A comparison of permutation and parametric testing for between group effective connectivity differences using DCM

Nia Goulden a,*, Shane McKie a, John Suckling b, Stephen Ross Williams c, Ian Muir Anderson a, John Francis William Deakin a, Rebecca Elliott a

a Neuroscience and Psychiatry Unit, University of Manchester, Stopford Building, Oxford Road, Manchester, M13 9PT, UK
b Brain Mapping Unit, Department of Psychiatry, University of Cambridge, UK
c Imaging Science and Biomedical Engineering, University of Manchester, UK

A R T I C L E   I N F O

Article history:
Received 30 August 2009
Revised 2 November 2009
Accepted 23 November 2009
Available online 11 December 2009

A B S T R A C T

Effective connectivity is becoming an increasingly popular technique for obtaining additional information from functional magnetic resonance imaging (fMRI) cognitive activation studies. It is potentially important for investigating psychiatric illnesses, which are thought to depend on disrupted connections and in observing the action of psychoactive drugs used to treat these disorders. If researchers are to apply these techniques confidently then it is important to establish the level of power that is available in an experiment. This study compares the level of power available when applying effective connectivity to test for differences between groups using parametric tests and permutation testing. Permutation testing has previously been shown to have superior sensitivity to parametric tests in fMRI studies. As an illustrative example, both the parametric t-test and equivalent permutation test were applied to a comparison between healthy controls and remitted depressed volunteers performing an emotional face processing task. Permutation testing was found to provide superior power compared with the nonparametric equivalent.

Introduction

Statistical power is an important element in study design and interpretation. Power is defined as the probability of rejecting the null hypothesis when it is false, i.e., the probability that the experiment will yield statistically significant results (Hallahan and Rosenthal, 1996). These calculations are generally carried out in two ways: either before data collection to ensure that an adequate number of participants are included to detect a given effect size or after data collection to assess the level of power (and thus predict the smallest detectable effect size) in the conducted experiment (Cohen, 1976; Bausell and Li, 2002). Power is dependent on: the type I error rate (i.e., the probability of mistakenly rejecting the null hypothesis or significance level); the effect size (ES), a standardised measure of the magnitude of the effect of interest; and the sample size. Power calculations to predict sample size thus assume an ES and significance level; the sample size required to achieve an acceptable level of power. Alternatively, following data acquisition, calculations of ES can be made from the known sample size and acceptable type I and type II error rates (where power is 1 – type II error rate).

Permutation testing is a nonparametric alternative to parametric tests. The procedure involves the appropriate reassignment of data labelling (depending on the study design) and recalculation of the test statistic. From multiple repeats, the distribution of the test statistic under simulated conditions of the null hypothesis is accrued. For a given p-value of interest, a critical value is then found above which a% of the differences lie. The critical value is then applied to the test statistic obtained with the correct labelling (i.e., the observed data). In the normal way, if the test statistic from the observed data exceeds the threshold, then it is denoted as significant (Good, 2005).

Permutation testing assumes exchangeability of data. This means that the overall distribution of the data is not affected by reassignment (Lindley and Novick, 1981) and thus requires randomisation of volunteers to groups or conditions so that each reassignment is equally likely to have occurred. The advantages of permutation testing arise mainly from the avoidance of strong assumptions on the form of the null distribution and may therefore be advantageous in circumstances in which distributional models of parametric tests are not satisfied. Permutation testing has been criticised in that the results of an experiment do not allow inference to the broader target population (Berger, 2000). However, parametric inference methods, such as random effects analyses, can only generalise to the population subtending the experimental sample and are thus constrained by the demographic, clinical, or behavioural characteristics of the sample. In other words, irrespective of the interference methods results cannot apply more to the general population than to the sample under immediate investigation.

Permutation testing provides superior sensitivity to detect signals in functional magnetic resonance imaging (fMRI) data when compared
to parametric alternatives (Hayasaka and Nichols, 2004; Suckling and Bullmore, 2004). Thirion et al. (2007) suggest at least 20 volunteers per group, preferably 27, are required for sufficient reliability of patterns of stimulus induced brain activation, and recommend nonparametric testing for fMRI studies. In the small samples typically seen in fMRI activation studies, deviations from the normal assumption have given rise to interest in permutation inference, with many techniques applied to fMRI activation paradigms reported in a literature of increasing volume (Brammer et al., 1997; Nichols and Holmes, 2002; Suckling and Bullmore, 2004; Suckling et al., 2006). However, to our knowledge, permutation testing has not yet been applied to effective connectivity analysis.

