Title

Pooling morphometric estimates: a statistical equivalence approach

Authors

Heath R. Pardoe[1], Gary R. Cutter[2], Rachel Alter[1], Rebecca Kucharsky Hiess[1], Mira Semmelroch[3], Donna Parker[3], Shawna Farquharson[3], Graeme D. Jackson[3], and Ruben Kuzniecky[1]

[1]Comprehensive Epilepsy Center, Department of Neurology, New York University School of Medicine, New York, USA

[2]School of Public Health, University of Alabama at Birmingham, Birmingham, Alabama, USA

[3]The Florey Institute of Neuroscience and Mental Health, Melbourne, Australia

Corresponding Author

Heath R. Pardoe

Comprehensive Epilepsy Center, NYU School of Medicine

223 East 34th St,

New York, NY, USA, 10016

Phone: +1 347 285 1126

Email: heath.pardoe@nyumc.org

Abstract

Changes in hardware or image processing settings are a common issue for large multi-center studies. In order to pool MRI data acquired under these changed conditions, it is necessary to demonstrate that the changes do not affect MRI-based measurements. In these circumstances classical inference testing is inappropriate because it is designed to detect differences, not prove similarity. We used a method known as statistical equivalence testing to address this limitation.

Equivalence testing was carried out on three datasets: (i) cortical thickness and automated hippocampal volume estimates obtained from healthy individuals imaged using different multi-channel head coils; (ii) manual hippocampal volumetry obtained using two readers; and (iii) corpus callosum area estimates obtained using an automated method with manual cleanup carried out by two readers. Equivalence testing was carried out using the "two one-sided tests" (TOST) approach. Power analyses of the two one-sided tests were used to estimate sample sizes required for well-powered equivalence testing analyses. Mean and standard deviation estimates from the automated hippocampal volume dataset were used to carry out an example power analysis.

Cortical thickness values were found to be equivalent over 61% of the cortex when different head coils were used ($q < 0.05$, FDR correction). Automated hippocampal volume estimates obtained using the same two coils were statistically equivalent (TOST $p = 4.28 \times 10^{-15}$). Manual hippocampal volume estimates obtained using two readers were not statistically equivalent (TOST $p = 0.97$). The use of different readers to carry out limited correction of automated corpus callosum segmentations yielded equivalent area estimates (TOST $p = 1.28 \times 10^{-14}$). Power analysis of simulated and automated hippocampal volume data demonstrated that the equivalence margin affects the number of subjects required for well-powered equivalence tests.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.808v2 | CC-BY 4.0 Open Access | rec: 29 Jul 2015, publ: 29 Jul 2015

We have presented a statistical method for determining if morphometric measures obtained under variable conditions can be pooled. The equivalence testing technique is applicable for analyses in which experimental conditions vary over the course of the study.

Keywords

MRI, statistics, morphometry, volumetrics

1. Introduction

Measuring morphometric properties of brain structures from an MRI scan can be used to assess neuroanatomical changes in disease and normal development. These measurements may be automated or require manual input from a human assessor. Common issues with MRI scanning, particularly in large multicenter studies, are changes in the collection or processing conditions over the duration of the study. These changes may include the use of multiple scanners, scanner hardware or software upgrades, or the use of different human readers for manual morphometric measurements such as hippocampal volumes. A number of studies have investigated whether these variable experimental conditions introduce differences, systematic or otherwise, in quantitative measurements [1-6]. The introduction of systematic differences is undesirable because it may reduce the ability to detect differences between groups, or increase the probability of making a false positive or false negative finding.

In this study we show how to apply a statistical method known as equivalence testing to MRI datasets that were acquired or analyzed under changed experimental conditions that may be commonly encountered in neuroimaging studies. Equivalence testing is an inference-based method for determining if measures obtained under variable conditions can be considered ‘equivalent’. The equivalence testing approach addresses a common misinterpretation of classical inference testing that a $p > 0.05$ provides statistical support for the absence of differences and thus equivalence. This interpretation is incorrect. With equivalence testing, the null hypothesis is formulated as there being differences between the groups or measures being compared, and evidence must be used to disprove this hypothesis. If there is enough evidence against the null hypothesis of a difference, we can reasonably conclude that the two groups are equivalent, and we can carry out prospective analyses with a high degree of confidence that the comparisons using pooled data are legitimate. Equivalence testing approaches are becoming increasingly used for a number of biomedical applications, a number of which are described in Walker and Nowacki [7].

