition to searching for the optimal lambda that provides the best results across all simulations, the definition of best lambda was also separately examined as a function of SNR, coupling strength, and source extent. Finally, for the sake of comparison, we also computed the optimal λ value obtained from the standard L-curve method [51]. The range of possible lambda values examined in our estimations was selected based on visual inspection in a subset of simulations and chosen to ensure it included the lambda obtained via the L-curve approach.

Notably, for the coupling analysis, we used the reconstructed time series at one of the two simulated dipoles as

reference point, and we calculated coherence with all other source time series across the brain. Spectral power (at any location r) and magnitude squared coherence (between two locations r_1 and r_2) were calculated as follows:

$$P(r,f) = \mathrm{Cs}(r,r,f),$$
$$\mathrm{Coh}(r_1,r_2,f) = \frac{|\mathrm{Cs}(r_1,r_2,f)|^2}{P(r_1,r_1,f)\,P(r_2,r_2,f)}, \quad (5)$$

where Cs is the cross-spectral density matrix and f is the frequency bin. Power and MS coherence were calculated via built-in standard MATLAB (Mathworks Inc., MA, USA) functions based on Welch's averaged and modified periodogram method [52].

2.5. Receiver Operating Characteristic (ROC) Curves. We evaluated the performance of each method using the area under the curve (AUC) from the ROC curves obtained by plotting the True Positive Fraction (TPF) versus the False Positive Fraction (FPF) at a given threshold α, calculated as follows:

$$\mathrm{TPF}(\alpha) = \frac{\mathrm{TP}(\alpha)}{\text{Number of simulated dipoles}},$$
$$\mathrm{FPF}(\alpha) = \frac{\mathrm{FP}(\alpha)}{\text{Total number of dipoles} - \text{Number of simulated dipoles}}, \quad (6)$$

where TP(α) represents the true positives (defined as the intersection between simulated sources and active sources at threshold α) and FP(α) represents the false positives (defined as all active sources but excluding the true positives at activation threshold α). By computing TPF and FPF repeatedly for successive values of activation thresholds α, we obtain a ROC curve from which we derive the AUC. The AUC is taken as a measure of performance; that is, the best regularization coefficient would be the one that yields the highest AUC. Statistical comparisons were performed using standard parametric two-tailed t-tests. Note that, for reference-based coherence reconstruction, computing true positives can be ambiguous if we consider the sources within the "reference patch" (location 1) to be true positives. To avoid this problem, we chose to quantify how well the distant coherent patch (location 2) was detected. In other words, we used the time series estimated at location 1 (as a seed) and considered only the simulated activity that make up the patch at location 2 to be the vertices we wish to detect. Therefore, the vertices within the reference patch were excluded from the ROC calculations for coherence evaluations.

3. Results

3.1. Optimal Tikhonov Regularization Coefficient Is Not the Same for Power and Coherence Reconstructions. Figure 1(a) shows the effect of lambda selection on quality of power and coherence detection across all simulated conditions and configurations (measured as mean AUC). Importantly, we found that the best mean value of lambda for power ($10e-7$) differs from the best mean lambda for coherence, which was two orders of magnitude smaller ($10e-9$). In comparison to the optimal value for power, the lower optimal value for coherence implies that a smaller residual should be allowed to have an optimal coherence reconstruction. The observations in Figure 1(a) also indicate that coherence reconstructions are more sensitive to the selection of an appropriate lambda value than power reconstructions (as AUC for coherence peaks at $1e{-}09$ and drops again, while AUC for power displays a flatter distribution for values neighboring $1e-07$).

Interestingly, Figure 1(b) shows a significant difference ($p < 0.001$, t-test) between the power reconstruction performances achieved using the optimal lambda for power compared to using the lambda determined to be optimal for coherence. Similarly, coherence reconstructions were also significantly better when using the coherence optimal ambda compared to the power optimal lambda. Simply put, our simulations demonstrate that, on average across 21,600 simulated pairs of sources, tuning lambda selection differently for coherence and for power analyses significantly impacts the results. In the next section, we examine how SNR, coupling strength, and source size individually affect these results.

3.2. Effect of SNR, Source Size, and Coupling Strength. Figure 2 depicts the effect of (a) SNR, (b) source size, and (c) coupling strength on optimal lambda. In each panel, the results are shown independently for power (upper row) and coherence (lower row).

Figure 2(a) shows that a decrease in SNR leads, as expected, to an overall decrease in the performance of MNE as measured by mean AUC. For power reconstructions, a peak (i.e., easily identifiable optimal lambda) seems to be limited to the higher SNR. As for coherence reconstructions, the lower row of Figure 2(a) indicates that stronger regularization is needed as SNR drops. Next, Figure 2(b) suggests that for both power and coherence detection, point-like sources require more (an order of magnitude higher) regularization than cortical patches. Furthermore, Figure 2(c) shows that stronger coupling leads to better detection performances (higher mean AUC). However, it also suggests that the optimal lambda values ($1e-07$ for power and $1e-09$ for coherence) do not vary with source coupling strength (Figure 2(c)).

3.3. Comparison with the L-Curve Method. A common and rather well-established heuristic to optimize λ in a data-driven manner is the L-curve. The L-curve is a plot of the norm of the regularization term versus the norm of the residuals, representing the trade-off between these two terms. The lambda with the best compromise between minimizing the norm of the current and that of the residual is chosen as the optimal lambda [51]. As it is a fast and widely used method, we decided to compare the optimal lambda given by the L-curve approach with the two average lambda values which we found to optimize either power or coherence.

The mean optimal lambda obtained with the L-curve was $1e-10$. Application of this Tikhonov coefficient led to suboptimal reconstructions of both power and coherence. For both cases, the mean AUC obtained was significantly

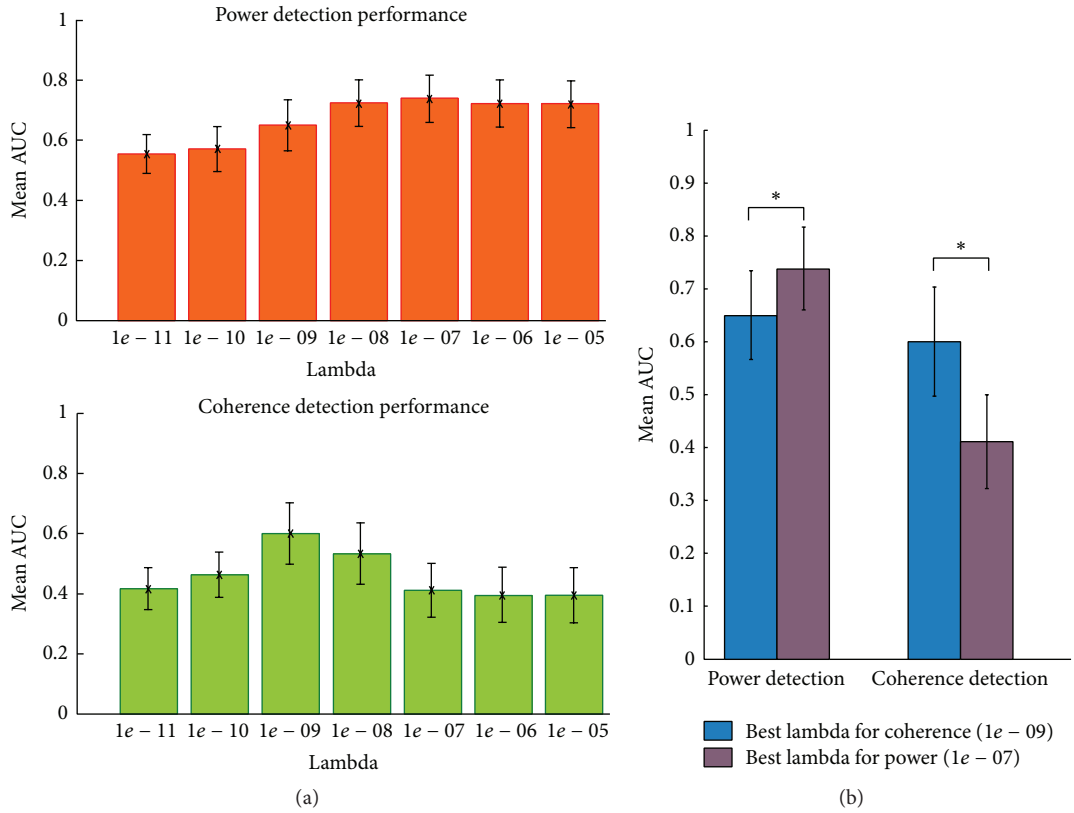


FIGURE 1: The averaged optimal regularization for power reconstruction is different from the one obtained for coherence detection. (a) Mean AUC for power and coherence reconstructions. (b) Mean AUC achieved with best lambda for power ($1e-07$) or best lambda for coherence ($1e-09$) when applied in each case both for power and for coherence. $*$ indicates statistically significant differences at $p < 0.001$, t-test.

smaller ($p < 0.001$, t-test) than the mean AUC observed with the optimal lambda values derived from the data (Figure 3(a)). In fact, the L-curve based estimation of optimal lambda turned out to be one and three orders of magnitude smaller than the optimal values we had found for coherence and power, respectively. Figures 3(b)–3(d) show an example of a simulated pair of sources (Figure 3(a)) and its reconstructions using the optimal lambdas found with the data-driven AUC-based optimization ($1e-7$ for power and $1e-9$ for coherence) and with the L-curve optimization ($1e-10$). This configuration illustrates a typical case where the regularization coefficient derived from the L-curve approach fails. It also illustrates how the application of regularization that is suitable for power detection ($1e-07$) can lead to very poor detection in the case of coherence, where less regularization ($1e-09$) provides acceptable detection. Note that the case shown here is based on an arbitrary selection of a source configuration from among 21600 simulations. Both poorer and nicer single examples can be found of course, but we chose to represent a realistic intermediate case that serves to illustrate the point made by our study.

4. Discussion

Overall, we have shown using extensive simulations of coupled pairs of sources that on average when using MNE, the best results are achieved by selecting separate regularization coefficients for power and for coupling estimations in source space. In particular, we found on average that the Tikhonov regularization coefficient that yields best coherence detection is substantially smaller than the optimal regularization coefficient for the detection of oscillatory power. This is largely due to the fact that increased smoothing (which arises from increased regularization) blows up the rate of false positives, a problem that appears to be amplified for corticocortical coupling.

Most MEG studies that apply MNE to estimate the source-level spatiotemporal dynamics do not fine-tune the regularization as a function of the proposed subsequent spectral analyses. Our simulations suggest that a regularization parameter that maximizes the detection of oscillatory source power may impair our ability to reliably reconstruct spectral connectivity. As a rule of thumb, we suggest using 1 to 2 orders of magnitude lower Tikhonov coefficient when searching for source coupling. Obviously, this suggestion stems from the simulations performed in this study and might not necessarily be the best choice if other minimum norm methods are used or if different source coupling configurations are present.

The variables that appear to have the strongest effect on the optimal lambda, according to our simulations, are the SNR and the spatial extent of the sources. In contrast, the strength of source coupling only minimally impacted the optimal choice for lambda. It is also noteworthy that,

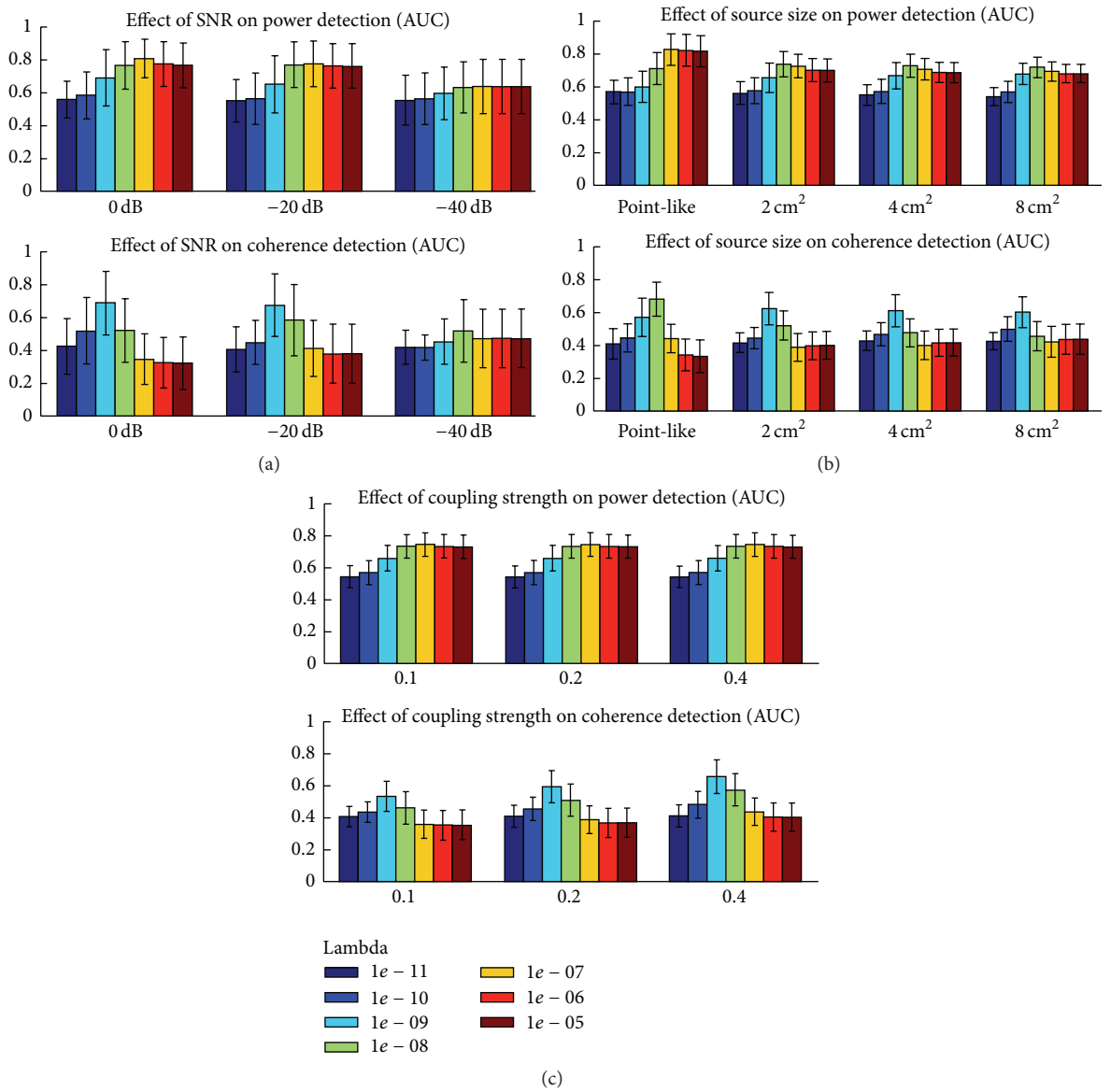


FIGURE 2: Power and coherence detection with MNE as a function of three simulation parameters: best lambda for power and coherence detection as a function of (a) SNR (0 dB, −20 dB, and −40 dB), (b) size of the sources (point-like, 2, 4, and 8 cm^2), and (c) coupling strength (0.1, 0.2, and 0.4). Power and coherence performances are depicted in the upper and lower rows of each panel, respectively.

in general, hitting the right lambda seems to be even more critical for coherence than for power: while the performances for power appeared to stabilize when sufficient regularization is applied, coherence detection appeared to drop again when regularization becomes too excessive (see Figure 1(a)).

For comparison, we also used a standard procedure known as the L-curve approach to identify an optimal Tikhonov regularization from our simulated data. This resulted in a lambda value that was three orders of magnitude smaller than the best choice for power analyses and one order of magnitude smaller than the best choice for coherence analysis. The fact that the L-curve does not yield the best results is not a major problem. The L-curve approach remains a useful approach in real data analysis with MNE where the ground truth is not known. Here, because we simulated the MEG data, we were in a position to compare the performance of the regularization coefficient lambda calculated using the L-curve approach to the results achieved by a wide range of lambda values. Previous reports have also indicated suboptimal results when applying L-curve to simulated data (e.g., [53]). Other methods for a data-driven selection of lambda (e.g., [44, 47]) and other regularization methods (such as SVD truncation) have been proposed. For instance, lambda selection can be derived from SNR as lambda = $\text{trace}(LRL^T)/[\text{trace}(Q)\times \text{SNR}^2]$ [47] which with prewhitening and appropriate selection of source covariance matrix R (such that $\text{trace}(LRL^T)/\text{trace}(Q) = 1$) yields lambda ~1/SNR, where SNR is (amplitude) signal-to-noise ratio. Approaches used to determine SNR values vary substantially. Exploring specific links between these formulations and the analyses performed

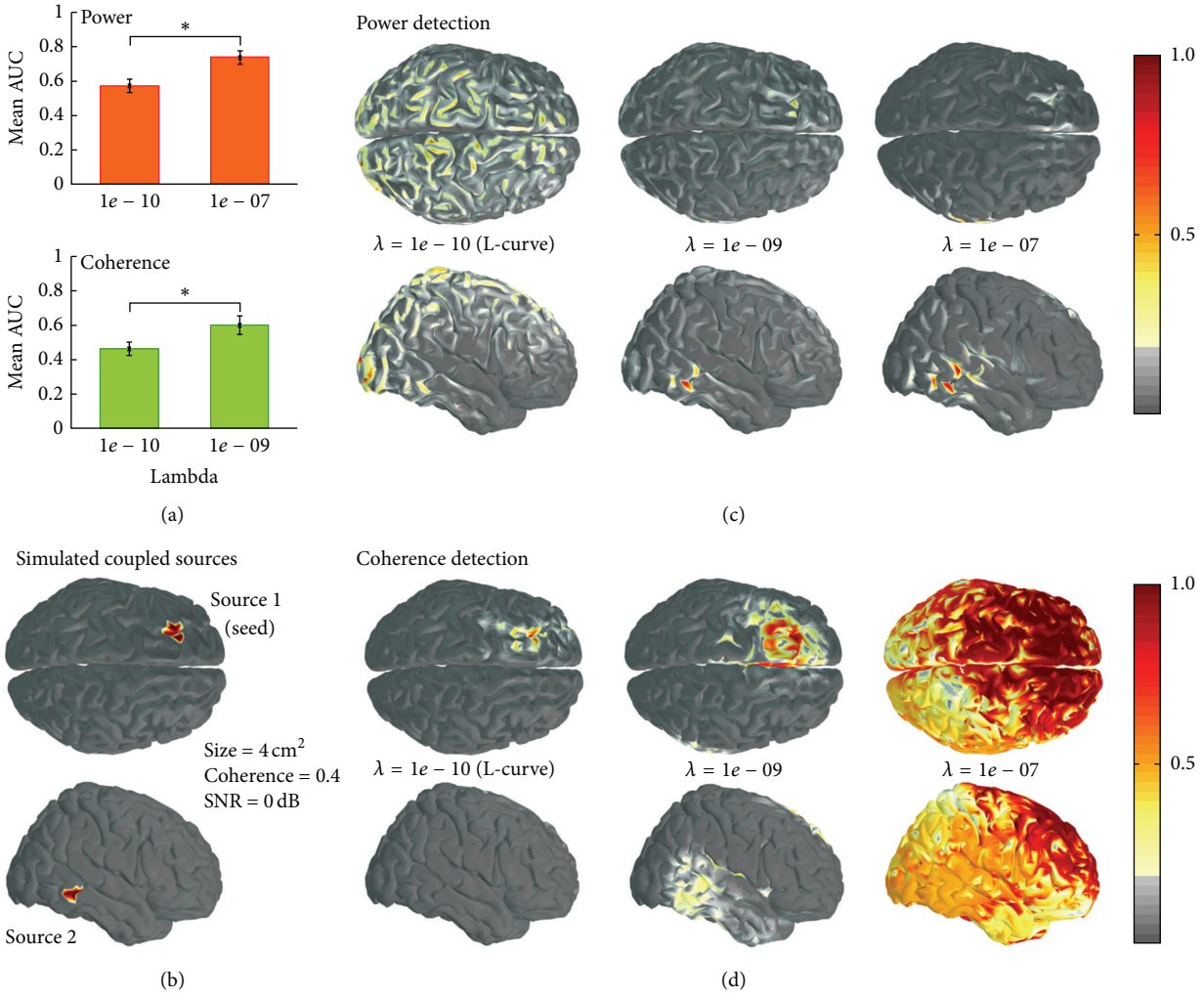


FIGURE 3: Differences in MNE performance when using the optimal lambda found using the L-curve approach and the best lambda values obtained from our simulations. (a) Difference in performance using the L-curve and the empirically derived lambda for power (upper row) and coherence (lower row) (∗ indicates statistically significant differences at $p < 0.001$, t-test). (b) Illustrative example of a simulated pair of coupled sources (alpha-band coherence = 0.4, patch size = 4 cm^2, and SNR = 0 dB). ((c) and (d)) Power and coherence reconstructions based on simulated data shown in panel (b) using three options for the definition of an optimal lambda value: $\lambda = 1e-10$ (obtained with the L-curve), $\lambda = 1e-09$ (mean optimal value for coherence detection), and $\lambda = 1e-07$ (mean optimal value for power detection). Note that the power and coherence maps are normalized with respect to the maximum on each map and subsequently thresholded at 20% of maximum amplitude.

here is of high interest, but it goes beyond the scope of the specific question we address, which focuses on differences in optimal lambda selection when switching between power and connectivity analysis.

Note that we focused here on MNE because it is a widely applied source estimation technique that is implemented in a number of toolboxes [49, 50, 54]. Besides, the classical minimum norm solution has been shown to be a valuable method whenever no reliable a priori information about source generators is available [30]. However, many other approaches exist (e.g., spatial filters) and the increased suitability of one method over another generally depends on the availability of a priori information and the validity, for the data at hand, of the theoretical assumptions that go into each method.

A prominent observation in the current study is that increases in the degree of regularization have a stronger impact on coherence than on power detection. While local power peaks remain fairly stable even with a high degree of smoothing, higher lambda values yield a drop in sensitivity in coherence analysis driven by an increase in false positive detections (spurious coupling). Since the smoothing effect is largely specific to MNE assumptions, one may ask whether coherence would be better estimated by using a different inverse approach based on different assumptions (e.g., spatial filters). While this could theoretically be the case, one should keep in mind that each inverse method comes with its own set of assumptions that are more or less suitable for the detection of coupled sources. For instance, from a theoretical point of view, the beamformer approach assumes that the sources are

not correlated. In practice, beamformers are able to detect interacting sources if the extent of coupling is sufficiently weak. To fully tackle the question in the context of our data, we would need to perform the same simulation study using other inverse solutions alongside MNE in order to directly compare performances of the methods using the same detection metrics. This is part of an ongoing study and goes beyond the specific scope of the current paper which is to compare the effect of regularization on power and coherence detections using MNE.

It is noteworthy that our evaluation of coherence detection was characterized by how accurately the second source is detected when using the first as a seed point in a brain-wide coherence analysis. In other words, we do not use the result of a source localization procedure to identify the seed. Although this comes with certain limitations, it also allows us to address the two questions separately. In addition, corticocortical coupling is not exclusively performed on sources that show significant activations in the source localization step. A selection of a source or a region of interest (ROI) based on the literature (or on a specific hypothesis) is also used as a preliminary step to seed-based source-space coupling analyses. Our results directly apply to such approaches.

Although we explored many source configurations, varying spatial location, source coupling strength, SNR, and source size (21,600 simulations in total), our simulations are of course not exhaustive. For example, it could be of interest to extend this framework to EEG data, or a combination of EEG and MEG simulations (here we focused on MEG since it is more commonly used for source-level connectivity analysis). In addition, exploring more realistic noise signals and head models could be beneficial. Furthermore, it could be of interest to examine the effect of the presence of a third noninteracting source on our findings. Also, whether the results would significantly change if other forms of min-norm estimators are used is an open question, although previous findings suggest that this is quite unlikely [31].

Standard and modified ROC analyses and the associated AUC metrics have often been used to assess and compare the performance of different inverse methods to reconstruct EEG/MEG with simulated neural sources [47, 53, 55–58]. An alternative approach to comparing detection performance is the use of precision-recall (PR) curve [59]. This approach, as well as some of the modified ROC/AUC metrics mentioned above, can be particularly useful in the case of a heavily imbalanced ratio of true positives and true negatives (e.g., [60–62]). It has been shown that when comparing performances, a curve that dominates in ROC space will also dominate in PR space. However, a method that optimizes the area under the ROC curve is not guaranteed to optimize the area under the PR curve [63]. In the future, it could therefore be of interest to extend the current framework to include other performance metrics such as the PR curve.

Moreover, we restricted our coupling simulations to generating magnitude squared coherence between two distant time series. Coherence estimation has known limitations, such as its sensitivity to field spread effects in MEG [2]. Other coupling measures can be implemented in our framework in future studies, but it is important to keep in mind that we actually generated true magnitude squared coherence between the sources, and we evaluate the effect of regularization on our ability to recover this specific type of coupling from sensor recordings (in comparison to local spectral power). Likewise, an interesting question for future research is to evaluate to which extent these results hold in the presence of more than two coupled sources. In addition, all the results reported here pertain to the robustness of detecting absolute coherence and absolute power. It could be of interest to explicitly test the implications for condition-based (e.g., task A versus task B) or baseline-based (e.g., task A versus prestimulus period) comparisons. This being said, our results readily apply to analysis of ongoing MEG signals, such as MEG resting-state studies.

5. Conclusions

In this study, we address a simple question that has generally been overlooked in the field of MEG source reconstruction using MNE: should one use a different amount of regularization depending on whether one is interested in estimating local activity or in detecting interareal connectivity? Our results based on Monte-Carlo simulations (21,600 source configurations) suggest that optimal results are obtained when setting separate regularization coefficient, to estimate the source time series with MNE, for the analysis of power and coherence. In particular, coherence estimation in source-space is enhanced by using less regularization than what is used for spectral power analysis. Furthermore, we found that SNR and the spatial extent of the source generally have a stronger impact on lambda selection than coupling strength. These results provide some practical guidelines for MNE users and suggest, in particular, that one should regularize one to two orders of magnitude less when performing connectivity analysis, compared to local spectral power estimations. Ideally, if one plans to run both power and coupling analysis based on MNE source estimates, two distinct inverse operators should be implemented. Whether the phenomenon observed here may have similar implications or concerns in the context of other inverse solutions is an interesting topic for future research.

Competing Interests

The authors declare they have no competing interests.

Acknowledgments

Ana-Sofía Hincapié was supported in part by a scholarship from Comisión Nacional de Ciencia y Tecnologíca de Chile (CONICYT), Colfuturo, Colombia, and by the French ANR project ANR-DEFIS 09-EMER-002 CoAdapt. The research was performed within the framework of the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon, within the program ANR-11-IDEX-0007. The authors would like to thank the Academy of Finland (personal grant to Jan Kujala), TES Foundation, KAUTE Foundation, and Oskar Öflund Foundation for financial support. Diego Cosmelli acknowledges support by FONDECYT Grant no. 1130758.

This research was also undertaken thanks to funding from the Canada Research Chairs program and a Discovery Grant (RGPIN-2015-04854) awarded by the Natural Sciences and Engineering Research Council of Canada to Karim Jerbi. The authors are also thankful to the computing center of the National Institute for Nuclear Physics and Particle Physics (CC-IN2P3-CNRS, Lyon).

References

[1] K. Friston, R. Moran, and A. K. Seth, "Analysing connectivity with Granger causality and dynamic causal modelling," *Current Opinion in Neurobiology*, vol. 23, no. 2, pp. 172–178, 2013.

[2] J.-M. Schoffelen and J. Gross, "Source connectivity analysis with MEG and EEG," *Human Brain Mapping*, vol. 30, no. 6, pp. 1857–1865, 2009.

[3] J. Gross, S. Baillet, G. R. Barnes et al., "Good practice for conducting and reporting MEG research," *NeuroImage*, vol. 65, pp. 349–363, 2013.

[4] S. Baillet, J. C. Mosher, and R. M. Leahy, "Electromagnetic brain mapping," *IEEE Signal Processing Magazine*, vol. 18, no. 6, pp. 14–30, 2001.

[5] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalography theory, instrumentation, and applications to noninvasive studies of the working human brain," *Reviews of Modern Physics*, vol. 65, no. 2, pp. 413–497, 1993.

[6] M. L. Schölvinck, D. A. Leopold, M. J. Brookes, and P. H. Khader, "The contribution of electrophysiology to functional connectivity mapping," *NeuroImage*, vol. 80, pp. 297–306, 2013.

[7] H. K. M. Meeren, B. de Gelder, S. P. Ahlfors, M. S. Hämäläinen, and N. Hadjikhani, "Different cortical dynamics in face and body perception: an MEG study," *PLoS ONE*, vol. 8, no. 9, Article ID e71408, 2013.

[8] I. Simanova, M. A. J. van Gerven, R. Oostenveld, and P. Hagoort, "Predicting the semantic category of internally generated words from neuromagnetic recordings," *Journal of Cognitive Neuroscience*, vol. 27, no. 1, pp. 35–45, 2015.

[9] F. Lopes da Silva, "EEG and MEG: relevance to neuroscience," *Neuron*, vol. 80, no. 5, pp. 1112–1128, 2013.

[10] K. Jerbi, J.-P. Lachaux, K. N'Diaye et al., "Coherent neural representation of hand speed in humans revealed by MEG imaging," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 18, pp. 7676–7681, 2007.

[11] A. Schnitzler and J. Gross, "Normal and pathological oscillatory communication in the brain," *Nature Reviews Neuroscience*, vol. 6, no. 4, pp. 285–296, 2005.

[12] G. C. O'Neill, M. Bauer, M. W. Woolrich, P. G. Morris, G. R. Barnes, and M. J. Brookes, "Dynamic recruitment of resting state sub-networks," *NeuroImage*, vol. 115, pp. 85–95, 2015.

[13] J. Cabral, H. Luckhoo, M. Woolrich et al., "Exploring mechanisms of spontaneous functional connectivity in MEG: how delayed network interactions lead to structured amplitude envelopes of band-pass filtered oscillations," *NeuroImage*, vol. 90, pp. 423–435, 2014.

[14] L. Marzetti, S. Della Penna, A. Z. Snyder et al., "Frequency specific interactions of MEG resting state activity within and across brain networks as revealed by the multivariate interaction measure," *NeuroImage*, vol. 79, pp. 172–183, 2013.

[15] J. F. Hipp, D. J. Hawellek, M. Corbetta, M. Siegel, and A. K. Engel, "Large-scale cortical correlation structure of spontaneous oscillatory activity," *Nature Neuroscience*, vol. 15, no. 6, pp. 884–890, 2012.

[16] F. de Pasquale, S. Della Penna, A. Z. Snyder et al., "A cortical core for dynamic integration of functional networks in the resting human brain," *Neuron*, vol. 74, no. 4, pp. 753–764, 2012.

[17] M. J. Brookes, M. Woolrich, H. Luckhoo et al., "Investigating the electrophysiological basis of resting state networks using magnetoencephalography," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 40, pp. 16783–16788, 2011.

[18] R. N. Henson, J. Mattout, K. D. Singh, G. R. Barnes, A. Hillebrand, and K. Friston, "Population-level inferences for distributed MEG source localization under multiple constraints: application to face-evoked fields," *NeuroImage*, vol. 38, no. 3, pp. 422–438, 2007.

[19] W.-T. Chang, I. P. Jääskeläinen, J. W. Belliveau et al., "Combined MEG and EEG show reliable patterns of electromagnetic brain activity during natural viewing," *NeuroImage*, vol. 114, pp. 49–56, 2015.

[20] C.-H. Cheng, S. Baillet, F.-J. Hsiao, and Y.-Y. Lin, "Effects of aging on the neuromagnetic mismatch detection to speech sounds," *Biological Psychology*, vol. 104, pp. 48–55, 2015.

[21] F.-J. Hsiao, H.-Y. Yu, W.-T. Chen et al., "Increased intrinsic connectivity of the default mode network in temporal lobe epilepsy: evidence from resting-state MEG recordings," *PLoS ONE*, vol. 10, no. 6, Article ID e0128787, 2015.

[22] M. S. Hämäläinen and R. J. Ilmoniemi, "Interpreting measured magnetic fields of the brain: estimates of current distributions," Tech. Rep., Helsinki University of Technology, Department of Technical Physics, 1984.

[23] Y. Attal and D. Schwartz, "Assessment of subcortical source localization using deep brain activity imaging model with minimum norm operators: a MEG study," *PLoS ONE*, vol. 8, no. 3, Article ID e59856, 2013.

[24] Y. Kanamori, H. Shigeto, N. Hironaga et al., "Minimum norm estimates in MEG can delineate the onset of interictal epileptic discharges: a comparison with ECoG findings," *NeuroImage: Clinical*, vol. 2, no. 1, pp. 663–669, 2013.

[25] O. Hauk and M. Stenroos, "A framework for the design of flexible cross-talk functions for spatial filtering of EEG/MEG data: DeFleCT," *Human Brain Mapping*, vol. 35, no. 4, pp. 1642–1653, 2014.

[26] S. Palva, S. Monto, and J. M. Palva, "Graph properties of synchronized cortical networks during visual working memory maintenance," *NeuroImage*, vol. 49, no. 4, pp. 3257–3268, 2010.

[27] F.-H. Lin, T. Witzel, S. P. Ahlfors, S. M. Stufflebeam, J. W. Belliveau, and M. S. Hämäläinen, "Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates," *NeuroImage*, vol. 31, no. 1, pp. 160–171, 2006.

[28] J. Mattout, M. Pélégrini-Issac, L. Garnero, and H. Benali, "Multivariate source prelocalization (MSP): use of functionally informed basis functions for better conditioning the MEG inverse problem," *NeuroImage*, vol. 26, no. 2, pp. 356–373, 2005.

[29] O. David, L. Garnero, D. Cosmelli, and F. J. Varela, "Estimation of neural dynamics from MEG/EEG cortical current density maps: application to the reconstruction of large-scale cortical

synchrony," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 9, pp. 975–987, 2002.

[30] O. Hauk, "Keep it simple: a case for using classical minimum norm estimation in the analysis of EEG and MEG data," *NeuroImage*, vol. 21, no. 4, pp. 1612–1621, 2004.

[31] O. Hauk, D. G. Wakeman, and R. Henson, "Comparison of noise-normalized minimum norm estimates for MEG analysis using multiple resolution metrics," *NeuroImage*, vol. 54, no. 3, pp. 1966–1974, 2011.

[32] R. Grave de Peralta Menendez, O. Hauk, S. Gonzalez Andino, H. Vogt, and C. Michel, "Linear inverse solutions with optimal resolution kernels applied to electromagnetic tomography," *Human Brain Mapping*, vol. 5, no. 6, pp. 454–467, 1997.

[33] A. K. Liu, J. W. Belliveau, and A. M. Dale, "Spatiotemporal imaging of human brain activity using functional MRI constrained magnetoencephalography data: Monte Carlo simulations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 15, pp. 8945–8950, 1998.

[34] K. Matsuura and Y. Okabe, "Selective minimum-norm solution of the biomagnetic inverse problem," *IEEE Transactions on Biomedical Engineering*, vol. 42, no. 6, pp. 608–615, 1995.

