******* Welcome to the LI-Toolbox *******

******* MOST IMPORTANT *******

Before doing anything else PLEASE check if the handedness assumed in this toolbox conforms to your local settings! Left and right are defined via LI_left and LI_right in the /data-directory, such that in LI-left, all voxels believed to define left are 1, all others are 0. You absolutely MUST make sure that this is also true for YOUR data. The function LI_test.m has been included to help you with this: call

```
LI_test
```

in Matlab and see if the left side is shown in white (a message box will explain that again :). You should also do test-runs with data of known laterality. Please note that as of version 1.2, the toolbox tries to gather the handedness from the data passed to it, but this does not mean nothing can go wrong anymore, particularly when using data from older spm-versions. For this reason, the toolbox now only works in spm8 and later versions! If things still don't seem to be in order, you can, as a last resort, rename LI_left to LI_right and vice versa (or make your own and pass it to the function as custom images, see below).

******* Introduction *******

This is the manual for the LI-toolbox, an integrated software package allowing for the investigation of laterality effects in imaging data. It is made available to the scientific community to explore aspects of hemispheric specialization. It was developed and is intended for use within spm8 or later (Wellcome Department of Imaging Neuroscience, University College London, UK: www.fil.ion.ucl.ac.uk/spm/) by Marko Wilke in cooperation with Karen Lidzba from the Department of Pediatric Neurology and Developmental Medicine at Tuebingen University Children's Hospital, and Vincent Schmithorst, Imaging Research Center, Cincinnati Children's Hospital Medical Center. For information on our disclaimer, see ReadMe.txt; for version information, see history.txt.

The lateralization index as implemented here relies on the basic computation LI = (Left - Right) / (Left + Right). Therefore, a negative value indicates a right hemispheric dominance and a positive value indicates a left hemispheric dominance.

******* Documentation *******

This file :) The primary paper on this is Wilke M & Lidzba K: LI-tool: *A new toolbox to assess lateralization in functional MR-data*, J Neurosci Meth, 2007, 163: 128-136; see below for a pre-print. The bootstrap approach was published in Wilke M & Schmithorst VJ: *A combined bootstrap/histogram*

analysis approach for computing a lateralization index from neuroimaging data, Neuroimage 2006, 33: 522-530; see below for a pre-print.

******* Installation & Compatibility *******

The toolbox is available from the spm-GUI if you unpack the files into (spm-dir)\toolbox\LI. The directory structure should thus be \spm[8|>8]\toolbox\LI\. Since a file with the same name is found there, you are able to call the toolbox by clicking on Toolboxes => LI. There is also a config file, allowing you to add LI-jobs to the spm8-batch editor (courtesy of Arnold Skimminge!). Within the LI-directory you should have a data folder. Alternatively, you can call it from the command line as "LI", which allows passing additional input arguments. Type "help LI" or see below for more on these options. This version was written to work with spm8 (when all updates are installed) and should work with later versions.

******* Interactive vs. Scripted mode *******

If you call the toolbox via the command line with LI or via the dropdown menu, it will start up in interactive mode, asking you what it needs to know (see below, "The prompts explained"). However, the toolbox can also be run from the command line. In order to do this, you need to supply an input argument as in LI(out). Note the special case that, "out" being the number 2 will "only" result in the toolbox output being saved to a file (see below, "Hints"). However, if "out" is a structure, then the following fields will be necessary: A (name and path of image(s) to be analyzed), B1 (which inclusive mask to use), C1 (which exclusive mask to use), and thr1 (which thresholding approach to use). In scripted mode, graphic windows are suppressed (but saved to a file), and a text file with the results will always be saved. Options are:

A	char array (path or paths to contrast images, one image per line)
B1	1 (frontal)
	2 (parietal)
	3 (temporal)
	4 (occipital)
	5 (cingulate)
	6 (central)
	7 (cerebellar)
	8 (gray matter)
	9 (all standard masks)
	10 (no mask)
	11 (custom, enables interactive selection [somewhat besides the point of scripting, but...])

	char array (path or paths to inclusive mask(s), one image per line)
C1	1 (standard exclusive mask, midline -5 mm)
	2 (standard exclusive mask, midline -10 mm)
	3 (no exclusive mask)
	4 (custom, enables interactive selection)
	char array (path or paths to exclusice mask(s), one image per line)
thr1	1 (use same threshold for all images; needs to be supplied below)
	0 (individual thresholding for all images; needs to be supplied below)
	-1 (adaptive thresholding
	-2 (rank-based thresholding)
	-3 (iterative thresholding, LI-curves)
	-4 (no threshold)
	-5 (bootstrap)

Such a minimal structure may be generated by using

```
out = struct('A',spm_select,'B1',1,'C1',1,'thr1',-5),
```

which will ask for an image to analyze. If you then type

```
LI(out)
```

this will analyze your selected image in the frontal lobe (B1 = 1), excluding the midline (C1 = 1), using the bootstrap approach (thr1 = -5).

There are also optional arguments which may or may not be supplied: default values will be used if they are missing. Note that even when scripting, you can still supply custom left and right mask images.

pre	0 (default)
	1 (enables preprocessing for custom masks, only necessary if B1 > 10 or char)
thr3	0 (default; meaningless if not thr1 = 1; if so, adapt value to your needs)
	[3 4 4 5] (supplies multiple thresholds if thr1 = 2)
op	1 (use optional data clustering; may not be allowed)
	2 (use optional variance weighting
	3 (use combined clustering and variance weighting; may not be allowed)
	4 (default; no optional steps)
vc	1 (use total voxel count; may not be allowed)
	0 (use total voxel values; default)

ni	0 (images not normalized)
	1 (images normalized; default)
outfile	string with name of a custom output file (defaults to 'li.txt')

As you can see, some options only make sense in a certain combination (for example, if you choose the ranking approach, smoothing is disallowed; if you still supply out.op = 1, this will be ignored). Also, note that the toolbox by default assumes the images that are supplied as a structure are normalized, so be mindful of the fact that they will only be normalized if you specify out.ni = 0.

******* The prompts explained *******

- *Select Contrast Image(s)*

Here, you can select any image you wish to explore with regard to laterality effects, but ideally this will be a spmT-image from an spm8 analysis, normalized to the template space bounding box (meaning 91x109x91 voxels at 2x2x2mm resolution, see also below). There is no limitation as to how many images can be selected.

- *Select thresholding method*

One threshold

one threshold will be applied to all contrast images you selected

Individual threshold

a threshold can be entered for each contrast image you selected

Adaptive threshold

computed average intensity of all positive voxels and used it as the threshold

Ranking

this will rank-transform and weight an spmT-image

Iterative (LI-curves)

this will compute LIs at up to 20 equally-spaced thresholds (0-max)

Bootstrap

this will employ a bootstrap procedure in addition to generating LI-curves. For what it's worth, I consider this the currently "best" compromise

- *Select INclusive mask*

This allows for the optional selection of an anatomically predefined mask (see below). If you want classical hemispheric lateralization, choose "none". Note that you can choose a subject's individual gray matter segment (or a standard gray matter mask) if you click on "Gray Matter...". If you want to use more then one mask, choose "all lobes" or "Custom", which allows for the selection of whatever masks as you like (needless to say, masks should be in alignment with the contrast image[s]). Since custom masks are not necessarily binary or symmetrical (see below), the algorithm will then ask if the masks should be

preprocessed accordingly. If the mask is binary already, you can click no, as slightly different mask sizes for each side will be accounted for (i.e. it does not have to be symmetrical). If you have only chosen standard masks from the LI\data directory, you can also say "no" here. There is no limitation as to how many masks can be selected: all possible mask-contrast combinations will be explored.

- Select EXclusive mask

By default, masking out the midline (+/-5mm) is offered to avoid flow artifacts in the large draining veins, but you can also choose to mask out +/-11mm, "none" or a custom one. There is no preprocessing option here since I have no clue what this custom mask of yours may be like; so, please make sure that your custom exclusion mask is in the correct format.

- Select optional steps

These are options with a potentially very severe impact on the data, so they are not usually recommended!

Clustering

spatial smoothing of 3*voxel dimensions (X,Y,Z); removes outliers (NB: this is not available for ranking); think long and hard before selecting this as it may lead to a very smooth dataset for which the mask you chose may not be appropriate anymore!

Variance weighting

uses ResMS.img to determine "reliability" of activation; this sounds meaningful enough but may have a profound (!) impact on your results. I recommend using this option only i fand when you know exactly why, and what, you are doing.

Both

combines the above (NB: this is not available for ranking)

None

no optional steps (default)

- Is the contrast image normalized?

Yes

select this if the analysis has been done on normalized images

No

if not normalized, the algorithm will do an affine matching. As the normalization employed here is rather rough, results will be better if you do it properly, i.e. before submitting data to the toolbox.

Note 1: this prompt will not appear if the images are found to be in the correct format, i.e. in template space and in 2x2x2mm resolution

Note 2: to avoid asking over and over again, the algorithm assumes that you DO NOT MIX normalized and un-normalized images; i.e., all contrast should be normalized or all should be in native space.

- Other prompts

Optional prompts may appear based on special situations, which (I hope) will be self-explaining.

- Prompts from LI_boot.m

The bootstrap algorithm can be told to only explore voxels above a certain threshold (e.g., a significance threshold determined before), what size (as percentage of the input data) the output data should be and what the maximal size of this sample should be. These values can be entered when asked here, they default to no lower threshold, 25% and 10.000, respectively. Defaults will also be used when scripting is used.

******* Outputs *******

In contrast to previous versions, the default has changed so that the toolbox will always save the information to a file. This is identical to calling the toolbox from the command line with LI(2); if you want to suppress file output, call it with LI(1); you can also or re-enable the manual selection of this option by re-enabling the explicit text-input in LI.m, at about line 93. This will generate a tab-delimited text-file called li.txt in your current work directory.

Some general information is given in the Matlab command window as a summary of inputs, options, and results. Please note warning messages about small clusters or insufficient numbers of voxels in the command window; for ease of ensuing processing in a spreadsheet program, these warnings are not included in the resulting text file!

Graphical information about the masking is shown in the spm-graphics window (top left: contrast image, top right: mask [or left mask, if no masking was chosen], bottom left: masked left contrast, bottom right: masked right contrast). Note that this window may not be „clickable“ anymore if the run is finished as the temporary files are deleted at the end of processing (to prevent this, set RA to 0 near the bottom of LI.m or LI_boot.m). However, a file called LI_masking.ps will also be saved, containing all the graphical masking information normally shown on the screen for some approaches. This may also be helpful when larger collections of contrast/mask-combinations are investigated in one run.

If you chose to do iterative thresholding, the LI-curves will be written to a file called LI_curves.ps, and the complete output from LI_boot will always go to LI_boot.ps. In these cases, li.txt may also contain information from all thresholding steps (depending on the setting in these files, see the version information).

******* Hints *******

Please make sure you have all the updates for spm installed!

Inclusion masks can be any image defining your area of interest, but remember that they should be binary and in register with the images to examine. Flipping and binarization will be applied to the mask if you click on "yes" if asked whether to preprocess masks. Note that when selecting an individual GM segment, this should not be modulated gm since the cutoff value used during binarization would be too high.

Exclusion masks should be 0 where tissue should be excluded and 1 everywhere else (see LI-midline_ex.img as a hint). If you want to mask out a lesion or tissue of no interest, be sure to investigate any interaction between in- and exclusion masks.

Custom output file names are possible by editing LI.m at around line 153. The default is 'li.txt' but you can enter whatever name suits you best, for example enable the inclusion of the date, as available as an alternative in the line below. Custom names can also be supplied in scripting mode.

******* Credits *******

As of version 1.2, the masks included in the LI\data directory are based on a population-based atlas as described in Hammers et al., Hum Brain Mapp 9: 224-247, 2003, and Gousias et al., Neuroimage 40: 672-684, 2008. They were made symmetrical, slightly smoothed (FWHM = 6mm), and binarized at 0.25 in order to render them slightly more inclusive. They have been included with kind permission by the author. The toolbox includes the script used to generate the masks, so if you approach Dr. Hammers for the atlas, you can recreate the masks using different compositions and/or settings.

The masks originally included in the toolbox were built based on Tzourio-Mazoyer et al, NeuroImage 15:273-89, 2002, used with kind permission by the author. They are still available within the LI\data directory in a zip-file.

The toolbox comes with a lot of modified functions, most of which are hacked spm-functions, changed to alter their naming conventions or output options. They are:

LI_FDR.m	based on FDRill.m (v1.2, Tom Nichols)
LI_global.m	based on calculate_global.m (v?, John Ashburner)
LI_imcalc.m	based on spm_imcalc_ui (v2.7, John Ashburner, Andrew Holmes)
LI_normalise.	m based on spm_normalise.m (v2.8, John Ashburner)
LI_write_sn.m	based on spm_write_sn.m (v2.17, John Ashburner)

Additionally, many of the routines, calls, or syntax incorporated in this toolbox have been inspired by numerous spm-functions, scripts or hints from the mailbase. I am most indebted to those of you who

really know how to program (mainly John Ashburner, Volkmar Glauche, Tom Nichols and many others) so I could steal, borrow and learn from you.

******* Disclaimer & License information *******

Please see the ReadMe.txt in the toolbox directory for disclaimer and license information; see the history.txt file in the toolbox directory for version information.

As we would like to keep track of who uses the toolbox for our strictly internal documentation purposes, we kindly ask you not to re-distribute the toolbox yourself but instead to refer interested scientists to our website (see ReadMe.txt). Alternatively, send us the name, institution and e-mail address of the new user and we will include him/her in our list.

The authors expect you to include a citation or acknowledgment if you present or publish results obtained using this toolbox. By installing and using this software, you agree to all the terms and conditions specified above and in the accompanying ReadMe.txt. If these conditions are not acceptable, do not use the software.

******* Support *******

This collection of files is not officially supported or developed further since we do not have the resources to provide this service on an acceptable and ongoing basis. Should there be the need for a public announcement, it will be posted to the SPM-mailing list (currently at www.jiscmail.ac.uk/lists/spm.html). Also, possible new versions or updates may be sent out to registered users.

We do, however, appreciate any comments and are open to suggestions. Please contact us at Marko.Wilke@med.uni-tuebingen.de

Cheers,
Marko

******* Files *******

Files in the main toolbox folder

LI	main function of the LI-toolbox
LI_FDR	stripped-down function from Tom Nichols; see FDRill.m for details
LI_boot	function to iteratively calculate bootstrapped lateralization indices

LI_boot_hf	bootstrap helper function
LI_imcalc	perform algebraic functions on images
LI_iter	iteratively calculates LIs and yields lateralization curves
LI_make_mask	re-create inclusion masks (not usually necessary, but can be used to adapt settings); backup masks before running this function!
LI_normalise	spatial (stereotactic) normalization
LI_rank	ranks t-values to avoid thresholding effects and to remove outliers
LI_mreslice	reslicing of an image in memory +/- disk
LI_test	left-right-helper function of the LI-toolbox:
LI_write_sn	writes out warped images.
LI_cfg_tbx	toolbox configuration file for spm8 and above
LI_manual	this file

- Files in the toolbox\data folder

LI-temporal-mask	temporal lobe mask
LI-central-mask	central GM structures: basal ganglia & thalamus
LI-cerebellar-mask	cerebellar mask
LI-cingulate-mask	cingulate mask
LI-frontal-mask	frontal lobe mask
LI-gray-matter-	mask standard gray matter mask
LI-occipital-mask	occipital lobe mask
LI-parietal-mask	parietal lobe mask
LI-right	mask defining the right side of the image; **CHECK IF THIS CONFORMS TO YOUR DATA!**
LI-left	mask defining the left side of the image; **CHECK IF THIS CONFORMS TO YOUR DATA!**
randbrain	random data in mask size to facilitate normalization
randmask	a mask for randbrain in order to allow for a finer affine matching
aalmasks.zip	previously-used LI masks based on the aal atlas
LI-midline_5_ex	default exclusive mask, masks out 10 mm in the midline
LI-midline_11_ex	additional exclusive mask, masks out 22 mm in the midline
demo.fig	helper figure for the LI_test function
demo.jpg	helper file for the LI_test function

******* Preprints *******

On the following pages, you will find preprints of the manuscripts this work is based on.