Defined as “the influence of one brain region on another” (Buchel and Friston, 1997; Friston et al., 2003), effective connectivity makes it possible to test networks of brain regions exploring the mechanisms contributing to changes in cognitive activations. One method to assess effective connectivity is dynamic causal modelling (DCM) (Friston et al., 2003), which specifies intrinsic connections, task-related inputs, and modulations due to these inputs. It has been argued previously that it may be difficult to make broad interpretations from effective connectivity analyses (Hornitz, 2003) due to individual variability and different researchers specifying different models and testing different hypotheses.

In the present study, the power that may be obtained with both parametric t-tests and the equivalent nonparametric permutation tests are presented to determine which is most appropriate for group comparisons of effective connectivity. To do this, both tests were applied to a group comparison of healthy control and remitted depressed volunteers performing an implicit face-emotion task. Specifically, we considered the differences in modulation associated with viewing happy faces. Remitted depressed volunteers were likely to show only subtle changes in face processing when compared to controls and therefore using this group as a “test case” allowed us to determine the sensitivity of the technique.

Effective connectivity analyses have been carried out for face emotion tasks previously (de Marco et al., 2006; Fairhall and Ishai, 2007). The model of Fairhall and Ishai was based on work by Haxby et al. (2000, 2002). Brain areas found to be important for carrying out the emotional face processing tasks include the primary visual cortex (V1), fusiform gyrus, the amygdala, and the orbitofrontal cortex. These were used as regions of interest in our DCM model.

Tractography and diffusion tensor imaging (DTI) studies have shown white matter tracts between V1 and temporal cortex (Rockland and Van Hoesen, 1994; Webster et al., 1994; Catani et al., 2003), temporal cortex and frontal cortex (Webster et al., 1994), temporal cortex and amygdala (Iwai and Yukie, 1987; Stefanacci and Amaral, 2000; Ghashghaei and Barbas, 2002), and amygdala and orbitofrontal cortex (Cavada et al., 2000; Idaka et al., 2001; Ghashghaei and Barbas, 2002). It is therefore reasonable to specify an effective connectivity model focused on these brain regions.

In the analysis presented here, the group means and standard deviations for all DCM connectivity measures were calculated for both models. A Jarque–Bera test of normality in both groups was also applied.

**Methods**

**Volunteers**

All volunteers were right-handed with normal or corrected-to-normal vision and no contraindications to functional magnetic resonance imaging (fMRI) scanning. Two groups of volunteers were recruited. The first group comprised 29 healthy control volunteers, 8 males and 21 females, mean age 31.1 years (SD 9.97 years). These volunteers had no personal or family history of psychiatric or neurological illness, substance abuse or significant medical illness. The second group comprised 30 remitted depressed volunteers, 8 males and 22 females, mean age 33.73 years (SD 10.69 years). These volunteers met criteria for major depression in full remission by SCID interview with a mean number of depressive episodes of 3.13 (SD 2.6) and were required to have been remitted for a minimum of 3 months, with scores of <12 on the Montgomery Asberg Depression Rating Scale (MADRS). Volunteers with past history of neurological disorder, substance abuse or axis 1 psychiatric disorder other than major depression or anxiety disorders were excluded. All participants gave written informed consent. The study was approved by the Central Manchester Research Ethics Committee.

**Emotional face processing task**

A series of faces with emotional expressions were presented for 3 seconds with a 1-second gap. There were five conditions, namely neutral (N), happy (H), sad (S), fear (F) and rest (R), during which a fixation cross was displayed.

During the task, 6 faces per block were displayed, 3 males and 3 females, using 80% intensity from a series of morphed faces from neutral to 100% intensity of emotion (Ekman and Friesen, 1976). Each block was 21 seconds and faces were presented in an NHNSNFRSNHNSFNSNHNR design. The duration of the task was therefore 7 minutes 42 seconds. Each time a face appeared on the screen the volunteers responded with a button press to indicate whether they thought the face was male or female; they were not told of any change in emotion of the faces and were not asked to focus on the emotion. Stimulus delivery was done using E-prime software (http://www.pstnet.com/products/e-prime/).