[35] A. M. Dale and M. I. Sereno, "Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach," *Journal of Cognitive Neuroscience*, vol. 5, no. 2, pp. 162–176, 1993.

[36] J.-Z. Wang, S. J. Williamson, and L. Kaufman, "Magnetic source imaging based on the minimum-norm least-squares inverse," *Brain Topography*, vol. 5, no. 4, pp. 365–371, 1993.

[37] M. S. Hämäläinen and R. J. Ilmoniemi, "Interpreting magnetic fields of the brain: minimum norm estimates," *Medical & Biological Engineering & Computing*, vol. 32, no. 1, pp. 35–42, 1994.

[38] K. Uutela, M. Hämäläinen, and E. Somersalo, "Visualization of magnetoencephalographic data using minimum current estimates," *NeuroImage*, vol. 10, no. 2, pp. 173–180, 1999.

[39] P. Hansen, M. Kringelbach, and R. Salmelin, *MEG: An Introduction to Methods*, Oxford University Press, New York, NY, USA, 2010.

[40] J. Sarvas, "Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem," *Physics in Medicine and Biology*, vol. 32, no. 1, pp. 11–22, 1987.

[41] A. Pascarella and A. Sorrentino, "Statistical approaches to the inverse problem," in *Magnetoencephalography*, E. W. Pang, Ed., InTech, Rijeka, Croatia, 2011.

[42] A. N. Tikhonov and V. I. Arsenin, *Solutions of Ill-Posed Problems*, Winston, 1977.

[43] M. Foster, "An application of the Wiener-Kolmogorov smoothing theory to matrix inversion," *Journal of the Society for Industrial and Applied Mathematics*, vol. 9, no. 3, pp. 387–392, 1961.

[44] C. Phillips, J. Mattout, M. D. Rugg, P. Maquet, and K. J. Friston, "An empirical Bayesian solution to the source reconstruction problem in EEG," *NeuroImage*, vol. 24, no. 4, pp. 997–1011, 2005.

[45] P. L. Nunez and R. Srinivasan, *Electric Fields of the Brain: The Neurophysics of EEG*, Oxford University Press, 2006.

[46] A. M. Dale, A. K. Liu, B. R. Fischl et al., "Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity," *Neuron*, vol. 26, no. 1, pp. 55–67, 2000.

[47] F.-H. Lin, T. Witzel, M. S. Hämäläinen, A. M. Dale, J. W. Belliveau, and S. M. Stufflebeam, "Spectral spatiotemporal imaging of cortical oscillations and interactions in the human brain," *NeuroImage*, vol. 23, no. 2, pp. 582–595, 2004.

[48] B. Fischl, "FreeSurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, 2012.

[49] F. Tadel, S. Baillet, J. C. Mosher, D. Pantazis, and R. M. Leahy, "Brainstorm: a user-friendly application for MEG/EEG analysis," *Computational Intelligence and Neuroscience*, vol. 2011, Article ID 879716, 13 pages, 2011.

[50] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, "FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," *Computational Intelligence and Neuroscience*, vol. 2011, Article ID 156869, 9 pages, 2011.

[51] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the L-curve," *SIAM Review*, vol. 34, no. 4, pp. 561–580, 1992.

[52] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[53] E. Küçükaltun-Yildirim, D. Pantazis, and R. M. Leahy, "Task-based comparison of inverse methods in magnetoencephalography," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 9, pp. 1783–1793, 2006.

[54] A. Gramfort, M. Luessi, E. Larson et al., "MNE software for processing MEG and EEG data," *NeuroImage*, vol. 86, pp. 446–460, 2014.

[55] F. Darvas, D. Pantazis, E. Kucukaltun-Yildirim, and R. M. Leahy, "Mapping human brain function with MEG and EEG: methods and validation," *NeuroImage*, vol. 23, supplement 1, pp. S289–S299, 2004.

[56] C. Grova, J. Daunizeau, J.-M. Lina, C. G. Bénar, H. Benali, and J. Gotman, "Evaluation of EEG localization methods using realistic simulations of interictal spikes," *NeuroImage*, vol. 29, no. 3, pp. 734–753, 2006.

[57] J. Mattout, C. Phillips, W. D. Penny, M. D. Rugg, and K. J. Friston, "MEG source localization under multiple constraints: an extended Bayesian framework," *NeuroImage*, vol. 30, no. 3, pp. 753–767, 2006.

[58] R. A. Chowdhury, J. M. Lina, E. Kobayashi, and C. Grova, "MEG source localization of spatially extended generators of epileptic activity: comparing entropic and hierarchical bayesian approaches," *PLoS ONE*, vol. 8, no. 2, Article ID e55969, 2013.

[59] T. Yu, "ROCS: receiver operating characteristic surface for class-skewed high-throughput data," *PLoS ONE*, vol. 7, no. 7, Article ID e40598, 2012.

[60] R. Bunescu, R. Ge, R. J. Kate et al., "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial Intelligence in Medicine*, vol. 33, no. 2, pp. 139–155, 2005.

[61] M. Goadrich, L. Oliphant, and J. Shavlik, "Learning ensembles of first-order clauses for recall-precision curves: a case study in biomedical information extraction," in *Inductive Logic Programming*, R. Camacho, R. King, and A. Srinivasan, Eds., vol. 3194 of *Lecture Notes in Computer Science*, pp. 98–115, Springer, Berlin, Germany, 2004.

[62] M. Craven and J. Bockhorst, "Markov networks for detecting overalpping elements in sequence data," in *Advances in Neural*

Information Processing Systems 17, L. K. Saul, Y. Weiss, and L. Bottou, Eds., pp. 193–200, MIT Press, 2005.

[63] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 233–240, New York, NY, USA, June 2006.

Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2016, Article ID 2094601, 10 pages
http://dx.doi.org/10.1155/2016/2094601

Research Article

How Many Is Enough? Effect of Sample Size in Inter-Subject Correlation Analysis of fMRI

Juha Pajula[1] and Jussi Tohka[2,3]

[1]*Department of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland*
[2]*Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganes, Spain*
[3]*Instituto de Investigacion Sanitaria Gregorio Marãnon, Calle de Doctor Esquerdo 46, 28007 Madrid, Spain*

Correspondence should be addressed to Juha Pajula; juha.pajula@tut.fi

Received 8 September 2015; Revised 9 December 2015; Accepted 14 December 2015

Academic Editor: Thomas DeMarse

Inter-subject correlation (ISC) is a widely used method for analyzing functional magnetic resonance imaging (fMRI) data acquired during naturalistic stimuli. A challenge in ISC analysis is to define the required sample size in the way that the results are reliable. We studied the effect of the sample size on the reliability of ISC analysis and additionally addressed the following question: How many subjects are needed for the ISC statistics to converge to the ISC statistics obtained using a large sample? The study was realized using a large block design data set of 130 subjects. We performed a split-half resampling based analysis repeatedly sampling two nonoverlapping subsets of 10–65 subjects and comparing the ISC maps between the independent subject sets. Our findings suggested that with 20 subjects, on average, the ISC statistics had converged close to a large sample ISC statistic with 130 subjects. However, the split-half reliability of unthresholded and thresholded ISC maps improved notably when the number of subjects was increased from 20 to 30 or more.

1. Introduction

Inter-subject correlation (ISC) [1, 2] is a widely used method for detecting and comparing activations in functional magnetic resonance imaging (fMRI) acquired during complex, multidimensional stimuli such as audio narratives, music, or movies [3–9]. Instead of trying to model the stimulus as in the standard general linear model (GLM) based fMRI analysis ISC computes voxel-by-voxel correlations of the subjects' fMRI time courses, assuming that the images have been registered to a common stereotactic space. The activation maps can then be formed by thresholding the average correlation coefficient values. The ISC method has been shown to produce activation maps closely matching those of the standard GLM based analysis when the stimuli are simple and can be modelled [10]. Note, however, that while not using a model time course of the stimulus, ISC expects that all the subjects are exposed to the same stimulus and it is not a method for an analysis of resting state fMRI.

A common challenge in any fMRI group analysis, including ISC analysis, is to define the required number of subjects in such a way that the analysis results are reliable and have enough statistical power, but the costs of the data acquisition are minimized. In principle, a larger sample size provides a more reliable analysis and more statistical power [11, 12]. Obviously, the sample size is not the only factor contributing to reliability (or the statistical power) of the study, but ideally the whole study design should be done to reach the desired limits of statistical power [13–15]. However, between-subject variability in fMRI data is generally much higher than within-subject variability and consequently choosing a large enough sample size is essential [16].

While there are no general methods for the optimal experimental design using naturalistic stimuli, the generalizability of the analysis results, necessarily with a limited sample size, to the population level is an important consideration. Particularly, it is important to know how many subjects are required for a reproducible (or reliable) analysis, so that small

variations in the subject sample do not cause too large variations in the analysis results. This is the question we ask in this paper and to our knowledge it has not been addressed previously in the context of the ISC analysis. Similar studies on the reliability of fMRI group studies with general linear model (GLM) analyses have been reported earlier in [16–18]. All of these studies have concluded that closer to 30 subjects should be included in a group level studies in fMRI data analysis. The sample size issue has been studied also with independent component analysis [19], where the reproducibility of the results was noticed to improve with an increased number of subjects. Critically, David et al. [20] reported that the average number of subjects in their meta-analysis was 13 and 94% of all studies were applied with less than 30 subjects, which suggests that typically fMRI group studies based on GLM might not reach the required level of reliability.

In this study, we examined how the number of subjects included in the study affects the reliability of the statistical ISC maps and the FDR corrected binary thresholded maps. We used a large 130-subject data set with a simple block design task and performed a split-half resampling based analysis (similar to [16]) while varying the number of subjects in each split-half. The resampling procedure was repeated 1000 times. This setup enables us to address the reproducibility of the studies with the maximum of 65 subjects. We compared the statistical ISC maps formed using independent subjects samples and also the thresholded ISC maps. In addition and similarly to [17] we compared statistical ISC maps with the subsets of 130 subjects with the statistical ISC map derived from the whole 130-subject data set.

2. Materials and Methods

2.1. fMRI Data. The fMRI data used in the preparation of this work were obtained from the ICBM database (https://ida.loni.usc.edu/login.jsp?project=ICBM) in the Image Data Archive of the Laboratory of Neuro Imaging. The ICBM project (Principal Investigator John Mazziotta, M.D., University of California, Los Angeles) is supported by the National Institute of Biomedical Imaging and BioEngineering. ICBM is the result of efforts of coinvestigators from UCLA, Montreal Neurologic Institute, University of Texas at San Antonio, and the Institute of Medicine, Juelich/Heinrich Heine University, Germany.

We selected all subjects from the ICBM database who had fMRI measurements with the verb generation (VG) task and the structural MR image available. This produced 132 subjects' data set. After a quality check by visual inspection two subjects were discarded due to clear artifacts in their fMRI data. This led to a final data set of 130 subjects: 61 males, 69 females; age range 19–80 years, mean 44.35 years; 117 were right-handed, 10 were left-handed, and 3 were ambidextrous. The data was acquired during the block design VG task (a language task with a visual input) from Functional Reference Battery (FRB) developed by the International Consortium for Human Brain Mapping (ICBM) [21]. The FRB holds a set of behavioral tasks designed to reliably produce functional landmarks across subjects and we have previously used fMRI data extracted from the ICBM FRB database for other experiments [10, 22]. The details of the data and VG task are provided in [10]. The VG task contained the largest number of subjects with fMRI measurements in the ICBM database among the five FRB tasks and therefore we selected it for this study.

The functional data was collected with a 3-Tesla Siemens Allegra fMRI scanner and the anatomical T_1 weighted MRI data was collected with a 1.5-Tesla Siemens Sonata scanner. The TR/TE times for the functional data were 4 s/32 ms, with flip angle 90 degrees, pixel spacing 2 mm, and slice thickness 2 mm. The parameters for the anatomical T_1 data were 1.1 s/4.38 ms, 15 degrees, 1 mm, and 1 mm, correspondingly.

2.2. Preprocessing. The preprocessing of the data was performed with FSL (version 5.0.2.2) from Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Oxford University, Oxford, UK [23]. The data preprocessing, which was identical to [10], included motion correction with FSL's MCFLIRT and the brain extraction for the functional data was done with FSL's BET [24]. The fMRI images were temporally high-pass filtered with a cutoff period of 60 s and the spatial smoothing was applied with an isotropic three-dimensional Gaussian kernel with the full-width half-maximum (FWHM) 5 mm in each direction. The brain extraction of the structural T_1 images was also performed by BET, but this was done separately from the main procedure for each T_1 weighted images as the parameters of BET required individual tuning for the images.

The image registration was performed with FSL Linear Registration Tool (FLIRT) [25, 26] in two stages. At the beginning, the skull-stripped functional images were aligned (6 degrees of freedom, full search) to the skull-stripped high-resolution T_1 weighted image of the same subject, and then the results were aligned to the standard (brain only) 2 mm ICBM-152 template (12 degrees of freedom, full search).

2.3. ISC Analyses. All of the ISC analyses were computed with ISCtoolbox for Matlab [2]. ISCtoolbox computes the ISC statistic by first computing Pearson's correlations between the corresponding time series of all subject-pairs. Then, to obtain the final multisubject test statistic, correlation values of all subject-pairs are combined into a single ISC statistic by averaging. This is the ISC statistical map.

The statistical inference was accomplished by a fully nonparametric voxel-wise resampling test implemented in the ISCtoolbox [27]. The resampling test constructs the null-distribution of the ISC values by circularly shifting the time series of each subject by a random amount. This test resembles the circular block bootstrap test [28] and it accounts for temporal correlations inherent to fMRI data. For a more detailed description of the test, we refer to [29]. For thresholding each ISC map, the resampling distribution was approximated with 10 000 000 realizations, sampling randomly across the brain voxels for each realization and generating a new set of time-shifts (one for each subject) for each realization. The resulting p-values were corrected voxel-wise over the whole brain using a false discovery rate (FDR) based multiple comparisons correction [30].

2.4. Experimental Procedure. We performed a split-half resampling type of the analysis for the ISC method. The process consisted of randomly drawing (without replacement) two independent subsets of $P = 10, 15, \ldots, 65$ subjects from the total pool of 130 subjects. Then, the full ISC analysis (including resampling distribution approximation and computation of corrected thresholds) was performed for both subsets and the full ISC analysis results from both sets were saved. This process was repeated 1000 times meaning that the ISC analysis was performed separately and independently 2000 times for each number of subjects $P = 10, 15, \ldots, 65$.

We compared the ISC statistical maps of the split-half analysis with the following criteria.

(1) Pearson's correlation coefficient C_n for comparing the nonthresholded statistical maps was defined as

$$C_n = \frac{1}{K-1}\sum_{k=1}^{K}\left(\frac{\bar{l}_k - \overline{L}}{s_{\bar{l}}}\right)\left(\frac{\bar{r}_k - \overline{R}}{s_{\bar{r}}}\right), \quad (1)$$

where K is the total number of brain voxels in the volume. $\bar{l}_k$ and $\bar{r}_k$ are the two ISC statistics of the kth voxel, respectively. $\overline{L}$ and $\overline{R}$ are the sample means of $\{\bar{l}_k\}$ and $\{\bar{r}_k\}$ across the brain volume, and $s_{\bar{l}}$ and $s_{\bar{r}}$ are the standard deviations of $\{\bar{l}_k\}$ and $\{\bar{r}_k\}$ across the brain volume. The final measure was computed by averaging the correlation measures C_n according to

$$C_{\text{avg}} = \frac{1}{N}\sum_{n=1}^{N} C_n, \quad (2)$$

where N is the number of resampling replications, which was 1000 in this study.

(2) The mean absolute error (MAE) between paired ISC maps was defined according to

$$M_n = \frac{1}{K}\sum_{k=1}^{K}\left|\bar{r}_k - \bar{l}_k\right|, \quad (3)$$

where K is the total number of brain voxels in the volume. $\bar{r}_k$ and $\bar{l}_k$ are the two ISC statistics of the kth voxel, respectively. The final measure was computed by averaging the MAE measures M_n according to

$$M_{\text{avg}} = \frac{1}{N}\sum_{n=1}^{N} M_n, \quad (4)$$

where $N = 1000$ is the number of resampling replications.

We used Dice index to compare the thresholded paired binary ISC activation maps [31]. The justification for the use of Dice index can be found in [10]. The Dice index between two sets (A_n and B_n, $n = 1, \ldots, 1000$ refers to resampling replication) of activated voxels was defined as

$$D_n = \frac{2\left|A_n \cap B_n\right|}{\left|A_n\right| + \left|B_n\right|} \quad (5)$$

and it takes values between 0 and 1. The tested thresholds were corrected with a false discovery rate (FDR) over the whole brain using $q = 0.05$, $q = 0.01$, and $q = 0.001$ (no correlation assumptions). The Dice indexes were computed for 1000 times for each number of subjects and the reported average Dice index was computed by averaging 1000 Dice indexes D_n in the same way as with correlation and MAE measures.

The Dice index defines the binary similarity between two binary images and it can be categorized with Landis and Koch categorization for Kappa coefficients [10]. According to [32] the categories are

(i) ≤0, no agreement,

(ii) 0–0.2, slight agreement,

(iii) 0.2–0.4, fair agreement,

(iv) 0.4–0.6, moderate agreement,

(v) 0.6–0.8, substantial agreement,

(vi) 0.8–1.0, almost perfect agreement.

As Landis and Koch themselves note these categories are highly subjective [32] but are maybe useful as a reference.

Similarly to [17], we considered how fast the statistic maps converge to a large sample statistic map with 130 subjects. For this, we repeated Pearson's correlation analyses described above by comparing statistic maps resulting from resampling to the statistic map obtained using all 130 subjects as in (1) and averaging over 2000 resampling iterations. More specifically, $\bar{r}$ and $\overline{R}$ in (1) were from the same statistic map with 130 subjects and in (2) N was then 2000. We computed also the sensitivity and specificity of thresholded ISC maps by using the thresholded 130 subjects ISC statistic with the corresponding threshold ($q = 0.05$, $q = 0.01$, and $q = 0.001$ with no correlation assumptions) as the ground truth. The final sensitivity and specificity (for each number of subjects) were averaged from 2000 sensitivity and specificity measures that resulted from 1000 split-half resampling replications.

2.5. Implementation. This study was computationally demanding. For each number of subjects, 2000 ISC analyses with 10 000 000 realizations for corrected thresholds were computed. This was repeated with 12 different numbers of subjects and the whole analysis required 24 001 ISC analyses (one extra analysis was for the whole data set of 130 subjects). For implementing the computations, parallel computing environment Merope of Tampere University of Technology, Finland, was used. It has nodes running on HP ProLiant SL390s G7 equipped with Intel Xeon X5650 CPU 2,67 GHz and minimum of 4 GB RAM/core. The used grid engine was Slurm. The equivalent computing time would have been 4.75 years if they had been computed with a single high end CPU.

3. Results

Figure 1 presents the thresholded (voxel-wise FDR corrected over the whole brain $q = 0.001$) results from the ISC analysis with the whole 130 subjects' data set. Significant ISC values were found around occipital and temporal lobes, lateral occipital cortex, and paracingulate gyrus as well as on

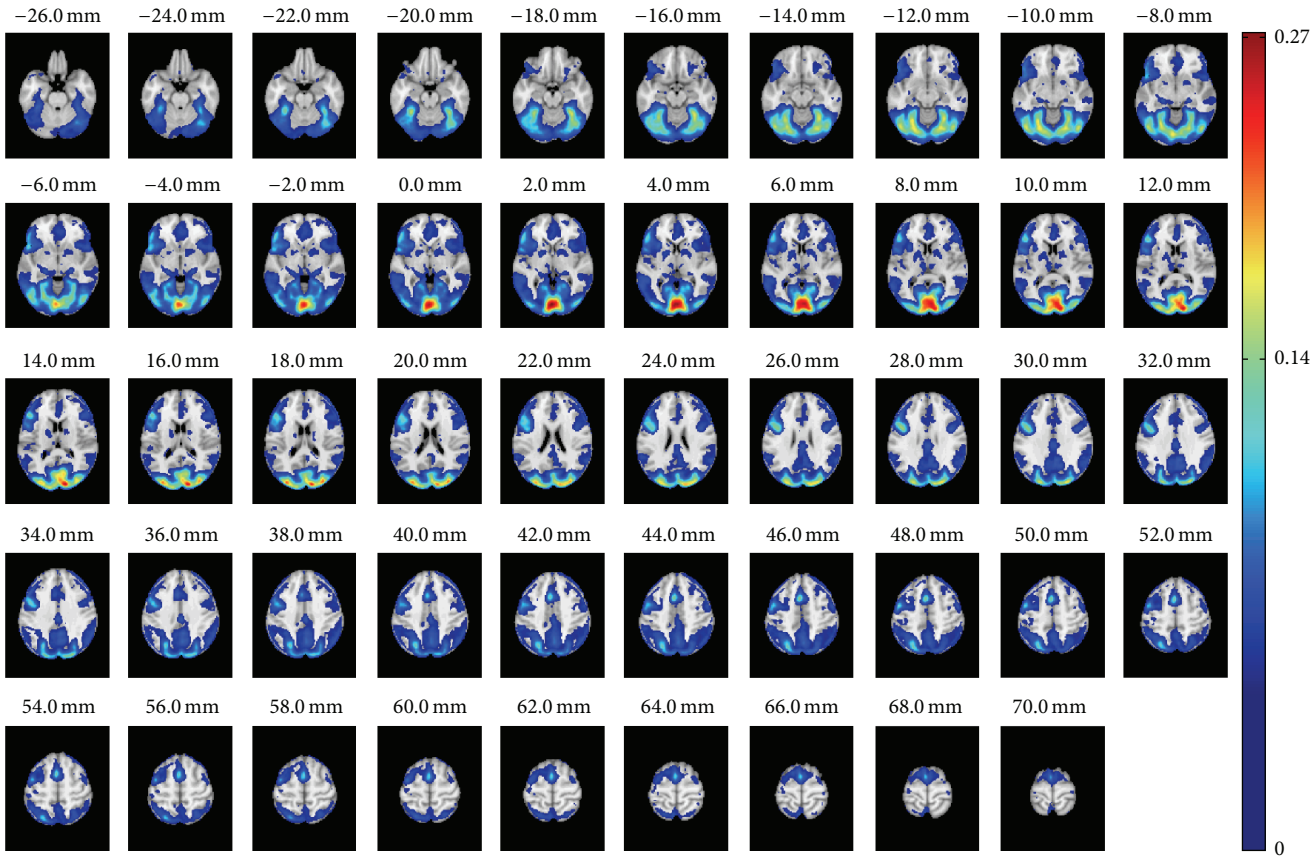


Figure 1: The ISC analysis based on 130 subjects. The figure presents the axial slices of the ISC analysis results of the whole 130 subjects' data set after applying FDR corrected $q = 0.001$ thresholding. The full statistical map is visible and available in NeuroVault: http://www.neurovault.org/collections/WTMVBEZP/images/11576/.

middle frontal and inferior frontal gyri. The 130-subject ISC map was highly similar to ISC map presented earlier with partially the same data but with smaller number of subjects ($P = 37$) [10]. The most noticeable difference compared with the 37-subject analysis was that with 130 subjects a larger number of voxels survived from the threshold and significant ISCs formed a more symmetric pattern over the hemispheres. One specific note concerning ISC map of Figure 1 is in order: There appears to be an artifact, which can be seen as a thin activation line in the left frontal cortex (e.g.) in the axial slice $z = 50$ mm. The investigation of the data at that location revealed a slight signal drop in time series of majority of subjects, buried under the noise in any single subject data, which increased ISC values with the large data set to level of statistical significance. The temporal location of the drop was in the middle of the time series ($t = 172$ s, while not counting the stabilization volumes). The statistical ISC map from 130 subjects is available in the NeuroVault service [33] at http://www.neurovault.org/collections/WTMVBEZP/images/11576/.

Figure 2 presents the correlation criteria resulting from the split-half resampling analysis. Figure 2(a) presents the average correlation C_{avg} (2) and Figure 2(b) presents the corresponding variance of C_n, $n = 1, \ldots, 1000$ (see (1)). As expected the average correlation between nonoverlapping samples increased when the number of subjects increased and, at the same time, the variance decreased. The average correlation curve was not linear with respect to the number of subjects and stabilized after 30 subjects finally reached the value of 0.95 as the number of the subjects reached the value of 65.

Figure 3 presents the MAE criteria resulting from the split-half resampling analysis. Figure 3(a) presents the average MAE (3) and Figure 3(b) presents the corresponding variance of M_n, $n = 1, \ldots, 1000$ (see (3)). Again, as expected, the average MAE between nonoverlapping samples decreased when the number of subjects increased and at the same time the variance decreased, largely replicating the correlation based curves in Figure 2. With 20 subjects the average MAE was 0.015 and with 30 subjects it was 0.011 indicating that, on average, ISC with 20 or 30 subjects already provided a high degree of reproducibility when averaged over the whole brain. However, this does not reveal whether there were variations in the reproducibility in voxel-wise ISC values across the brain. Figure 4 presents how the MAEs were distributed over the brain volume with 30 subjects. We note that the spatial shape of MAE distribution across the brain was highly similar to all numbers of subjects, and only the magnitude of the average MAE changed. Comparing Figure 4 with Figure 1 revealed that the highest

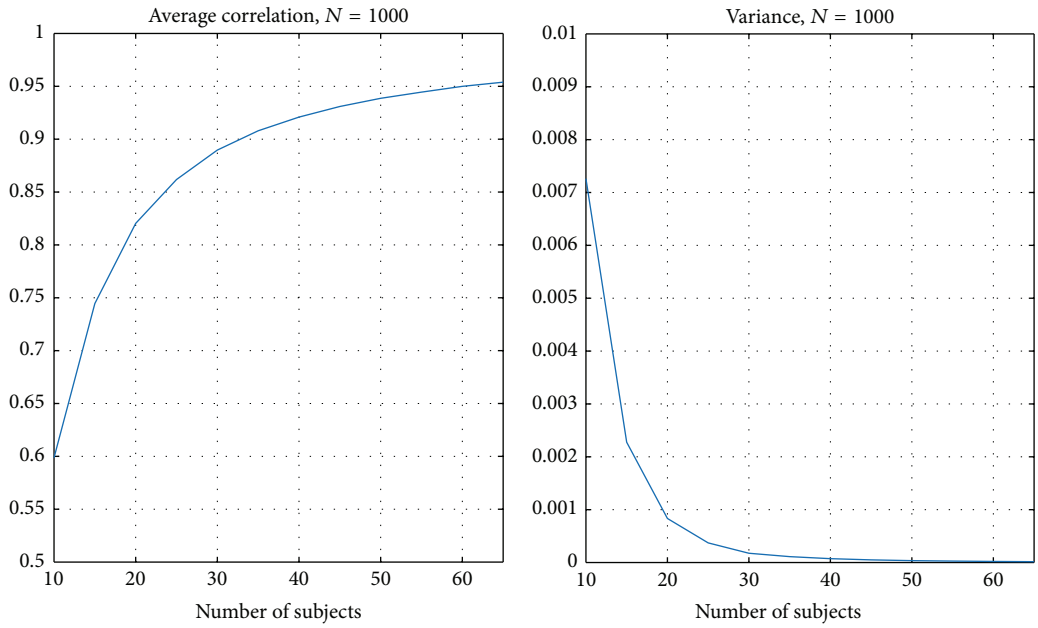


FIGURE 2: Average correlation C_{avg} over 1000 resampling replications. (a) presents the average correlation over and (b) the corresponding variance of C_n, $n = 1, \ldots, 1000$. The correlation increased when the sample size increased and at the same time the variance decreased.

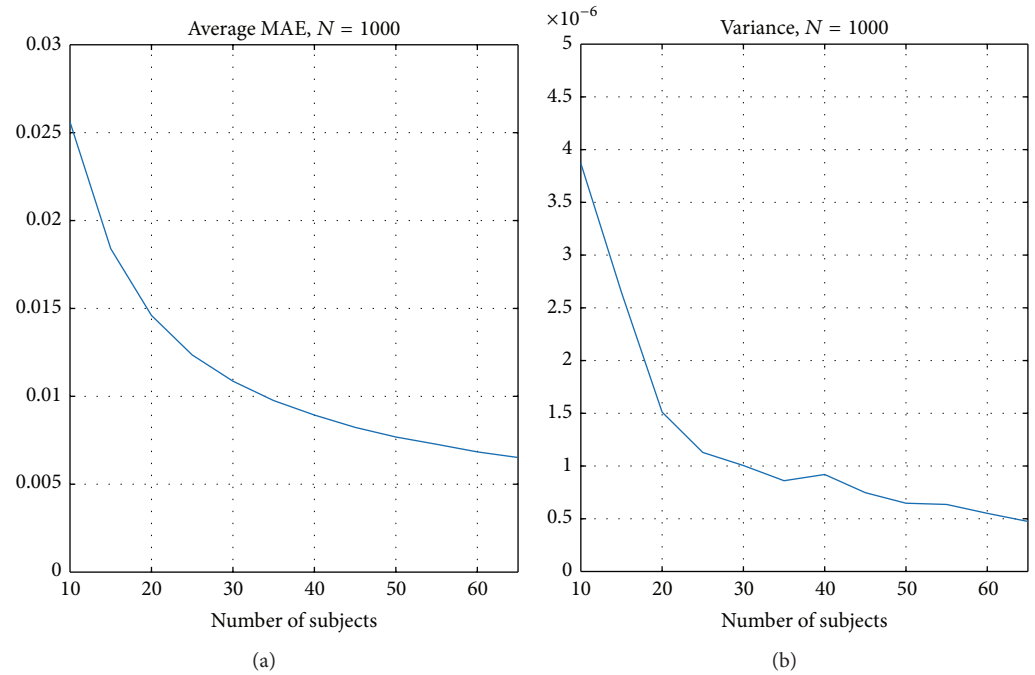


FIGURE 3: Average MAE M_{avg} over 1000 resampling replications. (a) presents the average MAE and (b) the corresponding variance of M_n, $n = 1, \ldots, 1000$.

variations in the ISCs coincided with the highest ISC values. The three-dimensional MAE maps with all numbers of subjects are available in the NeuroVault service [33] at http://www.neurovault.org/collections/WTMVBEZP/.

Figure 5 presents Dice indexes over the 1000 resampling replications. Figure 5(a) presents the average of Dice indexes D_n for three threshold levels (voxel-wise FDR corrected over the whole brain with $q = 0.05$ (blue), $q = 0.01$ (red), and $q = 0.001$ (yellow)). Figure 5(b) presents the corresponding variance of the Dice indexes D_n. Again, as expected the Dice similarity between thresholded ISC maps increased when the number of subjects increased and the variance of Dice

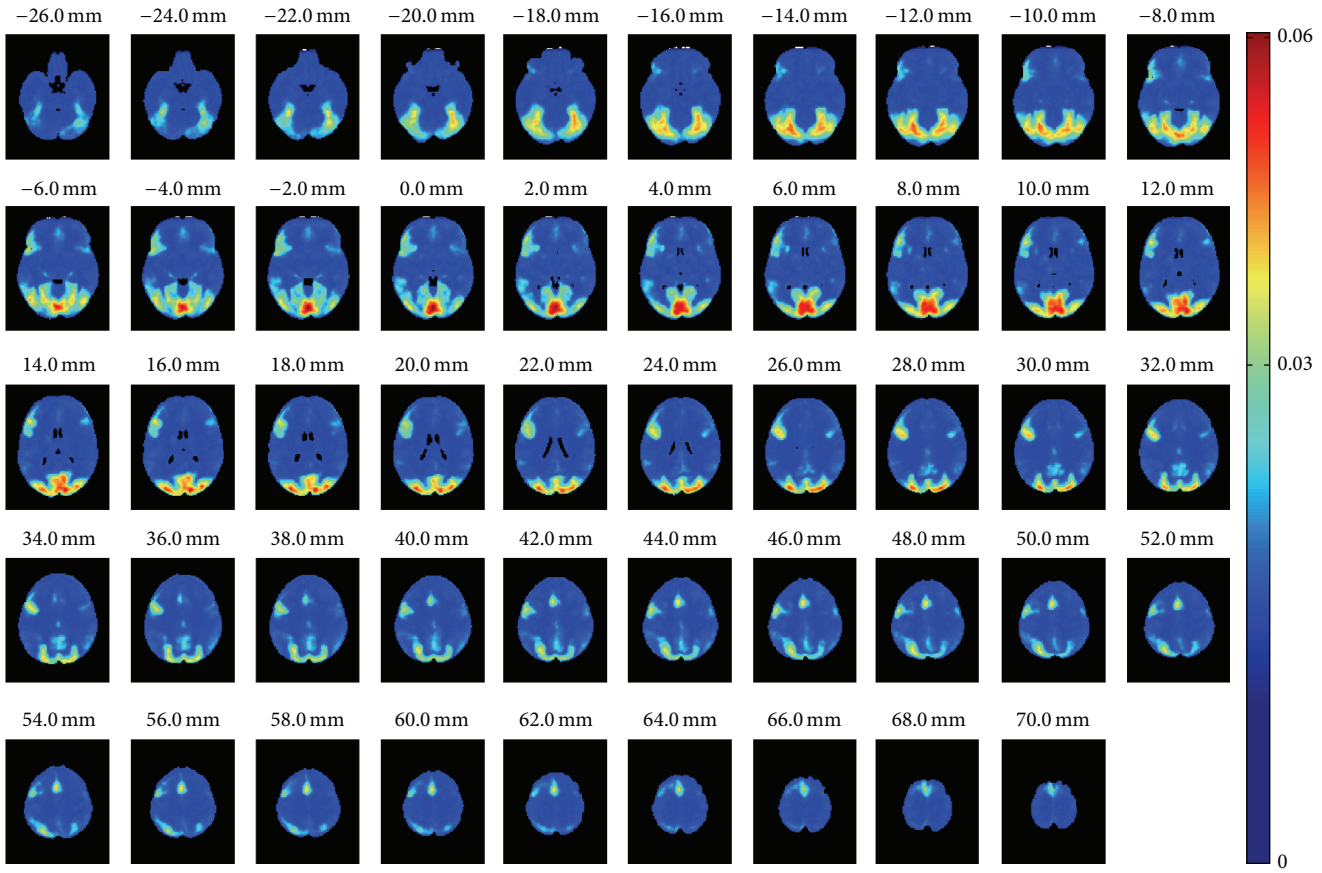


Figure 4: Average MAE computed voxel-wise over 1000 resampling replications with 30 subjects. The average voxel-wise MAE map had similar spatial shape with every tested number of subjects. The only clear difference was the magnitude of MAE values.

indexes decreased when the number of subjects increased. Based on Figure 5(a), it is noticeable that more conservative thresholds required slightly more subjects to stabilize. The most liberal threshold $q = 0.05$ had all average Dice indexes within the category "substantial agreement" but stays under the level of "almost perfect agreement" even with 65 subjects. The more conservative $q = 0.01$ reached the "almost perfect agreement" level with 45 subjects and $q = 0.001$ had the Dice index over the required 0.8 already with 35 subjects.