An important concept associated with equivalence testing analysis is the need for an 'equivalence margin'. The equivalence margin is a predefined interval that allows for differences to exist between sets of measurements obtained under different conditions that are of little practical consequence for the desired application. The equivalence margin typically is centered around zero, and is defined *a priori*. While few formal methods exist for determining the equivalence margin, typical values may be a fraction of a clinically important difference (such as ¼ or ⅓ of measures considered to be different), or a similar amount of meaningful biological change measured over some time interval. In the case of morphometric analyses a clinically important difference may not be well defined for specific neurological disorders, in which case previously reported effect sizes may serve as a proxy for a clinically important difference.

In our example herein, the first dataset consists of healthy individuals that were scanned in one session using an identical T1-weighted whole brain MPRAGE acquisition with a 20 channel and 32 channel receiver coil. We used equivalence testing to assess if automated measures of cortical thickness and hippocampal volume estimates from the two coils can be considered equivalent. If morphometric measures estimated from the paired MRI scans obtained using the two coils are shown to be equivalent, future studies can combine MRI scans from each coil with reasonable certainty that no systematic bias has been introduced.

The second dataset addresses inter-rater variability in manual hippocampal volumetry. Differences between readers are typically assessed using descriptive statistics such as percentage volume difference or intra-class correlation, or spatial overlap measures such as the Dice coefficient or Jaccard index. Intuitively people understand that a lower volume difference or higher overlap is better, however these methods are not inference based and so there is no accepted standard for these measures and thus, no formal way to say the measurements are comparable. Statistical equivalence testing is an inference based method and so applies a standard that is accepted by the scientific community. In the case of equivalence testing, this standard is often a false positive rate (p-value) p less than 0.05, where a false positive finding implies that samples are equivalent when in fact they are not.

We will carry out an equivalence analysis on a dataset of manual hippocampal volumes measured using two different readers. It is well known that different readers often obtain hippocampal volume estimates that are systematically different [8]; therefore we may reasonably expect these manual estimates to fail our test for equivalence for this type of difference between readers. Excess variability by either reader can also lead to failure to show equivalence. Finally we will investigate an automated method for estimating corpus callosum area [9] that occasionally requires manual correction for small segmentation errors. Previous experience with the software indicates these errors occasionally occur in images obtained from some MRI scanners.

In order to demonstrate equivalence between two sets of measurements, a study must be adequately powered. The main way a researcher can control the power of an equivalence study is by including an appropriate number of participants. An additional consideration for power analyses of equivalence tests (relative to traditional tests of difference) are that an equivalence test allows for differences between subject groups, as long as the confidence intervals of the mean difference are within the pre-specified equivalence margin. An adequately powered study with a non-zero difference between the sets of measures that is close to the equivalence margin will require more participants than two sets of measures with a mean difference closer to zero. We investigated how an equivalence margin influences sample size estimates for well-powered equivalence analyses using (i) simulated values and (ii) the automated hippocampal volume data described above as an example case.

The following specific hypotheses were tested in this study:

1. Automated vertex-wise cortical thickness measurements and hippocampal volumes measured using MRI data acquired with a 20-channel and 32-channel head coils are NOT statistically equivalent.
2. Manual hippocampal volumes of healthy controls, segmented using two different readers, are NOT statistically equivalent.
3. Corpus callosum area measured using a semi-automated method by two different readers, are NOT statistically equivalent.

Each of these three null hypotheses are to be rejected if equivalence is established. Code for carrying out the equivalence analyses presented in this paper is provided at https://sites.google.com/site/hpardoe/equivalence.

2. Methods

A common method for statistical equivalence testing is the two one-sided tests approach (TOST, described in [10]). The technique requires that an equivalence margin be defined *a priori*. For the analyses presented in this study we used an equivalence margin of 5% of the average value for each morphometric measure. The 5% value was chosen to set a consistent margin across different morphometric estimates and provide a demonstration of the method, and does not necessarily reflect a particular clinically useful important difference. Clinically important differences for morphometric estimates will depend on the neurological disorder or other potential applications. The equivalence margin θ, which is in the same units as the measure of interest (or defined as a suitable proportion of the mean values of the measure of interest), defines the limits of acceptability for defining 'equivalence'.