LI-tool: A new toolbox to assess lateralization in functional MR-data

Marko Wilke & Karen Lidzba

Department of Pediatric Neurology and Developmental Medicine, Children's Hospital, and

Section for Experimental MR of the CNS, Dept. of Neuroradiology,

University of Tübingen, Germany

- Preprint as accepted for publication in the Journal of Neuroscience Methods -

Corresponding author:

Marko Wilke, MD
Department of Pediatric Neurology and Developmental Medicine
Children's Hospital, University of Tübingen
Hoppe-Seyler-Str. 1
72076 Tübingen, Germany
Phone: + 49 7071 – 29 83416
Fax: + 49 7071 – 29 5473
e-mail: Marko.Wilke@med.uni-tuebingen.de

Abstract

A lateralization index (LI) is commonly computed to describe the asymmetry of activation as detectable by various functional imaging techniques, particularly functional magnetic resonance imaging (fMRI). In this article, we examine and compare different approaches that have been used in the past. For illustration purposes, 100 synthetic datasets and real fMRI-data from 12 subjects were evaluated. As shown before, the calculation of a lateralization index suffers from a number of drawbacks, namely vulnerability to statistical outliers, data sparsity, thresholding effects and lack of taking into account regional variability of activation. Optional processing steps investigated here seem to increase reliability of the such-calculated indices. To allow a more standardized, reproducible and accessible evaluation of laterality effects, current and new approaches have been implemented in a versatile toolbox running within the spm2 or spm5 software environment.

Keywords: lateralization index, functional magnetic resonance imaging, hemispheric specialization, language lateralization, LI-toolbox

Introduction

A lateralization index (LI) is commonly used to describe the asymmetry of functional activation in functional neuroimaging studies. While not formally statistically proving it, such lateralization or asymmetry indices (AI) serve to illustrate the brain's hemispheric specialization for a given task. From the very first descriptions of language as being a predominantly left-hemispheric function of the brain, this issue has been of special importance to the field of language research (Price, 2000; Hugdahl & Davison, 2002). One of the most commonly used approaches to describe such functional lateralization is to calculate a lateralization index based on

$$LI = \frac{\sum activation_{left} - \sum activation_{right}}{\sum activation_{left} + \sum activation_{right}} \qquad \text{(Equation 1)}$$

This will yield values between -1 and 1, with +1 being a purely LEFT and -1 a purely RIGHT activation. Of note, some authors have used the opposite notation (Staudt *et al.*, 2001; Liégeois *et al.*, 2004) or multiply the resulting value by 100 (Nagata *et al.*, 2001), as in a classical handedness score (Oldfield, 1971). However, there is no consensus on how to best compute the (sum of) activation that will enter this equation, with a large number of different approaches being used in the past (Holland *et al.*, 2001; Nagata *et al.*, 2001; Adcock *et al.*, 2003; Liégeois *et al.*, 2004). This multitude of approaches in effect disallows comparing different studies with regard to lateralization.

In the course of our work on hemispheric specialization, we opted to develop and explore new approaches that are aimed at minimizing the influence of confounding factors and the necessity of user interaction while allowing for maximum flexibility in terms of optional steps. In this article, we describe our approaches towards identifying and dealing with potential sources of error. These have resulted in the compilation of a

versatile toolbox for a widely-used MR-image analysis suite (spm2/5, Wellcome Department of Imaging Neuroscience, University College, London, UK), based on the MATLAB programming environment (MathWorks, Natick, MA, USA). Before describing some unresolved aspects of current approaches motivating this work, it should be stressed that all conclusions and solutions presented here results from an *ad hoc* problem-oriented approach. No attempt was made to prove or disprove, in a formal statistical sense, the superiority of any approach described here.

Voxel value vs. Voxel count: The simplest approach to computing a lateralization index is analyzing a statistical image volume and to count the number of voxels surviving a given threshold in each hemisphere (Liégeois *et al.*, 2004). This binary decision has the drawback of disregarding the value of a given voxel (typically representing the strength of its correlation with the task [Holland *et al.*, 2001]), but makes the procedure supposedly robust against (positive) statistical outliers (Nagata *et al.*, 2001). Using the sum of voxel values instead takes these correlations into account (Holland *et al.*, 2001; Adcock *et al.*, 2003). In the case of symmetrical masks (see below), taking the sum of voxel values is equivalent to calculating the average voxel value. Theoretically, it may be more sensitive to statistical outliers, but it also more adequately reflects the individual contribution of a given voxel.

Global vs. Regional inference: The human brain is a complicated arrangement of highly specialized neuronal subdivisions. It is immediately apparent that an index comparing the whole of the left with the whole of the right hemisphere is bound to miss important aspects of neuronal activation inherent in current high-resolution functional neuroimaging data. This is especially obvious for language applications (Holland *et al.*, 2001; Adcock *et al.*, 2003; Deblaere *et al.*, 2004; Liégeois *et al.*, 2004), where different

aspects of a given task yield distinctly localized activations (Price, 2000). Therefore, a spatially more specific approach to assessing laterality effects is essential.

Thresholding issues: Thresholding functional imaging data is necessary in order to ensure the significance of observed results. Different approaches favor specificity or sensitivity (like the family-wise error correction [FWE] or the false discovery rate [FDR], respectively). Extent-based approaches, non-parametrical methods, Bayesian inference, or combinations have also been used (Friston *et al*., 1994; Worsley *et al*., 1996; Nichols & Hayasaka, 2003; Penny *et al*., 2003; Hayasaka & Nichols, 2004). Currently, no method can be considered to be equally applicable to all scenarios (Marchini & Presanis, 2004). Therefore, different ways to compute "significance" will reach different results, and one method will declare voxels significant that were discarded by others (while both solutions are perfectly legitimate from a purely statistical point of view). As the lateralization index will be extremely dependant upon the chosen cutoff, these thresholds are of outmost concern (Gaillard *et al*., 2002; Adcock *et al*., 2003). Previous approaches included the suggestion to not threshold statistical maps at all (Holland *et al*., 2001) or to plot lateralization curves as a function of threshold (Deblaere *et al*., 2004).

Data sparsity and statistical outliers: Accompanying the thresholding problem is the fact that with higher thresholds, fewer voxels remain until ultimately, a single surviving voxel on one side will lead to a lateralization index of ±1 (see Figure 1). It is obvious that such a value is biologically as well as computationally not meaningful. Data sparsity will mainly arise when a single or very few voxels show exceedingly large values, brought about by either actual (biological) or artificial circumstances. Regardless of the reason, the domination of the resulting index by only a small number

of voxels is to be avoided, and caution is warranted when basing such decisions on only a small number of voxels.

Methods

Data

Null-data: Synthetic imaging data was generated using custom Matlab-scripts. Normally distributed noise, lacking any kind of systematic lateralization effects, was generated and saved as 100 individual image volumes, thus serving as null-data for the exploration of artifacts and the effect of suggested control mechanisms.

Subjects: For illustration purposes, imaging data from 12 healthy adults (5m, 7f, mean age 23.6 ± 2.3 years, see Table 1) was randomly drawn from an ongoing study addressing aspects of hemispheric specialization. Each subject performed either a left- or a right lateralizing task: right-hemispheric activation was induced using a visual search task (Lidzba et al., 2006); a language task (silent generation of word chains) was used to evoke left-hemispheric activation (Staudt *et al.*, 2001). Right-handedness was ensured using the Edinburgh handedness inventory (Oldfield, 1971). Subjects were required to be free of any current or past neurological or psychiatric disease. Procedures were in accordance with local institutional review board requirements; all subjects gave written informed consent.

Data Acquisition & processing: All data was acquired on a 1.5-T scanner (Siemens Sonata, Erlangen, Germany), using a standard EPI-sequence (TR = 3s, TE = 39ms, 28 axial slices of 4 mm thickness, 1 mm gap, in-plane matrix = 64 x 64, voxel size: 3 x 3 x 5 mm). For the left-lateralizing task, an interscan interval of 2s was introduced

for the application of auditory stimuli. A T1-weighted anatomical 3D-dataset was also obtained (128 contiguous sagittal slices, in-plane matrix 256 x 256). MRI data was processed using spm2, including the removal of the first five image volumes, a wavelet-based denoising scheme (Wink & Roerdink, 2004), and the removal of EPI-distortions and EPI*movement effects using an individually acquired fieldmap (data was motion-corrected in the same step; Andersson *et al*, 2001). Spatial normalization was achieved using default parameters (7 x 9 x 7 nonlinear basis functions, 16 iterations, final resolution 2 x 2 x 2 mm). Global image signal drifts were removed and the data was smoothed using a 12 mm (FWHM) isotropic Gaussian filter. For analysis, the framework of the general linear model was applied, using a box-car reference function convolved with the hemodynamic response function. This results in one individual t-map for each task which was used for further analyses.

Implementation of solutions and analysis approach

Voxel value vs. Voxel count: Lateralization curves from each subject using either method were generated. To assess the degree of similarity, a Spearman's correlation coefficient was calculated over the respective lateralization indices from all thresholds for the global and a local analysis.

Global vs. Regional inference: We decided to include an option to select anatomically-defined regions of interest within which laterality will be assessed; this data conforms to MNI-space (Mazziotta *et al.*, 2001), is publicly available (Tzourio-Mazoyer *et al.*, 2002) and has been used before to this effect (Wilke *et al.*, 2003; Tzourio-Mazoyer *et al.*, 2004; the masks are used here with kind permission by the author). For use as a standard mask, the mask is rendered symmetrical and binarized. To account for inter-

individual variability and the spatial smoothing commonly applied to functional imaging data, the masks were smoothed with a 6 mm Gaussian filter prior to binarization, rendering them slightly more inclusive. All cerebral lobes, the cingulate cortex, and the cerebellum are included as standard masks. In order to rule out influences of different mask sizes and to allow for the use of non-symmetrical custom masks, a mask weighting factor (mwf) is calculated, representing the relation of the volumes of the masks on the left and on the right. This weighting factor is then used to extend equation 1, thus preventing an artificially skewed lateralization index due to mask size influences:

$$LI = \frac{(\sum activation_{left}) / mwf - \sum activation_{right}}{(\sum activation_{left}) / mwf + \sum activation_{right}} \qquad \text{(Equation 2)}$$

Matching and interpolation errors will be minimal for images normalized to MNI-space (bounding box: 182 x 218 x 182 mm, voxel size: 2 x 2 x 2 mm). Native-space images will automatically be normalized prior to masking using an affine transformation (Ashburner *et al.*, 1997); normalized images with different dimensions/resolutions are rigidly matched to mask space (Cox & Jesmanowicz, 1999). Exclusive masks are also supplied (defining tissue volume to exclude from the analysis). As artifacts typically occur near the midline (Krings *et al.*, 1999), masking out midline structures (± 5 mm) is offered by default, but as with inclusive masks, no masking or using a custom/lesion mask is also possible.

Thresholding issues: To address thresholding issues, we have implemented a new, *adaptive* approach. Akin to earlier approaches (Knecht *et al.*, 2003), one simple assumption was made that “interesting” data will be of above-average intensity, and that therefore, the mean intensity of the voxels in the image can serve as an internal

threshold. While there is no statistical justification for this procedure, it has the merit of being data-driven as the lower threshold is non-interactively determined from each individual data set.

Data sparsity and statistical outliers: As a first approach, we decided to use *clustering* by applying a Gaussian smoothing filter, determined by a smoothing factor (sf), such that

$$FWHM_{[X,Y,Z]} = sf * VS_{[X,Y,Z]} \quad \text{(Equation 3)}$$

This results in a filter with a full width at half maximum (FWHM) of sf * voxel size (VS) in each dimension (X, Y, and Z), which will weight each voxel with respect to its immediate, but not remote, neighbors. This approach is mainly intended to address data sparsity by removing outliers.

The *variance weighting* approach makes use of the information on the voxel-wise variability as stored in the residual mean squares image (ResMS) of spm-analyzed results. It is defined as

$$i_{ResMS} = (n-1) * Var \quad \text{(Equation 4)}$$

so that the resulting voxel intensity i_{ResMS} is a function of the variance (Var) in that voxel, accounting for the number of scans (n). As the residual error of the model, it reflects the voxel-wise variability of the observed signal. For the purpose of including this information in the calculation of a lateralization index, an inverse modulation with the corresponding weighting image is performed. Consequently, voxels that show a high noise level (as expressed by a high intensity i_{ResMS}) are de-valued and will thus exert a lesser influence on the ensuing calculations. This approach is mainly intended to deal with artificial statistical outliers, e.g. “activation” in the region of the eyes. As suggested

before, the variance image is smoothed to achieve a local pooling and thus contains regional variability information (Nichols & Holmes, 2001). For a *combined* approach (both clustering and variance weighting), the smoothing width should be identical to ensure correct voxel-to-voxel correspondence. Different smoothing widths (sf = 3, 4, 6, 8, and 10) shall be explored.

Results & Discussion

Null data: The analysis of the null data sets (n = 100) demonstrates that, at high thresholds, an artificial impression of laterality is apparent even in complete random data (Figure 2, upper left panel). Requiring a minimum number of voxels (n = 5) is, in these artificial datasets lacking any spatial cohesion, not effective in preventing this effect (Figure 2, lower left panel). In contrast to this, the introduction of a minimum cluster size (n = 5) is highly effective in preventing such artificial skewing of the resulting index (Figure 2, upper right panel). While smoothing is effective by removing outliers (note compressed scale in Figure 2, lower right panel), by itself it does not effectively prevent the skewed indices towards higher values. We therefore suggest that information about the number of contributing voxels and their biological meaningfulness should routinely be taken into account. For example, fifty voxels in a highly significant cluster represent a (biologically and computationally) meaningful activation; in contrast to this, fifty voxels that are scattered over the whole hemisphere will not make a difference for the computation, but pose a biologically much less meaningful scenario. Based on these results, the toolbox will issue a warning if less then a critical number of voxels is available for computation (default is n = 10), and if a

minimum cluster size is not reached (default is n = 5). If less than 5 voxels survive thresholding, the algorithm aborts calculations.

Voxel value vs. Voxel count: When comparing the lateralization index as a function of whether a voxel value or a voxel count is used, surprisingly little differences are found. Very similar lateralization curves result with almost perfect correlations (Table 1). The theoretical advantage of a lesser vulnerability of the voxel counting approach towards statistical outliers could thus not be demonstrated here with these high-quality datasets. This points toward high data homogeneity and, conversely, could imply that large differences between these approaches are indicative for the presence of outliers.

Global vs. Regional inference: Our results clearly show that a global lateralization index is prone to be influenced by activation remote from the region of interest. For example, the focus of our fMRI tasks was activation in the frontal lobe, which, if analyzed separately, shows a consistently lateralizing trend even at low thresholds (Figure 3, lower panels) for all subjects and all tasks. In contrast to this, in the global indices, one subject each shows a switch in lateralization (* in upper panels in Figure 3), stressing that regionally restricted approaches are warranted to improve both sensitivity and specificity (Holland *et al.*, 2001; Liégeois *et al.*, 2004). *A priori* hypotheses can explicitly be tested, and the definition and exploration of functionally-defined regions of interest is feasible. Such data-driven approaches should be most appropriate in terms of anatomical/functional adequacy (Holland *et al.*, 2001). The exploration of cytoarchitectonically defined regions (Amunts *et al.*, 2003) is another option.