Figure 6 presents the average of correlation when ISC maps with resampled subsets of subjects were compared with the ISC map computed with the whole set of 130 subjects (average over 2000 resampling replications). In Figure 6, (a) presents the average correlation and (b) presents the corresponding variance. Again, the correlation increased when the number of subjects increased and the variance decreased when the number of subjects increased. The variance was close to zero and the correlation to the full 130-subject ISC map was 0.95 with 30 subjects. The sensitivity and specificity curves, using 130-subject thresholded ISC map as the ground truth, are presented in Figure 7. The sensitivity increased when the number of subjects increased and the specificity stayed close to 1 with all numbers of subjects. Figure 7 also shows that the more liberal the threshold the higher the sensitivity value at a slight expense of the specificity value.

4. Discussion

In this study, we evaluated the reliability of the ISC analysis for fMRI data and studied the effect of the sample size on the reliability of the ISC analysis. This was accomplished by using a split-half resampling based design, similar to that of [16]. We randomly sampled two nonoverlapping subsets of subjects from the 130-subject ICBM-fMRI data set with a verb generation task. We iterated the paired resampling procedure 1000 times for each number of subjects varying from 10 to 65 and compared the ISC analysis results obtained based on two nonoverlapping subsets of subjects. We compared both the raw ISC statistic maps and the thresholded statistical maps.

Previously, we have validated the ISC analysis against a gold standard set by GLM analysis in [10] and investigated the effect of smoothing to the ISC analysis results in [22]. Both of these studies used a relatively large fMRI data set of 37 subjects, which was larger than the data sets typically applied in the naturalistic stimulus experiments. Therefore, in addition to the question concerning the reliability of the ISC analysis, it was important to study how many subjects are needed for the ISC analysis in order for statistical maps to stabilize. When comparing the ISC results of our earlier study applied for 37 subjects [10] with the current study of 130 subjects, it is not surprising that the statistical power of

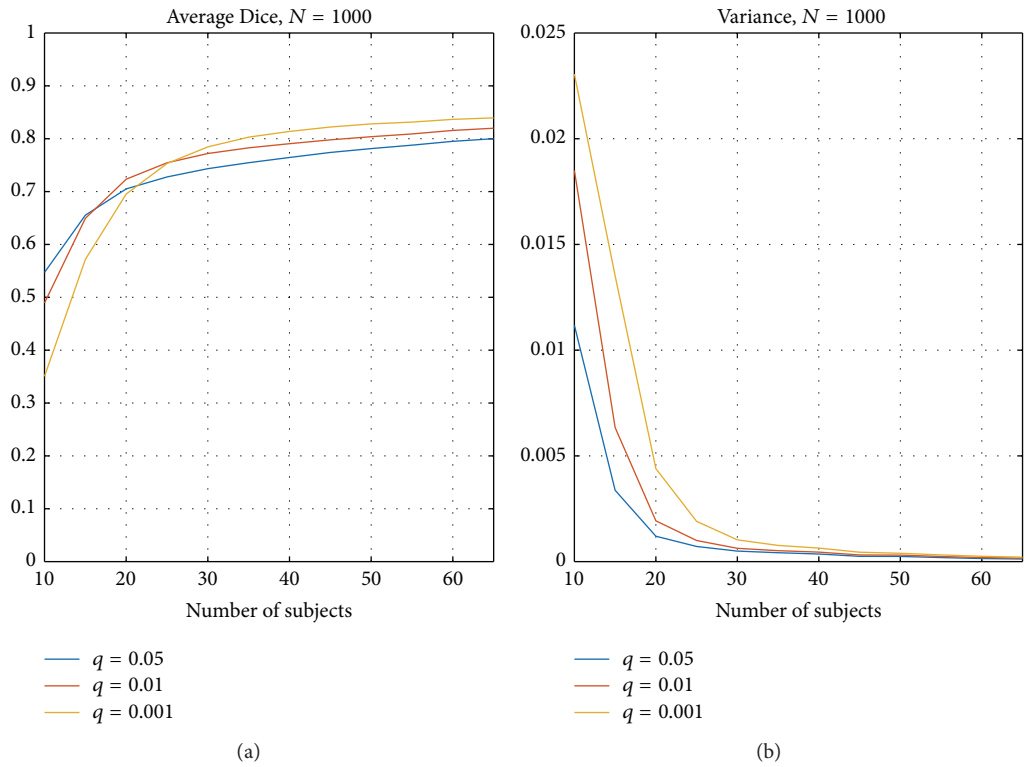


Figure 5: Average Dice index over 1000 resampling replications with three FDR levels: $q = 0.05$, $q = 0.01$, and $q = 0.001$. (a) presents the average Dice indexes D_n over 1000 replications and (b) presents the corresponding variance. The curve corresponding to the most conservative threshold $q = 0.001$ (yellow) shows that more subjects are required for greater similarity after applying the threshold to the data. The more liberal thresholds $q = 0.01$ (in red) and $q = 0.05$ (in blue) required fewer subjects to stabilize than the most conservative threshold $q = 0.001$ (yellow) but on the other hand the highest similarity was reached with the most conservative threshold.

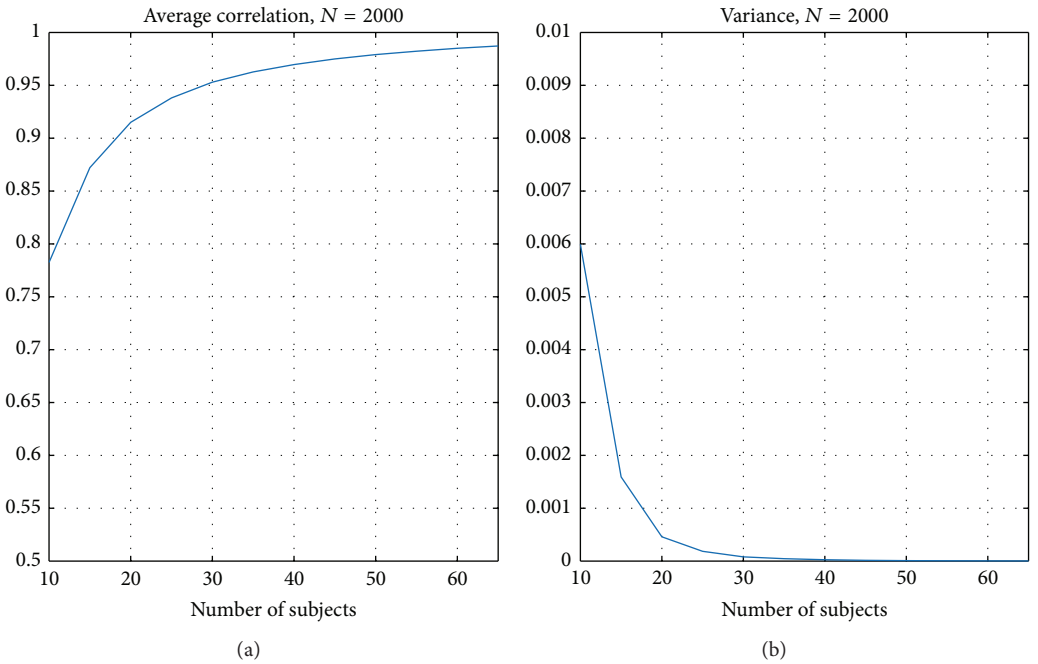


Figure 6: Average correlation comparing subsampled ISC maps with the ISC statistic map of the whole 130 subjects. (a) presents the average correlation over 2000 replications and (b) presents the corresponding variance. Again, the correlation increased when the number of subjects increased. With 30 subjects or more, the average correlation was greater than 0.95 and the variance was less than 0.0002.

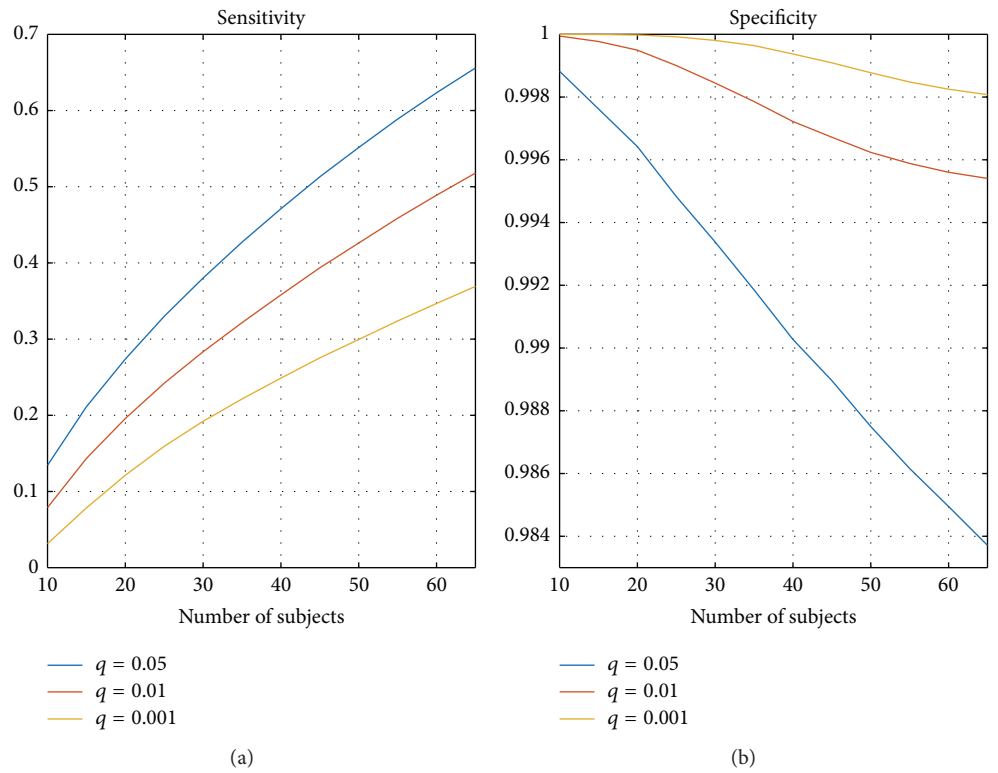


FIGURE 7: Average sensitivity and specificity from thresholded binary maps compared with the thresholded ISC statistic map of the full 130-subject sample size. (a) presents the average sensitivity over $N = 2000$ replications and (b) presents the corresponding specificity. The sensitivity increased when the number of subjects increased. The specificity was close to 1 with conservative thresholds and even with the most liberal threshold with $q = 0.05$ the specificity was over 0.98 with any number of subjects.

the analysis had been increased with the increased number of subjects; that is, the activated areas were larger with 130 subjects.

When examining the voxel-wise MAE values shown in Figure 4, it was clear that the largest MAE coincided with the strongest ISCs in Figure 1. This is an interesting phenomenon because purely technically the sample variance of the correlation coefficients decreases when the true correlation increases [34]. Thus, the increase in the voxel-wise MAE values with the average ISC means that subject-pair-to-subject-pair variability of ISC generally increases with increasing average ISC. We note that this phenomenon was independent of the applied sample size and particularly all the MAE maps, uploaded to http://www.neurovault.org/collections/WTMVBEZP/, were virtually identical except for the scale of MAE values.

The data in this study was based on a traditional block design stimulus while the ISC analysis is typically applied for fMRI data with naturalistic stimuli. This choice was made out of necessity since no large enough naturalistic stimulation studies exist. In principle, the block design data might have limitations not to reveal all sources of variation involved in the ISC analysis. In particular, the data involves the replication of the same task/stimulus pattern and therefore might lead to positively biased reliability measures for the naturalistic stimulation fMRI. On the other hand, we have shown that ISC is applicable to block design data [10, 22], which partially justifies the use of block design data. Also, it should be noted that the naturalistic stimuli themselves are highly varied and therefore using one type of naturalistic stimuli might have the same limitations as our use of the block design stimulus. Due to high computational demands of the analysis, we chose to only consider fMRI time series of certain length albeit the minimal length of the time series is an important consideration especially to the so-called time-window ISC analysis [2, 35]. To render the analysis more targeted towards the naturalistic stimulation studies, where one may stipulate that individual reactions to the used stimuli may differ more among the participants than with traditional fMRI setups, we included subjects with a wide age range spanning from 19 to 80 years to our analysis (see [36] for the age-effects on the verb generation task). We also included left-handed and ambidextrous subjects, which may be slightly controversial due to greater prevalence of right-lateralized language among the left-handed subjects (see [37] and references therein). However, most left-handers have left-lateralized language and there exist multiple other reasons not to exclude left-handers from neuroimaging studies [37].

The results of our split-half resampling analysis indicated that 20 subjects were the minimum number of subjects to achieve somehow reproducible ISC statistical maps, but for a good reproducibility it would be preferred to have 30 subjects or more. With 20 subjects, the correlation measure (C_{avg} (2)) was 0.82 (see Figure 2), the average MAE (M_{avg} (4)) was 0.015 (see Figure 3), and the average Dice coefficient was 0.71,

0.72, and 0.70 for $q = 0.05$, 0.01, and 0.001, respectively (see Figure 5). When the number of subjects was below 20, our analysis indicated weak reproducibility (see Figures 2, 5, and 3). The reproducibility improved clearly when the number of subjects was incremented from 20 to 30 (C_{avg} increased to 0.89, M_{avg} decreased to 0.010, and the average Dice coefficient increased to 0.74, 0.77, and 0.78, resp.), but adding more than 30 subjects did not improve the reproducibility so steeply any more. The average correlation between the subsample ISC statistical map and the whole sample ISC statistical map was 0.92 already with 20 subjects and 0.95 with 30 subjects indicating that ISC statistics maps converged rapidly towards the whole sample ISC maps. As seen in Figure 7, the average sensitivity of the ISC detection, when compared to the thresholded ISC map with 130 subjects, was not particularly high even with 30 subjects. However, the specificity of ISC detections was close to 1 indicating that nearly all voxels detected with small sample sizes were also detected in the full 130-subject sample. This is not surprising and largely replicates the findings for the GLM based analysis of the event related GO/NOGO task in [17]. Also, our results were in line with the studies on the reproducibility in the GLM based analysis [16] recommending that more than 20 or even more than 30 subjects should be used in fMRI group analysis. Obviously, how many subjects are required for a particular fMRI study ultimately depends on the experiment and the guidelines provided by this work may not be applicable for all experiments involving ISC analysis.

5. Conclusions

We studied the effect of sample size for ISC analysis to determine how many subjects are needed for a reliable ISC analysis. We also investigated how small sample is enough for the ISC statistic to converge to ISC statistic obtained with a large sample. We found that with 20 subjects the ISC statistics were converged close to a large 130 subjects' ISC statistic. However, the reliability of unthresholded and thresholded maps improved notably when the number of subjects was increased to 30 subjects, which indicated that with this data 30 subjects or more should be used with ISC analysis for truly reproducible results. Finally, we emphasize that the required number of subjects depends on the specific characteristic of the experiment, including the expected effect size.

Additional Material

Three-dimensional statistical maps are available in the NeuroVault service: http://www.neurovault.org/collections/WTMVBEZP/.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This project has received funding from Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for Research, Technological Development and Demonstration under Grant Agreement no. 600371, el Ministerio de Economía y Competitividad (COFUND2013-40258), and Banco Santander. Computing resources for the study was provided by the Signal Processing Department of Tampere University of Technology with the Merope computing cluster of Tampere University of Technology. Data collection and sharing for this project were provided by the International Consortium for Brain Mapping (ICBM; Principal Investigator: John Mazziotta, M.D., Ph.D.). ICBM funding was provided by the National Institute of Biomedical Imaging and BioEngineering. ICBM data are disseminated by the Laboratory of Neuro Imaging at the University of Southern California.

References

[1] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach, "Intersubject synchronization of cortical activity during natural vision," *Science*, vol. 303, no. 5664, pp. 1634–1640, 2004.

[2] J. P. Kauppi, J. Pajula, and J. Tohka, "A versatile software package for inter-subject correlation based analyses of fMRI," *Frontiers in Neuroinformatics*, vol. 8, article 2, 2014.

[3] R. Schmälzle, F. E. Häcker, C. J. Honey, and U. Hasson, "Engaged listeners: shared neural processing of powerful political speeches," *Social Cognitive and Affective Neuroscience*, vol. 10, no. 8, pp. 1137–1143, 2015.

[4] G. Bernardi, L. Cecchetti, G. Handjaras et al., "It's not all in your car: functional and structural correlates of exceptional driving skills in professional racers," *Frontiers in Human Neuroscience*, vol. 8, article 888, 2014.

[5] J. M. Lahnakoski, E. Glerean, I. P. Jääskeläinen et al., "Synchronous brain activity across individuals underlies shared psychological perspectives," *NeuroImage*, vol. 100, pp. 316–324, 2014.

[6] R. Schmälzle, F. Häcker, B. Renner, C. J. Honey, and H. T. Schupp, "Neural correlates of risk perception during real-life risk communication," *Journal of Neuroscience*, vol. 33, no. 25, pp. 10340–10347, 2013.

[7] D. A. Abrams, S. Ryali, T. Chen et al., "Inter-subject synchronization of brain responses during natural music listening," *European Journal of Neuroscience*, vol. 37, no. 9, pp. 1458–1469, 2013.

[8] I. P. Jääskeläinen, K. Koskentalo, M. H. Balk et al., "Inter-subject synchronization of prefrontal cortex hemodynamic activity during natural viewing," *The Open Neuroimaging Journal*, vol. 2, no. 1, pp. 14–19, 2008.

[9] Y. Golland, S. Bentin, H. Gelbard et al., "Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation," *Cerebral Cortex*, vol. 17, no. 4, pp. 766–777, 2007.

[10] J. Pajula, J.-P. Kauppi, and J. Tohka, "Inter-subject correlation in fMRI: method validation against stimulus-model based analysis," *PLoS ONE*, vol. 7, no. 8, Article ID e41196, 2012.

[11] J. Suckling, A. Barnes, D. Job et al., "Power calculations for multicenter imaging studies controlled by the false discovery rate," *Human Brain Mapping*, vol. 31, no. 8, pp. 1183–1195, 2010.

[12] J. E. Desmond and G. H. Glover, "Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses," *Journal of Neuroscience Methods*, vol. 118, no. 2, pp. 115–128, 2002.

[13] B. Maus, G. J. P. van Breukelen, R. Goebel, and M. P. F. Berger, "Optimal design of multi-subject blocked fMRI experiments," *NeuroImage*, vol. 56, no. 3, pp. 1338–1352, 2011.

[14] J. A. Mumford and T. E. Nichols, "Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation," *NeuroImage*, vol. 39, no. 1, pp. 261–268, 2008.

[15] S. Hayasaka, A. M. Peiffer, C. E. Hugenschmidt, and P. J. Laurienti, "Power and sample size calculation for neuroimaging studies by non-central random field theory," *NeuroImage*, vol. 37, no. 3, pp. 721–730, 2007.

[16] B. Thirion, P. Pinel, S. Mériaux, A. Roche, S. Dehaene, and J.-B. Poline, "Analysis of a large fMRI cohort: statistical and methodological issues for group analyses," *NeuroImage*, vol. 35, no. 1, pp. 105–120, 2007.

[17] K. Murphy and H. Garavan, "An empirical investigation into the number of subjects required for an event-related fMRI study," *NeuroImage*, vol. 22, no. 2, pp. 879–885, 2004.

[18] B. B. Zandbelt, T. E. Gladwin, M. Raemaekers et al., "Within-subject variation in BOLD-fMRI signal changes across repeated measurements: quantification and implications for sample size," *NeuroImage*, vol. 42, no. 1, pp. 196–206, 2008.

[19] N. W. Churchill, G. Yourganov, and S. C. Strother, "Comparing within-subject classification and regularization methods in fMRI for large and small sample sizes," *Human Brain Mapping*, vol. 35, no. 9, pp. 4499–4517, 2014.

[20] S. P. David, J. J. Ware, I. M. Chu et al., "Potential reporting bias in fMRI studies of the brain," *PLoS ONE*, vol. 8, no. 7, Article ID e70104, 2013.

[21] J. Mazziotta, A. Toga, A. Evans et al., "A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)," *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, vol. 356, no. 1412, pp. 1293–1322, 2001.

[22] J. Pajula and J. Tohka, "Effects of spatial smoothing on inter-subject correlation based analysis of FMRI," *Magnetic Resonance Imaging*, vol. 32, no. 9, pp. 1114–1124, 2014.

[23] S. M. Smith, M. Jenkinson, M. W. Woolrich et al., "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage*, vol. 23, supplement 1, pp. S208–S219, 2004.

[24] S. M. Smith, "Fast robust automated brain extraction," *Human Brain Mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[25] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Medical Image Analysis*, vol. 5, no. 2, pp. 143–156, 2001.

[26] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, no. 2, pp. 825–841, 2002.

[27] J.-P. Kauppi, I. P. Jääskeläinen, M. Sams, and J. Tohka, "Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency," *Frontiers in Neuroinformatics*, vol. 4, article 5, 2010.

[28] D. N. Politis and J. P. Romano, "A circular block-resampling procedure for stationary data," in *Exploring the Limits of Bootstrap*, R. LePage and L. Billard, Eds., pp. 263–270, John Wiley & Sons, New York, NY, USA, 1992.

[29] J. Tohka, "Non-parametric test for inter-subject correlations," Tech. Rep., 2015, https://www.nitrc.org/docman/view.php/947/2017/parametric-test-inter.pdf.

[30] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B: Methodological*, vol. 57, no. 1, pp. 289–300, 1995.

[31] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, p. 297, 1945.

[32] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[33] K. J. Gorgolewski, G. Varoquaux, G. Rivera et al., "NeuroVault.org: a repository for sharing unthresholded statistical maps, parcellations, and atlases of the human brain," *NeuroImage*, vol. 124, pp. 1242–1244, 2016.

[34] A. L. Bowley, "The standard deviation of the correlation coefficient," *Journal of the American Statistical Association*, vol. 23, pp. 31–34, 1928.

[35] L. Nummenmaa, E. Glerean, M. Viinikainen, I. P. Jääskeläinen, R. Hari, and M. Sams, "Emotions promote social interaction by synchronizing brain activity across individuals," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 24, pp. 9599–9604, 2012.

[36] J. Persson, C.-Y. C. Sylvester, J. K. Nelson, K. M. Welsh, J. Jonides, and P. A. Reuter-Lorenz, "Selection requirements during verb generation: differential recruitment in older and younger adults," *NeuroImage*, vol. 23, no. 4, pp. 1382–1390, 2004.

[37] R. M. Willems, L. van der Haegen, S. E. Fisher, and C. Francks, "On the other hand: including left-handers in cognitive neuroscience and neurogenetics," *Nature Reviews Neuroscience*, vol. 15, no. 3, pp. 193–201, 2014.

Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2016, Article ID 1068434, 22 pages
http://dx.doi.org/10.1155/2016/1068434

Research Article

Evaluation of Second-Level Inference in fMRI Analysis

Sanne P. Roels, Tom Loeys, and Beatrijs Moerkerke

Department of Data Analysis, Ghent University, H. Dunantlaan 1, 9000 Ghent, Belgium

Correspondence should be addressed to Sanne P. Roels; sanne.roels@ugent.be

Received 9 July 2015; Revised 21 August 2015; Accepted 4 October 2015

Academic Editor: Pierre L. Bellec

We investigate the impact of decisions in the second-level (i.e., over subjects) inferential process in functional magnetic resonance imaging on (1) the balance between false positives and false negatives and on (2) the data-analytical stability, both proxies for the reproducibility of results. Second-level analysis based on a mass univariate approach typically consists of 3 phases. First, one proceeds via a general linear model for a test image that consists of pooled information from different subjects. We evaluate models that take into account first-level (within-subjects) variability and models that do not take into account this variability. Second, one proceeds via inference based on parametrical assumptions or via permutation-based inference. Third, we evaluate 3 commonly used procedures to address the multiple testing problem: familywise error rate correction, False Discovery Rate (FDR) correction, and a two-step procedure with minimal cluster size. Based on a simulation study and real data we find that the two-step procedure with minimal cluster size results in most stable results, followed by the familywise error rate correction. The FDR results in most variable results, for both permutation-based inference and parametrical inference. Modeling the subject-specific variability yields a better balance between false positives and false negatives when using parametric inference.

1. Introduction

In cognitive neurosciences, functional Magnetic Resonance Imaging (fMRI) plays an important role to localize brain regions and to study interactions among those regions (resp., functional segregation and functional integration; see, e.g., [1]) The analysis of an fMRI time course in a single subject (first-level analysis) offers some insight into subject-specific brain functioning while group studies that aggregate results over individuals (second-level analysis) yield more generalizable results. In this paper, we focus on the mass univariate approach in which the brain is divided in small volume units or voxels, although alternatives exist (e.g., [2]). For each of these voxels, a general linear model (GLM) is used to model brain activation, at the first and the second level [3]. The activation is then judged at the voxel level, rather than based on topological features. The selection of activated voxels can be viewed as a sequence of different phases [4]. For first-level analyses, Carp [5] demonstrated the large variation in the choices made in each of these different phases which impacts results. In second-level analyses, to a lesser extent, different combinations of choices are possible too. We consider the following phases in the analysis of group studies: (1) aggregation of data over subjects, (2) inference, and (3) correction for multiple testing.

In two commonly used software programs to analyze fMRI data (i.e., SPM and FSL [5]), the expected activation in each voxel is modeled in a two-step approach [6]. In the first-level analysis, the evidence per subject is summarized in a linear contrast of the parameters, necessary to model the study design. These contrast images are then passed to the second-level analysis in which the evidence is weighted over subjects. To pool this information over subjects, one can either take into account subject-specific variability in constructing the voxelwise test statistics or only rely on the estimated contrasts and not take into account this subject-specific variability [7].

After pooling the data, one proceeds to the second phase, the inference phase. While parametric inference offers the advantage of closed-form null distributions that can be used to obtain p values, it depends on strong assumptions which are not easy to satisfy in practice [8] and have not been tested extensively [9]. An alternative is to use nonparametric

methods such as permutation-based inference to create an empirical null distribution conditional on the observed sample [9–11].

Third, inference must be corrected for the huge multiple testing that is induced by the mass univariate approach in which simultaneously over 100.000 tests are performed. As Bennett et al. [12] and Lieberman and Cunningham [13] discuss, there was (and yet is) no golden standard to address the choice for multiple testing corrections. We consider three different multiple testing procedures: controlling the False Discovery Rate (FDR), controlling the familywise error rate (FWE), and an approach based on uncorrected testing combined with a minimal cluster size. While FDR [14, 15] and FWE control (see, e.g., [8]) have a strong theoretical background with a focus, respectively, on the proportion of false positives among all selected voxels and on the probability to observe at least one false positive, the third approach is purely empirical in nature [13].

These three corrections are designed to control the multiple testing problem at the voxel level. Other popular alternatives that focus on topological features such as cluster size (i.e., the size of a neighboring collection of voxels) or cluster height exist as well. In a recent study, Woo et al. [16] advocate against the use of cluster-based inference and demonstrate its problematic use when studies are sufficiently powered. By definition, it is cumbersome to interpret the findings resulting from "significant clusters" because these may not reflect a set of significant constituting voxels (see also [9]). On the other hand, the third approach [13] resembles cluster-based testing but instead of setting a threshold for cluster size based on cluster significance, a fixed prespecified threshold for the minimum cluster size is set. For completeness, we therefore also extend the third approach by choosing the threshold as in cluster-based inference. However, it is important to point out that we do not intend to investigate cluster-based testing which is fundamentally different from the approach taken here and relies on different topological assumptions. Instead, we focus on voxelwise testing (for an elaborate investigation of cluster-based testing, we refer to [4]).

The choices made in each of the 3 phases of a second-level analysis is crucial steps in the analysis of fMRI data and may consequently influence results. The use of such second-level analyses or group studies is widespread [6, 10, 17, 18] but the impact of varying procedures at the different phases has not yet been extensively validated. One can distinguish three different aspects in the evaluation of methods [4]: validity, reliability, and stability. The validity can be assessed by verifying whether the false positive rate is controlled at a predefined, nominal level. Further, the balance between type I errors (false positives) and type II errors (false negatives) has long been the main interest in the validation of testing procedures (e.g., [8]). One has also acknowledged the importance of investigating the reliability of methods (e.g., [19, 20]). The extent to which a method is reliable can be measured through the overlap between activated brain regions over repeated measures, for example, in test-retest settings.

The concept of data-analytical stability, originally developed in genetics [21], was recently introduced into the context of fMRI data analysis [4]. This measure allows us to quantify reproducibility of results through the variability on different measures, for example, the variance on the number of selected voxels over replications (either in simulation studies with a known ground truth or through subsampling of real data). Stable methods are characterized by a low variability on the number of selected voxels. Data-analytical stability is thus a useful additional criterion to distinguish between methods. In this paper, we assess the influence of different choices made in the three phases on the reproducibility of results. We hereby focus on the balance between false positives and false negatives and on the stability as measures for reproducibility.

In Section 2 we give a brief overview of the different techniques. Next, we describe the details and the results of our simulation study. In Section 4, we present the results and the details from the real data application. In Discussion, we summarize our findings and end with some recommendations for the practitioner.

2. Methods

In this section we provide an overview on the different inferential techniques that we will consider in the simulation study and real data example. First, we describe the methods for pooling the evidence over subjects in the mass univariate GLM approach for fMRI data at the second level. Next, we summarize different multiple testing strategies that are frequently exploited in the fMRI literature, such as approaches that control the familywise error rate, approaches for control of the False Discovery Rate, and a two-step procedure based on an uncorrected threshold but requiring a minimum cluster size. Finally, we discuss the construction of test statistics under the null hypothesis that rely on parametric assumptions versus nonparametric approaches.

2.1. Voxel-Based GLM Approach to Analyzing fMRI Data at the Group Level. Group-level inference typically proceeds via a two-step procedure [6]. In the first step, an analysis is conducted at the voxel level for each subject m separately (with $m = 1, \ldots, M$), and an appropriate contrast of interest is constructed. In a second step, these contrast images are combined to weight evidence over the M subjects.

2.1.1. First-Level Analysis. For each subject m, the BOLD signal is sampled on T time points in every voxel v (with $v = 1, \ldots, V$) during an fMRI experiment. For every voxel v, a general linear model (GLM) is then used to relate the voxels' time course (i.e., the BOLD signal) $\mathbf{Y_v} = (Y_{v1}, \ldots, Y_{vt}, \ldots, Y_{vT})$ to the expected BOLD signal under brain activation in the experimental setup (the design matrix $\mathbf{X}$) (see, e.g., [22–25]):

$$\mathbf{Y_v} = \mathbf{X}\boldsymbol{\beta_v} + \boldsymbol{\varepsilon_v}. \quad (1)$$

The design matrix $\mathbf{X}$ is the product of a convolution of the stimulus onset function with a hemodynamic response function (HRF) (e.g., [26]). When fitting model (1), one needs to account for the residual correlation between consecutive time points. Let $\mathbf{A}\sigma_\varepsilon^2$ represent the variance-covariance matrix of $\boldsymbol{\varepsilon}_v$ in model (1). To deal with the temporal correlation, a matrix $\boldsymbol{\Sigma_d}$ is typically constructed such that $\boldsymbol{\Sigma_d}\mathbf{A}\boldsymbol{\Sigma_d^t} = \mathbf{I}$

holds. If $\mathbf{A}$ and $\mathbf{X}$ are correctly specified, $\boldsymbol{\beta}_{\mathbf{v}}$ can be unbiasedly estimated via a simple least squares approach. By relying on "decorrelated" or whitened outcome and predictor, that is, $\mathbf{Y}$ and $\mathbf{X}$ are premultiplied by $\boldsymbol{\Sigma}_{\mathbf{d}}^{-1}$, an unbiased estimator for the variance of the estimator for $\boldsymbol{\beta}_{\mathbf{v}}$ is obtained (see, e.g., [3, 27, 28]). Testing for specific differences between the activation in conditions for voxel v is then possible by testing the appropriate contrasts of the elements of $\boldsymbol{\beta}_v$ with a contrast vector $\mathbf{c}$, that is, test $H_0 : \mathbf{c}\boldsymbol{\beta}_v = 0$.

2.1.2. Second-Level Analysis. Next we focus on the group level analysis for a specific voxel v ($v = 1, \ldots, V$). For ease of notation, we will drop the voxel index v in the text below. For the contrast of interest, let $\mathbf{b} = [b_1, \ldots, b_M]^t$ denote $[\mathbf{c}\widehat{\boldsymbol{\beta}}_{\mathbf{1}}, \ldots, \mathbf{c}\widehat{\boldsymbol{\beta}}_{\mathbf{M}}]^t$, the estimated contrasts at the first level for subjects 1 to M. Obviously, those contrasts are not exactly known but estimated with some imprecision. Suppose for now that those contrasts are known and denoted by $\mathbf{c}\boldsymbol{\beta}$, then a GLM can be used to weight the group evidence (e.g., [18]):

$$\mathbf{c}\boldsymbol{\beta} = \mathbf{X}_{\mathbf{M}}\boldsymbol{\gamma} + \boldsymbol{\eta}, \tag{2}$$

where $\mathbf{X}_{\mathbf{M}}$ denotes the design matrix. In the simplest case where one is interested in knowing whether there is activation over all subjects, the design matrix $\mathbf{X}_{\mathbf{M}}$ equals a simple column matrix consisting of M elements 1. Alternatively, in the presence of between-subjects conditions or groups (e.g., one wants to know whether the activation is different between males and females), $\mathbf{X}_{\mathbf{M}}$ can take more complex forms with additional regressors. Furthermore $\boldsymbol{\eta}$ is the group error vector, with $\mathrm{Var}(\boldsymbol{\eta}) = \sigma_\eta^2\mathbf{I}_{\mathbf{M}}$ with $\mathbf{I}_{\mathbf{M}}$ the identity matrix of dimension M and σ_η^2 the between-subject variance.

In practice however $\mathbf{c}\boldsymbol{\beta}$ is unknown, and instead $\mathbf{b}$ is used as outcome:

$$\mathbf{b} = \mathbf{X}_{\mathbf{M}}\boldsymbol{\gamma} + \boldsymbol{\eta}^*, \tag{3}$$

with $\boldsymbol{\eta}^* = [\eta_1^*, \ldots, \eta_M^*]^t$ and $\boldsymbol{\eta}^* \sim N(0, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^*)$. Since $\boldsymbol{\eta}^* = \mathbf{c}\boldsymbol{\beta} - \mathbf{b} + \boldsymbol{\eta}$, it follows that the variance-covariance matrix Σ_η^* consists of the sum of two parts:

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^* = \mathrm{var}_M(\mathbf{b}) + \sigma_\eta^2\mathbf{I}_{\mathbf{M}}, \tag{4}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^* = \boldsymbol{\Sigma}_{\mathbf{M}} + \sigma_\eta^2\mathbf{I}_{\mathbf{M}}, \tag{5}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^* = \underbrace{\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \vdots & 0 \\ 0 & 0 & \sigma_M^2 \end{bmatrix}}_{\text{within-subject}} + \underbrace{\sigma_\eta^2\mathbf{I}_{\mathbf{M}}}_{\text{between-subject}}. \tag{6}$$

The first term in the right hand side of (4) is inherent to the uncertainty associated with the estimation of $\mathbf{c}\boldsymbol{\beta}_{\mathbf{m}}$, the within-subject variability, while the second term is related to the variability in the estimation of $\boldsymbol{\gamma}$, that is, the between-subjects variance.