**Imaging**

Images were acquired on a 1.5 T Phillips Intera scanner (Eindhoven, Netherlands) using a single-shot echo-planar (EPI) pulse sequence. Each volume comprised 29 contiguous axial slices (TR/TE = 2100/40 ms, 3.5 mm by 3.5 mm in-plane resolution, slice thickness 4.5 mm with a 0.5 mm gap). During the task, 218 volumes were acquired. A T1-weighted turbo fast echo structural scan (256 × 256 matrix, TR = 8.99 ms, TE = 4.2 ms, 160 axial slices, voxel size 0.875 × 0.875 × 1 mm) was acquired for each participant to be used in the normalization stage of the pre-processing.

**Image analysis**

All preprocessing was carried out in SPM5 (http://www.fil.ion.ucl.ac.uk/spm). Images were slice timing-corrected and then realigned to the first volume to correct for participant movement in the scanner. The anatomical image was co-registered to the mean functional image. The anatomical image was also segmented into grey matter, white matter, and CSF. The grey matter segmented image was normalized to the standard grey matter template supplied with SPM. The resulting normalisation matrix was applied to the functional images which were smoothed with a 3D Gaussian kernel with FWHM 7 mm × 7 mm × 10 mm. A high-pass filter equal to twice the stimulus repetition time was applied.
A second-level, random effects, two-sample t-test analysis was carried out to identify those voxels significantly activated by the task in response to a contrast comparing all faces to the rest (i.e., cross-hair fixation) across all participants in both groups. Further analysis identified those voxels differentially activated by the two groups.

Dynamic causal modelling

The following right-hemispheric brain areas were selected as nodes in the DCM model describing the network involved in facial processing (Fig. 1): primary visual cortex (V1), fusiform gyrus, amygdala, and orbitofrontal cortex. These regions of interest were delineated in each data set by a sphere of fixed size centred on the peak coordinate nearest the peak in the activation map averaged across the group.

Initially, the group maxima were found in each region of interest from the average group activation map. The regions of interest were selected based on previous literature showing that these brain regions are important for performing emotional face processing tasks. Local maxima within 14 mm of the group activation were found for each participant, and time series extracted from a 6 mm radius sphere region of interest centred on that location. By using this method of data extraction, these coordinates were independent of the group inference method.

The DCM models were specified using the toolbox in SPM5 using the slice timing correction tool. The model tested was for the right side only and is illustrated in Fig. 1. This DCM had a driving input of the contrast of all faces vs rest, all possible connections present between the regions of interest and all connections, except those from and to the V1 area, modulated by viewing happy faces (we chose only one emotion to test the methods). This model is therefore testing the effect of happy faces on these connections by assessing the changes in effective connectivity when happy faces are presented. Mean and standard deviations, as well as a Jarque–Bera test for normality, were performed on the predictions parameters for each connection in both groups separately.

Power analysis

The effect size (ES) is the difference between group means divided by the pooled between-subject standard deviation. It may be specified based on information from previous studies, if available or simply hypothesised. A convention commonly followed is that a small effect size corresponds to an ES of 0.2, a medium effect corresponds to an ES of 0.5 and a large effect size corresponds to an ES of 0.8 (Cohen, 1976). Using our data set, the power (expressed as a percentage) was evaluated for each connection and modulation using these theoretical medium and large effect sizes at different sample sizes for both parametric and nonparametric methods.

Power calculations were performed according to the following procedure: first, the critical value of the t-statistic at the assumed type I error rate needed to declare statistical significance, $t_{\text{hyp}}$, was calculated from the parametrically or nonparametrically derived null distribution.

The critical value of the $t$-statistic, $t_{\text{hyp}}$, that would result if the null hypothesis were false with an observed effect size, $ES_{\text{hyp}}$, was then calculated (Cohen, 1976; Bausell and Li, 2002), where $N$ is the sample size per group:

$$t_{\text{hyp}} = \frac{ES_{\text{hyp}}}{\sqrt{\frac{N}{2}}} \tag{1}$$

Power was then calculated according to:

$$\text{Power} = \int_{-\infty}^{t_{\text{hyp}} - t_{\text{cv}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx \tag{2}$$

That is, the area under the curve of the unit normal cumulative distribution function was integrated from $-\infty$ to $(t_{\text{hyp}} - t_{\text{cv}})$. These calculations were performed in Matlab.

Derivation of the null distributions

The parametric null distribution was the normal distribution with zero mean and unit variance. Thus, for example, at type I error rate of $p = 0.05$, $t_{\text{hyp}} = 1.96$.