Thresholding issues: As noted before, the lateralization index shows a very severe threshold dependency (Nagata *et al.*, 2001; Gaillard *et al.*, 2002; Adcock *et al.*, 2003;

Deblaere *et al.*, 2004; Liégeois *et al.*, 2004,). When applying very strict thresholds, in each case a lateralization index of either -1 or 1 is returned (see Figure 1). This does not only impair comparability between studies, but the choice between two equally legitimate thresholds can, in the extreme case, lead to opposite results within a single patient (Figure 4, see also below; note routinely included information on the number of voxels on each side). Here, laterality curves do provide a more comprehensive estimate of lateralization and offer additional information since they allow for the assessment of the trend (or lack thereof) towards laterality in the data over the whole range of thresholds.

The simple *adaptive* thresholding performs reasonably well when compared with the other approaches, although yielding consistently lower values (Table 2). This is not surprising since it does not specifically exclude noise in any way, other than expecting it to be below the calculated average intensity. Interestingly, this does not make the approach any more liable to false positives, as illustrated by the performance on the null-data (mean LI = 0.00167, SD = 0.0074, range = -0.0168 – 0.018). Inherently, this approach will always return a lateralization index, even in the complete absence of any significant activation. It may thus be useful for explorative analyses, but data quality (i.e., the contributing voxels) must always be checked carefully.

Data sparsity and statistical outliers: Data sparsity is illustrated in Figure 1: a strong decrease in the number of voxels is uniformly seen, with a logarithmic acceleration towards higher threshold values. A practical example is also illustrated in Figure 4 (subject 2, inclusion mask: temporal lobe, exclusion mask: midline [± 5 mm], voxel values). Here, the data-adaptive thresholding methods (adaptive approach and FDR-thresholding) find left-lateralized activation in this example. Upon strict thresholding

(applying the FWE-correction), an apparently very strong right-lateralized activation results, as expressed by a lateralization index of LI = -0.891. However, this is based on only 3 remaining voxels on the left side (and 33 voxels on the right). A voxel counting approach does not ameliorate the problem (LI_{FWE} = -0.833), demonstrating that in the case of statistical outliers leading to data sparsity, a voxel counting approach is not more robust than a voxel value approach. Of note, this implausible LI was accompanied by routine warnings on data sparsity ($\leq$ 10 voxels on one side) and minimum cluster size (no cluster $\geq$ 5 voxels on one side). With default settings, the algorithm would have aborted due to less than 5 voxels.

The effects of clustering and variance weighting are illustrated in Figure 5. The effect of *clustering* (Figure 5, top panel) is very prominent: large smoothing factors (6, 8, and 10 times the voxel dimension) completely preclude the change in laterality that can be seen in the original curve. Although smoothing does, in itself, not alter the overall intensity of an image (only its spectrum and spatial distribution [Wink & Roerdink, 2004]), the effect here is pronounced since smoothing is done prior to masking and thresholding.

The effect of *variance weighting* is considerable (Figure 5, middle panel), although altering the smoothing widths only has a weak effect. This indicates that the original change in laterality is brought about by voxels showing a high variance (which are subsequently devalued during variance weighting). Utilizing the residual mean squares as a regional error variance estimate has the effect of introducing additional constraints upon outliers (Nichols & Holmes, 2001). Thus, the influence of unwanted variance can be reduced, especially in regions where the data presents a poor fit to the model, as in the case of “activation” in the eyes. We therefore suggest that the exploration of a variance weighting scheme is warranted in situations where data-inherent uncertainties

seem to influence lateralization. If such effects are ameliorated when variance weighting is used, this argues in favor of the effects being artificial.

Finally, *combined smoothing and variance weighting* (Figure 5, bottom panel) seems to exert a "stabilizing influence" on the inherent trend to laterality in the data. However, while increasing filter widths seem to enhance this effect, it must be borne in mind that spatial specificity is progressively reduced. Of note, the same smoothing factor must be used for a combined approach to ensure correct voxel-to-voxel correspondence. Considering spatial specificity, we suggest to use the smallest effective factor (sf = 3) as a compromise.

Final software implementation: All steps described here are implemented as custom MATLAB scripts and functions, in part utilizing functionality available within spm2, especially regarding the graphical user interface (GUI) modules. The algorithm is integrated into the spm2-software environment as a toolbox, available via the graphical user interface. It can also be used within spm5. Command line usage is also implemented, allowing to pass additional arguments.

The typical steps are illustrated in Figure 6: the algorithm first asks for image(s) to be analyzed (Figure 6, a), optional mask(s) defining the volume to investigate (b) and an optional mask defining volume to be excluded (c). Due to the critical issue of handedness in the images and recent software changes in this regard (SpmWeb, 2003), it was decided to allow for site-specific customization by including default images defining the left and right hemisphere. A command line option was also implemented that allows overriding these defaults images and specifying custom ones. All necessary images (including the standard masks) are provided in a specific subdirectory.

All thresholding options described in this work are implemented in the final toolbox (d), including free input, adaptive approach and the output of laterality curves. Depending upon the chosen approach, the user can opt for clustering, variance weighting, both, or no optional steps, followed by the decision on voxel value or voxel count (f). The contrast images are thresholded and masked accordingly. The original contrast image, the mask image and the resulting masked contrast image are simultaneously shown in a graphic window to allow for visual assessment of masking (g), and the results are presented in the MATLAB command window (h) or, if specified, are additionally written to a tab-delimited text file (including all necessary identifier information). In the case of iterative thresholding, the lateralization curves are shown in a separate graphic window. If more than one contrast or mask image have been selected, the calculation and output procedure is iteratively repeated until all mask/contrast combinations have been calculated. There is no limitation as to how many contrast images or masks can be explored in one run. Finally, the user has the option to delete intermediate files (default) or to leave these for further inspections (i).

Following file preprocessing, all ensuing calculations are performed directly on the memory-mapped images. This is highly time-efficient and allows a 3-contrast/3-mask combination with combined weighting and clustering to be completed within 45 seconds on a standard PC workstation.

Of note, we have not tried to find or present the “best way” to calculate a lateralization index. Such an “optimal approach” will very likely depend on the data at hand and the question posed to it. While it would have been preferable to assess the correlation with results from an independent modality for mapping neuronal activation (and thus, lateralization) as the Wada test or intraoperative cortical mapping, such comparisons

would have been beyond the scope of the current manuscript. A systematic exploration of the different tools offered within this toolbox in such a context would of course be highly interesting and would shed further light on which approach yields the highest specificity and sensitivity when compared to the invasive gold standard. Instead, this paper is aimed at raising the researcher's awareness for certain pitfalls in calculating a lateralization index, and to point out possible solutions. To this effect, our toolbox will allow exhaustive, reproducible explorations of imaging data with regard to the presence or absence of laterality effects. We therefore hope that, with this new tool for the exploration of hemispheric specialization, problematic issues in this field may be addressed successfully.

References

Adcock JE, Wise RG, Oxbury JM, Oxbury SM, Matthews PM. Quantitative fMRI assessment of the differences in lateralization of language-related brain activation in patients with temporal lobe epilepsy. *NeuroImage*, 2003; 18: 423-38

Amunts K, Schleicher A, Ditterich A, Zilles K. Broca's region: cytoarchitectonic asymmetry and developmental changes. *J Comp Neurol, 2003;* 465: 72-89

Andersson JLR, Hutton C, Ashburner J, Turner R, Friston K. Modeling geometric deformations in EPI time series. *NeuroImage*, 2001; 13: 903-19

Ashburner J, Neelin P, Collins DL, Evans AC, Friston KJ. Incorporating Prior Knowledge into Image Registration. *NeuroImage*, 1997; 6: 344-52

Cox RW, Jesmanowicz A. Real-time 3D image registration for functional MRI. *Magn Res Med*, 199; 42: 1014-8

Deblaere K, Boon PA, Vandemaele P, Tieleman A, Vonck K, Vingerhoets G, Backes W, Defreyne L, Achten E. MRI language dominance assessment in epilepsy patients at 1.0 T: region of interest analysis and comparison with intracarotid amytal testing. *Neuroradiology*, 2004; 46: 413-20

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, Evans AC. Assessing the significance of focal activations using their spatial extent. *Hum Brain Mapp*, 1994; 1: 214-20

Gaillard WD, Balsamo L, Xu B, Grandin CB, Braniecki SH, Papero PH, Weinstein S, Conry J, Pearl PL, Sachs B, Sato S, Jabbari B, Vezina LG, Frattali C, Theodore W. Language dominance in partial epilepsy patients identified with an fMRI reading task. *Neurology*, 2002; 59: 256–65

Hayasaka S, Nichols T. Combining voxel intensity and cluster extent with permutation test framework. *NeuroImage*, 2004; 23: 54-63

Holland SK, Plante E, Byars A, Strawsburg RH, Schmithorst VJ, Ball WS Jr. Normal fMRI brain activation patterns in children performing a verb generation task. *NeuroImage*, 2001; 14: 837-43

Hugdahl K, Davison RJ. The Asymmetrical Brain, 2nd ed. MIT Press: Cambridge, MA, USA, 2002

Knecht S, Jansen A, Frank A, van Randenborgh J, Sommer J, Kanowski M, Heinze HJ. How atypical is atypical language dominance? *NeuroImage*, 2003; 18: 917-27

Krings T, Erberich SG, Roessler F, Reul J, Thron A. MR blood oxygenation level-dependent signal differences in parenchymal and large draining vessels: implications for functional MR imaging. *Am J Neuroradiol*, 1999; 20: 1907-14

Lidzba K, Staudt M, Wilke M, Grodd W, Krägeloh-Mann I. Organization of Non-verbal Functions in Lesion-induced Right Hemispheric Language. *Neuroreport*, 2006; 17: 929-33

Liégeois F, Connelly A, Cross JH, Boyd SG, Gadian DG, Vargha-Khadem F, Baldeweg T. Language reorganization in children with early-onset lesions of the left hemisphere: an fMRI study. *Brain*, 2004; 127: 1229-36

Marchini J, Presanis A. Comparing methods of analyzing fMRI statistical parametric maps. *NeuroImage*, 2004; 22: 1203-13

Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, Woods R, Paus T, Simpson G, Pike B, Holmes C, Collins L, Thompson P, MacDonald D, Iacoboni M, Schormann T, Amunts K, Palomero-Gallagher N, Geyer S, Parsons L, Narr K, Kabani N, Le Goualher G, Boomsma D, Cannon T, Kawashima R, Mazoyer B. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Phil Trans Royal Soc*, 2001; 356: 1293-322

Nagata SI, Uchimura K, Hirakawa W, Kuratsu JI. Method for quantitatively evaluating the lateralization of linguistic function using functional MR imaging. *Am J Neuroradiol*, 2001; 22: 985-91

Nichols T, Hayasaka S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat Meth Med Res*, 2003; 12: 419-46

Nichols TE, Holmes AP. Nonparametric Analysis of PET functional Neuroimaging Experiments: A Primer. *Hum Brain Mapp*, 2001; 15: 1-25

Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 1971; 9: 97-113

Penny W, Kiebel S, Friston K. Variational Bayesian inference for fMRI time series. *NeuroImage*, 2003; 19: 727-41

Price C. The anatomy of language: contributions from functional neuroimaging. *J Anat*, 2000; 197: 335-59

SpmWeb: SPM, 2003: Compatibility. *www.fil.ion.ucl.ac.uk/spm/spm2.html#Compat*

Staudt M, Grodd W, Niemann G, Wildgruber D, Erb M, Krägeloh-Mann I. Early left periventricular brain lesions induce right hemispheric organization of speech. *Neurology*, 2003; 57: 122–5

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 2002; 15: 273–89

Tzourio-Mazoyer N, Josse G, Crivello F, Mazoyer B. Interindividual variability in the hemispheric organization for speech. *NeuroImage*, 2004; 21: 422-35

Wilke M, Sohn JH, Weber Byars AM, Holland SK. Bright spots: correlations of gray matter volume with IQ in a normal pediatric population. *NeuroImage*, 2003; 20: 202–15

Wink AM, Roerdink JBTM. Denoising functional MR images: a comparison of wavelet denoising and Gaussian smoothing. *IEEE Trans Med Imag*, 2004; 23: 374-87

Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC. A unified statistical approach for determining significant signals in images of cerebral activation. *Hum Brain Mapp*, 1996; 4: 58-73

Acknowledgements

We would like to thank the participants for their time and willingness to contribute to this study. We also thank Professor Ingeborg Krägeloh-Mann and Professor Wolfgang Grodd for supporting this project, and Michael Erb for helpful discussions. Finally, we would like to gratefully acknowledge the help and advice of Thomas E. Nichols, PhD, Department of Biostatistics, University of Michigan, and of Scott K. Holland, PhD, Imaging Research Center, Cincinnati Children's Hospital Medical Center, OH.

This work has been supported by the *Deutsche Forschungsgemeinschaft* DFG (SFB550/C4). The toolbox is available free of charge, please contact the authors at Marko.Wilke@med.uni-tuebingen.de.

Figure Captions

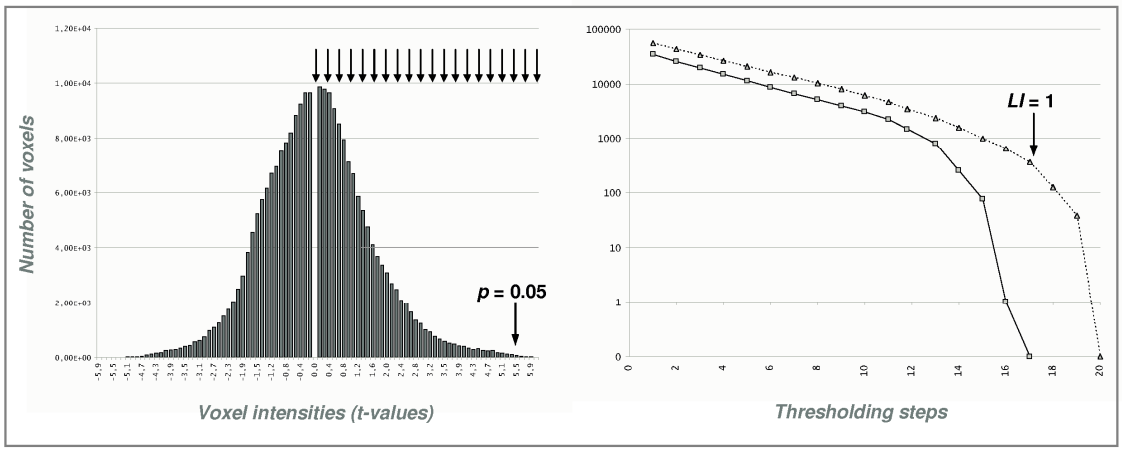

Figure 1: Left: Histogram of a typical statistical image (t-map) and illustration of the thresholding done for the lateralization curves (arrows). Right: corresponding number of voxels on the right (squares, solid line) and on the left (triangles, dotted line). Note logarithmic scaling and artificial lateralization index due to lack of voxels on one side (see text for details).