In the literature on multisubject fMRI data analysis, two ways of dealing Σ_m are frequently used. Below, we refer to these two approaches as the Ordinary Least Squares (OLS) approach and the Weighted Least Squares (WLS) approach, respectively.

OLS: The Homoscedastic Case. In the first case, described in Holmes and Friston [17], one assumes that within-subject variances do not differ over subjects and that the residual noise is homogeneous across all M subjects. Assume that $\sigma_1^2 = \cdots = \sigma_M^2$ simplifies the form of Σ_η^* (in model (6)) to

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}*} = \sigma_{\mathrm{OLS}}^2\mathbf{I}_{\mathbf{M}}. \tag{7}$$

This implies that the within- and between-subject variability cannot be disentangled.

Mumford and Nichols [18] demonstrate that $\boldsymbol{\gamma}$ in model (3) (p 1470, in (6)) can then be estimated as $\widehat{\boldsymbol{\gamma}}_{\mathrm{OLS}} = \mathbf{X}_{\mathbf{m}}^{-\mathbf{1}}\mathbf{b}$ while the residual error variance σ_{OLS}^2 is estimated as $(\mathbf{b} - \mathbf{X}_{\mathbf{m}}\widehat{\boldsymbol{\gamma}})'(\mathbf{b} - \mathbf{X}_{\mathbf{m}}\widehat{\boldsymbol{\gamma}})/(M - 1)$. Hence, this simply amounts to solving the normal equations in the simple linear regression case and inference proceeds as usual under the GLM [28]. This is implemented in FSL [29] under `OLS` while in SPM [30] this is the standard implementation. In AFNI [31] this is implemented under `3dttest++` (see also [32]).

WLS: Allowing for Heteroscedasticity. The WLS approach, or more generally the Generalized Least Squares (GLS) approach, explicitly models the two components of the variance-covariance of $\boldsymbol{\eta}^*$ in (6):

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^* = \begin{bmatrix} \sigma_1^2 + \sigma_\eta^2 & 0 & 0 \\ 0 & \vdots & 0 \\ 0 & 0 & \sigma_M^2 + \sigma_\eta^2 \end{bmatrix}. \tag{8}$$

More specifically, a weighting matrix W is constructed such that more variable estimates b_m are down-weighted in the estimation of $\boldsymbol{\gamma}$. In the special case where the design matrix $\mathbf{X}_{\mathbf{m}}$ only consists of a column of 1's, the closed form expression for the estimator of $\boldsymbol{\gamma}$ equals [18]

$$\widehat{\boldsymbol{\gamma}}_{\mathrm{WLS}} = \sum_{m=M}^{M} \frac{b_i}{\sigma_m^2 + \sigma_\eta^2} \left(\sum_{m=1}^{M} \frac{1}{\sigma_m^2 + \sigma_\eta^2} \right)^{-1}. \tag{9}$$

More generally, $\widehat{\boldsymbol{\gamma}}_{\mathrm{WLS}}$ equals

$$\left(\mathbf{X}_{\mathbf{m}}^{\mathbf{t}}\widehat{\mathbf{W}}\mathbf{X}_{\mathbf{m}}\right)^{-1}\mathbf{X}_{\mathbf{m}}^{\mathbf{t}}\widehat{\mathbf{W}}^{-\mathbf{1}}\mathbf{b} \tag{10}$$

with $\mathbf{W}$ the weighting matrix:

$$\mathbf{W} = \begin{bmatrix} \left(\sigma_1^2 + \sigma_\eta^2\right) & 0 & 0 \\ 0 & \vdots & 0 \\ 0 & 0 & \left(\sigma_M^2 + \sigma_\eta^2\right) \end{bmatrix}. \tag{11}$$

Inference for the variance components is more complex since no closed form solutions exist. Several (restricted) maximal likelihood approaches have been suggested in the

Table 1: Table of events for Null Hypothesis Significance Testing (NHST) in which evidence against a null hypothesis H_0 is evaluated in the direction of an alternative hypothesis H_1.

		Decision	
		Conclude H_0	Conclude H_1
Voxel	Active	False negative (FN)	*True positive* (TP)
	Inactive	*True negative* (TN)	False positive (FP)

literature (see, e.g., [32]). In practice, the within-subject variance is often set to the first-level variance estimates ([18], also in the FSL software package).

In FSL this is implemented under `Flame1` while in AFNI this is implemented under `3dMEMA` (see also [33]).

2.2. Dealing with the Multiple Testing Problem. It is well-known that the mass-univariate approach in which V ($V >$ 100.000) voxels are tested simultaneously is faced with huge multiple testing problem, even at the second level. Indeed, if 100.000 tests for which H_0 is true are conducted simultaneously, each at a significance level of $\alpha = 0.05$, then, by chance alone, 5000 voxels will be declared active. Hence, the number of false positives (FP, see Table 1) becomes unacceptably high. While the interest lies in minimizing both the number of FPs and false negatives (FNs), multiple testing procedures aim to control FP rates (type I error rates).

2.2.1. Familywise Error Rate (FWE). The FWE is the probability that at least one FP occurs among all tests performed (see, e.g., [8]). In order to control this error rate, one needs the null distribution of the maximum statistic over the V test statistics: $\max(T_v)$. Indeed, assuming that the global null (i.e., the null hypothesis holds for all voxels) holds, we have that

$$P\left(\text{FP} > 0 \mid \text{global } H_0\right) = P\left(\bigcup_{v=1}^{V} T_v > u \mid \text{global } H_0\right) = P\left(\max\left(T_v\right) > u \mid \text{global } H_0\right). \quad (12)$$

Hence, when u is chosen such that this probability is lower or equal to α, the FWE is controlled at level α. In fMRI data analysis, the most commonly used approach to controlling the FWE is based on Random Field Theory (RFT, see, e.g., [34]). Relying on parametric assumptions, RFT allows a closed form approximation of the upper tail of the null distribution of the maximum statistic. Alternatively, nonparametric methods for inference such as permutation-based testing may be used. In the latter case. This will be discussed more extensively in Section 2.3.2.

Note that the expressions in (12) imply weak control of the FWE as control is only guaranteed under the assumption that the null is true for all voxels. Nichols and Hayasaka [8, Section 2.3] argue that in imaging this weak control of FWE also entails strong control, that is, control for any subset of null voxels. This is essential to localize individual significant voxels.

Further note that the classical Bonferroni correction, in which the observed p value is multiplied with the number of tests and compared with to α, can also be used to control the FWE. The underlying assumption of independence when using the Bonferroni correction implies very conservative results in the fMRI context however and makes the Bonferroni correction relatively useless. While corrections for dependence exist, these are seldom used in the analysis of neuroimaging data [8].

2.2.2. False Discovery Rate (FDR). FWE is a very stringent error rate and controlling it leads to conservative corrections. Given that one is willing to accept more FPs, provided that this number is small relative to the total number of selected voxels, one can rely on a different error measure, the False Discovery Rate (FDR). The FDR equals $E(Q)$ with

$$Q = \begin{cases} \dfrac{\#\text{FP}}{\#\text{selected voxels}} = \dfrac{\#\text{FP}}{\#\text{FP} + \#\text{TP}} & \text{if \# selected voxels} > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

Genovese et al. [15] introduced a procedure to control the FDR in neuroimaging. Using the procedure of Benjamini and Hochberg [14], the FDR is considered at level q in the sense that

$$E\left(Q\right) \leq \frac{\#\text{FP} + \#\text{TN}}{V} q \leq q. \quad (14)$$

The algorithm is as follows [15]:

(1) Select a level q.

(2) Order all V original p values from smallest to largest. With ℓ_v representing the vth smallest p value, that is, $p_{\ell_v} = p_{(v)}$, the ordered p values are as follows:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(V)}. \quad (15)$$

(3) Define r such that it is the largest v for which $p_{(v)} \leq (v/V)q$ holds.

(4) Declare all voxels $\ell_1 \cdots \ell_r$ to be active.

Genovese et al. [15] argue that this procedure controls the FDR under the assumption of *positive dependence*; that is, noise is Gaussian with nonnegative correlation. This assumption is reasonable given that smoothing images imposes increased dependency between neighboring voxels (and thus tests).

2.2.3. Uncorrected Threshold with Minimum Cluster Size. Based on simulation studies, Lieberman and Cunningham [13] proposed a more ad hoc two-step procedure that aims for a better balance between FP and FN. In the first step, the test image is thresholded at u, corresponding to an uncorrected α of, for example, 0.005. In the second step, only those voxels belonging to a cluster with minimal cluster size of 10 are selected.

Relation with Cluster-Based Significance Testing. It should be noted that the method of an uncorrected threshold with

a minimum cluster size shows superficial resemblances with cluster-based significance testing procedures. Cluster-based significance testing is a popular method to detect activation [16]. It is however fundamentally different in nature from the procedures described above. Indeed, it uses topological features rather than purely voxel-based characteristics and therefore relies on different assumptions.

As suggested by the reviewers, we added this method to our comparison in the simulations for completeness (see Section 3). More specifically, we added the cluster size (S) based significance testing with FWE-corrected and FDR-corrected p values. This corresponds to the two-step procedure but the minimum cluster size S is obtained based on cluster significance instead of fixing it at 10. Similar to the two-step procedure, a first threshold α is chosen and only clusters that are sufficiently large are retained as significant. Without going into technical details for both permutation-based and parametrical inference (which can be found in, e.g., [16, 35, 36]), this procedure determines the significance of a cluster in order to obtain the minimum cluster size S. More specifically, in a first step, after having set a sufficiently high fixed first threshold (e.g., $\alpha = 0.001$), clusters are determined by a cluster-forming algorithm. In a second step, for each of these suprathreshold clusters, the probability to observe a cluster of size S under the null hypothesis of no activation can be determined. These cluster p values can be corrected to control either the FWE (further referred to as cluster-FWE) or the FDR (further referred to as cluster-FDR) at cluster level.

In the two-step procedure with a fixed cluster size of 10, the first threshold α can be varied (empirically). For cluster-based inference on the other hand, it is important to note that the null distribution of cluster sizes relies on the assumption that the first (cluster-forming) threshold remains fixed at a stringent α-level, typically of $\alpha = 0.001$. This implies that, in the simulations, it is the minimum cluster size S that is varied empirically for the cluster-based approach (by imposing different statistical thresholds for cluster sizes through varying the FWE or FDR) and not the cluster-forming threshold α.

2.3. Inference

2.3.1. Parametric Inference. If one is willing to make distributional assumptions for the test statistic of interest, one can easily derive the thresholds for inferential decision making. We first discuss such parametric inference for the FWE and next for the FDR and the two-step approach.

For the FWE correction, one can rely on Random Field Theory (RFT) to derive the null distribution of $\max(T_v)$. Using two essential approximations from *Gaussian* Random Field Theory (which we will not discuss in full detail here, more details can be found elsewhere, e.g., [8, 34]), we have that

$$\text{FWE} = P\left(\max\left(T_v\right) > u \mid \text{global } H_0\right) \tag{16}$$

$$\approx P\left(\chi_u > 0\right) \tag{17}$$

$$\approx E\left(\chi_u\right). \tag{18}$$

In expression (17), the FWE is approximated by the probability that the Euler Characteristic χ_u is larger than 0. χ_u basically counts the number of clusters under the null hypothesis, that is, a collection of neighboring voxels for which $T_v > u$ holds. If the cluster-forming threshold u is set sufficiently high the probability to observe more than 1 cluster is neglected and one can approximate the FWE with expression (18). The expected value of χ_u is estimated through a closed-form approximation that uses information about the smoothness of the image of test statistic [8, 34]. Not only does the method take into account the spatial character of the data through the smoothness, but also its computational efficiency is a major advantage [9]. It is challenging however to satisfy the main underlying assumptions needed for valid inference, that is, normally distributed noise, sufficient smoothing, and a sufficiently high threshold (see, e.g., [34, 37]).

For the FDR corrected inference and the two-step procedure, uncorrected p values that are based on the usual t distributions of the test statistics which rely on normally distributed noise, as obtained from the OLS and WLS approach, can simply be used.

2.3.2. Permutation-Based Inference. Although some tools exist to verify the distributional assumptions underlying the test statistic (e.g., [38]), there is no widespread tradition to check those assumptions in fMRI data analysis [39]. The parametric null distributions indeed often rely on strong assumptions, which are seldom entirely fulfilled [10]. Therefore one could alternatively use nonparametric approaches such as bootstrap (e.g., [40–42]) and permutation procedures (e.g., [11, 43, 44]). Using resampling techniques, the permutation approach, for example, guarantees (asymptotically) valid inference at nominal levels by creating a null distribution conditional on the observed data, but that advantage comes at the cost of increased computational effort.

Focusing on second-level analysis and the scenario where one simply wants to test for activation over all individuals (i.e., the design matrix $\mathbf{X_M}$ is a vector of 1's), permutation-based testing proceeds as follows:

(1) Define P, the number of permutations; the higher P, the higher the precision of the empirical null distribution. However, the computational burden also increases with increasing P.

(2) Compute for each voxel v the test statistic in the original sample: T_{v0} for each voxel.

(3) Create P new samples by randomly flipping the sign of some of the elements in $\mathbf{X_M}$; that is, for randomly chosen individuals the 1 is changed into -1 [10] (if the individuals belong to different groups or the study design is more complex, more appropriate schemes can be found in, e.g., [45]).

(4) For each of the P (with $p = 1, \ldots, P$) samples compute the test statistic T_{vp}.

(5) The permutation null distribution for voxel v is then defined as the empirical distribution of T_{vp}'s. Clearly, the smaller the number of permutations P is, the more discrete the null distribution will be.

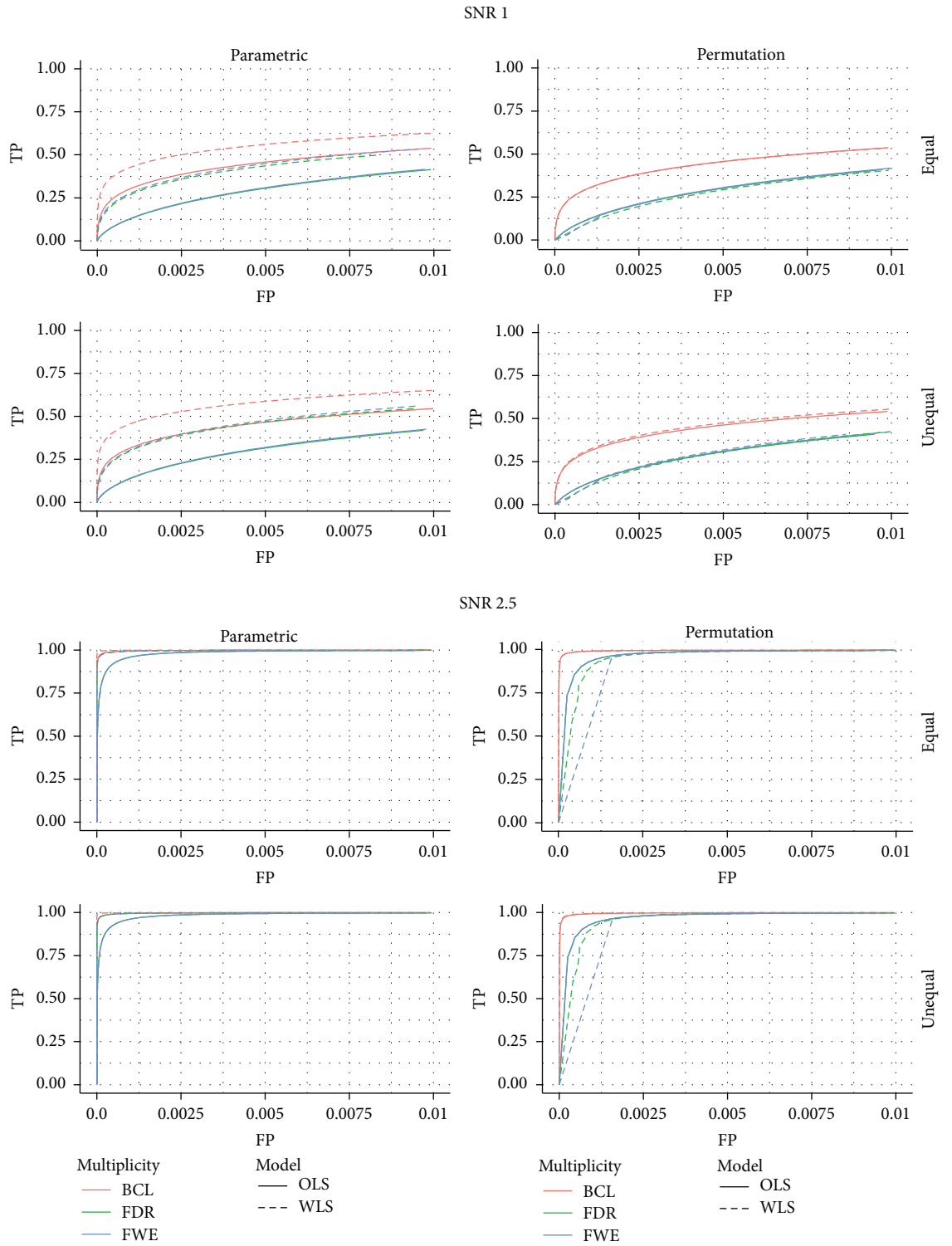


FIGURE 1: ROC for the low signal strength (SNR = 1) and for the high signal strength (SNR = 2.5); for differences in the subject-specific variability (*unequal*) or identical subject-specific variability (*equal*); for permutation-based inference and for parametric inference. FWE: familywise error correction, FDR: False Discovery Rate correction, and BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10. OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach.

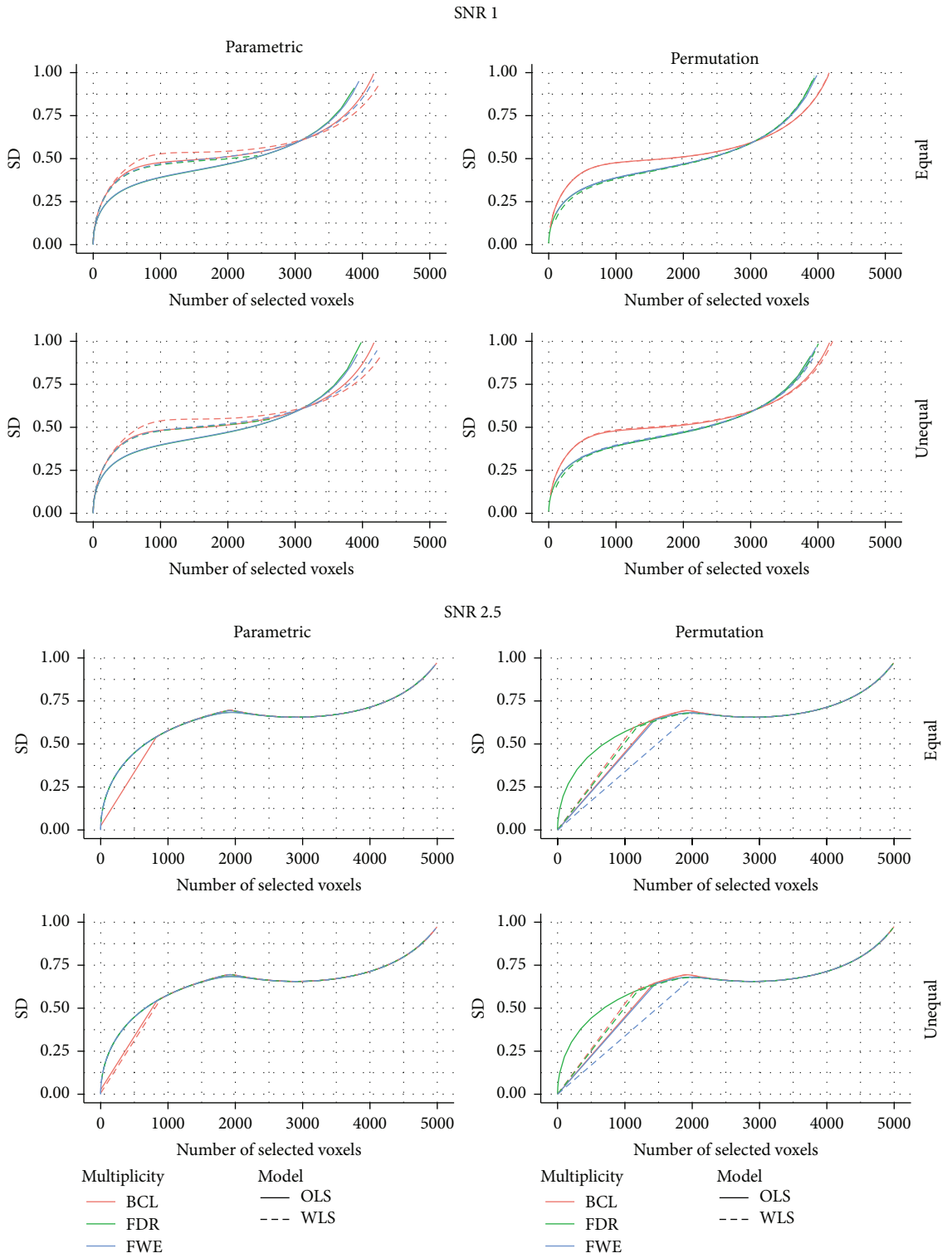


Figure 2: Matthews correlation coefficient (MCC) for the low signal strength (SNR = 1) and for the high signal strength (SNR = 2.5); for differences in the subject-specific variability (*unequal*) or identical subject-specific variability (*equal*); for permutation-based inference and for parametric inference. FWE: familywise error correction, FDR: False Discovery Rate correction, BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10, and OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach.

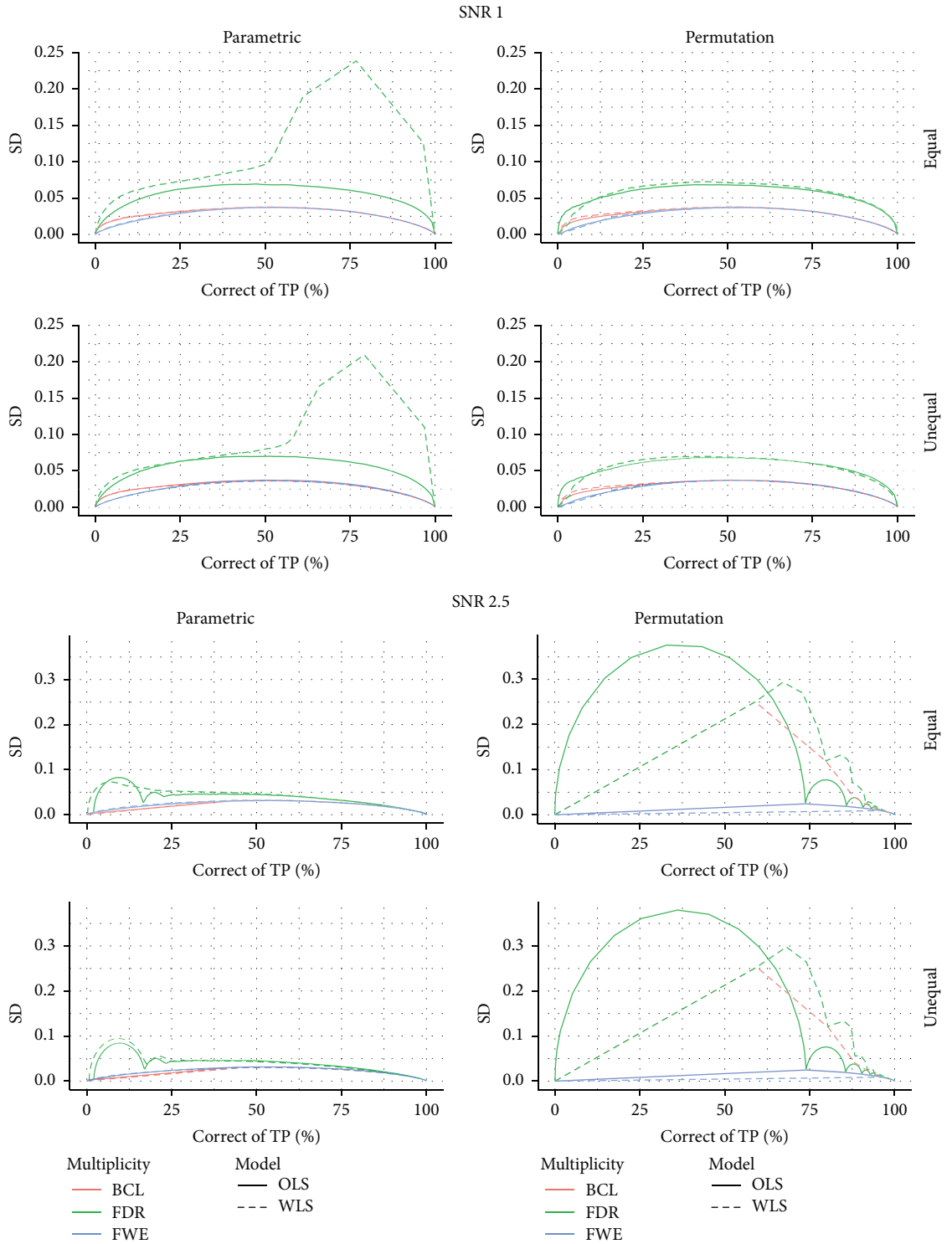


Figure 3: Stability plot for the number of correctly selected voxels in the simulation with low signal strength (SNR = 1) and for the high signal strength (SNR = 2.5); for differences in subject-specific variability (*unequal*) or identical subject-specific variability (*equal*); and for permutation-based inference and for parametric inference. FWE: familywise error correction, FDR: False Discovery Rate correction, and BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10. OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach.

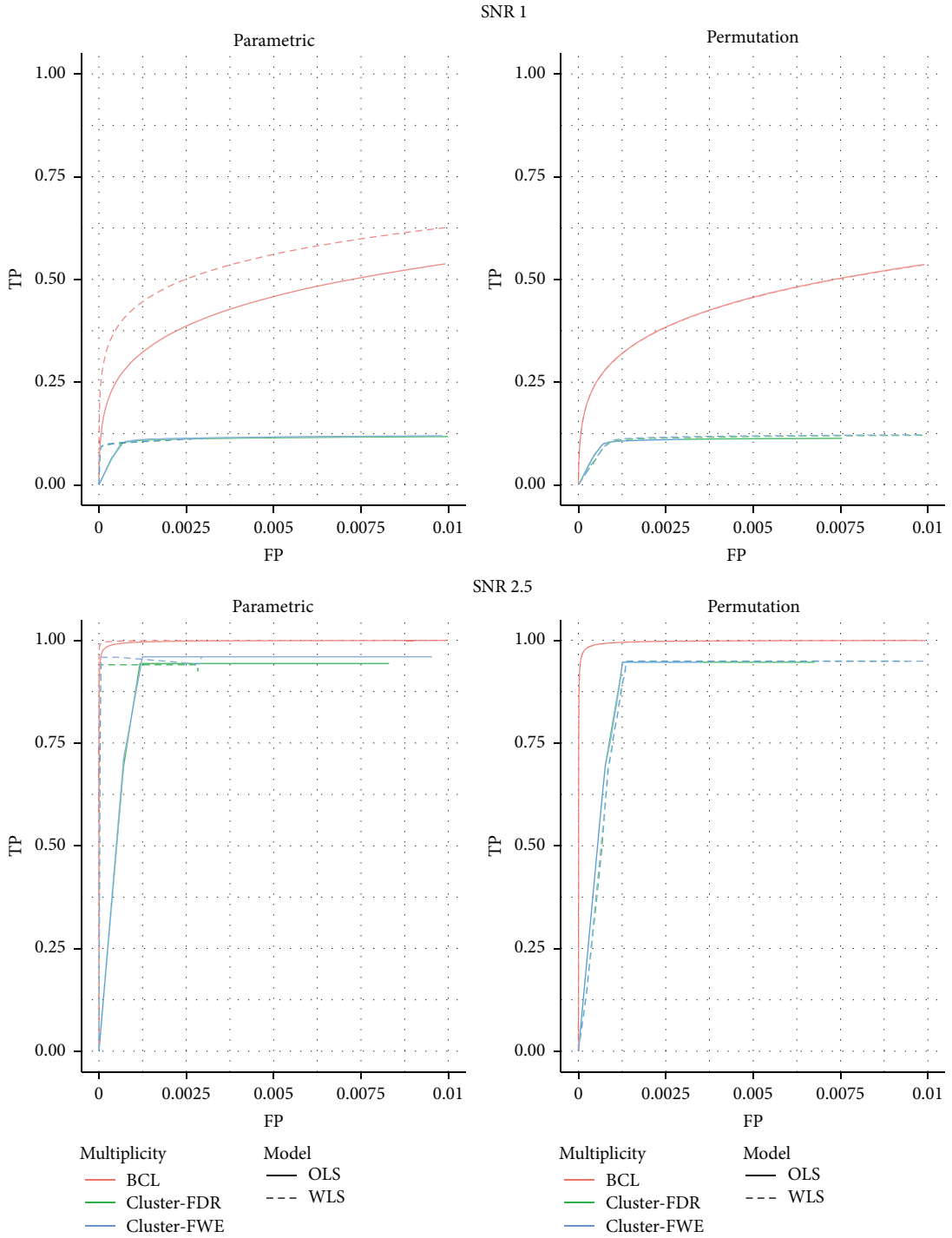


Figure 4: ROC for the low signal strength (SNR = 1) and for the high signal strength (SNR = 2.5) with identical subject-specific variability; for permutation-based inference and for parametric inference for cluster-based inference with $\alpha = 0.001$: cluster-FWE: familywise error correction based on cluster-size inference, cluster-FDR: False Discovery Rate correction based on cluster-size inference, and BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10. OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach.

Within a mass-univariate approach, empirical p values are obtained per voxel using $P(T_{pv} \geq T_{v0})$, the probability to observe a test statistic in the permutation null distribution that is at least as large as the test statistic observed in the sample at hand. The FDR correction and the two-step procedure are performed on these p values.

For the FWE correction, permutation based inference proceeds via the empirical sampling of the maximum statistic over all voxels to obtain the null distribution of the maximum statistic. This implies that in step (4) the maximum over the test statistic of all voxels is calculated: $T_p = \max(T_{pv})$ with $(v = 1, \ldots, V)$.

3. Simulations

3.1. Data Generation. For every subject $(m = 1, \ldots, 15)$ and for every voxel in a 3-dimensional space $(45 \times 45 \times 45)$, we generate a time series $\mathbf{y}$ for the signal on the first level using the following model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{d} + \boldsymbol{\epsilon}, \tag{19}$$

with $\boldsymbol{\beta} = [\beta_0, \beta_1]^t$ and with $\mathbf{X}$ the design matrix, consisting of a column for the intercept and a column describing the expected signal under a simple block design. $\mathbf{Z}$ is identical to $\mathbf{X}$, and $\mathbf{d}$ contains a random intercept d_0 and random slope d_1. The random intercept variance was set to zero, while a random slope d_1 is drawn from $N(0, \sigma^2_{d_1})$ for every subject to allow for heterogeneous effects of $\mathbf{X}$ on $\mathbf{y}$ between subjects. For every subject, voxel, and time point, $\boldsymbol{\epsilon}$ is drawn from $N(0, \sigma^2_m)$. In the simulation study no temporal correlation was induced as this unnecessarily might influence our variance estimates and consequent inference (see, e.g., [46], for an investigation of the impact of modeling the temporal autocorrelation in fMRI). We further define a signal-to-noise ratio (SNR) as the maximum amplitude $(\mathbf{x}\beta_1)$ divided by σ_{d_1} and focus on a simple contrast $\mathbf{c}\boldsymbol{\beta}$ with $\mathbf{c} = [0, 1]$.

The between-subjects standard deviation, σ_{d_1}, was set such that SNR = 1 (low signal strength) or SNR = 2.5. The variance σ^2_m is either constant or varying over the M subjects. To ensure comparability between both scenarios in terms of the average total amount of variability, the variance σ^2_m under the constant scenario is set to the average of all values under the varying scenario.

We use the neuRosim R package [47] and a canonical HRF to set up the first level activation [26] in (19). In total there are 1934 active voxels, distributed over two clusters, and 89191 inactive voxels in a $45 \times 45 \times 45$ volume (±2.5% of the voxels). The noise images that were added to the activation image were minimally smoothed in order to comply with the basic assumptions for RFT [3, 34, 39].

In total, 1000 simulations are performed for all 4 data generating mechanisms (2 SNR and constant versus varying σ^2_m).

3.2. Analysis and Evaluation Details

3.2.1. Analysis. We focus on the OLS and WLS approach to combining the individual evidence from the M subjects.

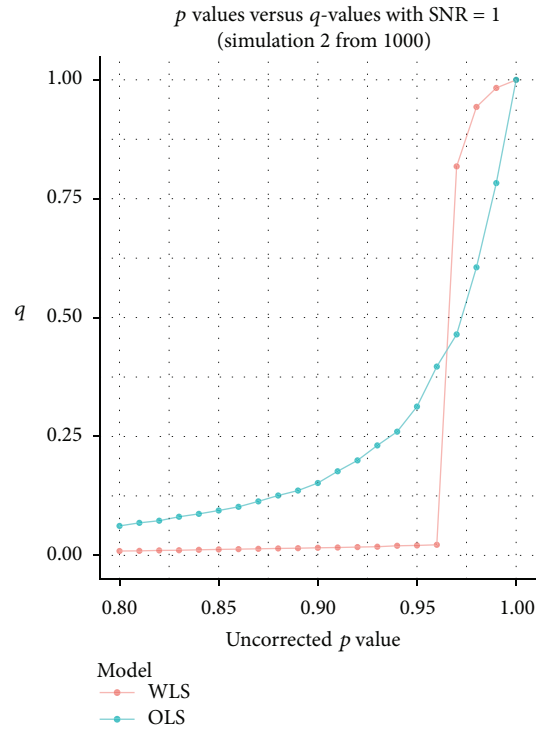


Figure 5: Uncorrected p values for the OLS and the WLS procedure, with their corresponding FDR corrected q-values based on one specific simulation under SNR = 1 with equal variance among subjects.

FSL (version 5.0.7, [29]), one of the most frequently used software packages to analyze fMRI data [5], has both methods implemented. First, the estimates $\mathbf{c}\widehat{\boldsymbol{\beta}}$ (see (1)) are obtained and next used for the second-level analysis. In the WLS approach, for every subject $m\sigma^2_m$ is estimated (see (6)) and then used to weight the evidence per subject as outlined in (11). For the parametrical inference in the OLS case, inference is based on the t distribution with $M - 1$ degrees of freedom. The WLS method uses an intrinsic Bayesian procedure that takes into account both the subject-specific variability and the variability on the estimation of $\mathbf{c}\boldsymbol{\beta}$. Further inference proceeds via a back-transformation of the posterior probability $P(\mathbf{c}\boldsymbol{\gamma} > 0 \mid \mathbf{b})$ (see (3) and [7]) to a Z-map.