The nonparametric (permutation) null distribution was derived by randomly reassigning the patient and control data labels to the data sets and recalculating the difference between the mean connection strengths between the two groups. This was repeated 5000 times to sample the distribution under the simulated null hypothesis of no group difference. The distribution was normalised by subtracting the mean and dividing by the variance to allow comparison with the alternative distribution in the power calculation (Eq. (2)). The value of $t_{\text{hyp}}$ was obtained empirically from the null distribution at a given type I error rate; i.e., $p = 0.05$.

Comparison of parametric and nonparametric techniques

Power calculations for null distributions derived from parametrically and nonparametrically were made for $N$ from 8 to 21, and for medium and large ES. For each sample size, the permutation null distribution was obtained by randomly selecting $N$ volunteers from each group before the repeated reassignment procedure.

Power calculations were also performed for measures of differences between the groups (controls and remitted depressed) using.

Table 1

<table>
<thead>
<tr>
<th>Region</th>
<th>BA</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left V1</td>
<td>17</td>
<td>−14</td>
<td>−98</td>
<td>0</td>
<td>12.54</td>
</tr>
<tr>
<td>Right V1</td>
<td>17</td>
<td>18</td>
<td>−95</td>
<td>0</td>
<td>11.01</td>
</tr>
<tr>
<td>Left fusiform</td>
<td>17</td>
<td>−35</td>
<td>−77</td>
<td>−20</td>
<td>8.02</td>
</tr>
<tr>
<td>Right fusiform</td>
<td>42</td>
<td>−60</td>
<td>−15</td>
<td>8.21</td>
<td></td>
</tr>
<tr>
<td>Left amygdala</td>
<td>25</td>
<td>−7</td>
<td>−15</td>
<td>4.33</td>
<td></td>
</tr>
<tr>
<td>Right amygdala</td>
<td>25</td>
<td>−4</td>
<td>−15</td>
<td>4.78</td>
<td></td>
</tr>
<tr>
<td>Left OFC</td>
<td>47</td>
<td>−42</td>
<td>21</td>
<td>−15</td>
<td>2.22</td>
</tr>
<tr>
<td>Right OFC</td>
<td>47</td>
<td>28</td>
<td>32</td>
<td>−15</td>
<td>2.10</td>
</tr>
</tbody>
</table>
the correct sample labellings with statistical thresholds obtained via either parametric or nonparametric null distributions. In doing so, the relative sensitivity between the two approaches was assessed.

For the parametric approach, the 95% confidence interval around the mean difference in sample means, \( \Delta \), was calculated using the well-known formulae:

\[
\Delta = \pm 1.96 \times \frac{SD}{\sqrt{N}}
\]

where the standard deviation, SD, was calculated from the sample. For the nonparametric test, confidence intervals were calculated via bootstrapping, where surrogate groups of the same size as the original sample were generated from the population, sampling with replacement, and the sample means recalculated from the DCM. This was repeated 1000 times and the symmetrical confidence interval was defined by the limits within which 95% of the difference in means occurred.

**Results**

**fMRI activations**

Maxima were found in the regions of interest in the faces-rest contrast at the group level. The regions of interest were the primary visual cortex (V1), the fusiform gyrus, the amygdala, and the orbitofrontal cortex (OFC). The maximum activations in these regions and the t-value of the activation are reported in Table 1. Right amygdala activity was found in 21 healthy control volunteers and 22 remitted depressed volunteers for the DCM analysis.

**Dynamic causal modelling**

The mean, standard deviation, and Jarque-Bera test p-value for each group and each effective connectivity measure are presented in Table 2. Some of the connectivity measures are not normally distributed in both groups (connection: V1 → fusiform gyrus), and many of the parameters for the remitted depressed volunteers were not normally distributed. The parametric test is likely to be conservative in the cases where the data are not normally distributed.

