Magnetic Resonance Imaging 33 (2015) 1299–1305

Contents lists available at ScienceDirect

Magnetic Resonance Imaging

journal homepage: www.mrijournal.com

Permutation and parametric tests for effect sizes in voxel-based morphometry of gray matter volume in brain structural MRI

David A. Dickie [a,b,c,*], Shadia Mikhael [a,b,c], Dominic E. Job [a,b,c], Joanna M. Wardlaw [a,b,c], David H. Laidlaw [d], Mark E. Bastin [a,b,c]

[a] Neuroimaging Sciences, University of Edinburgh, Edinburgh, United Kingdom
[b] Scottish Imaging Network–A Platform for Scientific Excellence (SINAPSE), 15 Redburn Avenue, Giffnock, United Kingdom
[c] Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom
[d] Computer Science Department, Brown University, Providence, RI, United States

ARTICLE INFO

Article history:
Received 23 April 2015
Revised 31 July 2015
Accepted 31 July 2015

Keywords:
Magnetic resonance imaging
Brain structure
Gray matter
Statistical inference
Aging

ABSTRACT

Permutation testing has been widely implemented in voxel-based morphometry (VBM) tools. However, this type of non-parametric inference has yet to be thoroughly compared with traditional parametric inference in VBM studies of brain structure. Here we compare both types of inference and investigate what influence the number of permutations in permutation testing has on results in an exemplar study of how gray matter proportion changes with age in a group of working age adults. High resolution T_1-weighted volume scans were acquired from 80 healthy adults aged 25–64 years. Using a validated VBM procedure and voxel-based permutation testing for Pearson product-moment coefficient, the effect sizes of changes in gray matter proportion with age were assessed using traditional parametric and permutation testing inference with 100, 500, 1000, 5000, 10000 and 20000 permutations. The statistical significance was set at $P < 0.05$ and false discovery rate (FDR) was used to correct for multiple comparisons. Clusters of voxels with statistically significant ($P_{FDR} < 0.05$) declines in gray matter proportion with age identified with permutation testing inference ($N \approx 6000$) were approximately twice the size of those identified with parametric inference ($N = 3221$ voxels). Permutation testing with 10000 ($N = 6251$ voxels) and 20000 ($N = 6233$ voxels) permutations produced clusters that were generally consistent with each other. However, with 1000 permutations there were approximately 20% more statistically significant voxels ($N = 7117$ voxels) than with ≥10000 permutations. Permutation testing inference may provide a more sensitive method than traditional parametric inference for identifying age-related differences in gray matter proportion. Based on the results reported here, at least 10000 permutations should be used in future univariate VBM studies investigating age related changes in gray matter to avoid potential false findings. Additional studies using permutation testing in large imaging databanks are required to address the impact of model complexity, multivariate analysis, number of observations, sampling bias and data quality on the accuracy with which subtle differences in brain structure associated with normal aging can be identified.

1. Introduction

Brain MRI data are often analyzed using parametric statistical methods, for example the general linear model (GLM) [1–3]. These methods make a number of assumptions about the generation and statistical distributions of these imaging data. Specifically, subject samples are assumed to have been acquired randomly from their population and distributions of data are assumed to be approximately statistically normal, or "Gaussian" [4–6]. Previous seminal work in voxel-based morphometry (VBM) has used voxel-wise smoothing, i.e. averaging, to circumvent the issue of statistical normality [1,2]. Permutation testing was proposed at a similar time [7], and has recently been widely implemented in VBM methods, for example FMRIB Software Library (FSL; http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Randomise), to address the assumptions of random samples and homoscedasticity [8,9]. Current implementations of permutation testing in VBM are optimized for *t*-tests and analysis of variance (ANOVA). These provide robust tools for assessing differences in, for example, the proportion of gray matter voxels between two or more groups. Reductions in gray matter volume are a commonly observed feature of normal aging [10], and are also seen in diseases such as amyotrophic lateral sclerosis [11], epilepsy [12], Alzheimer's disease [13] and schizophrenia [14]. However,

* Corresponding author at: Neuroimaging Sciences, The University of Edinburgh, Royal Infirmary of Edinburgh, Little France, Edinburgh, United Kingdom.
E-mail address: ddickie1@staffmail.ed.ac.uk (D.A. Dickie).

http://dx.doi.org/10.1016/j.mri.2015.07.014
0730-725X/

differences in tissue structure can be subtle and difficult to identify consistently between studies [14,15].

Effect size statistics, for example Cohen's *d* for two groups or Pearson product-moment coefficient (*r*) for continuous data such as age [16,17], may be a useful addition to imaging statistics derived from existing implementations of permutation testing. Measures of effect size provide standardized results that can be more easily compared across different studies and populations [16,17]. However, the influence of parametric versus permutation inference for effect sizes and the impact of the number of permutations on results have not yet been formally tested in VBM studies.

In the present study we therefore describe a framework for permutation testing of effect size in VBM studies of brain structural MRI data. We then compare parametric and permutation testing inference and assess the impact of the number of permutations on the latter in an exemplar study of changes in brain gray matter proportion with age in structural MRI data acquired from a cohort of healthy subjects with ages spanning normal working age adulthood.