For both the OLS and the WLS we use the permutation technique based on *sign-flipping*; see Section 2.3.2. The command line tool randomised allows for permutation based on the OLS method. For the WLS approach we followed the same protocol, but via an in-house R script with the test statistic as in (9). The permutation null distributions are based on 5000 permutations. On a standard laptop computer the computational time for the OLS permutation was less than 10 minutes compared to over about 40 minutes for the WLS permutation. We note that compared to the FSL implementation our in-house script was not fully optimized to speed up computational time.

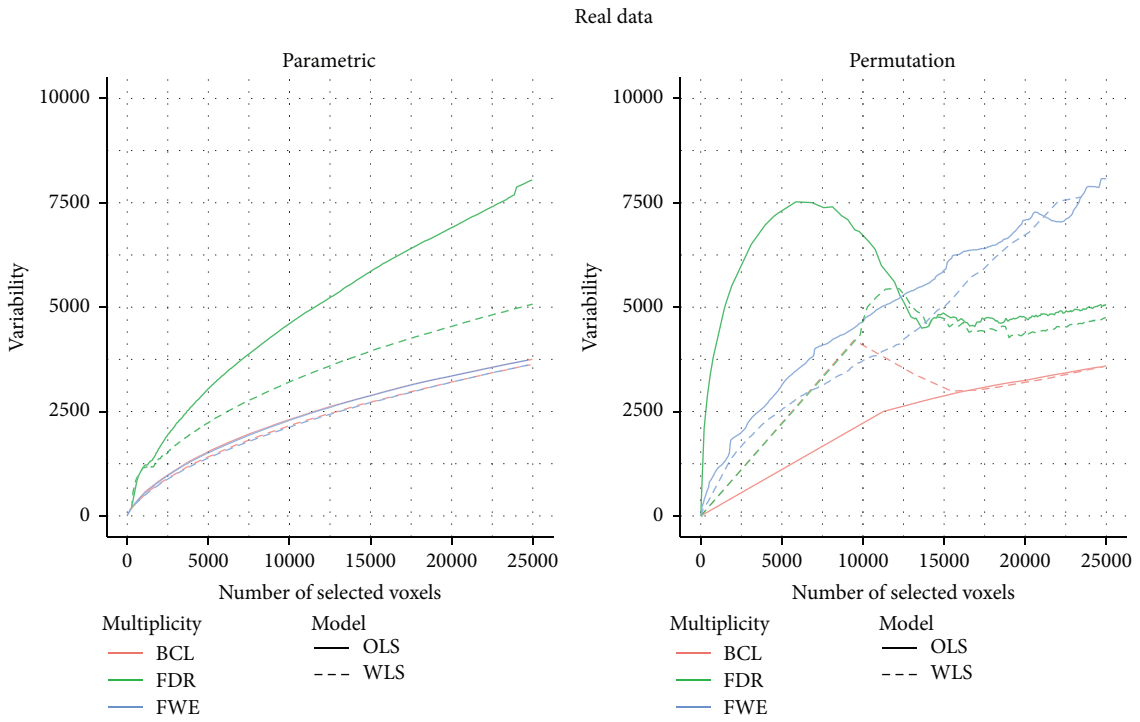


Figure 6: Stability plot for the number of selected voxels for $n = 15$ of the HPC dataset for permutation-based inference and for parametric inference. FWE: familywise error correction, FDR: False Discovery Rate correction, and BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10. OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach.

3.2.2. Evaluation. The performance of the different combination of techniques is evaluated based on the Receiving Operating Characteristics (ROC) curves. The ROC curves show the true positives (TP) rate in function of the false positives (FP) rate, with the FPs defined as voxels that are declared active but not in the true activation region and the TPs as the voxels that are declared active and in the true activation region.

ROC-curves provide a means to investigate the balance between the FP and TP rate; however, bias may be introduced for imbalanced data. As in fMRI, there are typically more true inactive than true active voxels; we also provide the Matthews correlation coefficient [48]. This measure takes into account the four cells as displayed in Table 1 and is therefore a more comprehensive measure for the quality of a test criterion, even for imbalanced data (see, e.g., [49], for an application in the genetical context). The Matthews correlation coefficient (MCC) is calculated as follows:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})\,(\text{TP} + \text{FN})\,(\text{TN} + \text{FP})\,(\text{TN} + \text{FN})}}. \quad (20)$$

Values close to 1 indicate more correct decisions, values close to 0 indicate random decisions, and values close to −1 indicate more incorrect decisions.

Furthermore we study stability through the variation on the number of correctly selected voxels. Stable methods are methods that do not induce much variability on the number of selected voxels. At last, from the above, it should be clear that all measures are defined in voxel-based way.

3.3. Results. In Figure 1 we present the ROC curves under each of the four data generating mechanisms (low versus high SNR in left versus right panel, equal versus unequal σ_m^2 in the upper versus lower panel). In total 12 ROC curves are presented, one for each of the $2 \times 2 \times 3$ combinations of selection procedures (OLS versus WLS, parametric versus nonparametric inference, FWE versus FDR versus 2-step procedure). We summarize the most important findings below.

First, we find that under all scenarios the two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10 (further denoted as BCL) has a better trade-off between FP and TP than the FWE-control or FDR-control.

Second, under both high and low signal strength, the ROC of the permutation-based method and the parametric inference have very similar shapes at almost the same height when focusing on the OLS approach. When considering the WLS approach, one finds that the ROC curves are substantially higher with permutation-based inference than with the parametric inference under both SNR (regardless of the type of control).

Third, in almost all panels of Figure 1 we find a good performance of the WLS versus the OLS method under the parametric approach, regardless of the type of multiplicity control. When permutation-based inference is used a similar performance of OLS and WLS is observed when the SNR is

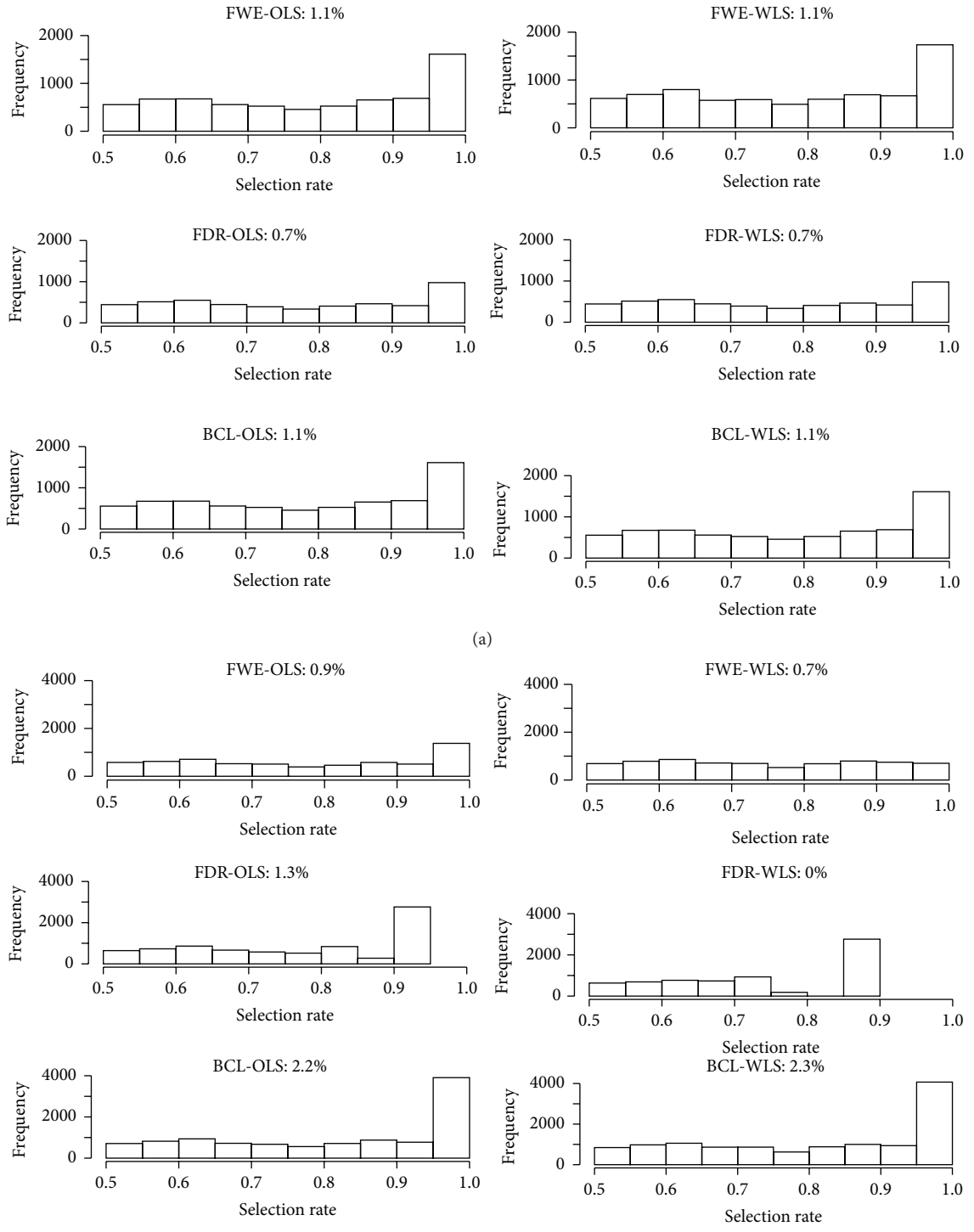
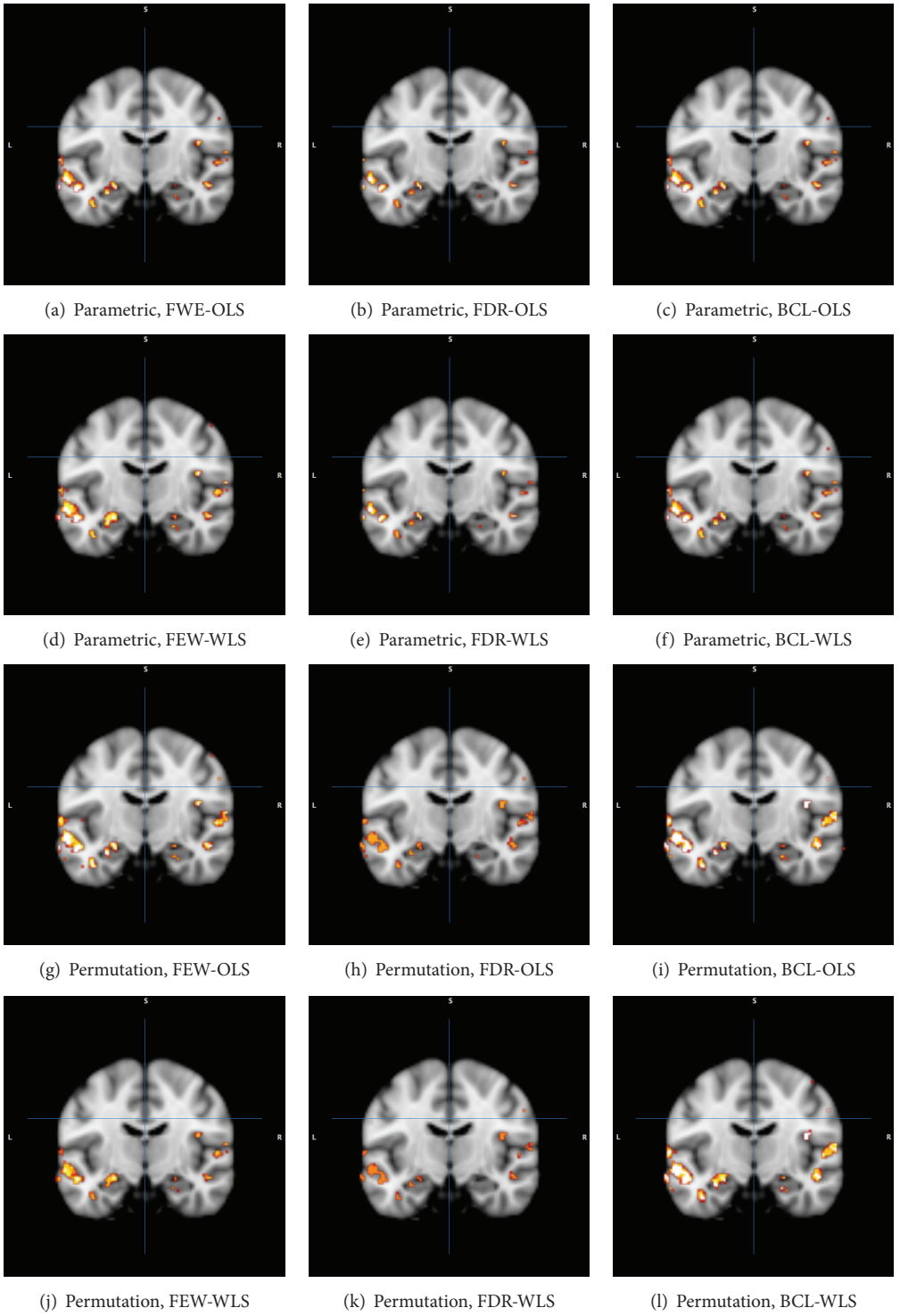


Figure 7: Plot with the reselection rates of the voxels that are larger than 0.5 over 100 bootstrap samples for real data for parametric inference (a) and for permutation-based inference (b). FWE: familywise error correction, FDR: False Discovery Rate correction, and BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10. OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach. The indicated percentage denotes the number of voxels that is declared active in more than 90% of the bootstrap cases.



(a) Parametric, FWE-OLS (b) Parametric, FDR-OLS (c) Parametric, BCL-OLS

(d) Parametric, FEW-WLS (e) Parametric, FDR-WLS (f) Parametric, BCL-WLS

(g) Permutation, FEW-OLS (h) Permutation, FDR-OLS (i) Permutation, BCL-OLS

(j) Permutation, FEW-WLS (k) Permutation, FDR-WLS (l) Permutation, BCL-WLS

Figure 8: Plot with the reselection rates that are larger than 0.75 for the HPC data for parametric inference (a–f) and for permutation-based inference (g–l). FWE: familywise error correction, FDR: False Discovery Rate correction, and BCL: Bonferroni-like first threshold and minimal cluster size. OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach. The average number of activated voxels was kept constant for all cases. Red/orange: closer to 0.75; white: closer to 1.

low, but the WLS seems to perform worse than OLS when the SNR is high. It should be noted that this is due to the discreteness of the permutation-based inference, which is mostly apparent when the signal is strong.

In Figure 2, the MCC is depicted for, respectively, a low and high signal strength with respect to the total number of selected voxels (FP + FN). While the findings based on the pattern of the ROC-curve are mostly confirmed in these figures, the differences under high SNR are somewhat less pronounced. This may indicate that under high SNR the decisions diverge less than when the SNR is lower for a same number of selected voxels.

Figure 3 shows the proportion of correctly selected voxels on the x-axis and its corresponding standard deviation on the y-axis. For all 4 data-generating mechanisms, we find that the FDR correction for multiple testing results in more variability than the other two procedures that correct for multiple testing. We also find that the FWE correction results in slightly more variable results than the BCL based corrections. Furthermore, this pattern is not altered by the choice for permutation-based inference or parametric inference. One exception is however observed. Indeed, we find that, for the WLS procedure, under the high SNR, the BCL procedure becomes more variable than the FWE procedure. We attribute this, again, to the discreteness of the permutation method and the high signal present in this simulation.

Figure 4 depicts the comparison between the BCL procedure and the pure cluster-size based inference in the ROC-curve in the simulations with no between-subject differences in the residual variability. The results for the case *with* differences in the within-subject variability and the results for the stability plots and the MCC are presented in Appendix B. We note that, due to the first fixed threshold in pure cluster-based testing, the maximum number of selected voxels is limited. For the ROC-curves and for the stability we find discrete patterns. These are a logical consequence of our simulation setup, in which two relatively large clusters are set active. Based on the ROC-curve we find a good trade-off between FP and TP for the cluster-based inference when the SNR is high, but not when the SNR is low. For the stability, it is hard to draw conclusions based on the observed results due to the above-mentioned limitations.

Finally note that, under the lowest signal strength, we find a peak in the variability for the WLS approach in combination with the FDR correction. Further inspection of the p values for the WLS approach reveals that this is due to more discreteness in the highest p values compared to the OLS procedure (Figure 5).

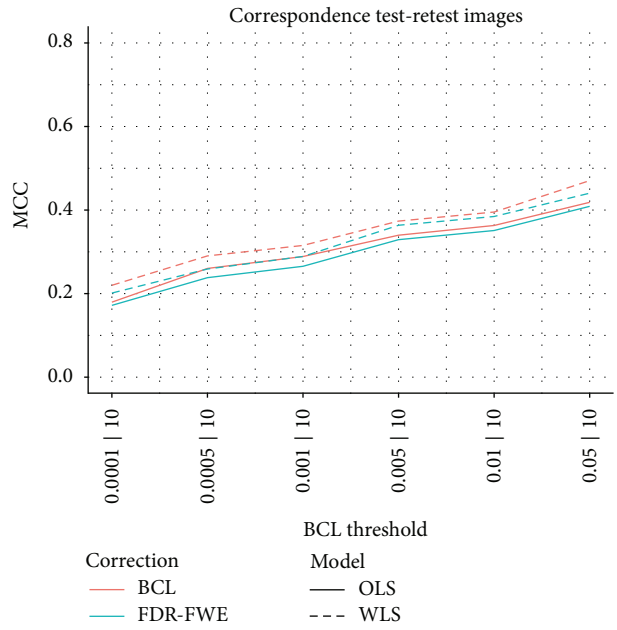


Figure 9: Test-retest correspondence measured trough the correspondence between two binary images (selected/nonselected voxels). Each BCL threshold corresponds to a specific number of selected voxels which may vary between images but not between methods.

4. Real Data Example

4.1. Human Connectome Project Dataset. To check the findings from the simulation study on real data, we use data from the Human Connectome Project (HCP, [50]). Those data are analyzed on the first level, using a standard protocol that is described elsewhere [51]. To mimic a typical fMRI study with about 15 subjects, we select the first 15 subjects (subject identifiers can be found in Appendix A.) from the HCP dataset with a focus on contrast 4, which entails the difference between a mathematical task and a story-telling task.

4.2. Stability of the Selected Voxels. For the HCP data, we determine the stability of the different proposed methods by bootstrapping subjects from the original sample, that is, drawing subjects with replacement from the original sample. In total, 100 bootstrap samples are taken. The number of active voxels at level 2 is determined in each of these bootstrapped datasets, using one of 12 the aforementioned combinations for inference at the second level. The stability on the number of selected voxels over bootstrap samples is further assessed by considering the *reselection rate* of a specific voxel, which is the proportion of bootstrap samples in which that voxel is declared active.

4.3. Results. In Figure 6, we find the same pattern as in the simulations when using parametric inference, that is, the FDR based correction for multiple testing results in more variability on the number of selected voxels. Also, we find that the FWE and the BCL correction result in similar variability. This finding holds for both the WLS and the OLS approach. In contrast to the simulation study, we find however that the WLS approach is always less variable than the OLS approach for a given type of multiplicity control.

For the permutation-based inference we find that when the number of selected voxels is relatively low (less than ±5% of the ±200.000 voxels) the FDR correction with the OLS is far more variable than all other combinations. We note again that the WLS suffers from the discreteness of p values in

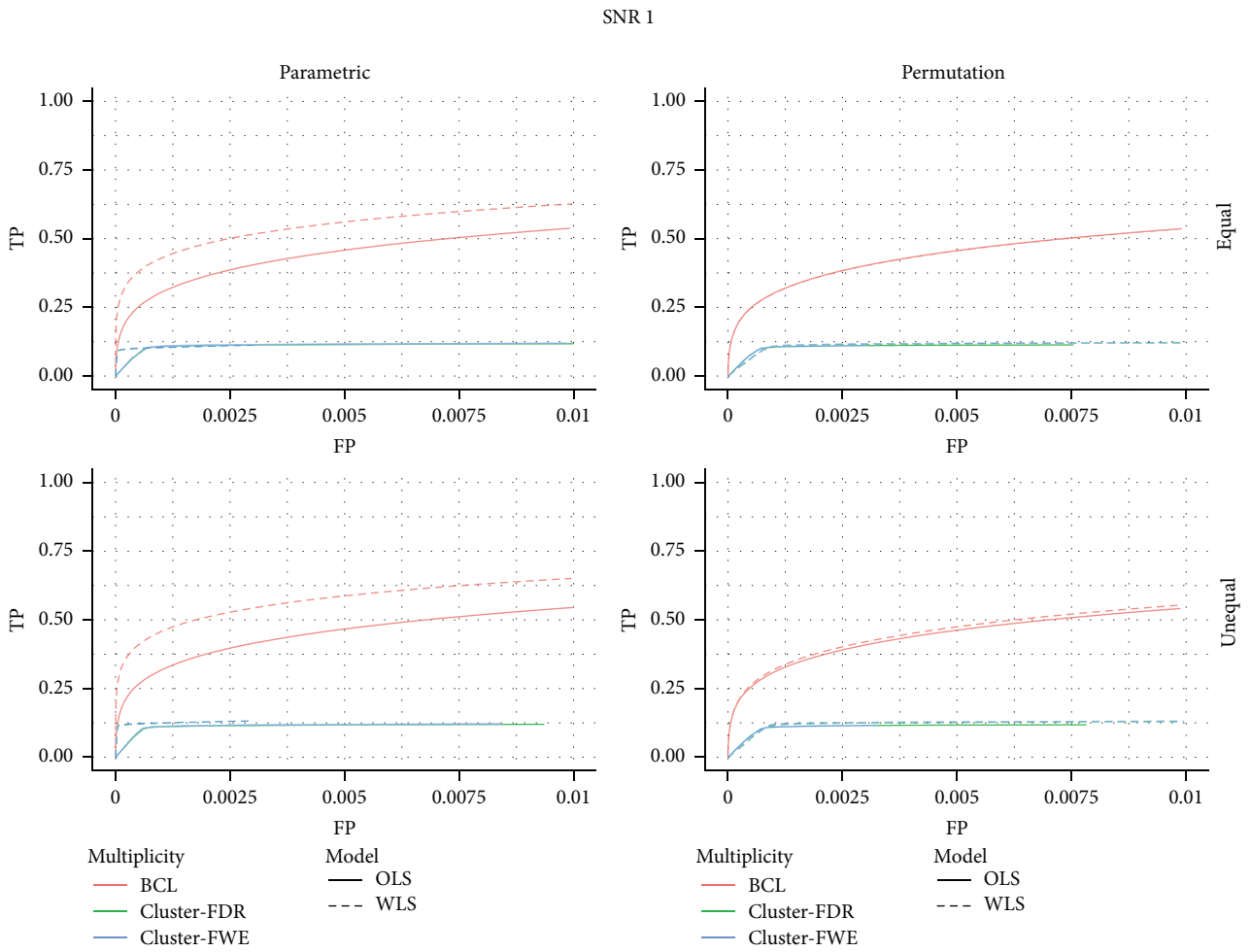


Figure 10: Receiver operating curve for a signal-to-noise ratio of 1 over the range [0; 0.01].

the permutation-based inference when the FDR correction is used. Due to this discreteness, several small original p values are converted to only one corrected q value, causing the straight line from the origin to the first point. For the two-step procedure, there is a similar artifact when using WLS. This can be attributed to the fact that the lower p values do not occur in clusters larger than 10, until these reach a certain threshold that results in a huge amount of activation. If more than 5% of the voxels are selected, the results are more variable if one uses the FWE correction for multiple testing, compared to the other methods.

Based on Figure 6, we next determine the thresholds for which 10.000 voxels are selected on average over the 100 bootstrap samples. These thresholds are then used to determine the reselection rate of each specific voxel over the 100 bootstrap samples. Figure 7 depicts the histograms of the reselection rates that are larger than 50%. The header of each histogram shows the percentage of voxels that are selected in more than 90% of the samples.

From Figure 7 we find the highest reselection rates when using the FWE or BCL multiplicity control in the parametric inference framework (i.e., the 6 upper panel histograms). In the permutation-based inference framework (i.e., the 6 lower panel histograms), we find that the FDR achieves higher reselection rates than the FWE if the OLS approach is used, but the highest reselection rates are found with the BCL multiplicity control with both the OLS and the WLS approach.

To take into account the localization of voxels that are frequently reselected, we also constructed brain images in Figure 8, where we identified all voxels that have a reselection rate of at least 75%. Although we acknowledge that the slice depicted is only exemplary, the above-described trends are clearly confirmed.

4.4. Test-Retest Correspondence. As suggested by one of the reviewers, stable methods should reflect more similar results using different real samples. To study this, we used an additional run for each of the 15 subjects in the HCP data. We exemplary demonstrate this test-retest similarity for the parametrical analysis. We matched the number of selected voxels per image in the FWE/FDR method by the respective numbers that are found using the two-step BCL procedure. Indeed, when selecting the N voxels with the N smallest p

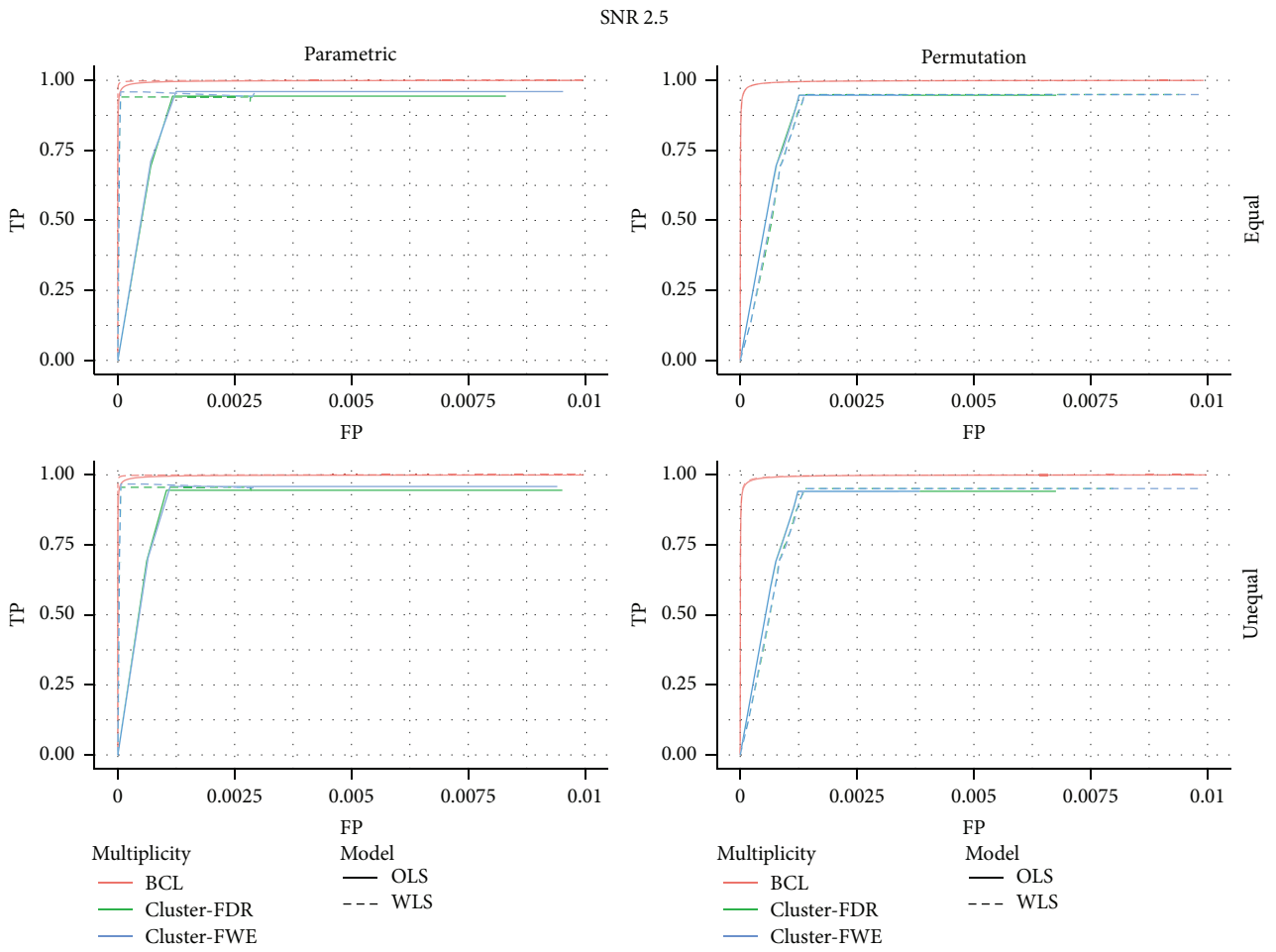


Figure 11: Receiver operating curve for a signal-to-noise ratio of 2.5 over the range [0; 0.01].

values, the FWE and FDR method results are identical. This matching on the number of selected voxels is motivated by the simulation findings that the larger number of selected voxels results in a higher MCC. In a test-retest setting, the MCC coincides with the correlation between two binary images (selected/nonselected voxels). In Figure 9 we see that indeed the BCL outperforms the FDR/FWE and that the WLS outperforms the OLS. We note however that this methods has a major drawback as it does not allow us to calculate the variability on these numbers and it requires a second sample.

5. Discussion

In this study we investigated both the balance between true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) and data-analytical stability of methodological choices in the second-level analysis of fMRI data. Following the traditional evaluation of techniques in the fMRI literature, we first focused on the balance between FP and TP, using ROC-curves, and on the Matthews correlation coefficient (MCC), a measure that takes all possible decisions into account. Aiming for more reproducible brain imaging research, we believe however that data-analytical stability is also an important criterion that offers an additional unique perspective on the behavior of methods. While studies using the criterion of data-analytical stability are sparse and mostly focused on the first-level inferential decisions (e.g., [4, 52], for, resp., a focus on mass univariate inference and topological inference), this study filled this gap through considering data-analytical stability of different methods at the second-level analysis. Unlike the NPAIRS framework [53, 54] that allows exploring overall stability, we furthermore focused on the *selected* voxels, obtained via thresholded images, when assessing the data-analytical stability.

More specifically, we assessed in this paper the impact of three different choices that the researcher has to make when analyzing fMRI data at the second level: (1) should one use a WLS-approach or an OLS-approach, (2) should one rely on parametric assumptions for the test statistic or rely on a nonparametric framework, such as permutation-based inference, and (3) which type of control should one use to limit the multiplicity issue. The impact of these choices was assessed from the ROC-curves, MCC, and the data-analytical stability perspective.

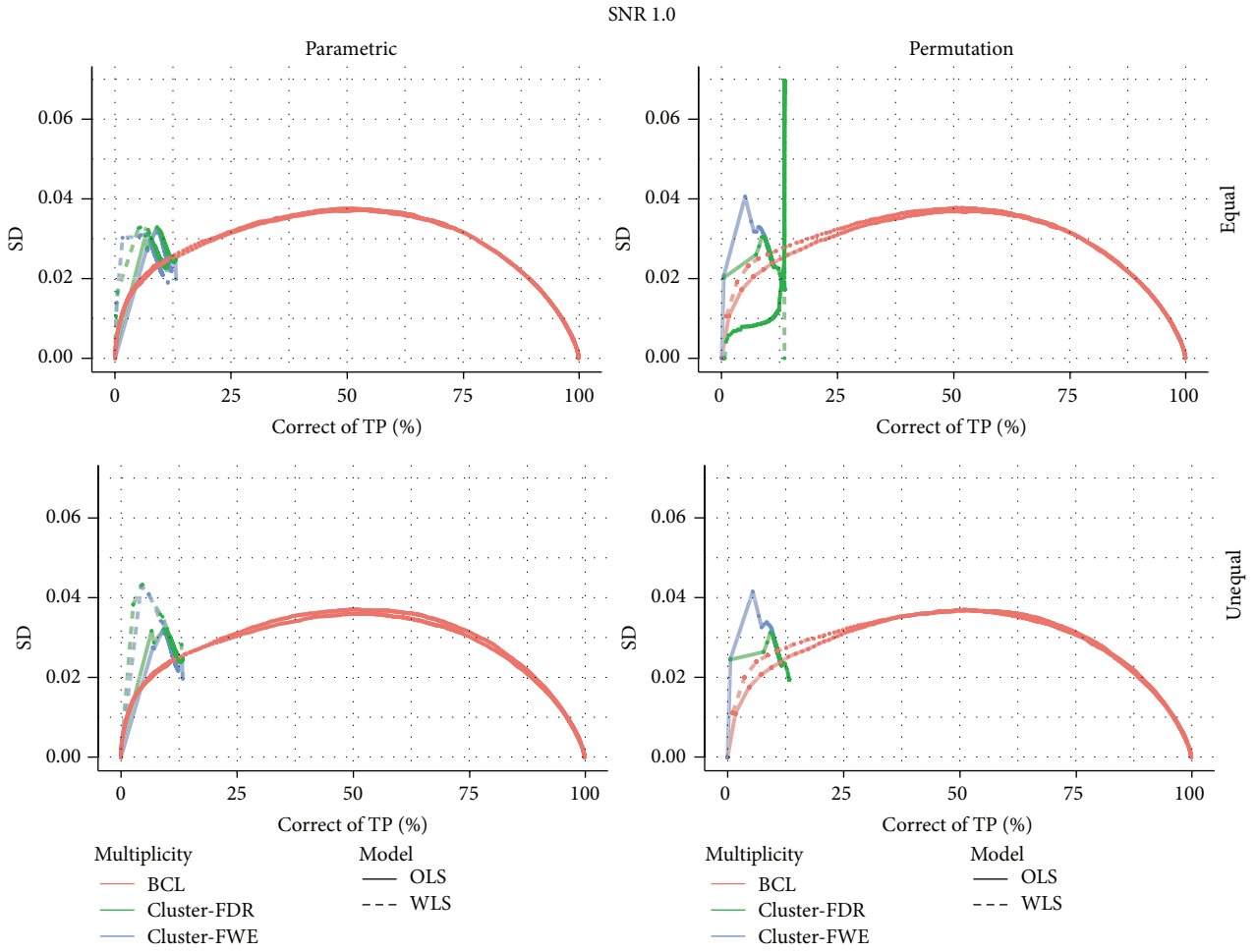


Figure 12: % of correctly activated voxels with their standard deviation for a signal-to-noise ratio of 1.

For the balance in the decision context, based on the ROC-curves and the MCC, results were pretty clear when parametric inference is used. Regardless of the choice of the multiple testing correction, we found that the WLS-method yields a better balance between FP and TP than when the OLS-method is used. While the MCCs confirmed most of the results based on the ROC-curves, they revealed the fact that differences are more obvious when the SNR was low. Under the high signal strength, the balance in the decision context did not diverge remarkably between methods. These findings on the balance between FP and TP are in line with Mumford and Nichols [18], although the magnitude of the difference between WLS and OLS was more pronounced, based on the ROC-curves, in our simulation study. When permutation-based inference is used, there were barely any differences between OLS and WLS. We found however that there were some effects of discreteness when permutation-based inference was used in combination with WLS. In the simulation settings this was associated with spiky patterns under a high SNR due to substantial jumps in the number of voxels that are selected. But also in the real data application, we found some evidence for discreteness with the WLS statistic when jumps in the activation occur. When comparing the parametric with the nonparametric approach, we found in contrast to Thirion et al. [43] no evidence for a better performance of permutation based inference. Note however that in all our simulation settings the basic assumptions of parametric inference were satisfied (Gaussian noise and sufficient smoothing). Upon inspection of the ROC-curves we also found in our simulation study that the two-step procedure, which ignores multiplicity first but requires a minimal cluster size next, outperforms the traditional FWE-control and FDR-control.