**Power calculations**

The predicted power for each connection vs. sample size for medium (0.5) and large (0.8) effect sizes are detailed in Tables 3–6. The power predictions where the permutation test outperforms the equivalent parametric test are shown in bold. As can be seen from both tables, the permutation test is associated with improved power in each connection and the improvement is more pronounced for

### Table 2

Mean, standard deviation, and Jarque-Bera test for normality for both groups. Jarque-Bera test p < 0.05 shown in bold.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Healthy controls</th>
<th>Remitted depressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connections</td>
<td>Mean</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>V1 → F</td>
<td>0.4120</td>
<td>0.3432</td>
</tr>
<tr>
<td>F → V1</td>
<td>-0.0335</td>
<td>0.1002</td>
</tr>
<tr>
<td>F → A</td>
<td>0.2438</td>
<td>0.1606</td>
</tr>
<tr>
<td>F → OFC</td>
<td>0.2069</td>
<td>0.3773</td>
</tr>
<tr>
<td>A → F</td>
<td>0.1386</td>
<td>0.1158</td>
</tr>
<tr>
<td>A → OFC</td>
<td>0.0935</td>
<td>0.1841</td>
</tr>
<tr>
<td>OFC → F</td>
<td>0.1112</td>
<td>0.1988</td>
</tr>
<tr>
<td>OFC → A</td>
<td>0.0866</td>
<td>0.1540</td>
</tr>
</tbody>
</table>

### Table 3

Permutation compared to parametric testing for medium effect sizes. Figures are percent power to detect large effect size difference in connection strength between the remitted depressed and control group. The permutation test outperforms the parametric test for every connection and sample size. N is sample size.

<table>
<thead>
<tr>
<th>N</th>
<th>Par test</th>
<th>Permutation testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1 → F</td>
<td>F → V1</td>
<td>F → A</td>
</tr>
<tr>
<td>8</td>
<td>12.61%</td>
<td>37.96%</td>
</tr>
<tr>
<td>9</td>
<td>14.47%</td>
<td>37.96%</td>
</tr>
<tr>
<td>10</td>
<td>16.28%</td>
<td>37.96%</td>
</tr>
<tr>
<td>11</td>
<td>18.05%</td>
<td>37.96%</td>
</tr>
<tr>
<td>12</td>
<td>19.79%</td>
<td>37.96%</td>
</tr>
<tr>
<td>13</td>
<td>21.50%</td>
<td>37.96%</td>
</tr>
<tr>
<td>14</td>
<td>23.19%</td>
<td>37.96%</td>
</tr>
<tr>
<td>15</td>
<td>24.85%</td>
<td>37.96%</td>
</tr>
<tr>
<td>16</td>
<td>26.50%</td>
<td>37.96%</td>
</tr>
<tr>
<td>17</td>
<td>28.12%</td>
<td>37.96%</td>
</tr>
<tr>
<td>18</td>
<td>29.73%</td>
<td>37.96%</td>
</tr>
<tr>
<td>19</td>
<td>31.31%</td>
<td>37.96%</td>
</tr>
<tr>
<td>20</td>
<td>32.88%</td>
<td>37.96%</td>
</tr>
<tr>
<td>21</td>
<td>34.42%</td>
<td>37.96%</td>
</tr>
</tbody>
</table>
small sample sizes. Graphs of sample size vs. power for large effect sizes for various connections are displayed in Fig. 2. These results illustrate that permutation testing is associated with superior power compared with the parametric test, with the largest benefit for small sample sizes and medium effect sizes. For smaller sample sizes, the increase in power with permutation testing is largest. With 8 volunteers, there was up to a 10% increase in power. For larger sample sizes, the increase was not as pronounced, with a 2–3% increase in power for a sample size of 21.

The results of the power calculations using the observed sample sizes (i.e., 21 vs 22) and the observed effect sizes for each connection/modulation are presented in Table 7. As can be seen from the table, all of the effect sizes were less than those used in the calculations above with only the intrinsic connection from fusiform→amygdala approaching the medium effect size of 0.5. However, the use of permutation testing improved the power in all connections and modulations when compared to the parametric equivalent except for the modulation in the OFC→amygdala connection. This was independent of effect size with the most marked improvement in the amygdala→fusiform connection. The lower bounds of the confidence intervals are higher for the nonparametric test compared with the parametric test and the confidence intervals are narrower. For connections where the data in both groups were normally distributed, e.g., fusiform→OFC, the increase in power was less than when the data are not normally distributed V1→fusiform. The power for detecting large effects with 21 volunteers is 71.94% and 73.43%, respectively (Table 5), compared with 71.61% for the parametric test.