Hypotheses to be tested using the TOST approach can also be stated as follows. These statements follow those described in [10], modified for the morphometric properties investigated in our study: m_1 represents the morphometric parameter of interest (eg. cortical thickness, hippocampal volume, corpus callosum cross-sectional area) measured under the first experimental conditions, and m_2 represents the same morphometric parameter measured under the second experimental conditions. All measurements presented in this study are obtained in the same subjects under variable conditions (ie. paired measurements); thus, for our analyses we investigated the equivalence of the mean difference between measures $\mu_{(m1-m2)}$ within individuals. The null and alternative hypotheses are formulated statistically as:

$$H_0: \mu_{(m1-m2)} \leq -\theta \text{ or } \mu_{(m1-m2)} \geq \theta \text{ (inferior)}$$

$$H_1: -\theta < \mu_{(m1-m2)} < \theta \text{ (equivalent)}$$

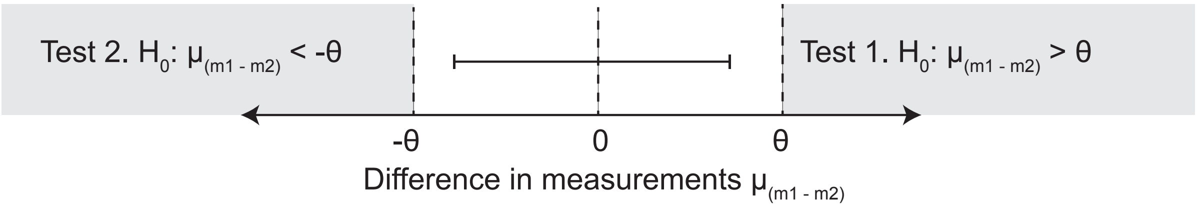
The null hypothesis H_0 states that the mean difference in paired morphometric estimates obtained under different conditions are outside our predefined equivalence margins and therefore are not equivalent. The alternative hypothesis H_1 states that the mean difference in paired measures obtained under different conditions lie within the margin, and so for practical purposes are equivalent. If we carry out an equivalence test and reject the null hypothesis, we reject both hypotheses $\mu_{(m1-m2)} \leq -\theta$ and $\mu_{(m1-m2)} \geq \theta$. We test each of these null hypotheses at the $\alpha = 5\%$ level and must reject both, thus, we obtain an experimentwise p-value less than each one sided test's alpha (typically equal to 0.05). If both tests are rejected, we conclude that measures obtained under two conditions are equivalent and may be pooled with a reasonable degree of certainty that we have not introduced systematic differences that impact on our ability to identify an effect.

In practice, the two one-sided tests consists of two sequential one-sided tests, with the two null hypotheses being (i) the mean difference in paired measurements is greater than θ and (ii) the mean difference in paired measurements is less than –θ.

$$\text{Test 1. } H_0: \mu_{(m1-m2)} \geq \theta \quad H_1: \mu_{(m1-m2)} < \theta$$

$$\text{Test 2. } H_0: \mu_{(m1-m2)} \leq -\theta \quad H_1: \mu_{(m1-m2)} > -\theta$$

The null hypothesis is only rejected if both null hypotheses are rejected for the separate tests. In our study, following the implementation in [11], the reported p-value for the two one-sided test procedure is the maximum p for the individual tests. As noted in [10], the two one-sided test procedure is the same as determining if the (1 - 2α) confidence interval is within the ±θ equivalence margin. Note that α = 0.05 means that the (1 - 2α) interval is a 90% confidence interval.



Test 2. H_0: $\mu_{(m1 - m2)}$ < -θ
Test 1. H_0: $\mu_{(m1 - m2)}$ > θ
-θ
0
θ
Difference in measurements $\mu_{(m1 - m2)}$

Figure 1. Equivalence testing depends on the prior definition of an equivalence margin [-θ, θ]. If the null hypotheses for Test 1 and Test 2 are rejected, we can conclude that measurements obtained under variable experimental conditions are equivalent. Applying these two one-sided tests is the same as determining if the (1 - 2α) confidence interval, shown as the capped line, is within the equivalence margin.

Graphically it can be seen that we are defining 2 overlapping intervals, such that the overlap defines the equivalence region, but we have shown than it is unlikely that the difference is greater than θ and that the difference is < - θ (Figure 1). If the difference is shown to fall within the equivalence margin, we interpret the differences as not being clinically meaningful.