From a data-analytical stability perspective, there were substantial differences between the three approaches we considered for multiple testing correction. In line with previous findings at the first level of analysis [21, 52], FDR-based corrections for multiple testing resulted in more variable selections. In both the simulation study and the real data application, we found that FWE based correction for multiple testing and a two-step procedure result in more stable results, as assessed by the variability on the number of selected voxels. This weaker performance of the FDR is observed, regardless of the WLS-approach versus OLS-approach, or the

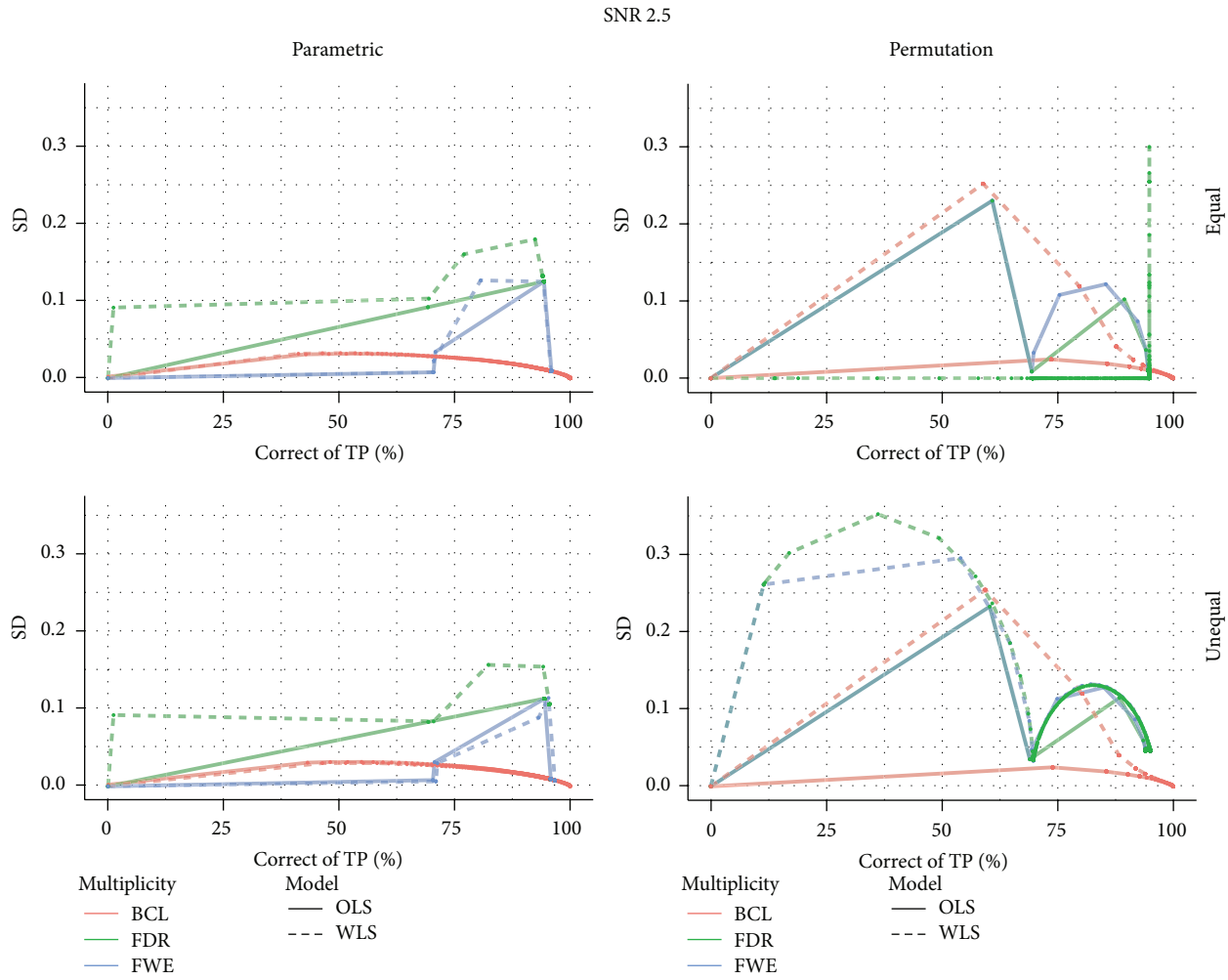


Figure 13: % of correctly activated voxels with their standard deviation for a signal-to-noise ratio of 2.5.

parametric versus nonparametric framework for inference. Interestingly, when we focused on the reselection rate of a specific voxel in the data application, we also found superior performance of the two-step procedure. As noted by one of the reviewers, the increased stability for the FWE and two-step procedures relying on parametrical inference might be attributed to the fact that these approaches exploit topological features of the data in contrast to the FDR.

While voxel-based inference is only one approach to controlling for multiple testing, several alternatives exist. Cluster-based inference (see, e.g., [35, 36]) is a very popular alternative that relies explicitly on topological features such as the cluster size and has been advocated because of the potential increase in power. However, Woo et al. [16] showed that the commonly used two-step procedure for cluster-based inference is nonrobust when too liberal first thresholds are used at the voxel level and that this results in unpractically large clusters when studies are sufficiently powered. This complicates the interpretation of the results as clusters could become as large as half of the hemisphere. In the same vein, Woo et al. [16] and Nichols [9] argue that the conceptual definition of a "significant cluster" is complicated by the fact that it is a randomly sized collection of voxels of which one can only claim that at least some are significant. We concur with Nichols [9] and Woo et al. [16] that voxel-wise inference remains a useful alternative and therefore opted for an extensive evaluation of commonly used voxel-based inference techniques.

The FP rates are evaluated only in a simulation study. While this might lack biological validity, this procedure allows us to have strict control on the ground truth and consequent determination of TN and TP. With an exhaustive simulation study (2 SNR and varying within-subject variability assumptions), we have covered some of the properties present in real data. Any simulation study comes naturally with the arbitrariness of these settings. However, compared to using real data to determine FP rates, simulation studies have the advantage to exclude unnecessary artifacts in

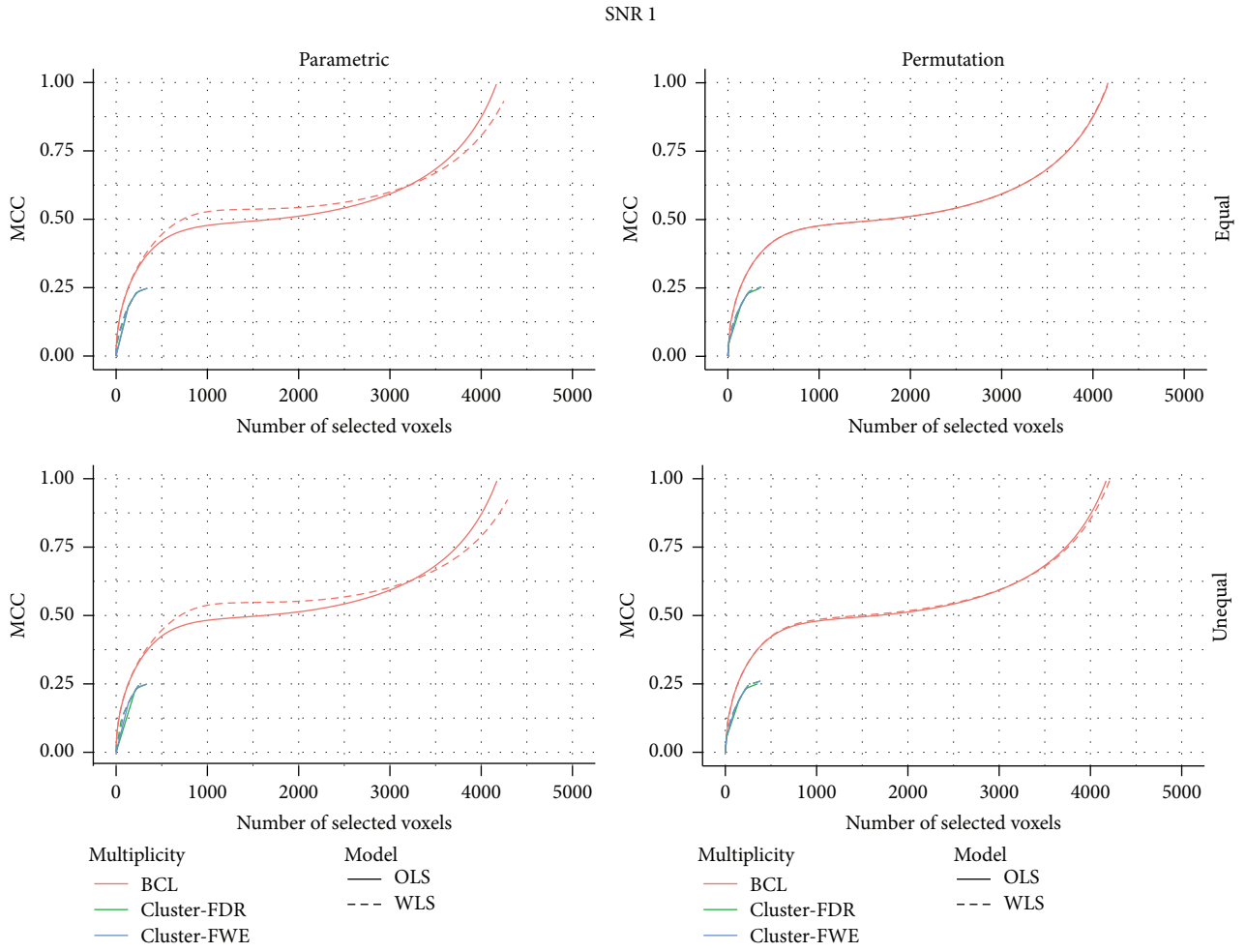


Figure 14: MCC for a signal-to-noise ratio of 1.

the procedure to determine the TP and the TN (see, e.g., [55], for differences in test errors based on the design) or its underlying assumptions.

Gathering all the above-described evidence, we would recommend the brain imaging researcher to use WLS at the second level in combination with the two-step procedure, hereby relying on the parametric framework for inference. Note that throughout the paper, we have assumed that all images at the first level are correctly normalized such that individuals are perfectly coregistered. It should be stressed that further exploration of the robustness against violations of the parametric assumptions is warranted. However, the proposed strategy in this paper to assess data-analytical stability of different methods on real data could be used in any future application and ultimately reveal the best choice from a data-analytical stability perspective in practice. Such validation on real data may also yield further insight into the appropriateness of the rather ad hoc but commonly used BCL-approach which lacks inferential justification.

Appendices

A. Additional Details HCP Dataset

Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research and by the McDonnell Center for Systems Neuroscience at Washington University.

The list of subject identifiers used in the real data application of this study can be found in

$$\begin{array}{lll} 100408 & 101915 & 103414 \\ 105115 & 106016 & 110411 \\ 111312 & 111716 & 113619 \\ 115320 & 117122 & 118730 \\ 118932 & 120111 & 122317 \end{array} \quad \text{(A.1)}$$

Subjects come from the 80 unrelated subjects dataset, release *Q*3 [50].

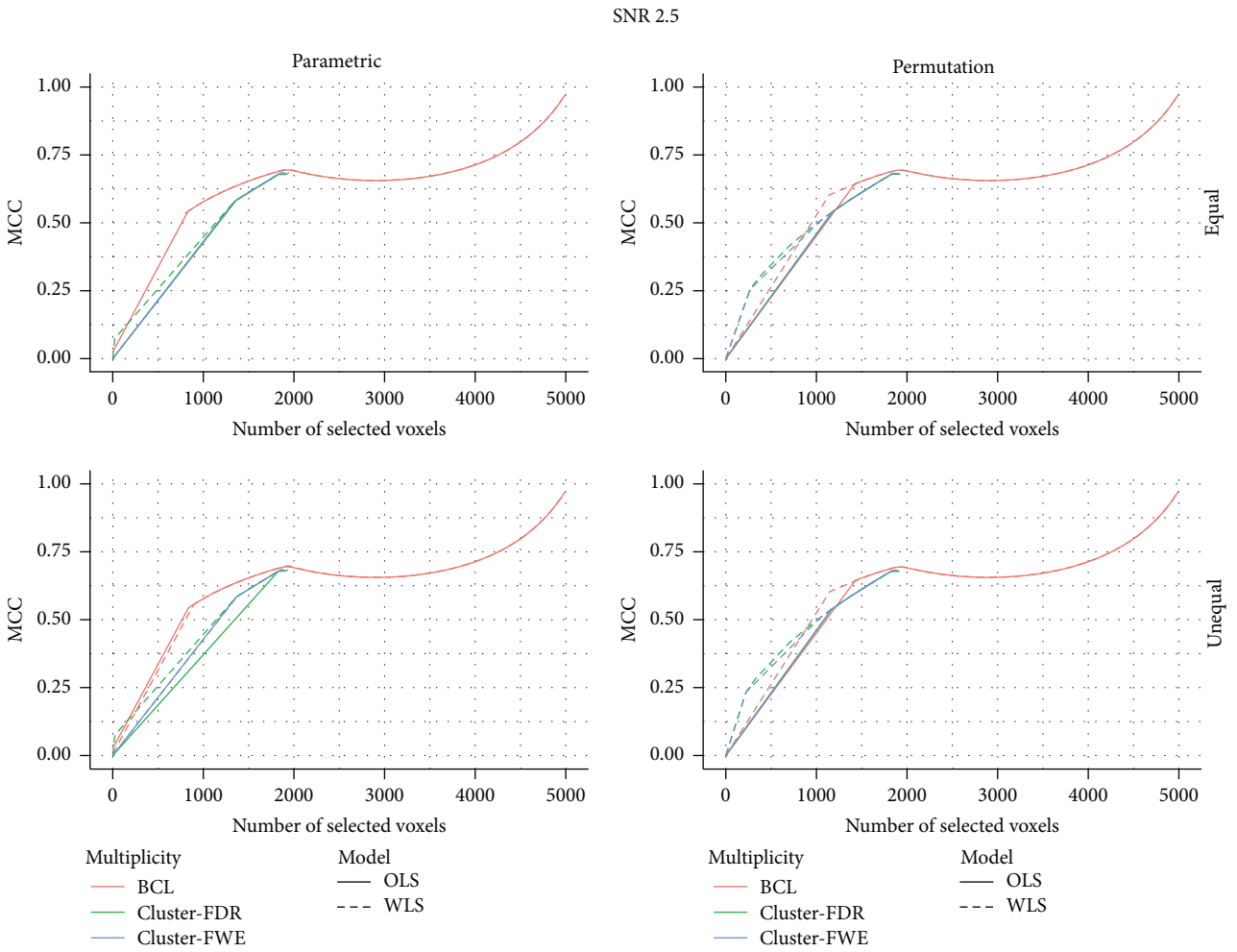


Figure 15: MCC for a signal-to-noise ratio of 2.5.

B. Additional Figures on the Relationship between Cluster-Based Inference and the Uncorrected Threshold Method with Minimal Cluster Size

This section contains the additional figures in which the BCL procedures are compared with cluster-based inference procedures. For all of the following pictures the following abbreviations are used: (1) SNR = 1: low signal strength, SNR = 2.5: high signal strength; (2) cluster-FWE: familywise error correction based on cluster-size inference, cluster-FDR: False Discovery Rate correction based on cluster-size inference, and BCL: two-step procedure with a Bonferroni-like first threshold and minimal cluster size of 10; (3) OLS: Ordinary Least Squares approach and WLS: Weighted Least Squares approach; (4) unequal: differences in the subject-specific variability, equal: identical subject-specific variability.

B.1. ROC-Curves. In Figures 10 and 11 the voxel-based ROC-curves are depicted.

B.2. Stability on the Percentage of TPs. In Figures 12 and 13 the voxel-based stability plots are depicted.

B.3. MCC. In Figures 14 and 15 the voxel-based stability plots are depicted.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation, and the Flemish Government department EWI.

References

[1] K. J. Friston, "Functional integration," in *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, K. J. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, Eds., chapter 36, Academic Press, Elsevier Science, 2007.

[2] S. Vahdat, M. Maneshi, C. Grova, J. Gotman, and T. E. Milner, "Shared and specific independent components analysis for between-group comparison," *Neural Computation*, vol. 24, no. 11, pp. 3052–3090, 2012.

[3] M. A. Lindquist, "The statistical analysis of fMRI data," *Statistical Science*, vol. 23, no. 4, pp. 439–464, 2008.

[4] S. P. Roels, H. Bossier, T. Loeys, and B. Moerkerke, "Data-analytical stability of cluster-wise and peak-wise inference in fMRI data analysis," *Journal of Neuroscience Methods*, vol. 240, pp. 37–47, 2015.

[5] J. Carp, "The secret lives of experiments: methods reporting in the fMRI literature," *NeuroImage*, vol. 63, no. 1, pp. 289–300, 2012.

[6] C. F. Beckmann, M. Jenkinson, and S. M. Smith, "General multilevel linear modeling for group analysis in FMRI," *NeuroImage*, vol. 20, no. 2, pp. 1052–1063, 2003.

[7] J. A. Mumford and T. Nichols, "Modeling and inference of multisubject fMRI data," *IEEE Engineering in Medicine and Biology Magazine*, vol. 25, no. 2, pp. 42–51, 2006.

[8] T. Nichols and S. Hayasaka, "Controlling the familywise error rate in functional neuroimaging: a comparative review," *Statistical Methods in Medical Research*, vol. 12, no. 5, pp. 419–446, 2003.

[9] T. E. Nichols, "Multiple testing corrections, nonparametric methods, and random field theory," *NeuroImage*, vol. 62, no. 2, pp. 811–815, 2012.

[10] A. P. Holmes, R. C. Blair, J. D. G. Watson, and I. Ford, "Nonparametric analysis of statistic images from functional mapping experiments," *Journal of Cerebral Blood Flow and Metabolism*, vol. 16, no. 1, pp. 7–22, 1996.

[11] T. E. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: a primer with examples," *Human Brain Mapping*, vol. 15, no. 1, pp. 1–25, 2002.

[12] C. M. Bennett, G. L. Wolford, and M. B. Miller, "The principled control of false positives in neuroimaging," *Social Cognitive and Affective Neuroscience*, vol. 4, no. 4, Article ID nsp053, pp. 417–422, 2009.

[13] M. D. Lieberman and W. A. Cunningham, "Type I and Type II error concerns in fMRI research: re-balancing the scale," *Social Cognitive and Affective Neuroscience*, vol. 4, no. 4, Article ID nsp052, pp. 423–428, 2009.

[14] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[15] C. R. Genovese, N. A. Lazar, and T. Nichols, "Thresholding of statistical maps in functional neuroimaging using the false discovery rate," *NeuroImage*, vol. 15, no. 4, pp. 870–878, 2002.

[16] C.-W. Woo, A. Krishnan, and T. D. Wager, "Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations," *NeuroImage*, vol. 91, pp. 412–419, 2014.

[17] A. Holmes and K. Friston, "Generalisability, random effects and population inference," *NeuroImage*, vol. 7, p. S754, 1998.

[18] J. A. Mumford and T. Nichols, "Simple group fMRI modeling and inference," *NeuroImage*, vol. 47, no. 4, pp. 1469–1475, 2009.

[19] M. Wilke, "An iterative jackknife approach for assessing reliability and power of fMRI group analyses," *PLoS ONE*, vol. 7, no. 4, Article ID e35578, 2012.

[20] K. J. Gorgolewski, A. J. Storkey, M. E. Bastin, I. Whittle, and C. Pernet, "Single subject fMRI test-retest reliability metrics and confounding factors," *NeuroImage*, vol. 69, pp. 231–243, 2013.

[21] X. Qiu, Y. Xiao, A. Gordon, and A. Yakovlev, "Assessing stability of gene selection in microarray data analysis," *BMC Bioinformatics*, vol. 7, article 50, 2006.

[22] S. Kiebel and A. P. Holmes, "The general linear model," in *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, K. J. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, Eds., chapter 8, Academic Press, Elsevier Science, 2007.

[23] J.-B. Poline and M. Brett, "The general linear model and fMRI: does love last forever?" *NeuroImage*, vol. 62, no. 2, pp. 871–880, 2012.

[24] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: a general linear approach," *Human Brain Mapping*, vol. 2, no. 4, pp. 189–210, 1994.

[25] K. J. Worsley, C. H. Liao, J. Aston et al., "A general statistical analysis for fMRI data," *NeuroImage*, vol. 15, no. 1, pp. 1–15, 2002.

[26] R. Henson and K. Friston, "Convolution models for FMRI," in *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, Eds., pp. 193–210, Academic Press, London, UK, 2007.

[27] D. Cochrane and G. H. Orcutt, "Application of least squares regression to relationships containing auto-correlated error terms," *Journal of the American Statistical Association*, vol. 44, no. 245, pp. 32–61, 1949.

[28] M. Kutner, C. Nachtsheim, J. Neter, and W. Li, *Applied Linear Models*, McGraw-Hill Irwin, New York, NY, USA, 2005.

[29] M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.

[30] Wellcome Trust Centre for Neuroimaging U.C.L, Spm8, 2010, http://www.fil.ion.ucl.ac.uk/spm/.

[31] R. W. Cox, "AFNI: software for analysis and visualization of functional magnetic resonance neuroimages," *Computers and Biomedical Research*, vol. 29, no. 3, pp. 162–173, 1996.

[32] G. Chen, Z. S. Saad, A. R. Nath, M. S. Beauchamp, and R. W. Cox, "FMRI group analysis combining effect estimates and their variances," *NeuroImage*, vol. 60, no. 1, pp. 747–765, 2012.

[33] G. Chen, Z. S. Saad, J. C. Britton, D. S. Pine, and R. W. Cox, "Linear mixed-effects modeling approach to FMRI group analysis," *NeuroImage*, vol. 73, pp. 176–190, 2013.

[34] M. Brett, W. Penny, and S. Kiebel, "Parametric procedures," in *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, K. J. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, Eds., chapter 8, Academic Press Inc, Elsevier Science, 2007.

[35] K. J. Friston, A. Holmes, J.-B. Poline, C. J. Price, and C. D. Frith, "Detecting activations in pet and fMRI: levels of inference and power," *NeuroImage*, vol. 4, no. 3, pp. 223–235, 1996.

[36] S. Hayasaka and T. E. Nichols, "Validating cluster size inference: random field and permutation methods," *NeuroImage*, vol. 20, no. 4, pp. 2343–2356, 2003.

[37] K. J. Worsley, A. C. Evans, S. Marrett, and P. Neelin, "A three-dimensional statistical analysis for CBF activation studies in human brain," *Journal of Cerebral Blood Flow and Metabolism*, vol. 12, no. 6, pp. 900–918, 1992.

[38] W.-L. Luo and T. E. Nichols, "Diagnosis and exploration of massively univariate neuroimaging models," *NeuroImage*, vol. 19, no. 3, pp. 1014–1032, 2003.

[39] M. M. Monti, "Statistical analysis of fMRI time-series: a critical review of the GLM approach," *Frontiers in Human Neuroscience*, vol. 5, article 28, 2011.

[40] O. Friman and C.-F. Westin, "Resampling fMRI time series," *NeuroImage*, vol. 25, no. 3, pp. 859–867, 2005.

[41] P. Bellec, P. Rosa-Neto, O. C. Lyttelton, H. Benali, and A. C. Evans, "Multi-level bootstrap analysis of stable clusters in resting-state fMRI," *NeuroImage*, vol. 51, no. 3, pp. 1126–1139, 2010.

[42] S. P. Roels, B. Moerkerke, and T. Loeys, "Bootstrapping fmri data: dealing with misspecification," *Neuroinformatics*, vol. 13, no. 3, pp. 337–352, 2015.

[43] B. Thirion, P. Pinel, S. Mériaux, A. Roche, S. Dehaene, and J.-B. Poline, "Analysis of a large fMRI cohort: statistical and methodological issues for group analyses," *NeuroImage*, vol. 35, no. 1, pp. 105–120, 2007.

[44] D. Adolf, S. Weston, S. Baecke, M. Luchtmann, J. Bernarding, and S. Kropf, "Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method," *Frontiers in Neuroinformatics*, vol. 8, article 72, 2014.

[45] A. M. Winkler, G. R. Ridgway, M. A. Webster, S. M. Smith, and T. E. Nichols, "Permutation inference for the general linear model," *NeuroImage*, vol. 92, pp. 381–397, 2014.

[46] B. Lenoski, L. C. Baxter, L. J. Karam, J. Maisog, and J. Debbins, "On the performance of autocorrelation estimation algorithms for fMRI analysis," *IEEE Journal on Selected Topics in Signal Processing*, vol. 2, no. 6, pp. 828–838, 2008.

[47] M. Welvaert and Y. Rosseel, "How ignoring physiological noise can bias the conclusions from fMRI simulation results," *Journal of Neuroscience Methods*, vol. 211, no. 1, pp. 125–132, 2012.

[48] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta—Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[49] M. Vihinen, "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis," *BMC genomics*, vol. 13, supplement 4, article S2, 2012.

[50] D. C. Van Essen, K. Ugurbil, E. Auerbach et al., "The human connectome project: a data acquisition perspective," *NeuroImage*, vol. 62, no. 4, pp. 2222–2231, 2012.

[51] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson et al., "The minimal preprocessing pipelines for the human connectome project," *NeuroImage*, vol. 80, pp. 105–124, 2013.

[52] J. Durnez, S. P. Roels, and B. Moerkerke, "Multiple testing in fMRI: an empirical case study on the balance between sensitivity, specificity, and stability," *Biometrical Journal*, vol. 56, no. 4, pp. 649–661, 2014.

[53] S. C. Strother, J. Anderson, L. K. Hansen et al., "The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework," *NeuroImage*, vol. 15, no. 4, pp. 747–771, 2002.

[54] S. Strother, S. La Conte, L. Kai Hansen et al., "Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics. I. A preliminary group analysis," *NeuroImage*, vol. 23, supplement 1, pp. S196–S207, 2004.

[55] A. Eklund, M. Andersson, C. Josephson, M. Johannesson, and H. Knutsson, "Does parametric fMRI analysis with SPM yield valid results? An empirical study of 1484 rest datasets," *NeuroImage*, vol. 61, no. 3, pp. 565–578, 2012.

Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2015, Article ID 813696, 16 pages
http://dx.doi.org/10.1155/2015/813696

Research Article

MRBrainS Challenge: Online Evaluation Framework for Brain Image Segmentation in 3T MRI Scans

Adriënne M. Mendrik,[1] Koen L. Vincken,[1] Hugo J. Kuijf,[1] Marcel Breeuwer,[2,3] Willem H. Bouvy,[4] Jeroen de Bresser,[5] Amir Alansary,[6] Marleen de Bruijne,[7,8] Aaron Carass,[9] Ayman El-Baz,[6] Amod Jog,[9] Ranveer Katyal,[10] Ali R. Khan,[11,12] Fedde van der Lijn,[7] Qaiser Mahmood,[13] Ryan Mukherjee,[14] Annegreet van Opbroek,[7] Sahil Paneri,[10] Sérgio Pereira,[15] Mikael Persson,[13] Martin Rajchl,[11,16] Duygu Sarikaya,[17] Örjan Smedby,[18,19] Carlos A. Silva,[15] Henri A. Vrooman,[7] Saurabh Vyas,[14] Chunliang Wang,[18,19] Liang Zhao,[17] Geert Jan Biessels,[4] and Max A. Viergever[1]

[1] *Image Sciences Institute, University Medical Center Utrecht, 3584 CX Utrecht, Netherlands*
[2] *Philips Healthcare, 5680 DA Best, Netherlands*
[3] *Faculty of Biomedical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, Netherlands*
[4] *Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, 3584 CX Utrecht, Netherlands*
[5] *Department of Radiology, University Medical Center Utrecht, 3584 CX Utrecht, Netherlands*
[6] *BioImaging Laboratory, Bioengineering Department, University of Louisville, Louisville, KY 40292, USA*
[7] *Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus MC, 3015 CN Rotterdam, Netherlands*
[8] *Department of Computer Science, University of Copenhagen, 2100 Copenhagen, Denmark*
[9] *Image Analysis and Communications Laboratory, Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, USA*
[10] *Department of Electronics and Communication Engineering, The LNM Institute of Information Technology, Jaipur 302031, India*
[11] *Imaging Laboratories, Robarts Research Institute, London, ON, Canada N6A 5B7*
[12] *Department of Medical Biophysics, Western University, London, ON, Canada N6A 3K7*
[13] *Signals and Systems, Chalmers University of Technology, 41296 Gothenburg, Sweden*
[14] *Applied Physics Laboratory, Johns Hopkins University, Laurel, MD 20723, USA*
[15] *Department of Electronics, University of Minho, 4800-058 Guimarães, Portugal*
[16] *Department of Computing, Imperial College London, London SW7 2AZ, UK*
[17] *Computer Science and Engineering Department, State University of New York at Buffalo, Buffalo, NY 14260-2500, USA*
[18] *Center for Medical Imaging Science and Visualization, Linköping University, 58185 Linköping, Sweden*
[19] *Department of Radiology and Department of Medical and Health Sciences, Linköping University, 58185 Linköping, Sweden*

Correspondence should be addressed to Adriënne M. Mendrik; a.m.mendrik@gmail.com

Received 10 July 2015; Accepted 19 August 2015

Academic Editor: Jussi Tohka

Many methods have been proposed for tissue segmentation in brain MRI scans. The multitude of methods proposed complicates the choice of one method above others. We have therefore established the MRBrainS online evaluation framework for evaluating (semi)automatic algorithms that segment gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) on 3T brain MRI scans of elderly subjects (65–80 y). Participants apply their algorithms to the provided data, after which their results are evaluated and ranked. Full manual segmentations of GM, WM, and CSF are available for all scans and used as the reference standard. Five datasets are provided for training and fifteen for testing. The evaluated methods are ranked based on their overall performance to

segment GM, WM, and CSF and evaluated using three evaluation metrics (Dice, H95, and AVD) and the results are published on the MRBrainS13 website. We present the results of eleven segmentation algorithms that participated in the MRBrainS13 challenge workshop at MICCAI, where the framework was launched, and three commonly used freeware packages: FreeSurfer, FSL, and SPM. The MRBrainS evaluation framework provides an objective and direct comparison of all evaluated algorithms and can aid in selecting the best performing method for the segmentation goal at hand.

1. Introduction

Multiple large population studies [1–3] have shown the importance of quantifying brain structure volume, for example, to detect or predict small vessel disease and Alzheimer's disease. In clinical practice, brain volumetry can be of value in disease diagnosis, progression, and treatment monitoring of a wide range of neurologic conditions, such as Alzheimer's disease, dementia, focal epilepsy, Parkinsonism, and multiple sclerosis [4]. Automatic brain structure segmentation in MRI dates back to 1985 [5] and many methods have been proposed since then. However, the multitude of methods proposed [6–13] complicates the choice for a certain method above others. As early as 1986, Price [14] stressed the importance of comparing different approaches to the same type of problem. Various studies have addressed this issue and evaluated different brain structure segmentation methods [15–19]. However, several factors complicate direct comparison of different approaches. Not all algorithms are publicly available, and if they are, researchers who use them are generally not as experienced with these algorithms as they are with their own algorithm in terms of parameter tuning, which could result in a bias towards their own method. This problem does not exist when researchers apply their own method to publicly available data. Therefore, publicly available databases like the "Alzheimer's Disease Neuroimaging Initiative" (ADNI) (http://adni.loni.usc.edu/), the "Internet Brain Segmentation Repository" (IBSR) (http://www.nitrc.org/projects/ibsr), the CANDI Share Schizophrenia Bulletin 2008 (https://www.nitrc.org/projects/cs_schizbull08) [20], and Mindboggle (http://www.mindboggle.info/) [21] are important initiatives to enable comparison of various methods on the same data. However, due to the use of subsets of the available data and different evaluation measures, direct comparison can be problematic. To address this issue, grand challenges in biomedical image analysis were introduced in 2007 [22]. Participants in these competitions can apply their algorithms to the provided data, after which their results are evaluated and ranked by the organizers. Many challenges (http://grand-challenge.org/All_Challenges/) have been organized since then, providing an insight into the performance of automatic algorithms for specific tasks in medical image analysis.

In this paper we introduce the MRBrainS challenge evaluation framework (http://mrbrains13.isi.uu.nl/), an online framework to evaluate automatic and semiautomatic algorithms that segment gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) in 3T brain MRI scans of older (mean age 71) subjects with varying degrees of atrophy and white matter lesions. This framework has three main advantages. Firstly, researchers apply their own segmentation algorithms to the provided data. Parameters are optimally tuned to achieve the best possible performance. Secondly, all algorithms are applied to the exact same data and the reference standard of the test data is unknown to the participating researchers. Thirdly, the evaluation algorithm and measures are the same for all evaluated algorithms, enabling direct comparison of the various algorithms. The framework was launched at the MRBrainS13 challenge workshop at the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference on September 26th in 2013. Eleven teams participated in the challenge workshop with a wide variety of segmentation algorithms, the results for which are presented in this paper and provide a benchmark for the proposed evaluation framework. In addition, we evaluated three commonly used freeware packages on the evaluation framework: FreeSurfer (http://surfer.nmr.mgh.harvard.edu/) [23, 24], FSL (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) [25], and SPM (http://www.fil.ion.ucl.ac.uk/spm/) [26].

2. Materials and Methods

2.1. Evaluation Framework. The MRBrainS evaluation framework is set up as follows. Multisequence (T1-weighted, T1-weighted inversion recovery, and T2-weighted fluid attenuated inversion recovery) 3T MRI scans of twenty subjects are available for download on the MRBrainS website (http://mrbrains13.isi.uu.nl/). The data is described in more detail in Section 2.1.1. All scans were manually segmented into GM, WM, and CSF. These manual segmentations are used as the reference standard for the evaluation framework. The annotation process for obtaining the reference standard is described in Section 2.1.2. For five of the twenty datasets the reference standard is provided on the website and can be used for training an automatic segmentation algorithm. The remaining fifteen MRI datasets have to be segmented by the participating algorithms into GM, WM, and CSF. For these fifteen datasets, the reference standard is not provided online. The segmentation results can be submitted on the MRBrainS website. With each submission, a short description of the segmentation algorithm has to be provided, which should at least describe the algorithm, the used MRI sequences, whether the algorithm is semi- or fully automatic, and the average runtime of the algorithm. The segmentation results are then evaluated (Section 2.1.3) and ranked (Section 2.1.4) by the organizers and the results are presented on the website. More information on how to use the evaluation framework is provided in the details section of the MRBrainS website (http://mrbrains13.isi.uu.nl/details.php).