### Discussion

Power calculations for the parametric independent-samples t-test and the equivalent permutation test have been performed. By carrying out calculations based on theoretical effect sizes rather than those from our data, the results are more generalisable to other effective connectivity studies. The results show that permutation testing provides superior power compared with the parametric equivalent with the largest benefit for smaller sample sizes (14 per group) and medium effect sizes. Increases in power of up to 10% were observed for detecting large effects with small sample sizes (Tables 5 and 6). In addition, the confidence intervals for permutation testing are typically narrower than for the parametric test and the lower bound of the confidence interval is higher. This suggests that the power predictions are more precise for the nonparametric test and thus that the level of power is generally greater for the nonparametric test compared to the parametric test under the normal assumption. Permutation testing improving the power with small sample sizes may be attributable to the distribution of the data being non-normal. Normal distributions are assumed for parametric, but not nonparametric testing. These results support the view of Ludbrook and Dudley (1998) that permutation tests should be preferred to parametric t or F tests. These data suggest that particularly for patient groups, permutation testing would be useful as the data may not be normally distributed.

The actual effect sizes in our data set were much less than those used in the initial power calculations. As hypothesized, the differences in connectivity between the two groups are small and subtle and these data therefore provide a “test” data set on which to determine the sensitivity of the effective connectivity measures. However, even with these smaller effect sizes, the results indicated clearly that permutation testing improved power in all connections. Our results therefore suggest that if data are obtained which are not normally distributed, then permutation testing provides superior results to an equivalent parametric test for DCM.

Care is needed with regard to the number of permutations applied in these analyses as an insufficient number may give a critical value that is not adequately robust. If R permutations are carried out, then the significance level needs to be a multiple of 1/R to obtain an exact test. If a large number of permutations are possible then a suitable subset is usually applied. This test is called approximate since all of the possible permutations have not been carried out, and therefore, all of...
the information regarding the distribution of the data is not available. It has been shown that an efficient approximate test may be carried out with as few as 1000 permutations (Edgington, 1969). For this analysis, 5000 permutations were applied to obtain good approximate p-values.

It is important to note that in this data set the standard deviation of the data for the remitted depressed group was larger than the healthy control group for the majority of parameters. The power may be overestimated if the standard deviation of the healthy control group is used. For this work, DCM has been applied in an exploratory manner (Bitan et al., 2006; Bitan et al., 2007; Cao et al., 2008); the power may be different for models with more specific hypotheses, due to differences in the effect size which is observable within the data. Another important caveat of our results is that permutation testing has been applied to a data set without random assignment of volunteers to either groups or conditions. Despite this, there was still an increase in power when permutation testing was compared with the parametric test. The benefits may be even larger in data sets where volunteers were randomly assigned to groups or conditions, e.g., randomly assigning healthy volunteers to receive stimulus A or stimulus B. Applying nonparametric tests to the fMRI data also could increase the power for detecting activations and therefore potentially increase the usefulness of effective connectivity. For simplicity and consistency with previous studies, the parametric test was applied to the fMRI data in this study. The within-subject standard errors are typically not stated for fMRI analysis due to bias arising from the correlation of the time series. Permutation testing can overcome this by preserving the correlational properties of the time series.

Our results demonstrate that it is possible to confidently test for differences in effective connectivity between groups. Interpreting the results in terms of pathophysiology of depression is beyond the scope of this paper. A full DCM analysis in this group, modelling left and right hemisphere connections for all emotions, will be presented in a separate paper. We used these patients to illustrate the greater power of permutation testing. Since remitted depressed volunteers have recovered from depression, it seems plausible that any residual differences from healthy controls will be relatively subtle and therefore detection of a difference provides important validation for the permutation testing approach.

**Conclusion**

This work has demonstrated that permutation testing improves statistical power when testing for group differences in effective...
connectivity measures and should be considered as an alternative to the equivalent parametric test. Studies with small samples would particularly benefit from a permutation test approach to inference. The better performance of permutation testing is attributable to significant deviations of connectivity parameters from the normal distribution. It is important to note that for parametric or nonparametric tests to be useful and generalisable careful sampling of the populations under investigation is needed to provide a representative sample. Power calculations for permutation testing in factorial experimental designs and for longitudinal comparisons, for example, of drug treatments, would be useful.

Acknowledgments

This work contributes to the NEWMOOD EU Integrated Programme LSHM-CT-2004-503474. We would also like to thank the staff at the Wellcome Trust Clinical Research Facility, Manchester, UK, for assistance with scanning. Thank you to Emma Thomas, Darragh Downey, and Gabriella Juhasz who carried out the scanning.

References