The TOST procedure provides a p-value that indicates the probability that our finding is a false positive (i.e. the measurements are not equivalent, but we say they are). In the case of equivalence testing, a p-value less than α (typically 0.05) provides supporting evidence that morphometric estimates obtained under variable conditions are equivalent. As with the more widely used tests of difference (Student's T-test and related methods), increasing the number of inferences increases the likelihood that a false positive finding will occur. In the case of equivalence testing, a false positive finding occurs when measures are declared equivalent based on the testing procedure when in reality they are not equivalent. Some form of multiple comparisons correction should be used for mass univariate analyses, such as the vertex-wise comparisons carried out when assessing cortical thickness equivalence over the cortex. In this study we used false discovery rate method to adjust p-values to account for the number of tests [12]. The false discovery rate procedure controls the proportion of expected false positive findings from a set of p-values.

2.1 MRI acquisition and image processing

2.1.1 Cortical thickness estimates from MRI data acquired using different coils.

The first dataset consists of 16 healthy controls (8 female, age 32.4 ± 6.1 years) who were scanned twice on the same MRI scanner in the same imaging session, using (a) a 20-channel receive head coil and (b) a 32-channel receive head coil. Images were acquired on a 3T Siemens Skyra MRI scanner, using a T1-weighted whole-brain 3D MPRAGE acquisition, sagittal slice prescription, 0.9 mm isotropic voxel size, TR

= 1900 ms, TE = 2.49 ms, TI = 900 ms, FA = 9°. Pixel bandwidth was 180 Hz/Px for the 20 channel coil and 230 Hz/Px for the 32 channel coil.

MRI scans were processed using Freesurfer version 5.1.0. Vertex-wise cortical thickness and hippocampal volume estimates were derived using the standard cross-sectional processing stream [13]. Cortical maps were coregistered to the common space “fsaverage” template to allow comparison across subjects. 10 mm full-width half-maximum (FWHM) smoothing was applied to the cortical thickness maps. Thickness values were then read into R using the package “cortex” [14]. The “fsaverage” template consists of 163 842 vertices for each hemisphere (left and right), resulting in 327 684 vertices in total; this corresponds to the number of tests carried out for the vertex-wise analyses and thus requires some form of correction for multiple testing.

Paired TOST inference tests of cortical thickness data were carried out vertex-wise with alpha = 0.05, using an equivalence margin of 5% of the vertex-wise mean cortical thickness. P-values for each vertex were mapped and thresholded at $p < 0.05$ to allow visualization of regions which may reasonably be concluded to be equivalent. False discovery rate thresholding was used to correct for multiple comparisons ($q < 0.05$). In addition to the TOST testing procedure, vertex-wise paired t-tests were carried out using the t.test function provided as part of the R stats package [15]. P-values were recorded and mapped in a similar manner to the TOST procedure for comparison.

The results of the cortical thickness analyses were summarized by recording the number of suprathreshold vertices as a percentage of the total number of vertices for both the equivalence testing procedure and vertex-wise paired t-tests. Non-cortex vertices along the medial hemispheric surfaces were excluded using the fsaverage “lh.cortex.label” and “rh.cortex.label” files provided with the Freesurfer distribution. Uncorrected ($p < 0.05$) and FDR corrected ($q < 0.05$) measures of coverage were measured.

2.1.2 Automated hippocampal volumes estimated from MRI data acquired using different coils.

A similar analysis was carried out with hippocampal volume estimates from the automated subcortical segmentations provided with Freesurfer. The longitudinal processing stream was used to estimate hippocampal volumes for images acquired using each coil, and paired TOST tests were carried out with alpha = 0.05, and an equivalence margin = 5% of the average hippocampal volume. Left and right hippocampal volumes were tested independently. Paired T-tests were performed with alpha = 0.05.

2.1.3 Manual hippocampal segmentations measured using different readers.

The second dataset consisted of hippocampal volumes measured from manual segmentations carried out by two readers. The dataset consists of 40 healthy controls (20 female, age 30.5 ± 8.8 years). Images were acquired on a 3T Siemens TIM Trio MRI scanner, using a T1-weighted whole-brain 3D MPRAGE acquisition, 0.9 mm isotropic voxel size, TR = 1900 ms, TE = 2.6 ms, TI = 900 ms, flip angle = 9°.