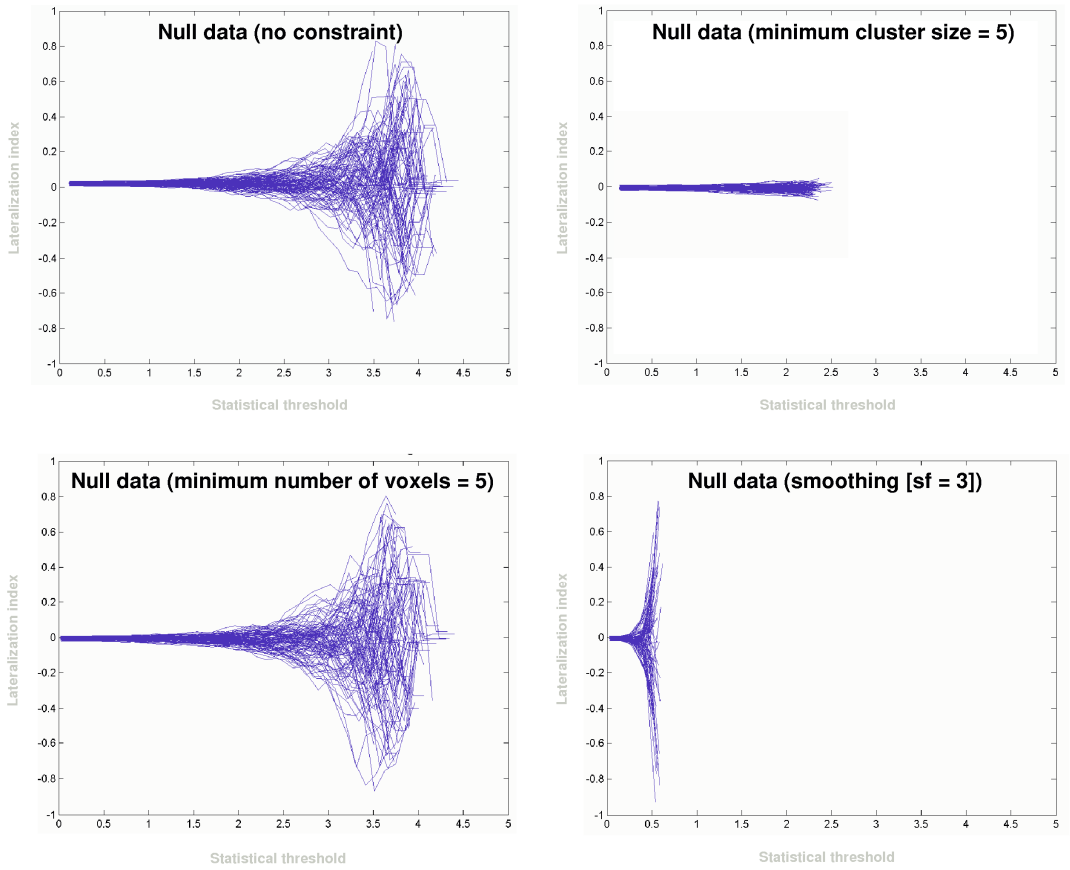

Figure 2: Iterative lateralization curves for synthetic null data (n = 100), with no constraint (top left panel), requiring a minimum number of 5 voxels (bottom left), requiring a minimum cluster size of 5 voxels (top right) and using smoothing (bottom right) before calculating a lateralization index. Note very effective control of false-positive extreme values by the cluster size criterion.

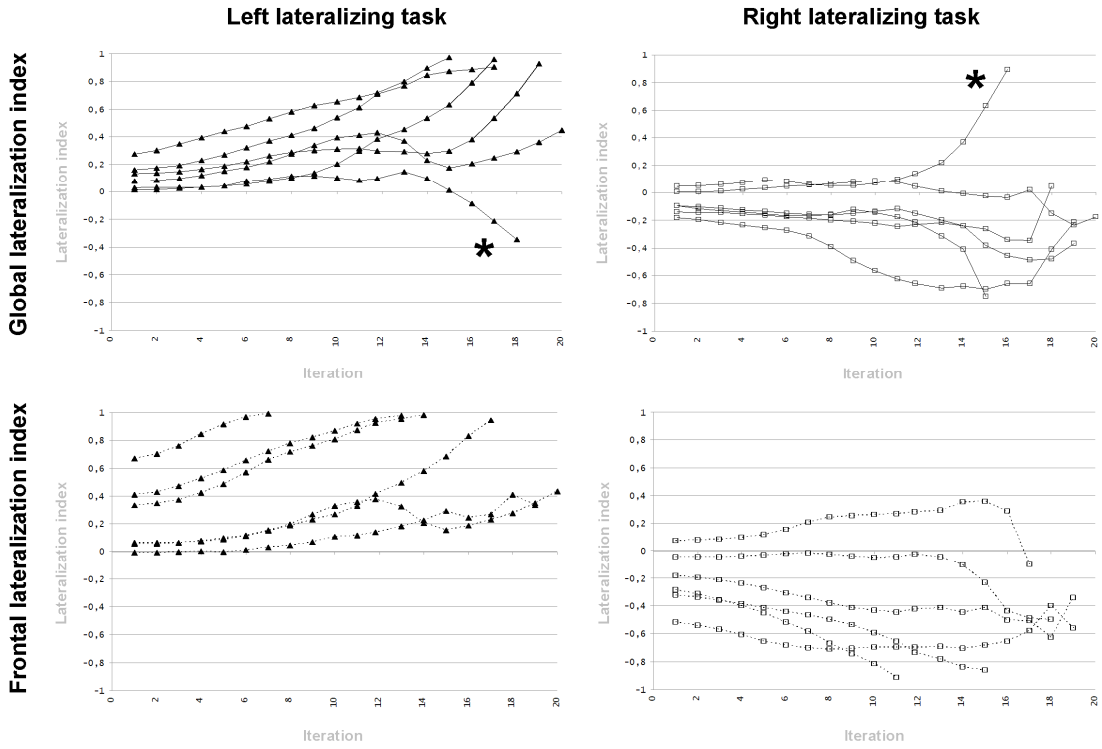

Figure 3: Iterative lateralization curves for all subjects and all tasks: global lateralization indices (top panels, solid lines) versus frontal lateralization indices (bottom panels, dashed lines) for the right- and left-lateralizing tasks. Note lack of artificially shifted lateralization (*) for the regional lateralization index.

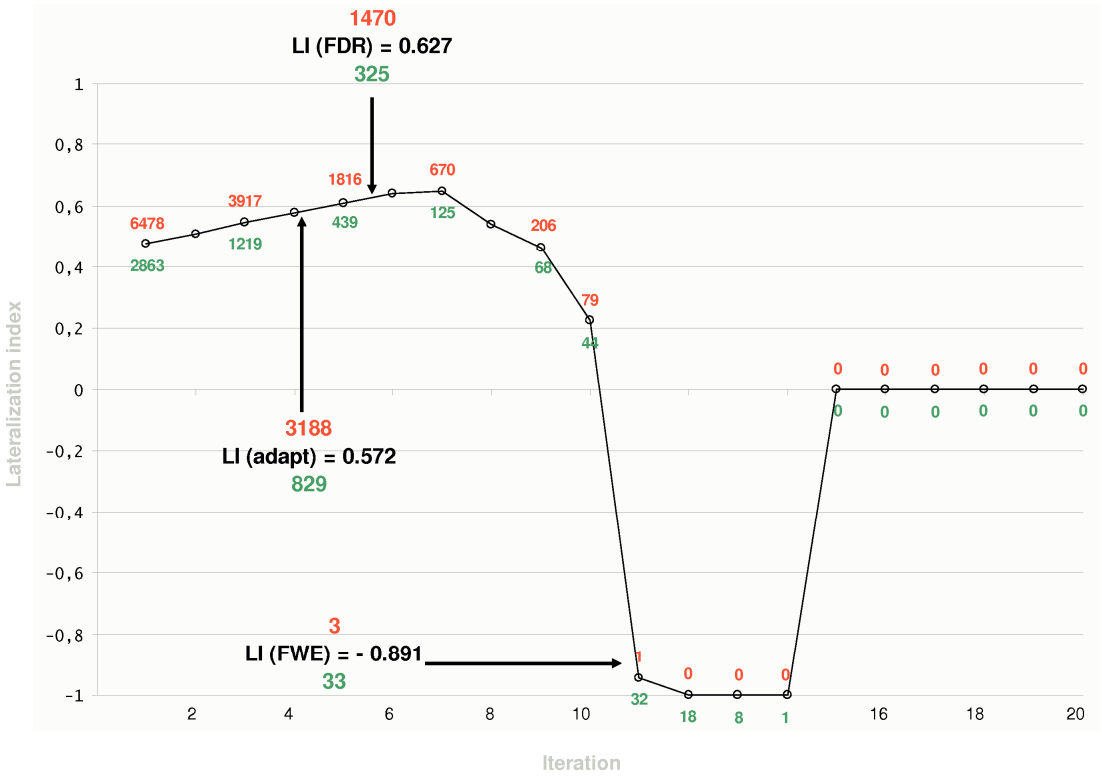

Figure 4: Adverse effect of applying different thresholds: for each iteration, the number of contributing voxels on the left (top) and right (bottom) is included. Note severely skewed lateralization index when using the FWE-correction, which includes only very few voxels (see text for details).

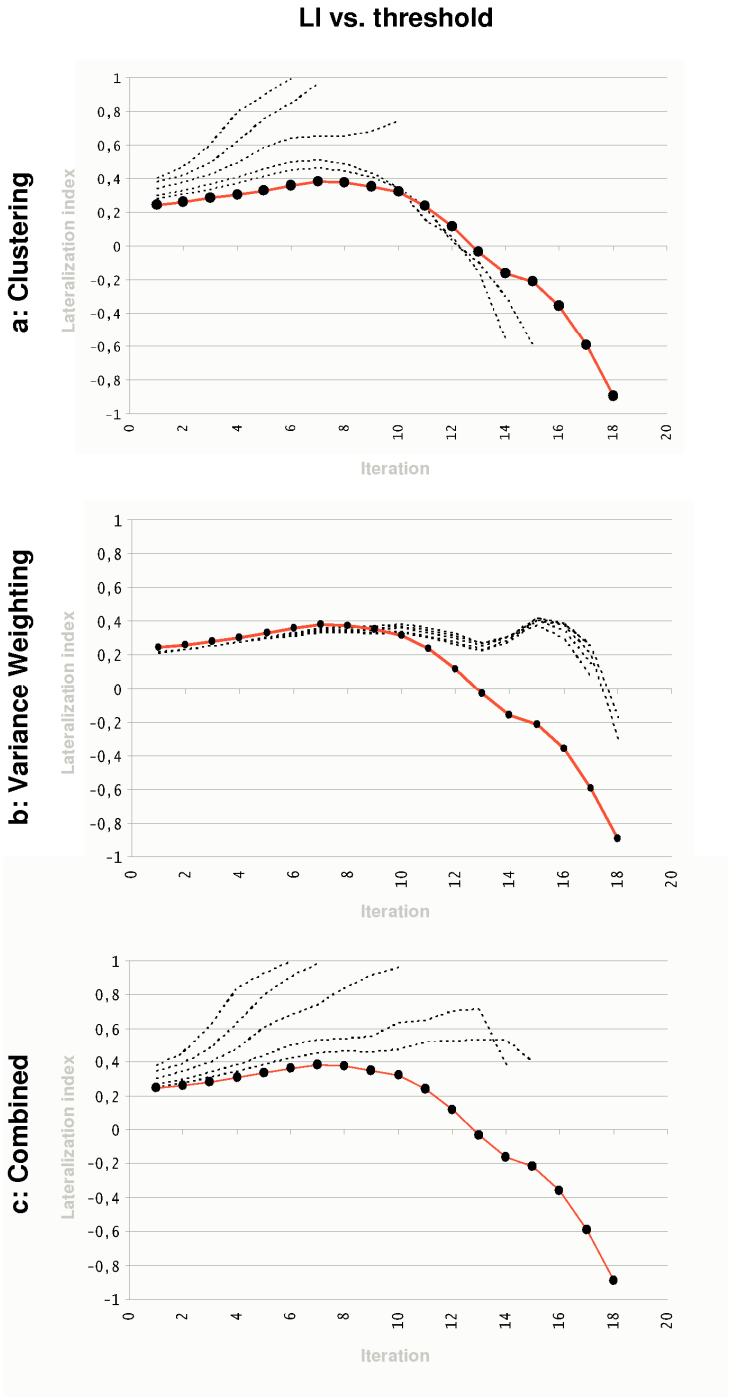

Figure 5: Effect of optional steps: clustering (a, top), variance weighting (b, middle), and combined effect (c, bottom) on the lateralization index; original solid lines (no optional steps) and effect of clustering/variance weighting as set of interrupted lines, corresponding to the increasing filter widths (sf = 3, 4, 6, 8, 10).

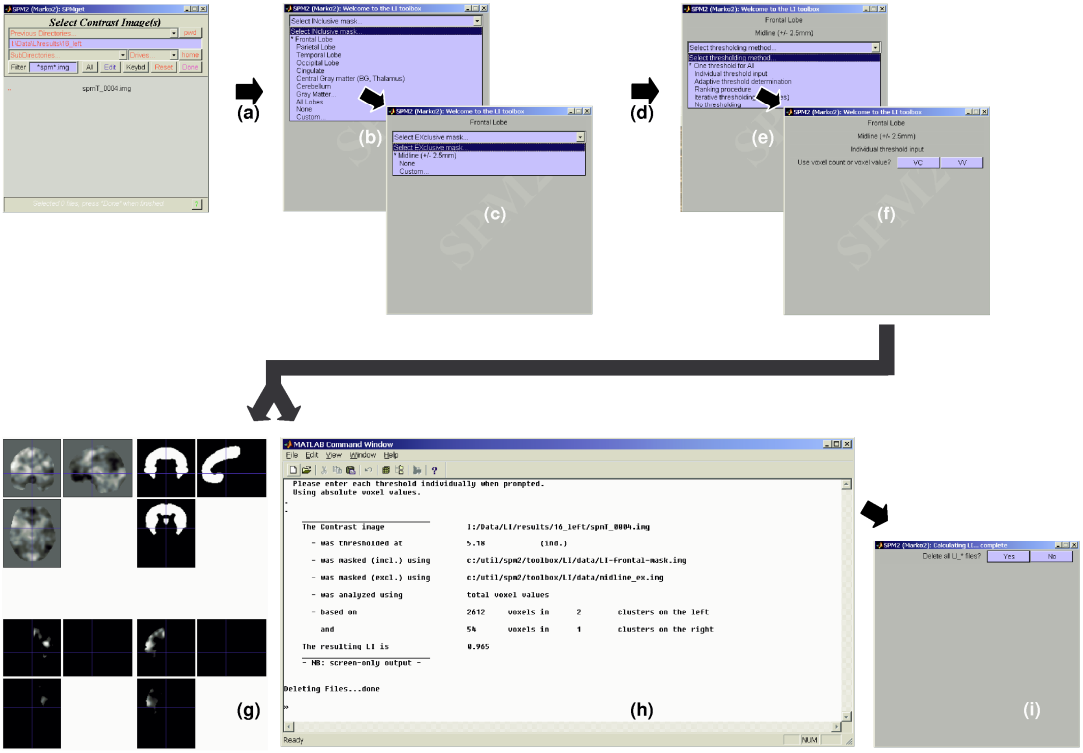

Figure 6: Illustration of the different steps when using the actual toolbox; see text for details.

Table 1

ID	Gender	IQ	EHI	Age	r VV/VC (global)	r VV/VC (frontal)
RT 1	m	115	100	22.31	0.997	0.996
RT 2	f	100	100	27.21	0.996	0.998
RT 3	m	137	100	24.49	0.999	0.994
RT 4	f	137	40	27.04	0.994	0.988
RT 5	m	110	100	20.08	0.999	1.000
RT 6	f	98	100	24.27	0.999	0.994
LT 1	m	112	100	23.21	0.982	0.998
LT 2	f	114	100	19.43	0.998	0.991
LT 3	f	107	100	23.90	0.991	0.990
LT 4	m	138	50	25.87	0.891	0.983
LT 5	f	126	100	22.59	0.985	0.974
LT 6	f	112	100	23.34	0.987	0.991

Table 1: Demographic properties of subjects and correlations between voxel value and voxel counting approaches (see text for details); RT/LT: subject performing a right/left lateralizing task; EHI: Edinburgh handedness inventory; r: Spearman's correlation coefficient; VV: voxel-value; VC: voxel count.