2. Materials and methods

2.1. Subjects

Eighty clinically normal, right-handed, healthy volunteers (40 males, 40 females) aged 25–64 (median 43, IQR 17) years were recruited by advertisement from staff working at the University of Edinburgh, the Western General Hospital and the Royal Infirmary, Edinburgh, United Kingdom. All subjects gave written informed consent. Health status was assessed using medical questionnaires and all structural MRI scans were reported by a fully qualified neuroradiologist. To aid identification of age-related differences in brain volumes, the cohort was divided into four 10-year age bands as detailed in Table 1.

2.2. MRI acquisition

All brain MRI data were acquired using a GE Signa Horizon HDxt 1.5 T clinical scanner (General Electric, Milwaukee, WI, USA) equipped with a self-shielding gradient set (33 mT m^{-1} maximum gradient strength) and manufacturer supplied 8-channel phased-array head coil. The imaging protocol consisted of whole brain axial T_2-, T_2*- and FLAIR-weighted structural sequences, and a high resolution 3D T_1-weighted inversion-recovery-prepared fast spoiled gradient-echo (FSPGR) volume scan acquired in the coronal plane with 180 contiguous 1.3 mm thick slices resulting in voxel dimensions of 1 × 1 × 1.3 mm.

2.3. Voxel-based morphometry

The T_1-weighted volume scans were first converted from DICOM to NIfTI–1 format (http://nifti.nimh.nih.gov/nifti-1) using MRIcron's "dcm2nii" tool (http://www.nitrc.org/projects/mricron). A modified FSL-VBM pipeline was then employed to process these imaging data and produce gray matter proportion volumes for each subject. The first step in this pipeline consisted of randomly selecting a subject for manual, slice-by-slice, brain extraction. This subject was then non-linearly registered to all other subjects to produce initial brain masks for the whole cohort [18]. These initial brain masks were manually edited slice-by-slice and applied to the raw imaging data to produce brain extracted T_1-weighted volumes for each subject. These brain extracted T_1-weighted volumes were then processed using the standard FSL-VBM pipeline [19]. Briefly, each subject's T_1-weighted scan was segmented into gray matter, white matter and cerebrospinal fluid volumes using signal intensity and spatial information [20]. These gray matter volumes contained the proportion of gray matter tissue within each voxel in native space. No subject had white matter hyperintensities on FLAIR-weighted MRI (hypointense on T_1-weighted MRI) which might confound the gray matter segmentations. After segmenting these three tissue types, all data were aligned to Montréal Neurological Institute (MNI) standard space. A study specific atlas was created by registering all subjects to the initial average of all subjects aligned in MNI space. The gray matter proportion volumes were then smoothed using a 3 mm Gaussian kernel in standard space. There are currently no standard optimal parameters for Gaussian kernels [21], and our reasoning for choosing 3 mm smoothing was that, based on visual assessment of the imaging data, it provided a reasonable middle ground between removing noise and maintaining the underlying anatomy. Finally, a 4D volume of voxel-wise gray matter proportions was created by concatenating all individual gray matter volumes together in the axial direction in standard space; effect sizes and *P*-values were then calculated using this cohort 4D volume.

Table 1
Demographics of the cohort.

Age group (years)	Number
25–34	21
35–44	23
45–54	24
55–64	12
25–64	**80**

2.4. Permutation testing for effect sizes

We provide the Pearson product-moment coefficient (*r*) as a measure of effect size. This was proposed as a measure of effect size by Cohen [16] and is valid for continuous variable data. Absolute effect sizes of approximately ±0.1 are considered small, approximately ±0.3 medium and approximately ±0.5 large [16]. Effect size *r* was calculated using Eq. (1)

$$\frac{\sum_{i=1}^{n} \frac{x_i-\bar{x}}{\sigma_x} \times \frac{y_i-\bar{y}}{\sigma_y}}{n-1} \quad (1)$$

where *n* is the number of pair-wise observations, $\bar{x}$ is the mean of variable *x*, $\bar{y}$ is the mean of variable *y*, σ_x is the standard deviation (SD) of variable *x* and σ_y is the SD of variable *y*. In the present study, *x* is age and *y* is gray matter proportion in each voxel.

Permutation testing is a very simple concept. For *i* permutations (for example 1000), the order of independent variables is randomly shuffled and the test statistic of interest (in this case, effect size) is calculated in each random permutation (see Fig. 1). This is supposedly equal to producing 1000 pseudo random samples and the *P*-value of the effect size is defined as the number of times this effect size could be produced by chance, i.e. in each random permutation of the data (see Fig. 1).

We report both parametrically defined *P*-values and non-parametric permutation testing *P*-values for effect sizes with the latter assessed using 100, 500, 1000, 5000, 10000 and 20000 permutations. For 20000 permutations the smallest achievable *P*-value is 0.00005, a value twenty times smaller than that used in previous "extensive simulations" [9]. False discovery rate (FDR) was used to correct for multiple comparisons [22–24], and we provide 1-P_{FDR} corrected and 1-*P* uncorrected volumes as outputs. Alpha (*P*-value cut off) and lambda (FDR corrected *P*-value cut off) were set at 0.05.