2.1.1. Data. The focus was on brain segmentation in the context of ageing. Twenty subjects (mean age ± SD = 71 ± 4 years, 10 male, 10 female) were selected from an ongoing

cohort study of older (65–80 years of age) functionally independent individuals without a history of invalidating stroke or other brain diseases [27]. This study was approved by the local ethics committee of the University Medical Center Utrecht (Netherlands) and all participants signed an informed consent form. To be able to test the robustness of the segmentation algorithms in the context of ageing-related pathology, the subjects were selected to have varying degrees of atrophy and white matter lesions. Scans with major artefacts were excluded. MRI scans were acquired on a 3.0 T Philips Achieva MR scanner at the University Medical Center Utrecht (Netherlands). The following sequences were acquired and used for the evaluation framework: 3D T1 (TR: 7.9 ms, TE: 4.5 ms), T1-IR (TR: 4416 ms, TE: 15 ms, and TI: 400 ms), and T2- FLAIR (TR: 11000 ms, TE: 125 ms, and TI: 2800 ms). Since the focus of the MRBrainS evaluation framework is on comparing different segmentation algorithms, we performed two preprocessing steps to limit the influence of different registration and bias correction algorithms on the segmentation results. The sequences were aligned by rigid registration using Elastix [28] and bias correction was performed using SPM8 [29]. After registration, the voxel size within all provided sequences (T1, T1 IR, and T2 FLAIR) was $0.96 \times 0.96 \times 3.00\,\text{mm}^3$. The original 3D T1 sequence (voxel size: $1.0 \times 1.0 \times 1.0\,\text{mm}^3$) was provided as well. Five datasets that were representative for the overall data (2 male, 3 female, varying degrees of atrophy and white matter lesions) were selected for training. The remaining fifteen datasets are provided as test data.

2.1.2. Reference Standard. Manual segmentations were performed to obtain a reference standard for the evaluation framework. All axial slices of the 20 datasets ($0.96 \times 0.96 \times 3.00\,\text{mm}^3$) were manually segmented by trained research assistants in a darkened room with optimal viewing conditions. All segmentations were checked and corrected by three experts: a neurologist in training, a neuroradiologist in training, and a medical image processing scientist. To perform the manual segmentations, an in-house developed tool based on MeVisLab (MeVis Medical Solutions AG, Bremen, Germany) was used, employing a freehand spline drawing technique [30]. The closed freehand spline drawing technique was used to delineate the outline of each brain structure starting at the innermost structures (Figure 1(a)), working outward. The closed contours were converted to hard segmentations, and the inner structures were iteratively subtracted from the outer structures to construct the final hard segmentation image (Figure 1(b)). The following structures were segmented and are available for training: cortical gray matter (1), basal ganglia (2), white matter (3), white matter lesions (4), peripheral cerebrospinal fluid (5), lateral ventricles (6), cerebellum (7), and brainstem (8). These structures can be merged into gray matter (1, 2), white matter (3, 4), and cerebrospinal fluid (5, 6). The cerebellum and brainstem are excluded from the evaluation. All structures were segmented on the T1-weighted scans that were registered to the FLAIR scans, except for the white matter lesions (WMLs) and the CSF outer border (used to determine the intracranial volume). The WMLs were segmented on the FLAIR scan by the neurologist in training and checked and corrected by the neuroradiologist in training. The CSF outer border was segmented using both the T1-weighted and the T1-weighted IR scan, since the T1-weighted IR scan shows higher contrast at the borders of the intracranial volume. The CSF segmentation includes all vessels (including the superior sagittal sinus and the transverse sinuses) and nonbrain structures such as the cerebral falx and choroid plexuses.

2.1.3. Evaluation. To evaluate the segmentation results we use three types of measures: a spatial overlap measure, a boundary distance measure, and a volumetric measure. The Dice [31] coefficient is used to determine the spatial overlap and is defined as

$$D = \frac{2\,|A \cap G|}{|A| + |G|} \cdot 100, \tag{1}$$

where A is the segmentation result, G is the reference standard, and D is the Dice expressed as percentages. The 95th-percentile of the Hausdorff distance is used to determine the distance between the segmentation boundaries. The conventional Hausdorff distance uses the maximum, which is very sensitive to outliers. To correct for outliers, we use the 95th-percentile of the Hausdorff distance, by selecting the Kth ranked distance as proposed by Huttenlocher et al. [32]:

$$h_{95}(A, G) = {}^{95}K^{\text{th}}_{a \in A} \min_{g \in G} \|g - a\|, \tag{2}$$

where ${}^{95}K^{\text{th}}_{a \in A}$ is the Kth ranked minimum Euclidean distance with $K/N_a = 95\%$, A is the set of boundary points $\{a_1, \ldots, a_{N_a}\}$ of the segmentation result, and G is the set of boundary points $\{g_1, \ldots, g_{N_g}\}$ of the reference standard. The 95th-percentile of the Hausdorff distance is defined as

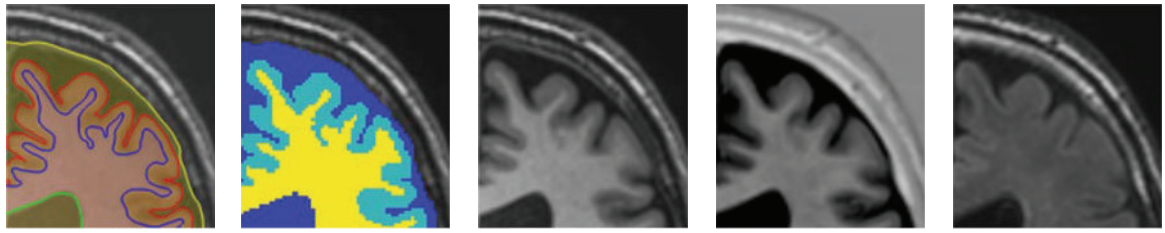
$$H_{95}(A, G) = \max\left(h_{95}(A, G), h_{95}(G, A)\right). \tag{3}$$

The third measure is the percentage absolute volume difference, defined as

$$\text{AVD} = \frac{\left|V_a - V_g\right|}{V_g} \cdot 100, \tag{4}$$

where V_a is the volume of the segmentation result and V_g is the volume of the reference standard. These measures are used to evaluate the following brain structures in each of the fifteen test datasets: GM, WM, CSF, brain (GM + WM), and intracranial volume (GM + WM + CSF). The brainstem and cerebellum are excluded from the evaluation.

2.1.4. Ranking. To compare the segmentation algorithms that participate in the MRBrainS evaluation framework, the algorithms are ranked based on their overall performance to segment GM, WM, and CSF. Each of these components ($C = \{\text{GM}, \text{WM}, \text{CSF}\}$) is evaluated by using the three evaluation measures ($M = \{D, H_{95}, \text{AVD}\}$) described in Section 2.1.3.



(a) (b) (c) (d) (e)

Figure 1: Example of the manually drawn contours (a), the resulting hard segmentation map (GM: light blue, WM: yellow, and CSF: dark blue) that is used as the reference standard (b), the T1-weighted scan (c), the T1-weighted inversion recovery (IR) scan (d), and the T2-weighted fluid attenuated inversion recovery (FLAIR) scan (e).

For each component $c \in C$ and each evaluation measure $m \in M$, the mean and standard deviation are determined over all 15 test datasets. The segmentation algorithms are then sorted on the mean D value in descending order and on the mean H_{95} and AVD value in ascending order. Each segmentation algorithm receives a rank (r) between 1 (ranked best) and n (number of participating algorithms) for each component c and each evaluation measure m. The final ranking is based on the overall score of each algorithm, which is the sum over all ranks, defined as

$$s = \sum_{m=0}^{|M|} \sum_{c=0}^{|C|} r_{mc}, \tag{5}$$

where r_{mc} is the rank of the segmentation algorithm for measure m of component c. For the final ranking r, the overall scores s are sorted in ascending order and ranked from 1 to n. In case two or more algorithms have equal scores, the standard deviation over all 15 test datasets is taken into account to determine the final rank. The segmentation algorithms are then sorted on the standard deviation in ascending order and ranked for each component c and each evaluation measure m. The overall score is determined using (5) and the algorithms are sorted based on this score in ascending order and ranked from 1 to n. The algorithms that have equal overall scores based on the mean value are then ranked based on this standard deviation rank.

2.2. Evaluated Methods. The evaluation framework described in Section 2.1 was launched at the MRBrainS13 challenge workshop at the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference on September 26th in 2013. For the workshop challenge, the test datasets were split into twelve off-site and three on-site test datasets. For the off-site part, teams could register on the MRBrainS website (http://mrbrains13.isi.uu.nl/) and download the five training and twelve test datasets. A time slot of eight weeks was available for teams to download the data, train their algorithms, segment the test datasets, and submit their results on the website. Fifty-eight teams downloaded the data, of which twelve submitted their segmentation results. The evaluation results were reported to the twelve teams and all teams submitted a workshop paper to the MRBrainS13 challenge workshop at MICCAI. Eleven teams presented their results at the workshop and segmented the three on-site test datasets live at the workshop within a time slot of 3.5 hours. These algorithms provide a benchmark for the proposed evaluation framework and are briefly described in Sections 2.2.1–2.2.11 in alphabetical order of teams' names. The teams' names are used in the paper to identify the methods. For a full description of the methods we refer to the workshop papers [33–43]. In Section 2.2.12 we describe the evaluated freeware packages.

2.2.1. BIGR2 [37]. This multifeature SVM [37] method classifies voxels by using a Support Vector Machine (SVM) classifier [44] with a Gaussian kernel. Besides spatial features and intensity information from all three MRI sequences, the SVM classifier incorporates Gaussian-scale-space features to facilitate a smooth segmentation. Skull stripping is performed by nonrigid registration of the masks of the training images to the target image.

2.2.2. Bigr_neuro [41]. This auto-kNN [45] method is based on an automatically trained kNN-classifier. First, a probabilistic tissue atlas is generated by nonrigidly registering the manually annotated atlases to the subject of interest. Training samples are obtained by thresholding the probabilistic atlas and subsequently pruning the feature space. White matter lesions are detected by applying an adaptive threshold, determined from the tissue segmentation, to the FLAIR sequence.

2.2.3. CMIV [43]. A statistical-model-guided level-set method is used to segment the skull, brain ventricles, and basal ganglia. Then a skeleton-based model is created by extracting the midsurface of the gray matter and defining the thickness. This model is incorporated into a level-set framework to guide the cortical gray matter segmentation. The coherent propagation algorithm [46] is used to accelerate the level-set evolution.

2.2.4. Jedi Mind Meld [42]. This method starts by preprocessing the data via anisotropic diffusion. For each 2D slice of a labeled dataset, the canny edge pixels are extracted, and the Tourist Walk is computed. This is done for axial, sagittal, and coronal views. Machine learning is used with these features to automatically label edge pixels in an unlabeled dataset.

Finally, these labels are used by the Random Walker for automatic segmentation.

2.2.5. LNMBrains [35]. The voxel intensities of all MRI sequences are modelled as a Gaussian distribution for each label. The parameters of the Gaussian distributions are evaluated as maximum likelihood estimates and the posterior probability of each label is determined by using Bayesian estimation. A feature set consisting of regional intensity, texture, spatial location of voxels, and the posterior probability estimates is used to classify each voxel into CSF, WM, GM, or background by using a multicategory SVM classifier.

2.2.6. MNAB [38]. This method uses Random Decision Forests to classify the voxels into GM, WM, and CSF. It starts by a skull stripping procedure, followed by an intensity normalization of each MRI sequence. Feature extraction is then performed on the intensities, posterior probabilities, neighborhood statistics, tissue atlases, and gradient magnitude. After classification, isolated voxels are removed by postprocessing.

2.2.7. Narsil [34]. This is a model-free algorithm that uses ensembles of decision trees [47] to learn the mapping from image features to the corresponding tissue label. The ensembles of decision trees are constructed from corresponding image patches of the provided T1 and FLAIR scans with manual segmentations. The N3 algorithm [48] was used for additional inhomogeneity correction and SPECTRE [49] was used for skull stripping.

2.2.8. Robarts [39]. Multiatlas registration [50] with the T1 training images was used to propagate labels to generate sample histograms in a log-likelihood intensity model and probabilistic shape priors. These were employed in a MAP data term and regularized via computation of a hierarchical max-flow [51]. A brain mask from registration of the T1-IR training images was used to obtain the final results.

2.2.9. S2_QM [36]. This method [52] is based on Bayesian-based adaptive mean shift and the voxel-weighted K-means algorithm. The former is used to segment the brain into a large number of clusters or modes. The latter is employed to assign these clusters to one of the three components: WM, GM, or CSF.

2.2.10. UB VPML Med [40]. This method creates a multiatlas by registering the training images to the subject image and then propagating the corresponding labels to a fully connected graph on the subject image. Label fusion then combines the multiple labels into one label at each voxel with intensity similarity based weighted voting. Finally the method clusters the graph using multiway cut in order to achieve the final segmentation.

2.2.11. UofL BioImaging [33]. This is an automated MAP-based method aimed at unsupervised segmentation of different brain tissues from T1-weighted MRI. It is based on the integration of a probabilistic shape prior, a first-order intensity model using a Linear Combination of Discrete Gaussians (LCDG), and a second-order appearance model. These three features are integrated into a two-level joint Markov-Gibbs Random Field (MGRF) model of T1-MR brain images. Skull stripping was performed using BET2 [40] followed by an adaptive threshold-based technique to restore the outer border of the CSF using both T1 and T1-IR; this technique was not described in [33], due to a US patent application [53], but is described in [54]. This method was applied semiautomatically to the MRBrainS test data, due to per scan parameter tuning.

2.2.12. Freeware Packages. Next to the methods evaluated at the workshop, we evaluated three commonly used freeware packages for MR brain image segmentation: FreeSurfer (http://surfer.nmr.mgh.harvard.edu/) [23, 24], FSL (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) [25], and SPM12 (http://www.fil.ion.ucl.ac.uk/spm/) [26]. All packages were applied using the default settings, unless mentioned otherwise. FreeSurfer (v5.3.0) was applied to the high resolution T1 sequence. The mri_label2vol tool was used to map the labels on the thick slice T1 that was used for the evaluation. FSL (v5.0) was directly applied to the thick slice T1 and provides both a pveseg and a seg file as binary output. We evaluated both of these files. The fractional intensity threshold parameter "f" of the BET tool that sets the brain/nonbrain intensity threshold was set according to [55] at 0.2 (Philips Achieva 3T setting). SPM12 was directly applied to the thick slice T1 sequence as well. However, it also provides the option to add multiple MRI sequences. Therefore we evaluated SPM12 not only on the thick slice T1 sequence but added the T1-IR and the T2-FLAIR scan as well and tested various combinations. The number of Gaussians was set according to the SPM manual to two for GM, two for WM, and two for CSF.

2.3. Statistical Analysis. All evaluated methods were compared to the reference standard. In summary of the results, the mean and standard deviation over all 15 test datasets were calculated per component (GM, WM, and CSF) and combination of components (brain, intracranial volume) and per evaluation measure (Dice, 95th-percentile Hausdorff distance, and absolute volume difference) for each of the evaluated methods. Boxplots were created using R version 3.0.3 (R project for statistical computing (http://www.r-project.org/)). Since white matter lesions should be segmented as white matter, the percentage of white matter lesion voxels segmented as white matter (sensitivity) was calculated for each algorithm over all 15 test datasets to evaluate the robustness of the segmentation algorithms against pathology.

3. Results

Table 1 presents the final ranking (r) of the evaluated methods that participated in the workshop, as well as the evaluated freeware packages. During the workshop team UofL BioImaging ranked first and BIGR2 ranked second with one point difference in the overall score s (5). However, adding the results of the freeware packages resulted in an equal score for UofL BioImaging and BIGR2. Therefore the standard deviation rank was taken into account and BIGR2 is ranked first with standard deviation rank four and UofL BioImaging is ranked second with standard deviation rank eight. Table 1 further presents the mean, standard deviation, and rank for each evaluation measure (D, H_{95}, and AVD) and component (GM, WM, and CSF), as well as the brain (WM + GM) and intracranial volume (WM + GM + CSF). Team BIGR2 scored best for the GM, WM, and brain segmentation and team UofL BioImaging for the CSF segmentation. Team Robarts scored best for the intracranial volume segmentation. The boxplots for all evaluation measures and components are shown in Figures 2–4 and include the results of the freeware packages. Figure 5 shows an example of the segmentation results at the height of the basal ganglia (slice 22 of test subject 9). The sensitivity of the algorithms to segment white matter lesions as WM and examples of the segmentation results in the presence of white matter lesions (slice 31 of test subject 3) are shown in Figure 6. Team UB VPML Med scores the highest sensitivity of white matter lesions segmented as white matter and is therefore most robust in the presence of this type of pathology.

4. Discussion

In this paper we proposed the MRBrainS challenge online evaluation framework to evaluate automatic and semiautomatic algorithms for segmenting GM, WM, and CSF on 3T multisequence (T1, T1-IR, and T2-FLAIR) MRI scans of the brain. We have evaluated and presented the results of eleven segmentation algorithms that provide a benchmark for algorithms that will use the online evaluation framework to evaluate their performance. Team UofL BioImaging and BIGR2 have equal overall scores, but BIGR2 was ranked first based on the standard deviation ranking. The evaluated methods represent a wide variety of algorithms that include Markov random field models, clustering approaches, deformable models, and atlas-based approaches and classifiers (SVM, KNN, and decision trees). The presented evaluation framework provides an insight into the performance of these algorithms in terms of accuracy and robustness. Various factors influence the choice for a certain method above others. We provide three measures that could aid in selecting the method that is most appropriate for the segmentation goal at hand: a boundary measure (95th-percentile Hausdorff distance H_{95}), an overlap measure (Dice coefficient D), and a volume measure (absolute volume difference AVD). All three measures are taken into account for the final ranking of the methods. This ranking was designed to get a quick insight into how the methods perform in comparison to each other. The best overall method is the method that performs well for all three measures and all three components (GM, WM, and CSF). However, which method to select depends on the segmentation goal at hand. Not all measures are relevant for all segmentation goals. For example, if segmentation is used for brain volumetry [4], the overlap (D) and volume (AVD) measures of the brain and intracranial volume (used for normalization [56]) segmentations are important to take into account. On the other hand, if segmentation is used for cortical thickness measurements, the focus should be on the gray matter boundary (H_{95}) and overlap (D) measures. Therefore the final ranking should be used to get a first insight into the overall performance, after which the performance of the measures and components that are most relevant for the segmentation goal at hand should be considered. Besides accuracy, robustness could also influence the choice for a certain method above others. For example, team UB VPML Med shows a high sensitivity score for segmenting white matter lesions as white matter (Figure 6) and shows a consistent segmentation performance of gray and white matter over all 15 test datasets (Figures 2–4). This could be beneficial for segmenting scans of populations with white matter lesions but is less important if the goal is to segment scans of young healthy subjects. In the latter case, the most accurate segmentation for gray and white matter (team BIGR2) is more interesting. If a segmentation algorithm is to be used in clinical practice, speed is an important consideration as well. The runtime of the evaluated methods is reported in Table 1. However, these runtimes are merely an indication of the required time, since academic software is generally not optimized for speed and the runtime is measured on different computers and platforms. Another relevant aspect of the evaluation framework is the comparison of multi- versus single-sequence approaches. For example, most methods struggle with the segmentation of the intracranial volume on the T1-weighted scan. There is no contrast between the CSF and the skull, and the contrast between the dura mater and the CSF is not always sufficient. Team Robarts used an atlas-based registration approach on the T1-IR scan (good contrast between skull and CSF) to segment the intracranial volume, which resulted in the best performance for intracranial volume segmentation (Table 1, Figures 2–4). Most methods add the T2-FLAIR scan to improve robustness against white matter lesions (Table 1, Figure 6). Although using only the T1-weighted scan and incorporating prior shape information (team UofL BioImaging) can be very effective also, the freeware packages support this as well. Since FreeSurfer is an atlas-based method, it uses prior information and is the most robust of all freeware packages to white matter lesions. However, adding the T2_FLAIR scan to SPM12 increases robustness against white matter lesions as well, as compared to applying SPM12 to the T1 scan only (Figure 6). In general SPM12 with the T1 and the T2-FLAIR sequence performs well in comparison to the other freeware packages (Table 1 and Figures 2–4) on the thick slice MRI scans. Although adding the T1-IR scan to SPM increases the performance of the CSF and ICV segmentations as compared to using only the T1 and T2-FLAIR sequence, it decreases the performance of the GM and WM segmentations. Therefore adding all sequences to SPM12 did not result in a better overall performance.

Table 1: Results of the 11 evaluated algorithms presented at the workshop and the evaluated freeware packages on the 15 test datasets. The algorithms are ranked (r) based on their overall score (s) by using (5). This score is based on the ranks of the gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) segmentation and the three evaluated measures: Dice coefficient (D in %), 95th-percentile Hausdorff distance (H_{95} in mm), and the absolute volume difference (AVD in %). The rank r_{mc} denotes the rank based on the mean (μ) over all 15 test datasets for each measure m (0: D, 1: H_{95}, and 2: AVD) and component c (0: GM, 1: WM, 2: CSF, 3: brain (WM + GM), and 4: intracranial volume (ICV = WM + GM + CSF)). Teams BIGR2 and UofL BioImaging, and FreeSurfer and Jedi Mind Meld have equal scores based on the mean (μ); therefore the ranking based on the standard deviation (σ) is taken into account to determine the final rank (BIGR2: σ rank 4, UofL BioImaging: σ rank 8, FreeSurfer: σ rank 13, and Jedi Mind Meld: σ rank 17). Columns 2 and 3 present the average runtime t per scan in seconds (s), minutes (m), or hours (h) and the scans (T1: T1-weighted scan, 3D T1: 3D T1-weighted scan, IR: T1-weighted inversion recovery (IR), and F: T1-weighted FLAIR) that are used for processing.

				GM			WM			CSF				Brain			ICV		
r	Team	t	Scans	r_{00}	r_{10}	r_{20}	r_{01}	r_{11}	r_{21}	r_{02}	r_{12}	r_{22}	s	r_{03}	r_{13}	r_{23}	r_{04}	r_{14}	r_{24}
				D	H_{95}	AVD	D	H_{95}	AVD	D	H_{95}	AVD		D	H_{95}	AVD	D	H_{95}	AVD
				μ (σ)	μ (σ)	μ (σ)	μ (σ)	μ (σ)	μ (σ)	μ (σ)	μ (σ)	μ (σ)		μ (σ)	μ (σ)	μ (σ)	μ (σ)	μ (σ)	μ (σ)
				1	2	4	2	2	4	4	5	14		1	2	12	8	4	12
1	BIGR2	35 m	T1, IR, F	84.7	1.9	6.1	88.4	2.4	6.0	78.3	3.2	23	38	95.1	2.7	3.2	96.0	3.9	5.2
				(1.3)	(0.4)	(3.3)	(1.2)	(0.5)	(5.1)	(5.0)	(0.8)	(17)		(0.5)	(0.8)	(1.6)	(1.3)	(1.1)	(3.0)
				5	1	9	4	1	13	2	2	1		2	1	12	5	2	5
2	UofL BioImaging*	6 s	T1	83.0	1.7	8.6	87.9	2.2	8.7	78.9	2.7	9.7	38	94.9	2.4	3.9	96.7	3.4	1.8
				(1.5)	(0.3)	(5.4)	(2.0)	(0.6)	(6.6)	(4.2)	(0.5)	(10)		(0.6)	(0.5)	(2.0)	(0.8)	(0.6)	(2.0)
				6	7	5	5	3	10	3	3	8		5	7	2	4	3	11
3	CMIV	3 m	T1, F	82.4	2.7	6.8	87.7	2.4	7.3	78.6	3.0	14	50	94.5	3.8	2.6	96.8	3.8	4.9
				(1.4)	(0.4)	(4.0)	(1.6)	(0.4)	(3.8)	(3.1)	(0.4)	(5.9)		(0.5)	(1.1)	(2.2)	(0.8)	(1.3)	(2.3)
				4	4	2	1	5	7	8	13	17		4	4	1	10	11	15
4	UB VPML Med	30 m	T1, IR, F	83.3	2.1	5.9	88.6	2.7	7.1	74.8	4.3	31	61	94.6	2.8	2.4	94.8	6.6	7.7
				(1.3)	(0.3)	(5.3)	(1.7)	(0.4)	(3.8)	(7.1)	(1.7)	(19)		(0.6)	(0.4)	(1.8)	(2.0)	(2.0)	(4.1)
				7	13	3	6	6	9	6	4	10		7	10	10	7	5	8
5	Bigr_neuro	2 h	T1, F	81.5	3.7	5.9	87.3	3.0	7.3	78.2	3.2	16	64	94.0	4.6	3.6	96.3	3.9	3.5
				(1.7)	(0.9)	(4.2)	(1.4)	(0.4)	(3.8)	(4.7)	(0.6)	(14)		(0.8)	(1.4)	(2.4)	(1.2)	(0.9)	(2.7)
				11	3	15	8	8	6	1	1	13		16	3	18	1	1	1
6	Robarts	16 m	3D T1, IR	79.7	2.0	9.8	86.2	3.1	7.1	80.3	2.7	20	66	93.1	2.8	7.9	97.9	2.6	0.9
				(2.4)	(0.1)	(7.3)	(1.3)	(0.4)	(6.2)	(4.1)	(0.5)	(13)		(1.6)	(0.5)	(3.6)	(0.3)	(0.4)	(0.7)
				3	5	1	7	11	2	17	18	7		3	5	3	16	18	9
7	Narsil	2 m	T1, F	83.5	2.3	5.5	87.1	3.3	5.8	66.6	13.3	14	71	94.8	2.9	2.9	92.5	24	3.7
				(1.8)	(0.4)	(4.4)	(1.3)	(0.9)	(5.3)	(2.4)	(5.4)	(9.5)		(0.5)	(0.5)	(2.0)	(0.5)	(8.9)	(1.7)
				8	9	16	9	7	1	9	14	2		9	14	14	6	14	4
8	SPM_T1_F	3 m	T1, F	81.2	2.9	10	86.0	3.0	5.2	74.1	4.6	10	75	93.9	5.8	5.3	96.6	8.2	1.5
				(2.2)	(0.3)	(8.5)	(1.5)	(0.1)	(3.8)	(3.4)	(0.6)	(4.7)		(1.0)	(2.2)	(3.8)	(0.2)	(3.0)	(1.0)
				12	11	7	16	12	5	5	10	3		8	11	7	2	8	2
9	SPM_T1_IR	3 m	T1, IR	79.4	3.0	7.2	83.5	3.6	6.3	78.3	4.0	10	81	93.9	4.6	3.4	97.7	6.5	1.0
				(2.1)	(0.4)	(6.3)	(2.1)	(0.3)	(4.6)	(3.8)	(0.6)	(5.7)		(0.8)	(1.2)	(2.8)	(0.2)	(1.3)	(0.8)
				2	8	12	3	4	11	15	15	16		6	9	11	15	12	17
10	MNAB	15 m	T1, IR, F	83.9	2.8	9.1	88.0	2.7	7.8	68.1	4.9	29	86	94.5	4.5	3.8	92.5	7.1	9.7
				(2.1)	(0.9)	(6.5)	(1.2)	(0.8)	(4.0)	(4.0)	(2.2)	(21)		(1.0)	(2.0)	(3.2)	(1.1)	(4.2)	(4.7)
				9	10	6	11	10	3	11	16	15		10	8	6	9	13	13
11	SPM_T1	3 m	T1	80.3	3.0	6.9	85.6	3.1	6.0	70.7	5.3	23	91	93.9	4.4	3.2	95.3	8.1	5.5
				(2.4)	(0.5)	(6.8)	(1.7)	(0.1)	(4.1)	(3.8)	(1.5)	(15.7)		(0.9)	(1.6)	(2.9)	(0.9)	(3.7)	(3.7)
				13	16	10	10	13	14	12	6	5		13	13	4	12	6	6
12	FSL_Seg	10 m	T1	78.7	4.3	8.6	86.0	3.7	11.5	69.9	3.4	12	99	93.3	5.5	3.0	94.2	5.3	3.4
				(2.2)	(1.2)	(6.3)	(2.6)	(0.8)	(6.3)	(2.8)	(0.2)	(10.3)		(0.8)	(1.4)	(1.5)	(0.8)	(1.1)	(1.5)

Table 1: Continued.

r	Team	t	Scans	GM			WM			CSF			s	Brain			ICV		
				r_{00}	r_{10}	r_{20}	r_{01}	r_{11}	r_{21}	r_{02}	r_{12}	r_{22}		r_{03}	r_{13}	r_{23}	r_{04}	r_{14}	r_{24}
				D	H_{95}	AVD	D	H_{95}	AVD	D	H_{95}	AVD		D	H_{95}	AVD	D	H_{95}	AVD
				$\mu(\sigma)$	$\mu(\sigma)$	$\mu(\sigma)$	$\mu(\sigma)$	$\mu(\sigma)$	$\mu(\sigma)$	$\mu(\sigma)$	$\mu(\sigma)$	$\mu(\sigma)$		$\mu(\sigma)$	$\mu(\sigma)$	$\mu(\sigma)$	$\mu(\sigma)$	$\mu(\sigma)$	$\mu(\sigma)$
				10	12	18	15	14	12	7	11	6		12	15	13	3	15	3
13	SPM_T1_IR_F	4 m	T1, IR, F	80.1	3.0	13.9	83.6	3.8	8.4	76.9	4.1	12	105	93.6	5.9	5.1	97.7	8.2	1.2
				(2.4)	(0.2)	(9.6)	(2.1)	(0.5)	(5.2)	(3.1)	(0.5)	(6.0)		(1.1)	(1.8)	(3.6)	(0.2)	(1.8)	(0.9)
				15	15	8	13	15	17	13	7	4		11	16	9	13	7	7
14	FSL _PVSeg	10 m	T1	77.7	4.3	8.4	84.8	3.8	19.7	69.5	3.4	11	107	93.6	6.1	3.5	94.2	5.3	3.4
				(2.6)	(1.3)	(6.5)	(3.2)	(0.9)	(10)	(2.2)	(0.3)	(5.8)		(0.9)	(1.6)	(3.0)	(0.8)	(1.1)	(1.5)
				16	6	17	12	9	8	18	12	18		17	6	16	17	10	18
15	FreeSurfer	1 h	3D T1	77.4	2.3	12.1	85.2	3.1	7.2	65.8	4.3	50	116	92.3	3.6	6.3	92.4	6.6	10
				(2.0)	(0.6)	(6.0)	(2.2)	(0.5)	(4.6)	(3.7)	(0.6)	(19.6)		(0.8)	(0.6)	(3.4)	(0.9)	(0.4)	(4.5)
				14	14	11	17	16	15	10	8	11		15	12	8	11	16	14
16	Jedi Mind Meld	27 s	T1, IR, F	77.8	3.9	8.9	80.6	4.5	13.7	73.3	3.7	18	116	93.2	5.4	3.5	94.3	20	6.8
				(5.9)	(2.8)	(7.6)	(9.6)	(3.6)	(12)	(3.5)	(0.8)	(9.0)		(1.7)	(4.8)	(3.0)	(1.7)	(3.7)	(3.1)
				17	17	13	14	17	18	16	9	9		14	17	15	14	9	10
17	S2_QM	1.5 h	T1, IR, F	76.4	5.5	9.3	83.9	4.9	24.8	67.9	3.8	14	130	93.3	7.0	5.5	93.7	6.5	4.0
				(3.4)	(3.0)	(6.4)	(3.4)	(3.2)	(10)	(2.3)	(0.6)	(9.9)		(1.4)	(3.5)	(4.5)	(1.1)	(1.4)	(2.2)
				18	18	14	18	18	16	14	17	12		18	18	17	18	17	16
18	LNMBrains	5 m	T1, IR, F	72.8	6.8	9.5	78.3	6.8	15.3	68.8	7.7	20	145	88.5	8.6	7.4	89.2	20.4	7.9
				(5.3)	(2.3)	(7.6)	(6.4)	(3.5)	(13)	(6.6)	(2.4)	(13)		(4.5)	(2.8)	(7.7)	(4.8)	(3.7)	(8.3)

*Semiautomatic due to per scan parameter tuning.

Besides the advantages of the MRBrainS evaluation framework, there are some limitations that should be taken into account. The T1-weighted IR and the T2-weighted FLAIR scan were acquired with a lower resolution ($0.96 \times 0.96 \times 3.00\,mm^3$) than the 3D T1-weighted scan ($1.0 \times 1.0 \times 1.0\,mm^3$). To be able to provide a registered multisequence dataset, the 3D T1-weighted scan was registered to the T2-weighted FLAIR scan and downsampled to $0.96 \times 0.96 \times 3.00\,mm^3$. The reference standard is therefore only available for this resolution. The decreased performance of the FreeSurfer GM segmentation as compared to the other freeware packages might be due to the fact that we evaluate on the thick slice T1 sequence instead of the high resolution T1. Performing the manual segmentations to provide the reference standard is very laborsome and time consuming. Instead of letting multiple observers manually segment the MRI datasets or letting one observer manually segment the MRI datasets twice, much time and effort was spent on creating one reference standard that was as accurate as possible. Therefore we were not able to determine the inter- or intraobserver variability. Finally, we acknowledge that our evaluation framework is limited to evaluating the accuracy and robustness over 15 datasets for segmenting GM, WM, and CSF on 3T MRI scans acquired on a Philips scanner of a specific group of elderly subjects. Many factors influence segmentation algorithm performance, such as the type of scanner (vendor, field strength), the acquisition protocol, the available MRI sequences, and the type of subjects. Participating algorithms might have been designed for different types of MRI scans. Therefore the five provided training datasets are important for participants to be able to train their algorithms on the provided data. Some algorithms are designed to segment only some components, such as only GM and WM, instead of all three components, and use freely available software such as the brain extraction tool [57] to segment the outer border of the CSF (intracranial volume). We have chosen to base the final ranking on all three components, but it is therefore important to assess not only the final ranking, but the performance of the individual components as well.

Despite these limitations, the MRBrainS evaluation framework provides an objective and direct comparison of segmentation algorithms. The reference standard of the test data is unknown to the participants, the same evaluation measures are used for all evaluated algorithms, and participants apply their own algorithms to the provided data.