Table 2

Thr (FWE)	**LI** (VV)	**LI** (VC)	**LI** (cl)	**LI** (vw)	**LI** (cl+vw)	**Thr** (FDR)	**LI** (VV)	**LI** (VC)	**LI** (cl)	**LI** (vw)	**LI** (cl+vw)	**Thr** (adapt)	**LI** (VV)	**LI** (VC)	**LI** (cl)	**LI** (vw)	**LI** (cl+vw)
5.18	0.86	0.84	0.96	0.88	0.97	**2.76**	0.58	0.53	0.64	0.59	0.65	**1.55**	0.39	0.29	0.43	0.41	0.46
4.74	-0.36	-0.33	-040	-0.34	-0.39	**2.53**	-0.31	-0.3	-0.34	-0.32	-0.35	**1.78**	-0.3	-0.28	-0.32	-0.3	-0.34
5.16	0.19	0.23	0.15	0.29	0.51	**2.59**	0.35	0.35	0.44	0.3	0.41	**1.73**	0.3	0.28	0.36	0.27	0.33
4.73	-0.09	-0.08	-0.08	-0.13	-0.13	**2.37**	-0.08	-0.07	-0.08	-0.1	-0.11	**3.16**	-0.09	-0.09	-0.09	-0.12	-0.13
5.14	0.47	0.47	N/A	0.17	N/A	**3.15**	0.43	0.43	0.46	0.41	0.4	**1.17**	0.3	0.26	0.34	0.31	0.36
4.71	-0.07	-0.07	-0.08	0.03	0.02	**2.29**	-0.06	-0.06	-0.08	0.01	0	**2.83**	-0.06	-0.06	-0.08	0.02	0

Table 2: Overview of basic statistical properties and comparison of different ways to determine a lateralization index for 6 representative statistical maps (spm_T-images). adapt: adaptive thresholding; cl: clustering; FWE: family-wise correction for multiple comparisons; FDR: false discovery rate-based correction for multiple comparisons; thr: threshold; VC: voxel count; VV: voxel-value; VW: variance weighting (see text for details). All values are global lateralization indices with the exclusion of midline structures (± 5 mm).

A bootstrap approach for assessing lateralization in functional imaging data

Marko Wilke [1,2] & Vincent J. Schmithorst [3]
[1] Pediatric Neurology, Children's Hospital and [2] Section of Experimental MR of the CNS, University of Tübingen, Germany
[3] Imaging Research Center, Cincinnati Children's Hospital Medical Center, University of Cincinnati, OH, USA
Marko.Wilke@med.uni-tuebingen.de

Introduction

The calculation of a lateralization index LI = $(\Sigma_{left} - \Sigma_{right}) / (\Sigma_{left} + \Sigma_{right})$ is a commonly-used approach to assess laterality effects in neuroimaging data. However, while a single number is convenient, its strong threshold dependency lead to the implementation of lateralization curves, exploring the whole range of intensity values in an image [1]. Still unresolved, though, is the fact that even a laterality curve does not allow to assess the influence of (statistical or artifactual) outliers, severely hampering the reliability of such an approach.

Here, we present a new approach to calculate LI's and assess the homogeneity of the underlying data, using a bootstrap approach. This method makes no assumption on the distribution of the underlying data but instead iteratively re-samples, with replacement, the dataset to assess its structure.

Methods & Results

Bootstrap approach: The concept of threshold-dependent laterality curves was adopted, thresholding the input (e.g., a statistical image volume) at regular intervals (default: 20). At each threshold, a bootstrap procedure was employed to generate 100 re-samples from each side, with the following characteristics: re-sample size (default: 25% of input size), minimum re-sample size (default: 5 voxels), maximum re-sample size (default: 10.000 voxels). From these 100 bootstrapped re-samples from each side, all possible lateralization indices (10.000) were calculated at each threshold. The algorithm was written in Matlab (Mathworks, Natick, MA, USA) and implemented in spm2 (FIL, UCL, UK).

Specificity: In order to emphasize robustness and specificity, a trimmed mean $_{25}$ was used to derive a mean LI at each threshold, disregarding the upper/lower 25% of the LI-matrix. This will effectively exclude outliers. In order to gain regionally specific information, standard anatomical masks [2] were implemented to restrict analyses to pre-defined regions of interest.

Sensitivity: To derive a single overall LI value, a mean from all trimmed mean values was obtained. Emphasizing sensitivity, a weighted mean was used here, weighting the mean value from each step with the corresponding threshold height. This will over-proportionally weight LI's from higher thresholds.

Outlier detection: Even very few voxels with extreme values will severely influence lateralization index calculations. In order to detect such outliers, the bootstrap algorithm was fine-tuned to only sample a subset of the initial input sample (default: 25%), pronouncing the influence of such extreme values even more. The resulting LI-matrix was then converted to normalized z-scores and plotted as a histogram, allowing to detect outliers. Additionally, the minimum and maximum values from each step were plotted in the laterality curves to further illustrate data homogeneity.

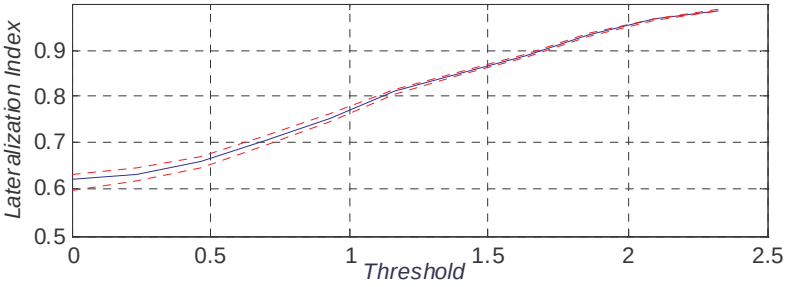

***Fig. 1**: results from the vowel-identification task: consistent left-dominant lateralization within the frontal lobe. Note narrow spread of minimum and maximum values (red lines). Over all thresholds, the resulting standard LI is .8 ± .13; the weighted LI is .88*

Data: Data from a real-life fMRI experiment was used to illustrate the results from the algorithm, from a subject performing a left-lateralizing language task [subject #2 in 3]. Additionally, phantom data was generated in the form of a random noise dataset; to illustrate the outlier detection capabilities, 2 voxels were designated to be outliers with 10 times the maximum image value.

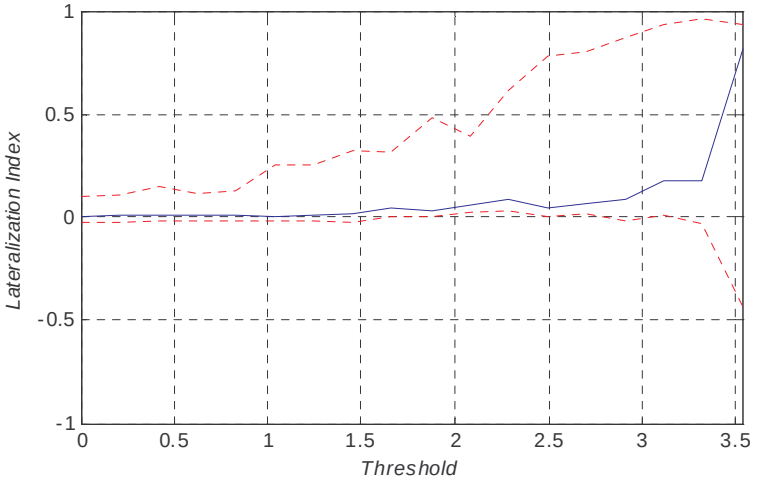

***Fig. 2**: results from the random dataset with 2 outliers on the left: terminal outlier-induced left-dominant lateralization within the frontal lobe. Note very wide and asymmetrical spread between minimum and maximum values (red) and relative robustness of the trimmed mean (blue line). Over all thresholds, the standard LI is .13 ± .25; the weighted LI is .16*

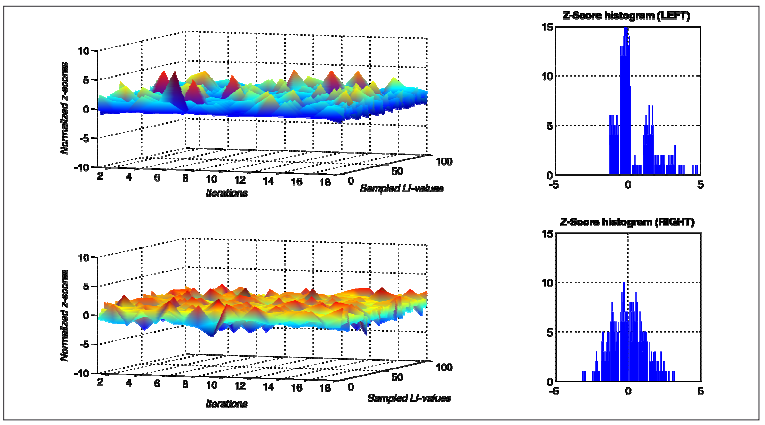

***Fig. 3**: Histogram analyses of the same dataset as in Figure 2, designed to detect the effect of outliers. Left: After transforming all lateralization indices from all thresholding steps to z-scores, outliers are clearly visualized (prominent blue peaks on the left). Right: the same effect is evident in the shift of the 2D-histogram; note normally distributed values on the right (bottom graphs)*

Discussion

Based on the concept of threshold-dependent laterality curves, we applied a bootstrapping approach to calculating lateralization indices from neuroimaging data. At each thresholding step, a LI-matrix of 10.000 values is generated, from which a trimmed mean is calculated (emphasizing specificity). These trimmed mean values from each threshold are then used to generate an overall mean, weighting each value by the corresponding threshold value. This takes into account the fact that lateralization indices from higher values should be more meaningful as more noise is excluded; such a procedure emphasizes sensitivity.

The bootstrap approach allows to explore the available data in such a way that, for the first time, confidence intervals can be attached to a given lateralization index, allowing to routinely assess the underlying data quality (see Figure 1). In a random dataset, even single outliers are detected (see Figures 2 and 3) We therefore believe that this approach is useful in exploring laterality effects in (functional) neuroimaging data.

Literature: [1] Deblaere *et al.*, Neuroradiology 2004; [2] Tzourio-Mazoyer *et al.*, NeuroImage 2002; [3] Wilke *et al.*, NeuroImage 2006

This study was supported by the DFG (SFB550/C4). The code is freely available, please contact us at marko.wilke@med.uni-tuebingen.de

A combined bootstrap / histogram analysis approach for computing a lateralization index from neuroimaging data

Marko Wilke [1,2] and Vincent J. Schmithorst [3, 4]

[1] *Department of Pediatric Neurology and Developmental Medicine, Children's Hospital, and*

[2] *Section for Experimental MR of the CNS, Dept. of Neuroradiology,*

University of Tübingen, Germany

[3] *Department of Pediatrics, University of Cincinnati, and* [4] *Imaging Research Center, Cincinnati*

Children's Hospital Medical Center, Cincinnati, OH

Running title: Bootstrapped lateralization index

Corresponding author:

Marko Wilke, MD
Department of Pediatric Neurology and Developmental Medicine
Children's Hospital, University of Tübingen
Hoppe-Seyler-Str. 1
72076 Tübingen, Germany
Phone: + 49 7071 – 29 83416
Fax: + 49 7071 – 29 5473
e-mail: Marko.Wilke@med.uni-tuebingen.de

Abstract

Cerebral hemispheric specialization has traditionally been described using a lateralization index (LI). Such an index, however, shows a very severe threshold dependency and is prone to be influenced by statistical outliers. Reliability of this index thus has been inherently weak, and the assessment of this reliability is as yet not possible as methods to detect such outliers are not available. Here, we propose a new approach to calculating a lateralization index on functional magnetic resonance imaging data, by combining a bootstrap procedure with a histogram analysis approach. Synthetic and real functional magnetic resonance imaging data was used to assess performance of our approach. Using a bootstrap algorithm, 10.000 indices are iteratively calculated at different thresholds, yielding a robust mean, maximum and minimum LI and thus allowing to attach a confidence interval to a given index. Taking thresholds into account, an overall weighted bootstrapped lateralization index is calculated. Additional histogram analyses of these bootstrapped values allow to judge reliability and the influence of outliers within the data. We conclude that the proposed methods yield a robust and specific lateralization index, sensitively detect outliers and allow to assess the underlying data quality.

Introduction

Hemispheric specialization of the brain has been the focus of a large number of studies, mainly using imaging methods like positron emission tomography (PET) or functional magnetic resonance imaging (fMRI; for review, see Cabeza & Nyberg, 2000, and Hugdahl & Davison, 2002). In this note, we propose to apply the bootstrap concept to calculating lateralization indices from neuroimaging data.

Background: bootstraps and lateralization indices

The term “bootstrap” reportedly comes from the legendary German figure Baron of Münchhausen, who dragged himself out of a swamp by pulling on his bootstraps. In statistics, the term describes a technique that tries to find the sampling distribution of a sample, by repeatedly re-sampling, with replacement, the original sample. In other words, several resamples are generated from a given original sample in order to estimate a bootstrap distribution that allows approximating the “real” distribution of the original sample. It is important to note that a bootstrap does not add or replace data from the original distribution, but only uses multiple resamples of the original (Davison & Hinkley, 1997; Hesterberg *et al.*, 2005; Janssen & Pauls, 2003, Moore *et al.*, 2002). This is illustrated in equation 1

$$\begin{array}{cccccccccccc} 1 & & 2 & & 5 & & 4 & & 5 & & 1 & \\ 2 & & 5 & & 4 & & 1 & & 5 & & 5 & \\ 3 & \Rightarrow & 5 & \& & 3 & \& & 3 & \& & 3 & \& 1 & \& \ldots \\ 4 & & 1 & & 3 & & 5 & & 4 & & 4 & \\ 5 & & 3 & & 3 & & 5 & & 3 & & 4 & \\ & & & & & & & & & & & \\ i & & r_1 & & r_2 & & r_3 & & r_4 & & r_5 & \ldots\ r_n \end{array} \qquad \text{(Equation 1)}$$

It is obvious that both computationally and statistically, the procedure is mainly influenced by the size of the resample (size $_r$) and the number or resamples (n). This size $_r$ can be variable: while it is in most cases identical to size $_i$, it can be less than that (the special case of it being size $_{i-1}$ is called a jackknife procedure; Davison & Hinckley, 1997). The number of resamples is mainly limited by the computational demand when dealing with larger sample sizes; typically, several hundred to thousand resamples are used (Davison & Hinkley, 1997; Hesterberg *et al.*, 2005). Bootstrap approaches have been used before in neuroimaging, mainly for fMRI (Auffermann, Ngan & Hu, 2002; Prohovnik *et al.*, 2004) and diffusion tensor imaging (Jones & Pierpaoli, 2005; Lazar & Alexander, 2005).

Hemispheric specialization is a common question in functional and structural neuroimaging as well as in the cognitive neurosciences in general (for review, see Hugdahl & Davison, 2002). Approaches to describe this asymmetry mostly aim at presenting a single number in order to allow for the comparison of results. Akin to classical approaches to describing handedness (Oldfield, 1971), lateralization has traditionally been described using a lateralization index LI, computed as

$$LI = \frac{\sum activation_{left} - \sum activation_{right}}{\sum activation_{left} + \sum activation_{right}} \qquad \text{(Equation 2)}$$

Resulting from such an equation is a value between 1 (complete left-lateralization) and -1 (complete right lateralization). However, issues concerning such an index include vulnerability to outliers, a strong threshold dependency and the lack of immediate inference on data quality (Adcock *et al.*, 2003; Deblaere *et al.*, 2004; Gaillard *et al.*, 2002; Holland *et al.*, 2001).

Approach

In this manuscript, we describe the application of the bootstrapping concept to the calculation of lateralization indices in (functional) neuroimaging data. The main aims are (1) to allow for the assessment of data quality by detecting (statistical or artifactual) outliers; (2) to increase the stability of a given index by broadening the base of underlying information and by restricting outlier influence; (3) software implementation should allow easy usage within a publicly available imaging analysis tool, and (4) to remove the necessity of defining a cutoff threshold for interpreting lateralization in functional neuroimaging data.

Regarding No. 1, factors possibly making the algorithm more sensitive to outliers shall be systematically explored. Moreover, parameters allowing for the detection of outliers should routinely be derived from the analysis of the data and presented as part of the results.

In order to increase stability as defined in No. 2, the bootstrap algorithm is ideally suited to broaden the available data basis by providing multiple resamples. Our approach for robustly excluding outliers from this data is described below.

Considering No. 3, available software suites and the underlying solutions, we implemented our algorithms in MATLAB (The Mathworks, Natick, MA) and designed a graphical user-interface based on routines available within the spm-software environment (SPM2, Wellcome Department of Imaging Neuroscience, University College London, UK).

As to No. 4, we decided to adopt the concept of threshold-dependent laterality curves, iteratively exploring increasing thresholds (Deblaere *et al.*, 2004), as the basis for further calculations. The image under investigation (for example, a t-map; Figure 1) is

thresholded at regularly-spaced intervals. The original implementation of the laterality curves submitted the surviving voxels on the left and the right to equation 1 to yield a lateralization index for each threshold. The resulting diagrams show the obtained lateralization index (in y-direction) versus the threshold (in x-direction).

For our approach here, we used the surviving voxels as input samples for a bootstrap procedure in such a way that from the original single input sample, a multitude of bootstrapped re-samples was generated. From these resulting *n* samples from each side, all possible lateralization indices were calculated. This procedure is repeated during each thresholding step. See also Figure 1 for an overview of the steps.

Issues: speed, specificity, accuracy & outlier detection

Speed: As fMRi deals with large datasets, the computational steps easily become very time consuming. Limiting steps within the framework of the bootstrap are not only the size of the input sample ($size_i$), but also the number of resamples (*n*) and the size of the resample ($size_r$). We believe that limiting both the upper and the lower size of this resample makes sense in this setting, although for different reasons. First, when analyzing large input samples (typically occurring at lower thresholds), the stability of the resulting values should be very high, owing to the large number of contributing voxels. We therefore postulated that it is justified to specify an upper size limit *max* for the resulting bootstrap sample in order to speed up processing at low thresholds. Samples smaller than this will be sampled completely (if complete sampling is chosen, see below). Secondly, very small samples pose additional dangers: a single remaining voxel on one side will lead to a lateralization index of +/-1, which is not a plausible scenario, biologically, statistically, or computationally. We therefore suggest specifying a lower boundary, i.e. the algorithm aborts if a minimum number *min* of

voxels is not found. As a further safeguard against scattered single voxels, a warning is issued if a certain minimum cluster size is not met (default is ET = 5 voxels).

Sensitivity vs. Specificity: In order to punish outliers and restrain their influence on the ensuing results, we used a "trimmed mean" value when analyzing the resulting LI-values from each threshold (along the y-direction of a laterality curve diagram, i.e. for each single iteration). A trimmed mean$_{25}$ only uses the mean 50% of data while disregarding the upper and lower 25% of datapoints (Hesterberg *et al.*, 2005). In a sample skewed by outliers, such a trimmed mean is a more representative measure of the "true" center of the distribution. On the other hand, as the range of resulting value is an important indicator of data homogeneity, we opted to retain this information by plotting the minimum and maximum LI-values from each step. These values reflect the range of results obtained at this threshold, which will be small (in the case of homogenous data) or large (in the case of inhomogeneous data).

Regarding the specificity of obtained results, (functional) neuroimaging data is usually thresholded to ensure the significance of obtained results. Classically, a lateralization index is computed from these "significant" voxels only, which are obtained after additionally accounting for multiple comparisons. However, a number of different correction methods exist which will, while statistically equally legitimate, yield different thresholds (Marchini & Presanis, 2004). Among them are approaches favoring specificity (family-wise error correction, FWE) or sensitivity (false discovery rate, FDR; Nichols & Hayasaka, 2003); the definition of voxels to exclude is therefore not straightforward. Previous attempts did not threshold the statistical images at all prior to lateralization analyses (Holland *et al.*, 2001). Consequently, however, many voxels showing no correlation with the task will also be included, which invariably introduces noise in the calculations. We therefore opted to use an

approach that allows to account for the “meaningfulness” of values obtained from different thresholds by attributing different weights to them. While no attempt was made to adopt any given scheme for determining significance, it is immediately obvious that a voxel showing a higher correlation with the task should have a greater impact on the ensuing results. We therefore decided to employ a *weighted mean* (along the x-direction of a laterality curve diagram, i.e. over the results from all thresholds). A weighted mean computes a mean of a given sample ($x_1 - x_n$), but takes into account a weighting factor w_i for each datapoint x:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i * x_i}{\sum_{i=1}^{n} w_i} \qquad \text{(Equation 3)}$$

In our case, the obvious choice for such a weighting factor is the threshold at which the image was thresholded in order to generate the value of x, which will result in a progressively stronger weighting of the lateralization index values obtained at higher thresholds. Note that, with a constant weighting factor w, a weighted mean is equivalent to the standard arithmetic mean. Therefore, a weighted mean of the results from all thresholding results (i.e., along the x-direction of a laterality curve diagram) can be based on the trimmed means obtained at each threshold, thus combining stability and specificity. Accordingly, all trimmed mean values (together with the respective thresholds as weighting factors) are submitted to equation 3 to yield an overall weighted mean.

Accuracy & outlier detection: As mentioned above, the size $_r$ of the resample usually is equivalent to the size $_i$ of the input sample (see equation 1). In this scenario, each bootstrapped resample r is a complete resample (with replacement) of i, and with large sample sizes, only a limited number of datapoints will not be sampled. However,

(Equation 4)

when decreasing $size_r$ in relation to $size_i$, the number of datapoints not sampled will increase accordingly, as shown in equation 4:

$$size_r = k * size_i$$

Obviously, modifying the resample ratio *k* within a range of 0-1 will result in a resample *r* that is smaller than the input sample *i*. Such a scenario is illustrated in Figure 2 (left panel): a sample of 100 points (Y-axis, left diagram) is randomly sampled 100 times (X-axis, left diagram) with a frequency of $k = .25$. This results in 100 samples of 25 datapoints each, and the resulting random frequency of each original datapoint in the resamples is illustrated in the histogram (Figure 2, right panel): on *average*, each point occurs 25 times in the resamples. While the bootstrapped sample is still based on 2500 datapoints as opposed to the original 100, the resulting samples will, on the whole, be much more likely to detect data inhomogeneity than a sample based on $k = 1$. This is simply because potential outliers will not be present in each sample, thus widening the spread of resulting averages. Considering the example of datapoint 50 (circled, Figure 2, left panel) being an outlier, it will only be present in 20/100 resamples (right panel). In order to increase an algorithms' sensitivity to outliers, the value of *k* can thus be systematically decreased. This might result in a decrease of specificity as the correspondence between each individual resample *r* and the input sample *i* will naturally also decrease. On the other hand, this is outweighed by a large number *n* of resamples. We therefore hypothesized that varying *k* should have discernible effects on the ability to detect outliers while, due to the large number of resamples and the outlier protection implemented above, accuracy should not be severely affected.

Methods

In order to compare the results from our algorithm, we randomly picked imaging data from a previous study on language lateralization (Wilke *et al.*, 2006; see also for details on data acquisition and processing). Data from 5 healthy subjects (3 boys, 2 girls, mean age 12.8 ± 2 years, right-handed with an average Oldfield-score of .77, range .64-1 [Oldfield, 1971]), performing a robust left-lateralized language task (Fernandez *et al.*, 2001) were selected. The resulting t-maps from each subject were analyzed with regard to lateralization within the frontal lobe, defined using masks based on anatomical definitions (Tzourio-Mazoyer *et al.*, 2002).

The bootstrapping algorithm makes use of the random number generation capabilities of the Matlab software environment. For example, consider a 3-dimensional image volume which, after thresholding and masking, consists of 500 non-zero voxels; it is then converted to a 1 x 500-vector and analyzed with a re-sample ratio of $k = .25$. Consequently, 125 random integer numbers are generated in the range from 1-500 which are then used to randomly, with replacement, sample the input data. This procedure is repeated 100 times, resulting in 100 bootstrapped samples from each side.

User interaction is completely GUI-driven, using a toolbox plug-in for spm2. Results from our approach should be compared to the classical approach to only assess voxels surviving a given threshold of significance, either defined using the family-wise correction for multiple comparisons or the false-discovery rate as implemented in spm2. To this effect, fMRI-data from the earlier study (Wilke *et al.*, 2006) is analyzed using all approaches, and results are assessed qualitatively.

To illustrate the influence of outliers on lateralization index calculations, the value of one voxel in a t-map (subject #1 in table 1) was iteratively varied to be between 1 and 50 times the maximum value in the individual image (outlier weight *ow* = 1-50). This serves to investigate robustness of the ensuing lateralization indices. To further assess stability of results from the bootstrap algorithm, 100 re-runs of the same (unaltered) dataset were analyzed, with different resample ratio settings ($k = 1$, $k = 0.5$ and $k = 0.25$).

Presets

We used a default of 20 thresholding intervals (equally-sized steps from 0 to the maximum value in the masked image) and to generate 100 bootstrap samples for each side (coded to be five times the number of thresholding steps to allow for a combined more exhaustive assessment). Owing to the open nature of MATLAB-script files, both values can easily be adjusted within the code. From these 100 samples from each side, a total of 10.000 lateralization indices for each thresholding step results, and an overall maximum of 200.000 indices (with default values: doubling the number of iterations will lead to an overall matrix of 1.6 million LIs). When default settings are used, results are stored in a matrix of 20 (thresholds, in the X-dimension) by 10.000 (lateralization indices, in the Y-dimension.)

By default, the upper sample size limit was set to $max = 1000$ voxels, i.e., from samples larger than this, only 1000 voxels will randomly be drawn. As discussed above, a lower boundary should also be set in order to avoid a lateralization index based on a very small number of voxels: we suggest using $min = 5$ voxels as the minimum bootstrap resample size. Of note, when using a smaller resample size in relation to the input $size_i$ ($k < 1$), in order to keep the bootstrapped sample $size_r$

constant, *k* needs to be taken into account by extending the minimum number of voxels to *min* = *min*/*k*, such that, with a sampling size of 50% (*k* = .5) the minimum input sample size would be 10 voxels.

The histogram of the overall lateralization index-matrix is plotted to allow assessing the resulting distribution at each threshold, giving an impression of how normally distributed each underlying sample is. Additionally, the matrix of lateralization indices from each side was converted to normalized z-scores according to

$$z_{LI} = \frac{LI - Mean_{LI}}{SD_{LI}} \qquad \text{(Equation 5)}$$

These were plotted as a function of iteration in order to allow for the visual assessment of data homogeneity over all thresholding steps. A histogram of these z-scores was also generated to further assess the (normal or skewed) distribution of the z-scores.

Results

For the fMRI-data of the 5 healthy subjects, our bootstrapped results and lateralization indices from significant voxels only are shown in Table 1. It is apparent that no results are obtained from some “classical” calculations if a minimum number of voxels is required (as in our algorithm; if this would be tolerated, a value of 1 resulted). Our lateralization indices are concordant with the results from the other approaches, with the weighted mean indicating stronger lateralization than the overall and trimmed mean. The excellent agreement for unaltered fMRI data is also evident in the stepwise comparison of results from a single dataset (subject #1), comparing results from a classical lateralization curve with our bootstrap approaches (Table 2).

Overall lateralization is clearly influenced by even a single outlier (Figure 3): from an outlier weight $ow \geq 7$, all negative values result in the classical lateralization curves, and with $ow > 8$, the result is almost uniformly -1 (average of weighted means: -.99). This effect becomes more pronounced with increasing outlier weights and increasing thresholds (as voxel numbers decline). An abated effect is seen in the bootstrapping approach with complete resampling ($k = 1$): an average weighted mean lateralization index of -.3 results if $ow > 8$. The outlier influence strongly decreases when using a smaller resample size ($k = .5$ and $k = .25$). With $k = .5$ and $ow > 8$, an average weighted lateralization index of .4 is returned. In the last case, the correct left lateralization is retained in all 50 runs (average weighted lateralization index for $ow > 8$: .68). Note absence of minimum/maximum bounds in these graphs for better accessibility.

When repeating our bootstrapping procedure 100 times on actual fMRI-data (Figure 4, upper panels), the stability of the resulting mean lateralization indices is

extremely high. In all cases, the sampling ratio *k* has no discernible influence on either the mean or the detected range if high-quality fMRI-data is examined (Figure 4, lower panels).

The actual output of the algorithm for a single dataset (with 2 outliers, *ow* = 10, on the right, *k* = 1) is shown in Figure 5 & 6: note much wider range of detected lateralization indices on the "artifacted" (right) side (Figure 5, top panel). The histogram of all lateralization indices shows several smaller, irregular peaks from outlier-influenced samples (Figure 5, lower panel). The z-score histogram for the left side (Figure 6, upper panels) shows an evenly distributed pattern over the whole range of thresholds, and the z-scores are nicely normally distributed. For the right side (Figure 6, lower panels), containing the outliers, a much less homogenous z-score sample results, with extreme values (in "colder" colors) present over all thresholds and a consecutively skewed z-score histogram.

Discussion

In this work, we applied the bootstrapping concept to the calculation of lateralization indices. Special emphasis was put on robust calculations and on sensitive outlier detection in order to both avoid and detect possible outlier influences.

When comparing our results with the “classical” approaches, Table 1 shows no results for the FWE-corrected approaches in 3/5 cases as no voxels survive thresholding on the right. While this will result in a very “clear-cut” lateralization index of LI = 1, it seems (mathematically and biologically) dangerous to accept no voxels on one side of the equation. Due to our minimum size criterion, our algorithm will abort here, as suggested before (Deblaere *et al.*, 2004). Our weighted mean results, designed to combine robustness and specificity, are in every case concordant with results from classical approaches. The false discovery rate only aims at controlling the rate of false positives, it is therefore more sensitive than the stricter family-wise error correction (Nichols & Hayasaka, 2003). With our values consistently indicating stronger lateralization than the values obtained from the FDR-corrected thresholds, a more specific analysis could be postulated, without reaching the rigidity of the FWE-correction. Of note, the user-defined optional input of a lower threshold (e.g., an FDR-corrected cutoff-value) will also allow to individually explore “significant voxels” only with our algorithm (see below).

The calculation of a lateralization index even in the form of a lateralization curve is strongly susceptible to outliers (Figure 3). If only a single voxel has a value of more than 7 times the maximum image value, the expression of an opposite laterality

artifactually results. While our simulation of outliers may not be a very realistic scenario, it was designed to demonstrate the algorithm's outlier detection capabilities. Only this sensitive outlier detection enables an informed choice with regard to whether a dataset is usable or not. In the case of our bootstrap approaches, a sampling ratio of $k = 1$ already shows an increased stability; lower values of k (.5 and .25, respectively), further improve stability against these outliers, with the smallest sampling ratio yielding the best results: in all cases, correct strong left lateralization is retained despite the influence of the outlier. Three additional points seem worth mentioning: one, in every case the widening spread of minimum/maximum LI-values and the consecutively skewed z-score histograms would have alerted the user as to the presence of strong data inhomogeneity even when a "nice-looking" mean curve results. This is a decisive advantage over the classical lateralization curve approach. The effect is not shown in Figure 3, but is illustrated in Figures 5 & 6. Two, it is interesting to notice that the outlier effect in the bootstrapped samples consistently only becomes apparent after the second or third iteration, while it is present from the first iteration on in the classical lateralization curves. As it is independent from the sampling ratio, it must be a consequence of the maximum sample size restriction: only then (approaching the upper bootstrap size limit of $max = 1000$ voxels) is the input sample sampled completely (in the case of $k = 1$) and outliers start to influence results. This further argues in favor of using smaller resamples that are, individually, less likely to contain a small number of outliers in many samples. Third, the amazing stability of the last approach (with $k = .25$) is likely to be enhanced by the "trimmed mean$_{(25)}$" we employed: if an outlier is only present in 25% of samples and the upper and lower 25% of the data are not used, the combined robustness against outliers must be expected to be very high.