In comparison to the online validation engine proposed by Shattuck et al. [58], the MRBrainS evaluation framework uses 3T MRI data instead of 1.5T MRI data and evaluates not only brain versus nonbrain segmentation, but also segmentation of gray matter, white matter, cerebrospinal fluid, brain, and intracranial volume. The availability of many different types of evaluation frameworks will aid in the development of more generic and robust algorithms. For example, in the NEATBrainS (http://neatbrains15.isi.uu.nl/) challenge, researchers were challenged to apply their algorithms to data

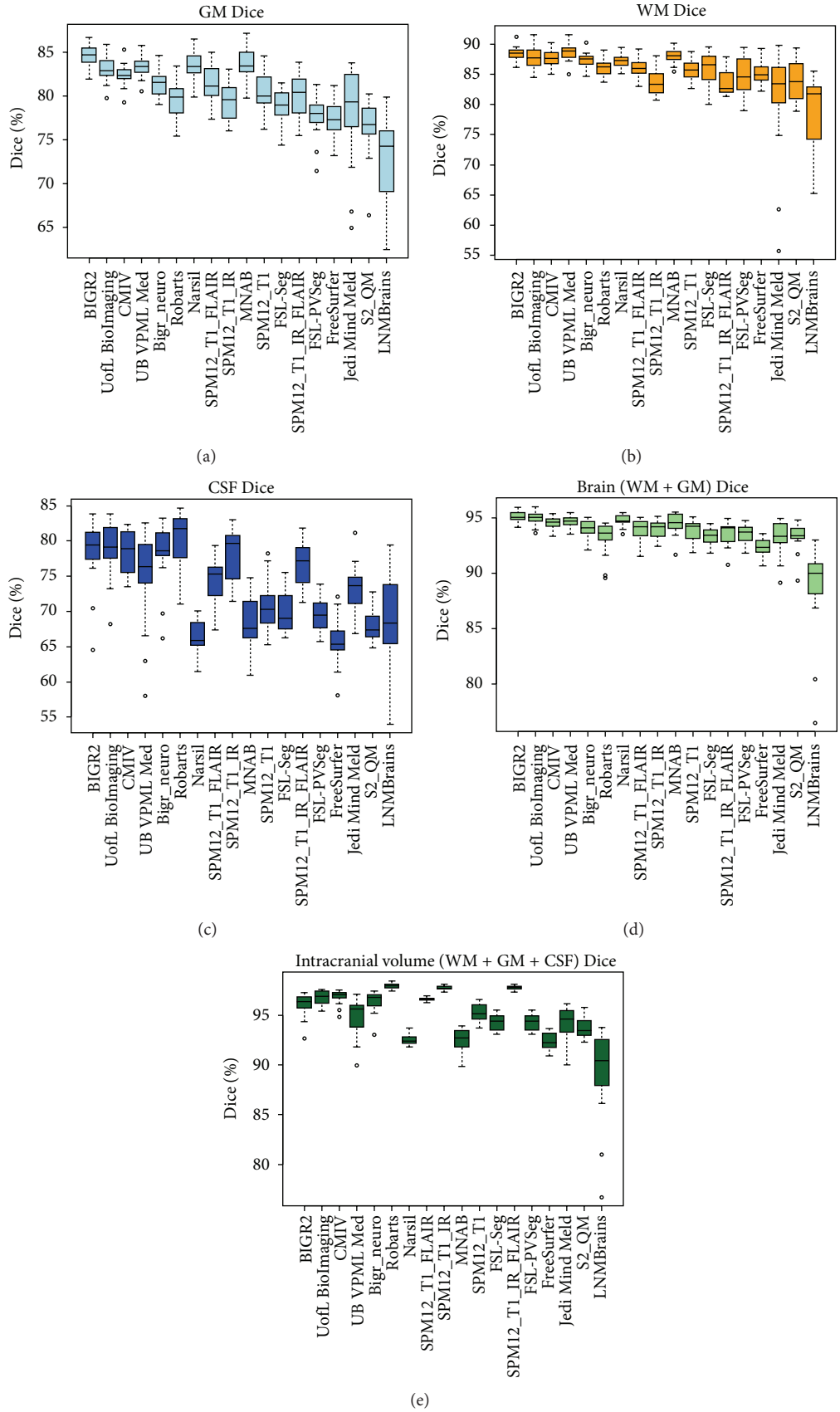


Figure 2: Boxplots presenting the evaluation results for the Dice coefficient (1) of the gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), and brain and ICV segmentations for each of the participating algorithms and freeware packages over all 15 test datasets.

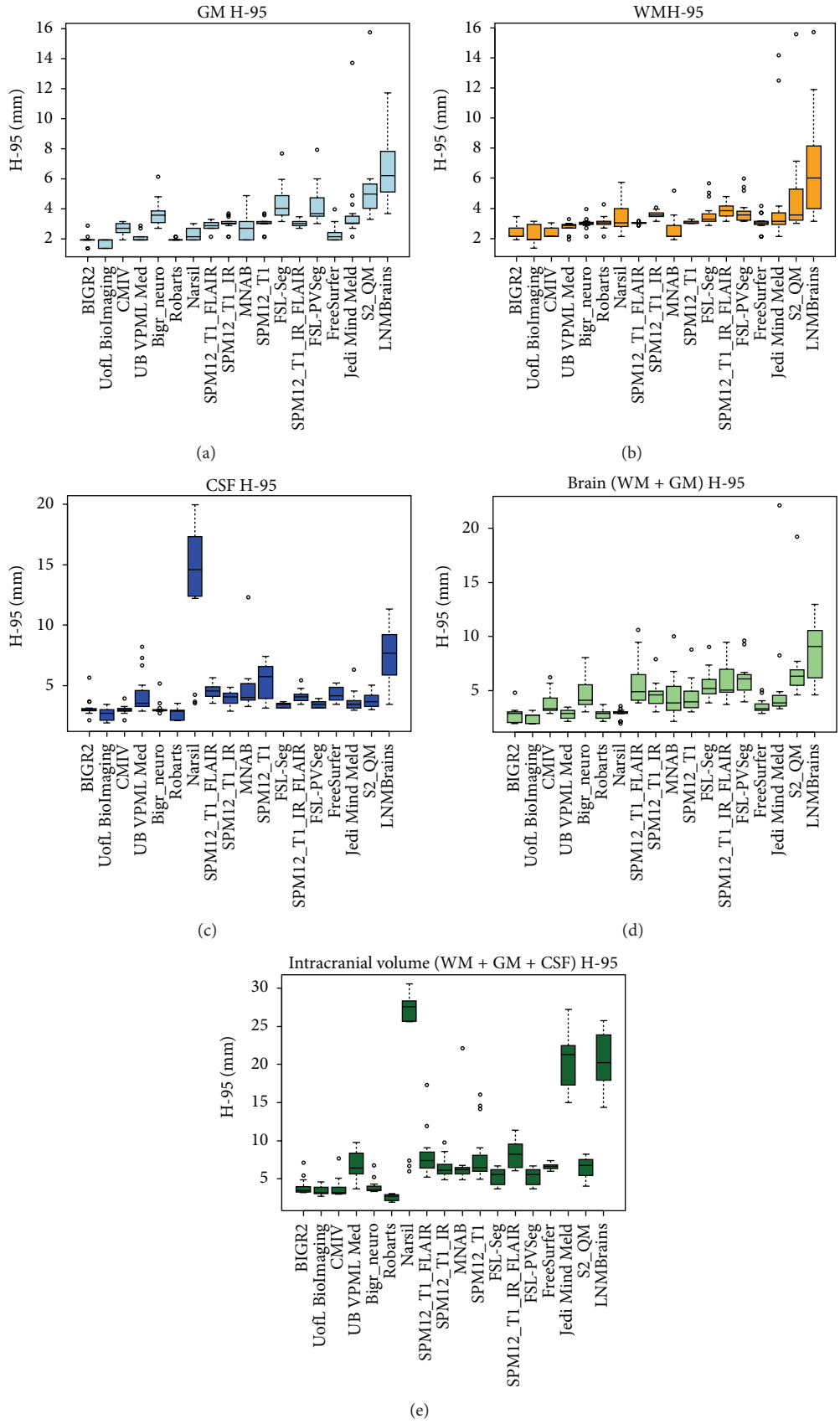


Figure 3: Boxplots presenting the evaluation results for the 95th-percentile Hausdorff distance (3) of the gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), and brain and ICV segmentations for each of the participating algorithms and freeware packages over all 15 test datasets.

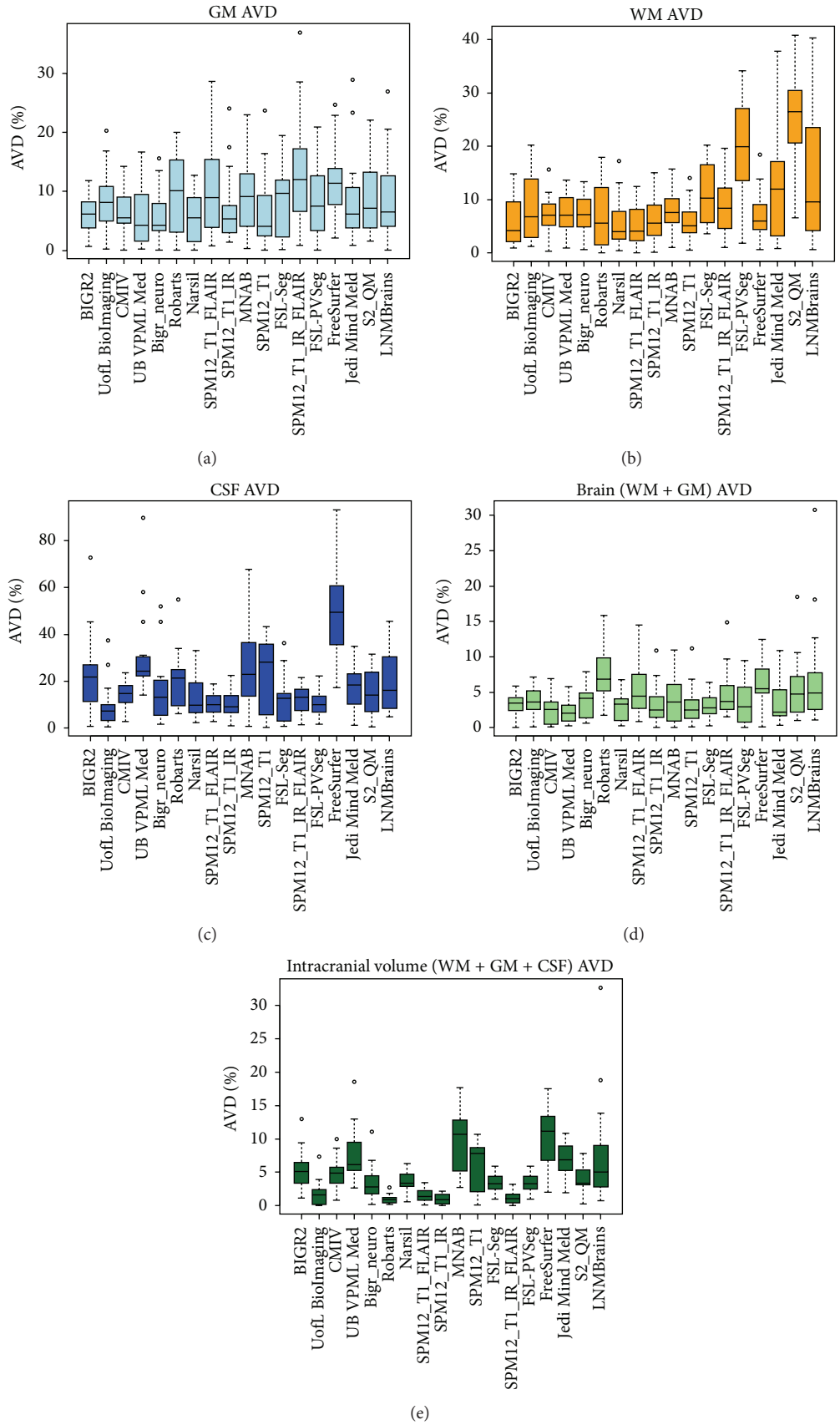


Figure 4: Boxplots presenting the evaluation results for the absolute volume difference (4) of the gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), and brain and ICV segmentations for each of the participating algorithms and freeware packages over all 15 test datasets.

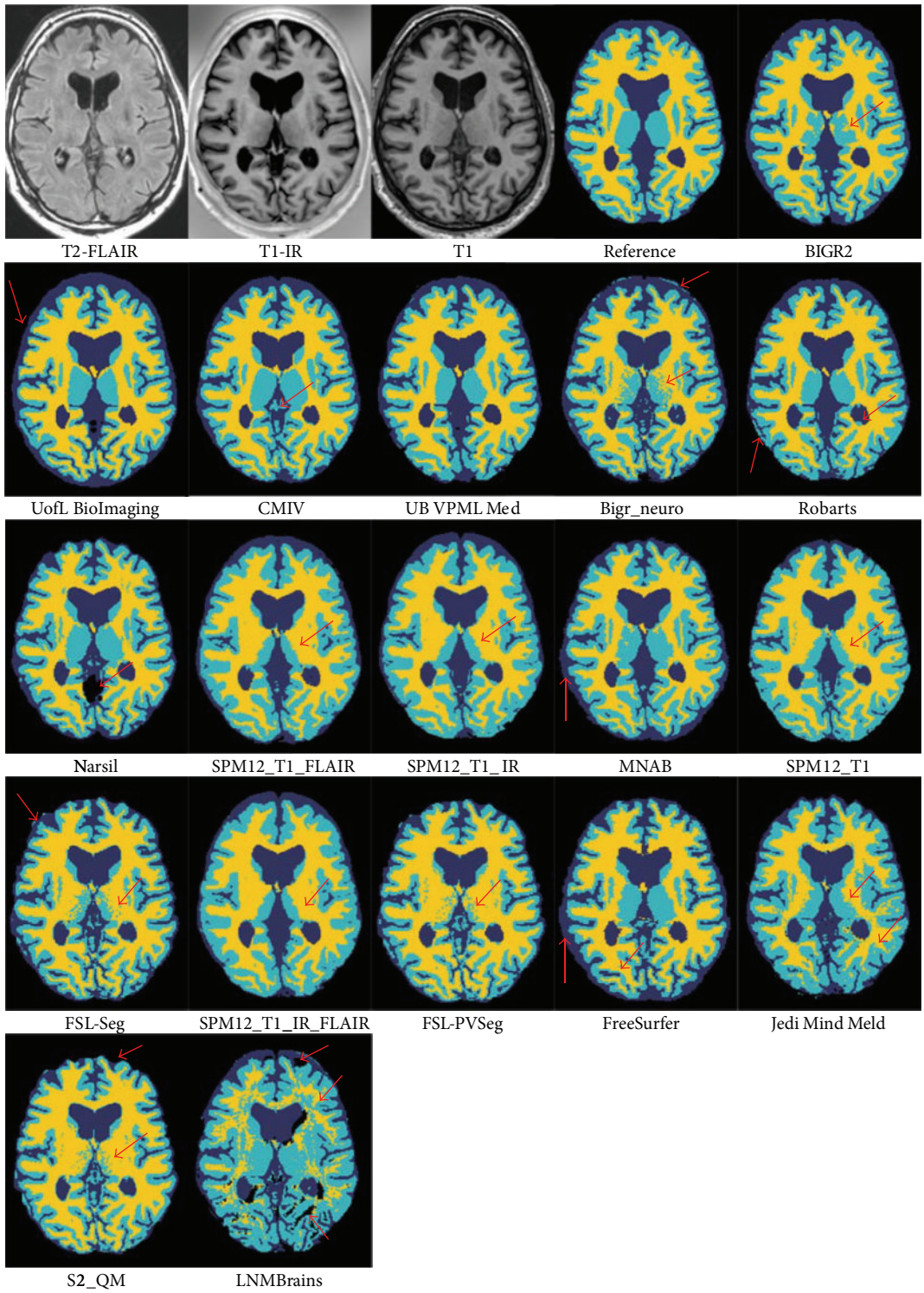


Figure 5: Illustration of the segmentation results at the height of the basal ganglia (test subject 9, slice 22). The basal ganglia should be segmented as gray matter. The first three images show the three MRI sequences. The fourth image (reference) is the manually segmented reference standard (yellow: white matter, light blue: gray matter, and dark blue: cerebrospinal fluid). The results of the participating segmentation algorithms are ordered from the best overall performance (BIGR2) to the worst overall performance (LNMBrains). Major differences are present in the segmentation of the basal ganglia and the outer border of the cerebrospinal fluid. The arrows indicate example locations where the segmentation results differ from the ground truth.

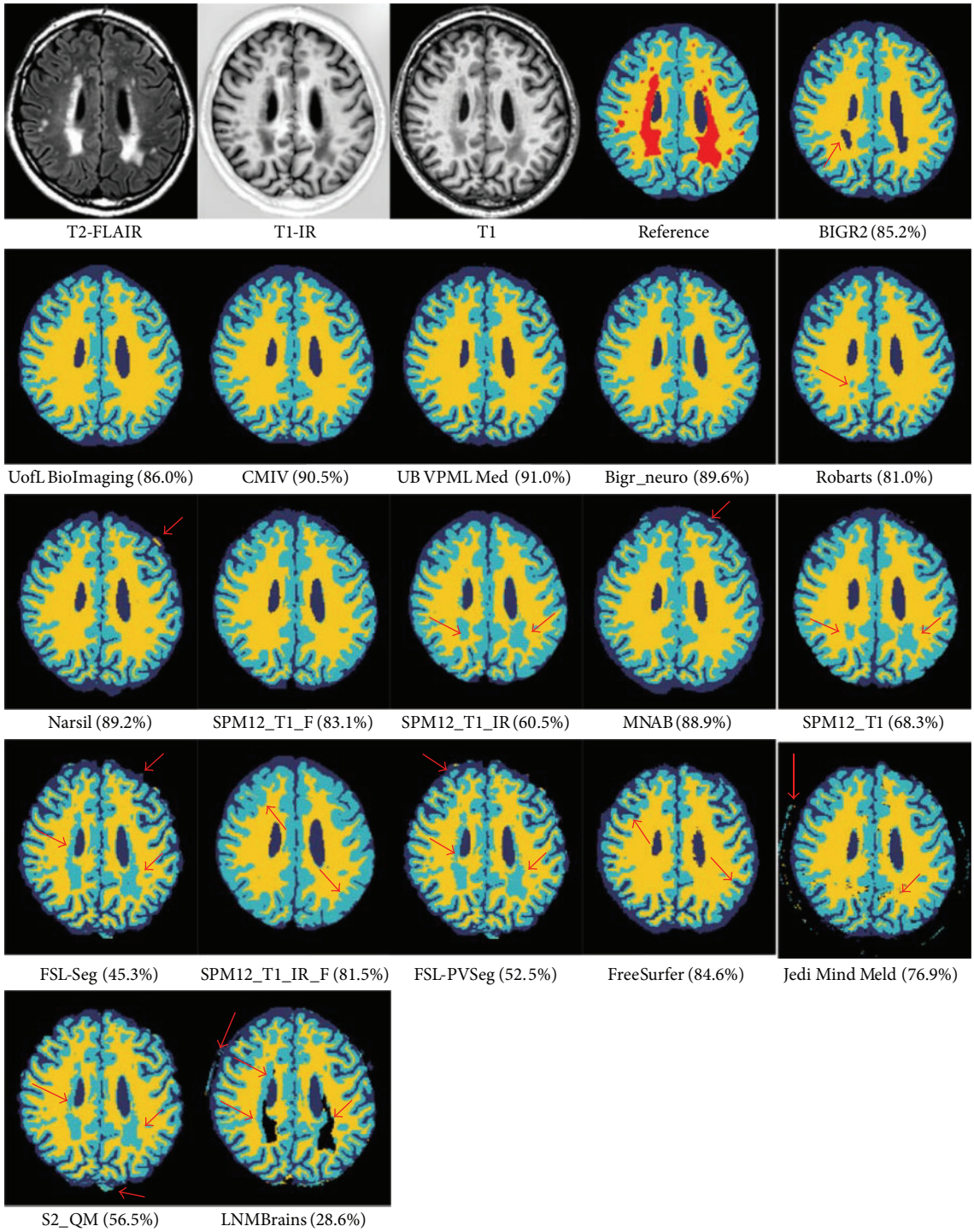


Figure 6: Illustration of the segmentation results in the presence of white matter lesions (test subject 3, slice 31). White matter lesions (WMLs) should be labeled as WM; the sensitivity (percentage of WML voxels is labeled as WM) over all 15 datasets is presented between brackets after the team name. The first three images show the three MRI sequences: T2-weighted fluid attenuated inversion recovery, T1-weighted inversion recovery, and T1-weighted scan. The fourth image (reference) is the manually segmented reference standard (red: white matter lesions, yellow: white matter, light blue: gray matter, and dark blue: cerebrospinal fluid). The results of the segmentation algorithms are ordered from the best overall performance (BIGR2) to the worst overall performance (LNMBrains). The arrows indicate example locations where the segmentation results differ from the ground truth.

from both the MRBrainS and the NeoBrainS [59] (brain tissue segmentation in neonates) challenge. Two methods [60–62] specifically designed for neonatal brain tissue segmentation showed a high performance for tissue segmentation on the MRBrainS data. Applying algorithms to different types of data has the potential to lead to new insights and more robust algorithms. The MRBrainS evaluation framework remains open for new contributions. At the time of writing, 21 teams had submitted their results on the MRBrainS website (http://mrbrains13.isi.uu.nl/results.php).

5. Conclusion

The MRBrainS challenge online evaluation framework provides direct and objective comparison of automatic and semiautomatic methods to segment GM, WM, CSF, brain, and ICV on 3T multisequence MRI data. The first eleven participating methods are evaluated and presented in this paper, as well as three commonly used freeware packages (FreeSurfer, FSL, and SPM12). They provide a benchmark for future contributions to the framework. The final ranking provides a quick insight into the overall performance of the evaluated methods in comparison to each other, whereas the individual evaluation measures (Dice, 95th-percentile Hausdorff distance, and absolute volume difference) per component (GM, WM, CSF, brain, and ICV) can aid in selecting the best method for a specific segmentation goal.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This study was financially supported by IMDI Grant 104002002 (Brainbox) from ZonMw, the Netherlands Organisation for Health Research and Development, within kind sponsoring by Philips, the University Medical Center Utrecht, and Eindhoven University of Technology. The authors would like to acknowledge the following members of the Utrecht Vascular Cognitive Impairment Study Group who were not included as coauthors of this paper but were involved in the recruitment of study participants and MRI acquisition at the UMC Utrecht (in alphabetical order by department): E. van den Berg, M. Brundel, S. Heringa, and L. J. Kappelle of the Department of Neurology, P. R. Luijten and W. P. Th. M. Mali of the Department of Radiology, and A. Algra and G. E. H. M. Rutten of the Julius Center for Health Sciences and Primary Care. The research of Geert Jan Biessels and the VCI group was financially supported by VIDI Grant 91711384 from ZonMw and by Grant 2010T073 of the Netherlands Heart Foundation. The research of Jeroen de Bresser is financially supported by a research talent fellowship of the University Medical Center Utrecht (Netherlands). The research of Annegreet van Opbroek and Marleen de Bruijne is financially supported by a research grant from NWO (the Netherlands Organisation for Scientific Research). The authors would like to acknowledge MeVis Medical Solutions AG (Bremen, Germany) for providing MeVisLab. Duygu Sarikaya and Liang Zhao acknowledge their Advisor Professor Jason Corso for his guidance. Duygu Sarikaya is supported by NIH 1 R21 CA160825-01 and Liang Zhao is partially supported by the China Scholarship Council (CSC).

References

[1] I. Driscoll, C. Davatzikos, Y. An et al., "Longitudinal pattern of regional brain volume change differentiates normal aging from MCI," *Neurology*, vol. 72, no. 22, pp. 1906–1913, 2009.

[2] K. Franke and C. Gaser, "Longitudinal changes in individual *BrainAGE* in healthy aging, mild cognitive impairment, and Alzheimer's disease," *GeroPsych*, vol. 25, no. 4, pp. 235–245, 2012.

[3] M. A. Ikram, H. A. Vrooman, M. W. Vernooij et al., "Brain tissue volumes in the general elderly population. The Rotterdam Scan Study," *Neurobiology of Aging*, vol. 29, no. 6, pp. 882–890, 2008.

[4] A. Giorgio and N. de Stefano, "Clinical use of brain volumetry," *Journal of Magnetic Resonance Imaging*, vol. 37, no. 1, pp. 1–14, 2013.

[5] M. W. Vannier, R. L. Butterfield, D. Jordan, W. A. Murphy, R. G. Levitt, and M. Gado, "Multispectral analysis of magnetic resonance images," *Radiology*, vol. 154, no. 1, pp. 221–224, 1985.

[6] M. A. Balafar, A. R. Ramli, M. I. Saripan, and S. Mashohor, "Review of brain MRI image segmentation methods," *Artificial Intelligence Review*, vol. 33, no. 3, pp. 261–274, 2010.

[7] J. C. Bezdek, L. O. Hall, and L. P. Clarke, "Review of MR image segmentation techniques using pattern recognition," *Medical Physics*, vol. 20, no. 4, pp. 1033–1048, 1993.

[8] L. P. Clarke, R. P. Velthuizen, M. A. Camacho et al., "MRI segmentation: methods and applications," *Magnetic Resonance Imaging*, vol. 13, no. 3, pp. 343–368, 1995.

[9] J. S. Duncan, X. Papademetris, J. Yang, M. Jackowski, X. Zeng, and L. H. Staib, "Geometric strategies for neuroanatomic analysis from MRI," *NeuroImage*, vol. 23, supplement 1, pp. S34–S45, 2004.

[10] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation," *Annual Review of Biomedical Engineering*, vol. 2, no. 2000, pp. 315–337, 2000.

[11] N. Saeed, "Magnetic resonance image segmentation using pattern recognition, and applied to image registration and quantitation," *NMR in Biomedicine*, vol. 11, no. 4-5, pp. 157–167, 1998.

[12] J. S. Suri, S. Singh, and L. Reden, "Computer vision and pattern recognition techniques for 2-D and 3-D MR cerebral cortical segmentation (Part I): a state-of-the-art review," *Pattern Analysis & Applications*, vol. 5, no. 1, pp. 46–76, 2002.

[13] A. P. Zijdenbos and B. M. Dawant, "Brain segmentation and white matter lesion detection in MR images," *Critical Reviews in Biomedical Engineering*, vol. 22, no. 5-6, pp. 401–465, 1994.

[14] K. Price, "Anything you can do, I can do better (No you can't)...," *Computer Vision, Graphics and Image Processing*, vol. 36, no. 2-3, pp. 387–391, 1986.

[15] R. de Boer, H. A. Vrooman, M. A. Ikram et al., "Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods," *NeuroImage*, vol. 51, no. 3, pp. 1047–1056, 2010.

[16] J. de Bresser, M. P. Portegies, A. Leemans, G. J. Biessels, L. J. Kappelle, and M. A. Viergever, "A comparison of MR based segmentation methods for measuring brain atrophy progression," *NeuroImage*, vol. 54, no. 2, pp. 760–768, 2011.

[17] M. B. Cuadra, L. Cammoun, T. Butz, O. Cuisenaire, and J.-P. Thiran, "Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images," *IEEE Transactions on Medical Imaging*, vol. 24, no. 12, pp. 1548–1565, 2005.

[18] F. Klauschen, A. Goldman, V. Barra, A. Meyer-Lindenberg, and A. Lundervold, "Evaluation of automated brain MR image segmentation and volumetry methods," *Human Brain Mapping*, vol. 30, no. 4, pp. 1310–1327, 2009.

[19] H. Zaidi, T. Ruest, F. Schoenahl, and M.-L. Montandon, "Comparative assessment of statistical brain MR image segmentation algorithms and their impact on partial volume correction in PET," *NeuroImage*, vol. 32, no. 4, pp. 1591–1607, 2006.

[20] J. A. Frazier, S. M. Hodge, J. L. Breeze et al., "Diagnostic and sex effects on limbic volumes in early-onset bipolar disorder and schizophrenia," *Schizophrenia Bulletin*, vol. 34, no. 1, pp. 37–46, 2008.

[21] A. Klein and J. Tourville, "101 labeled brain images and a consistent human cortical labeling protocol," *Frontiers in Neuroscience*, vol. 6, article 171, 2012.

[22] B. Van Ginneken, T. Heimann, and M. Styner, "3D segmentation in the clinic: a grand challenge," in *Proceedings of the 10th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI '07)*, pp. 7–15, Brisbane, Australia, October-November 2007.

[23] A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical surface-based analysis. I. Segmentation and surface reconstruction," *NeuroImage*, vol. 9, no. 2, pp. 179–194, 1999.

[24] B. Fischl, M. I. Sereno, and A. M. Dale, "Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system," *NeuroImage*, vol. 9, no. 2, pp. 195–207, 1999.

[25] M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.

[26] J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage*, vol. 26, no. 3, pp. 839–851, 2005.

[27] Y. D. Reijmer, A. Leemans, M. Brundel, L. J. Kappelle, and G. J. Biessels, "Disruption of the cerebral white matter network is related to slowing of information processing speed in patients with type 2 diabetes," *Diabetes*, vol. 62, no. 6, pp. 2112–2115, 2013.

[28] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[29] K. J. Friston, J. T. Ashburner, S. J. Kiebel, T. E. Nichols, and W. D. Penny, Eds., *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, Academic Press, 2011.

[30] F. Ritter, T. Boskamp, A. Homeyer et al., "Medical image analysis: a visual approach," *IEEE Pulse*, vol. 2, no. 6, pp. 60–70, 2011.

[31] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[32] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[33] A. Alansary, A. Soliman, F. Khalifa et al., "MAP-based framework for segmentation of MR brain images based on visual appearance and prior shape," in *Proceedings of the MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS '13)*, The Midas Journal, September 2013.

[34] A. Jog, S. Roy, J. L. Prince, and A. Carass, "MR brain segmentation using decision trees," in *Proceedings of the MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS '13)*, The Midas Journal, Nagoya, Japan, September 2013.

[35] R. Katyal, S. Paneri, and M. Kuse, "Gaussian intensity model with neighborhood cues for fluid-tissue categorization of multi-sequence MR brain images," in *Proceedings of the MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS '13)*, The Midas Journal, Nagoya, Japan, September 2013.

[36] Q. Mahmood, M. Alipoor, A. Chodorowski, A. Mehnert, and M. Persson, "Multimodal MR brain segmentation using Bayesian-based adaptive mean-shift (BAMS)," in *Proceedings of the MICCAI Workshops—The MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS '13)*, The Midas Journal, September 2013.

[37] A. van Opbroek, F. van der Lijn, and M. de Bruijne, "Automated brain-tissue segmentation by multi-feature SVM classification," in *Proceedings of the MICCAI Workshops—The MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS '13)*, The Midas Journal, September 2013.

[38] S. Pereira, J. Festa, J. A. Mariz, N. Sousa, and C. A. Silva, "Automatic brain tissue segmentation of multi-sequence MR images using random decision forests," in *Proceedings of the MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS '13)*, The Midas Journal, Nagoya, Japan, September 2013.

[39] M. Rajchl, J. Baxter, J. Yuan, T. Peters, and A. Khan, "Multi-atlas-based segmentation with hierarchical max-flow," in *Proceedings of the MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS '13)*, The Midas Journal, Nagoya, Japan, September 2013.

[40] D. Sarikaya, L. Zhao, and J. J. Corso, "Multi-atlas brain MRI segmentation with multiway cut," in *Proceedings of the MICCAI Workshops—The MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS '13)*, The Midas Journal, September 2013.

[41] H. Vrooman, F. van der Lijn, and W. Niessen, "Auto-kNN: brain tissue segmentation using automatically trained k-nearest-neighbor classification," in *Proceedings of the MICCAI Workshops—The MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS '13)*, The Midas Journal, September 2013.

[42] S. Vyas, P. Burlina, D. Kleissas, and R. Mukherjee, "Automated walks using machine learning for segmentation," in *Proceedings of the MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS '13)*, The Midas Journal, Nagoya, Japan, September 2013.

[43] C. Wang and Ö. Smedby, "Fully automatic brain segmentation using model-guided level sets and skeleton-based models," in *Proceedings of the MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS '13)*, The Midas Journal, Nagoya, Japan, September 2013.

[44] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[45] H. A. Vrooman, C. A. Cocosco, F. van der Lijn et al., "Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification," *NeuroImage*, vol. 37, no. 1, pp. 71–81, 2007.

[46] C. Wang, H. Frimmel, and Ö. Smedby, "Fast level-set based image segmentation using coherent propagation," *Medical Physics*, vol. 41, no. 7, Article ID 073501, 2014.

[47] A. Jog, S. Roy, A. Carass, and J. L. Prince, "Magnetic resonance image synthesis through patch regression," in *Proceedings of the IEEE 10th International Symposium on Biomedical Imaging (ISBI '13)*, pp. 350–353, April 2013.

[48] J. G. Sied, A. P. Zijdenbos, and A. C. Evans, "A non-parametric method for automatic correction of intensity non-uniformity in MRI data," *IEEE Transactions on Medical Imaging*, vol. 17, no. 1, pp. 87–97, 1998.

[49] A. Carass, J. Cuzzocreo, M. B. Wheeler, P.-L. Bazin, S. M. Resnick, and J. L. Prince, "Simple paradigm for extra-cerebral tissue removal: algorithm and analysis," *NeuroImage*, vol. 56, no. 4, pp. 1982–1992, 2011.

[50] M. Modat, G. R. Ridgway, Z. A. Taylor et al., "Fast free-form deformation using graphics processing units," *Computer Methods and Programs in Biomedicine*, vol. 98, no. 3, pp. 278–284, 2010.

[51] M. Rajchl, J. Yuan, J. A. White et al., "Interactive hierarchical-flow segmentation of scar tissue from late-enhancement cardiac mr images," *IEEE Transactions on Medical Imaging*, vol. 33, no. 1, pp. 159–172, 2014.

[52] Q. Mahmood, A. Chodorowski, A. Mehnert, and M. Persson, "A novel Bayesian approach to adaptive mean shift segmentation of brain images," in *Proceedings of the 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS '12)*, pp. 1–6, June 2012.

[53] A. El-Baz, M. Nitzken, A. Alansary, A. Soliman, and M. Casanova, "Brain Segmentation Method for Young Children and Adults," US Provisional Patent Application #13100.

[54] A. Alansary, A. Soliman, M. Nitzken et al., "An integrated geometrical and stochastic approach for accurate infant brain extraction," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '14)*, pp. 3542–3546, October 2014.

[55] V. Popescu, M. Battaglini, W. S. Hoogstrate et al., "Optimizing parameter choice for FSL-Brain Extraction Tool (BET) on 3D T1 images in multiple sclerosis," *NeuroImage*, vol. 61, no. 4, pp. 1484–1494, 2012.

[56] J. L. Whitwell, W. R. Crum, H. C. Watt, and N. C. Fox, "Normalization of cerebral volumes by use of intracranial volume: implications for longitudinal quantitative MR imaging," *American Journal of Neuroradiology*, vol. 22, no. 8, pp. 1483–1489, 2001.

[57] S. M. Smith, "Fast robust automated brain extraction," *Human Brain Mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[58] D. W. Shattuck, G. Prasad, M. Mirza, K. L. Narr, and A. W. Toga, "Online resource for validation of brain segmentation methods," *NeuroImage*, vol. 45, no. 2, pp. 431–439, 2009.

[59] I. Išgum, M. J. N. L. Benders, B. Avants et al., "Evaluation of automatic neonatal brain segmentation algorithms: the NeoBrainS12 challenge," *Medical Image Analysis*, vol. 20, no. 1, pp. 135–151, 2015.

[60] P. Moeskops, M. J. N. L. Benders, S. M. Chiţă et al., "Automatic segmentation of MR brain images of preterm infants using supervised classification," *NeuroImage*, 2015.

[61] P. Moeskops, M. A. Viergever, M. J. N. L. Benders, and I. Išgum, "Evaluation of an automatic brain segmentation method developed for neonates on adult MR brain images," in *Medical Imaging: Image Processing*, vol. 9413 of *Proceedings of SPIE*, International Society for Optics and Photonics, Orlando, Fla, USA, February 2015.

[62] L. Wang, Y. Gao, F. Shi et al., "LINKS: learning-based multi-source IntegratioN frameworK for Segmentation of infant brain images," *NeuroImage*, vol. 108, pp. 160–172, 2015.