However, as a decrease in *k* will require a greater input sample size $_I$ if *k* is constant, there is a tradeoff on how low *k* can be without having to prematurely abort iterations (note 11 iterations in the classical lateralization curve and only 10 iterations with $k = .25$, in Figure 3). Therefore, we suggest to use $k = .25$ by default, with lower values to be used when exploring inhomogeneous datasets (see below for user options). To avoid premature abortion of iterations when using low resample ratios, the algorithm now automatically adjusts the resample ratio when encountering low voxel counts in order to keep the minimum bootstrap sample size *min* constant (e.g., if a dataset with size $_i = 9$ voxels is explored with $k = .5$, $min = 5$ is not met; therefore, *k* is adjusted to be $k = .55$ such that size $_i * k \geq min$; this adaptation is naturally restricted by $k \leq 1$). Even using very low *k*-values (e.g., an exploratory $k = .05$), results in an unaltered dataset remain virtually indistinguishable from a straightforward lateralization index (Table 2). It should also be noted that even with such low resample ratios, the input data will still be oversampled (if $k = .05$, an input sample of size $_i = 100$ will still yield 100 resamples of size $_r = 5$, totaling 500 datapoints).

For unaltered fMRI-data the variability of the resulting lateralization indices reaches a maximum of .2%, strongly decreasing with increasing thresholds (Figure 4, lower panels). The range of detected values (difference between minimum and maximum LI) shows the same pattern, at low thresholds a maximum range of 17% is detected. This indicates that, for “normal” fMRI data, virtually identical results can be expected even with lower resample ratios ($k < 1$; Table 2). Interestingly, this range is higher at lower thresholds, which is again most likely due to the upper size limit we exposed on the bootstrap sample size in order to speed up processing (in these calculations

max = 1000 voxels). Alternatively, a higher data inhomogeneity at lower thresholds could be responsible for this effect (as both low and high values enter the sampling, instead of only high values at higher thresholds). Ultimately, as overall processing of a typical fMRI dataset is completed in less than a minute on a standard PC workstation, this upper limit can be adjusted. Based on these results, the algorithm now uses a maximum sample size of *max* = 10.000 voxels. Of note, this suggestion is based on the simulations conducted here, using typical fMRI data. For high-resolution data, a higher limit or even no limit may be more adequate (to this effect, "inf" can be entered when prompted for the upper bootstrap size)

We used a trimmed mean for averaging the lateralization indices at each threshold in order to only assess the central and most representative parts of the lateralization index-matrix from the bootstrapped samples, emphasizing robustness by being much less vulnerable to outliers (Hesterberg *et al.*, 2005). The excellent agreement in all cases with straightforward LI-calculations (Table 2) constitutes a validation of the approach. The positive effect on stability is also apparent in Figure 5: although the outliers severely influence the range of obtained LI-values on the side of the artifacts, the resulting trimmed mean is much closer to the upper limit of detected lateralization indices. Such an imbalance between the trimmed mean and the distance to the upper and lower bounds is a further criterion for uneven data homogeneity between the two sides. While it is currently only a visible indicator, a mathematical marker could be derived and implemented (e.g., the difference between a trimmed $_{(25)}$ and an arithmetic [regular] mean); at this point, however, we believe that an additional parameter would only over-complicate the already complex graphical output. As to the rigid control of upper and lower bounds, relaxing the criterion of how much data to be "trimmed"

may be an option to increase sensitivity (e.g., using a trimmed mean$_{(10)}$ instead). However, a detailed examination of this effect was not done here and remains a question for future research.

The disadvantage of using a trimmed mean along the x-direction (i.e., to assess an overall lateralization index over all thresholds) is that the higher values obtained at higher thresholds will also be discarded. This is counterproductive as voxels surviving a higher threshold in functional MR-imaging data do this due to their stronger (and ultimately, significant) correlation with the task at hand (Holland *et al.*, 2001; Marchini & Presanis, 2004; Nichols & Hayasaka, 2003). In this case, therefore, it seems justified to give more weight to lateralization indices obtained from such voxels, without specifying a hard cut-off. To this effect, we implemented a weighted mean (see equation 3) to calculate an overall lateralization index from the (trimmed) mean values obtained at all thresholds, therefore increasing specificity. Ultimately, the decision on which value to use is again one of sensitivity versus specificity: a straight mean will weigh all lateralization indices the same way, whether they come from high or low thresholds. A trimmed mean will effectively exclude outliers, but will, if applied to all values, exclude low as much as high values in order to yield a robust mean. Lastly, a weighted mean (based on the trimmed means from all thresholds) will be rather immune to outliers and will give proportionately more weight to values obtained from higher thresholds. Considering the multitude of possible scenarios in which the assessment of lateralization is of interest, the decision on which value to choose cannot be expected to be the same for all cases (therefore, all values are reported, see Figure 5). We believe that, if no indicator suggests significant outlier influence, a weighted mean over all thresholds is a good compromise.

A number of variables influence the results from this algorithm: the lower threshold cutoff, the resample ratio, the minimum and the maximum bootstrap sample size. While we suggest to explore the whole range of thresholds in an image, a lower cutoff may be specified by the user, so that the option to only investigate “significant” voxels, however defined, remains. To allow flexible explorations, the user is requested to confirm the defaults (lower cutoff = 0, k = .25, minimum bootstrap sample size = 5, maximum bootstrap sample size = 10.000) or to specify his own settings. Additionally, the open nature of Matlab script files allows changing all other relevant settings within the (thoroughly documented) code.

To summarize, our algorithm not only supplies a lateralization curve, describing lateralization at different thresholds, but also a comprehensive, single lateralization index based on large body of data. At the same time, several parameters allow for the assessment of the underlying data quality, thereby offering a decisive advantage over previous approaches. We therefore conclude that the application of the bootstrapping concept to calculating lateralization indices from imaging data is fast, robust, and powerful.

References

Adcock JE, Wise RG, Oxbury JM, Oxbury SM, Matthews PM (2003)
Quantitative fMRI assessment of the differences in lateralization of language-related brain activation in patients with temporal lobe epilepsy
NeuroImage 18: 423-438

Auffermann WF, Ngan SC, Hu X, 2002
Cluster significance testing using the bootstrap
NeuroImage 17: 583-591

Cabeza R, Nyberg L, 2000
Imaging cognition II: An empirical review of 275 PET and fMRI studies
J Cogn Neurosci 12: 1-47

Davison AC, Hinkley DV, 1997
In Davison AC, Hinkley DV: Bootstrap Methods and their Application, 1st Edition, Cambridge University Press, Cambridge

Deblaere K, Boon PA, Vandemaele P, Tieleman A, Vonck K, Vingerhoets G, *et al*., 2004
MRI language dominance assessment in epilepsy patients at 1.0 T: region of interest analysis and comparison with intracarotid amytal testing
Neuroradiology 46: 413-420

Fernandez G, de Greiff A, von Oertzen J, Reuber M, Lun S, Klaver P, *et al.*, 2001
Language mapping in less than 15 Minutes: real-time functional MRI during routine clinical investigation
NeuroImage 14: 585–594

Gaillard WD, Balsamo L, Xu B, Grandin CB, Braniecki SH, Papero PH, *et al.*, 2002
Language dominance in partial epilepsy patients identified with an fMRI reading task
Neurology 59: 256–265

Hesterberg T, Moore DS, Monaghan S, Clipson A, Epstein R, 2005
Bootstrap Methods and Permutation Tests. In: Moore DS, McCabe GP (eds.): Introduction to the Practice of Statistics, 5th Ed., WH Freeman & Co, 14.1-70

Holland SK, Plante E, Byars A, Strawsburg RH, Schmithorst VJ, Ball WS Jr. (2001)
Normal fMRI brain activation patterns in children performing a verb generation task
NeuroImage 14: 837-843

Hugdahl K, Davison RJ, 2002
The Asymmetrical Brain, 2nd ed.
MIT Press

Janssen A, Pauls T, 2003
How do bootstrap and permutation tests work?
Ann Statist 31: 768–806

Jones DK, Pierpaoli C, 2005

Confidence mapping in diffusion tensor magnetic resonance imaging tractography using a bootstrap approach

Magn Reson Med 53: 1143-1149

Lazar M, Alexander AL, 2005

Bootstrap white matter tractography (BOOT-TRAC)

NeuroImage 24: 524-532

Marchini J, Presanis A, 2004

Comparing methods of analyzing fMRI statistical parametric maps

NeuroImage 22: 1203-1213

Moore DS, McCabe GP, Duckworth WM II, Sclove SL, 2002

In: Moore DS, McCabe GP, Duckworth WM II, Sclove SL (eds.): The Practice of Business Statistics: Using Data for Decisions, 1st Ed., WH Freeman & Co, 18.1-73

Nichols T, Hayasaka S, 2003

Controlling the familywise error rate in functional neuroimaging: a comparative review

Stat Methods Med Res 12: 419-446

Oldfield RC, 1971

The assessment and analysis of handedness: the Edinburgh inventory

Neuropsychologia 9: 97-113

Prohovnik I, Skudlarski P, Fulbright RK, Gore JC, Wexler BE, 2004

Functional MRI changes before and after onset of reported emotions

Psychiatry Res 132: 239-250

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, *et al.*, 2002

Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain

NeuroImage 15: 273–289

Wilke M, Lidzba K, Staudt M, Buchenau K, Grodd W, Krägeloh-Mann I, 2006

An fMRI task battery for assessing hemispheric language dominance in children

NeuroImage (in press)

Acknowledgements

We would like to thank Ingeborg Krägeloh-Mann, MD, PhD, Wolfgang Grodd, MD, PhD, and Scott K. Holland, PhD, for their continued support. This work has been supported by the *Deutsche Forschungsgemeinschaft* DFG (SFB550/C4). The algorithm is part of our LI-toolbox and is available free of charge for scientific use. Interested Researchers are encouraged to contact the authors at Marko.Wilke@med.uni-tuebingen.de

Subject	***Classical lateralization indices***		***Bootstrapped lateralization indices (k = 1)***		
	FWE	**FDR**	**Overall mean**	**Trimmed Mean $_{(25)}$**	**Weighted Mean**
1 (m, 12y)	0,65	0,4	0,44	0,41	0,5
2 (m, 15y)	N/A	N/A	0,74	0,76	0,85
3 (m, 9y)	N/A	N/A	0,87	0,86	0,92
4 (f, 8y)	N/A	0,59	0,53	0,46	0,67
5 (f, 13y)	0,97	0,56	0,58	0,55	0,73

Table 1: Demographic details and lateralization indices from the unaltered fMRI example datasets (for details, see Wilke *et al.*, 2006).

Threshold	Classical lateralization index	Bootstrapped lateralization indices			
		Mean [k = 1] (Range)	Mean [k = .5] (Range)	Mean [k = .25] (Range)	Mean [k = .05] (Range)
0	.355	.353 (.299-.409)	.358 (.306-.410)	.354 (.293-.414)	.355 (.299-.412)
.29	.361	.360 (.313-.401)	.361 (.296-.405)	.361 (.310-.401)	.358 (.313-.427)
.58	.376	.375 (.339-.411)	.375 (.336-.419)	.376 (.335-.406)	.377 (.338-.422)
.87	.395	.394 (.361-.428)	.394 (.365-.431)	.394 (.362-.426)	.394 (.349-.434)
1.16	.410	.411 (.385-.433)	.410 (.384-.432)	.410 (.387-.435)	.410 (.372-.444)
1.45	.410	.410 (.388-.431)	.410 (.386-.430)	.410 (.391-.430)	.410 (.372-.438)
1.74	.380	.380 (.358-.398)	.380 (.360-.397)	.381 (.362-.400)	.381 (.34-.414)
2.03	.344	.344 (.329-.361)	.345 (.329-.361)	.346 (.324-.363)	.346 (.315-.373)
2.32	.307	.306 (.291-.322)	.307 (.294-.322)	.307 (.288-.324)	.307 (.279-.34)
2.61	.293	.293 (.279-.306)	.293 (.281-.308)	.293 (.280-.306)	.294 (.256-.322)
2.9	.331	.331 (.321-.343)	.331 (.319-.341)	.332 (.319-.347)	.331 (.296-.358)
3.19	.396	.396 (.387-.404)	.396 (.386-.405)	.396 (.378-.408)	.395 (.365-.421)
3.48	.467	.467 (.460-.475)	.467 (.457-.476)	.467 (.454-.482)	.467 (.439-.493)
3.77	.562	.562 (.555-.568)	.562 (.552-.570)	.562 (.548-.572)	.562 (.53-.589)
4.06	.608	.609 (.603-.615)	.608 (.600-.616)	.609 (.594-.621)	.609 (.585-.637)
4.35	.614	.614 (.607-.620)	.614 (.605-.624)	.614 (.602-.626)	.614 (.582-.641)
4.64	.645	.645 (.637-.653)	.645 (.635-.654)	.645 (.627-.658)	N/A
4.93	.703	.703 (.695-.710)	.703 (.693-.713)	N/A	N/A
5.22	.497	.497 (.485-.507)	N/A	N/A	N/A

Table 2: Comparison of results form a classical lateralization index calculation from different thresholds with bootstrapped results (k = 1/.5/.25/.05). N/A: not available with constant k due to decreasing voxel numbers.

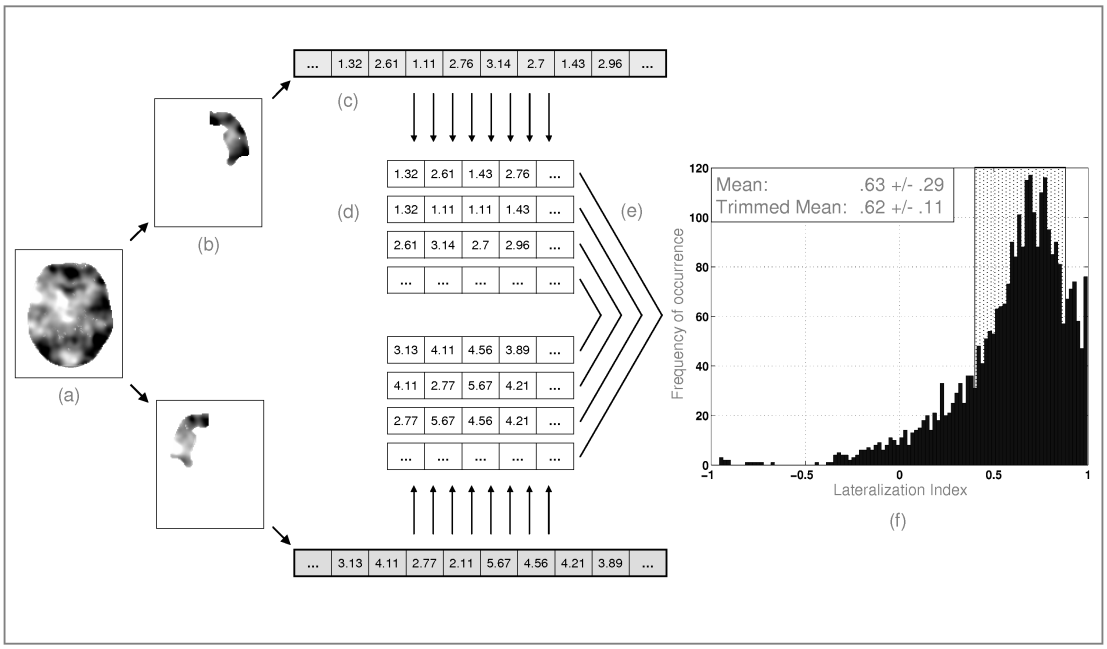

Figure 1: Overview of the bootstrapping procedure: an input image (a) is thresholded and masked, yielding data for the left and the right side (b), which is then converted to a vector containing all voxel values (c). From this, *n* bootstrapped resamples are generated (d; default: 100) from which all possible lateralization index combinations are calculated (e; default: 10.000). All resulting values are then plotted in a histogram (f), from which only the central 50% are used (shaded area), "trimming" the upper and lower 25% of datapoints. This procedure is repeated at each thresholding step.

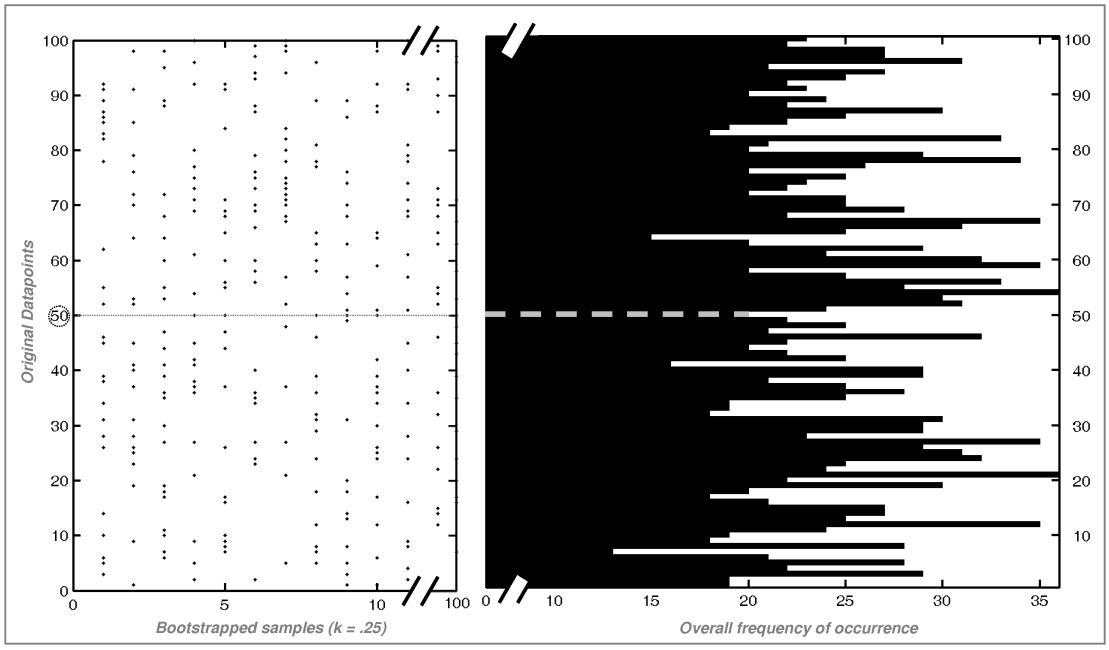

Figure 2: Left panel: illustration of the bootstrap approach: an input sample with 100 datapoints is randomly sampled, with replacement, 100 times with a resample ratio of $k = .25$. Right panel: resulting histogram of the presence of each point in the resulting resamples.

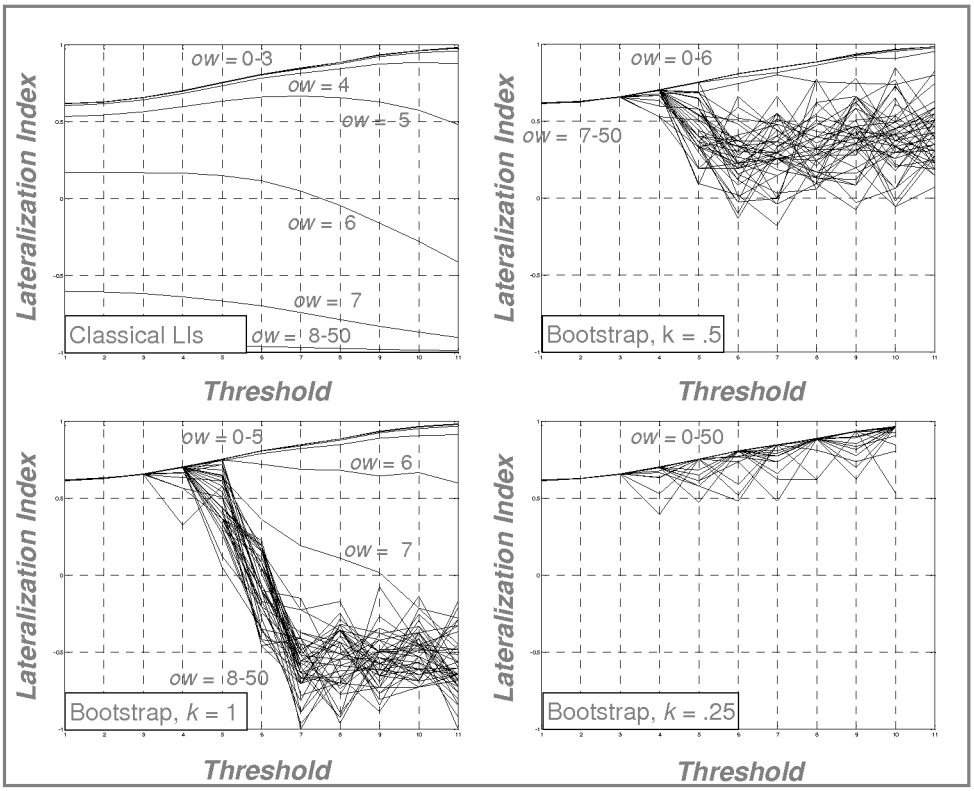

Figure 3: Effect of outliers of different values (outlier weight *ow*) on classical lateralization curves (top left) and on bootstrapped lateralization curves with different resample ratios. Note increasing stability towards outliers with decreasing resample ratio.

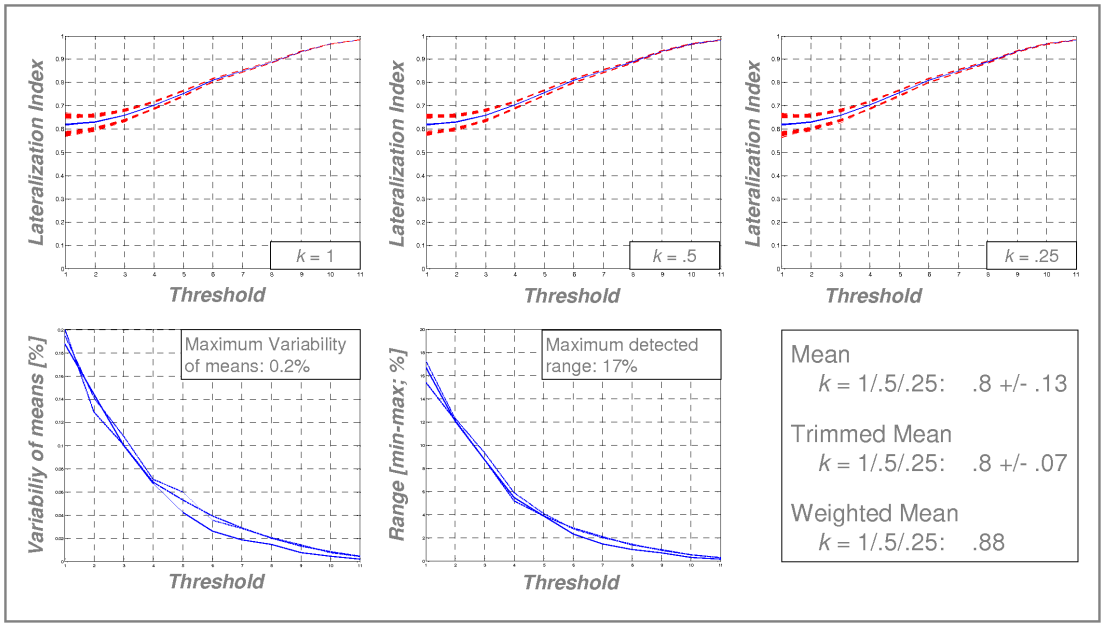

Figure 4: Upper panels: illustration of re-run stability of the bootstrapped lateralization curves: 100 re-runs of the same dataset with different resample ratios *k*. Solid lines: mean lateralization indices; slashed lines: minimum/maximum detected values. Note virtually identical results independent of *k*. Lower panels: Variability and range for different resample ratios (solid line: $k = 1$; slashed line: $k = .5$; dotted line: $k = .25$). Lower right panel: identical overall results independent of *k*.

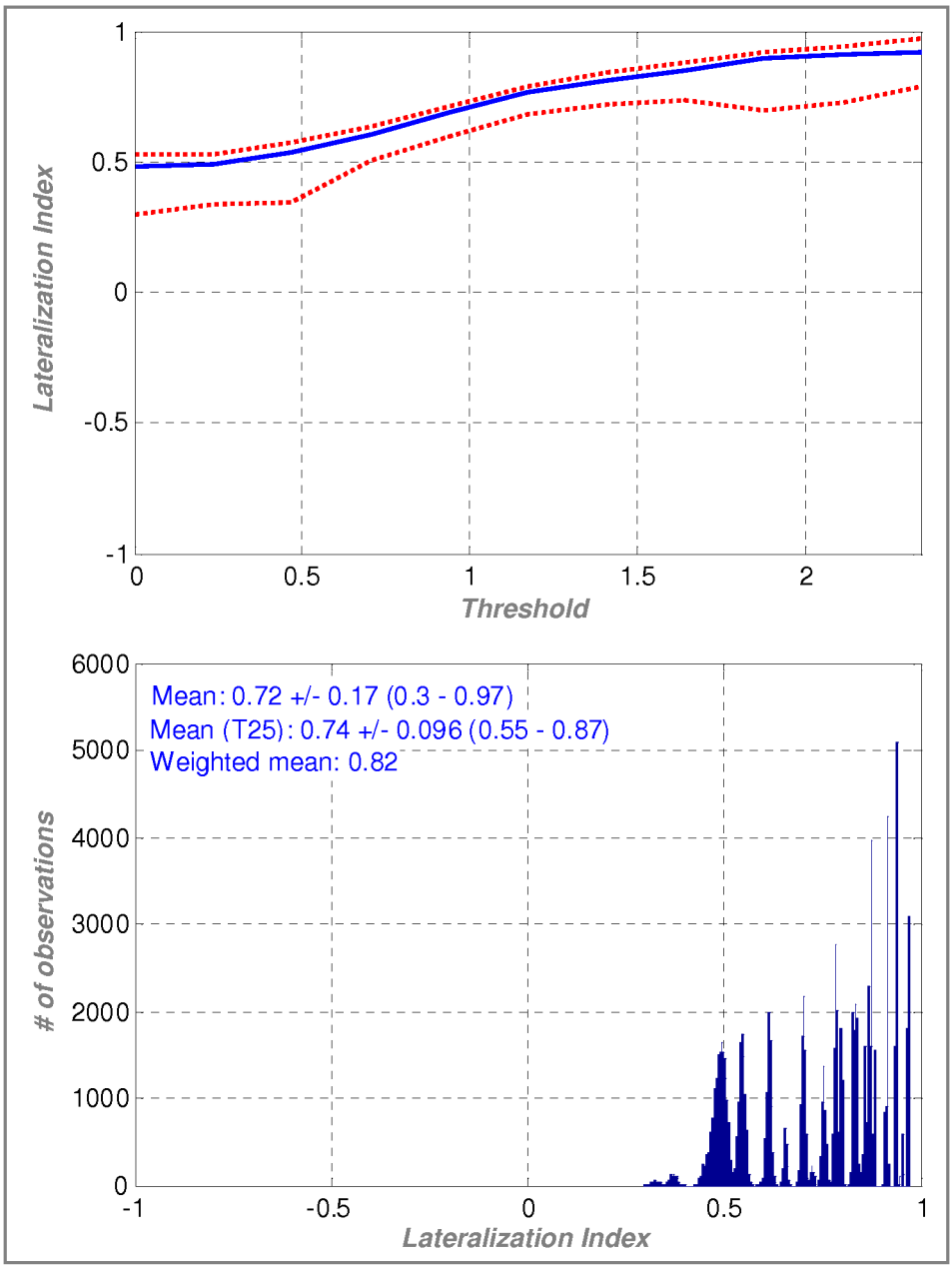

Figure 5: Actual algorithm output for a single dataset (lateralization within the frontal lobe), containing two outliers on the right (p1/2, see also Figure 6). Upper panel: note much wider spread of detected minimum lateralization indices on the right side (lower dotted line) and closer adherence of the trimmed mean (solid line) to the detected maximum values (upper dotted line). Also note absence of routinely included number of voxels at each threshold for better visibility. Lower panel: histogram of the overall lateralization index-matrix: note several small, irregularly shaped histogram contributions from outliers.

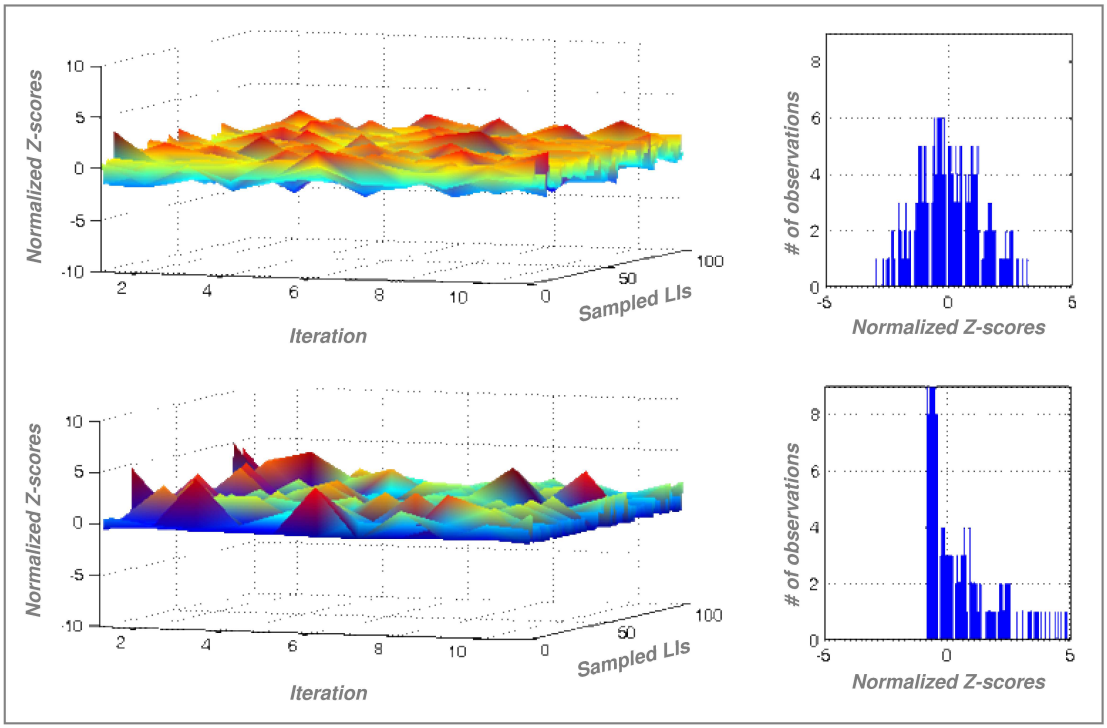

Figure 6: Actual algorithm output for a single dataset (lateralization within the frontal lobe), containing two outliers on the right (p2/2, see also Figure 5). Upper panels: surface representation of all z-scores from all iterations and the resulting histogram for the LEFT side; lower panels: corresponding data from the RIGHT side, clearly showing the outlier influence in both the surface plot and the skewed histogram (see text for details)