Paired TOST analyses were carried out with alpha = 0.05 and an equivalence margin of 5% of the overall average hippocampal volume (across left and right hippocampi and both readers). Left and right hippocampal volumes were measured independently. Paired T-tests were performed with alpha = 0.05.

2.1.4 Semi-automated corpus callosum segmentations measured using different readers.

The third dataset consists of corpus callosum area measurements measured using an automated software package “yuki” developed by Ardekani et al [9]. There are occasionally minor segmentation errors that require manual editing to obtain accurate estimates of corpus callosum area. In this study we compare corpus callosum area measurements assessed using two readers for a single site from the Autism Brain Imaging Data Exchange (ABIDE) study (http://fcon_1000.projects.nitrc.org/indi/abide/, site University of Michigan Sample 1 [16, 17]), in order to determine if we could pool corpus callosum area measurements obtained using different readers to assess the entire ABIDE dataset (consisting of 1000+

MRI scans). Whole brain T1-weighted SPGR MRI was obtained on a 3 T GE Signa scanner, 1.2 mm slice thickness, 1 mm^2 in-plane resolution, TE = 1.8 ms, Prep Time = 500 ms, flip angle = 15 degrees.

Paired TOST analyses were carried out with alpha = 0.05 and an equivalence margin of 5% of the average corpus callosum area. Paired T-tests were performed with alpha = 0.05.

2.1.5 Sample size estimates for well-powered equivalence analyses

Sample size estimates were calculated using the “power.t.test” function provided as part of the standard R distribution (R version 3.1.2, [15]). Two separate analyses were conducted, described below.

The first analysis used simulated values to demonstrate how distance from the equivalence margins and variance in the difference between paired measurements affected sample size estimates. An equivalence interval [-θ, θ] = [-0.1, 0.1] was used. The simulated mean difference between paired measurements $\mu_{(m1 - m2)}$ was systematically varied from -0.09 to 0.09. The standard deviation of the measurements was set at 0.05, 0.1 and 0.15. For each simulated mean difference in the range -0.09 to 0.09, sample size was calculated using power = 0.8, alpha = 0.05, using a one-sample, one-sided test. The effect size was the distance between the simulated mean difference and each equivalence margin. Each one-sided test yielded a sample size estimate (for the parameters described above); the maximum for each pair of one-sided tests was chosen as the target sample size.

The second analysis used the automated hippocampal volume estimates from the two head coils (section 2.1.2) to demonstrate how sample sizes could be calculated in practice. The equivalence margin [-θ, θ] was systematically varied with θ = 40 to 200; these values were chosen retrospectively based on the outcomes of the previously described equivalence analysis of automated hippocampal volume estimates derived from different head coils (see section 2.1.2). The mean difference and standard deviation of the difference between paired measures was calculated. The following parameters were used for the sample size calculation: power = 0.8, alpha = 0.05, and one sample, one-sided T-tests.

Results

Cortical thickness values measured using two different coils were determined to be equivalent over 60.4% of the left and 62.4% of the right hemisphere cortical surfaces when tested vertex-wise using the TOST approach (Figure 2, $q < 0.05$ FDR correction for multiple comparisons). The 5% equivalence margin corresponds to a median cortical thickness difference of 0.13 mm when measured across all vertices. If FDR correction is not applied, 68.9% of the left hemisphere and 69.7% of the right hemisphere were found to be equivalent (TOST alpha < 0.05). No significant differences were observed in mean vertex-wise cortical thickness values if tested using a paired t-test approach when FDR correction was applied to account for multiple comparisons; without FDR correction, 10.4% and 11.7% of vertices were identified to have different means (paired T-test alpha < 0.05).

Automated hippocampal volume estimates derived from MRI scans acquired on different coils were determined to be equivalent using the TOST approach (equivalence margin = 230 mm^3, TOST $p = 4.28 \times 10^{-15}$, 90% confidence interval = [-35.5,18.9]). Paired T-tests of differences between the volumes did not show significant differences (-8.31 mm^3 volume difference, $t(31) = -0.51$, $p = 0.61$). In contrast to this finding, when assessing hippocampal volumes measured using manual segmentation with two different readers the null hypothesis was not rejected, indicating the two readers were not equivalent (equivalence margin = 132 mm^3, TOST $p = 0.97$, 90% confidence interval = [142.7,249.2]). Systematic differences between the two readers were identified (214.9 mm^3 volume difference, paired T-test $t(39) = 5.36$, $p = 3.94 \times 10^{-6}$).

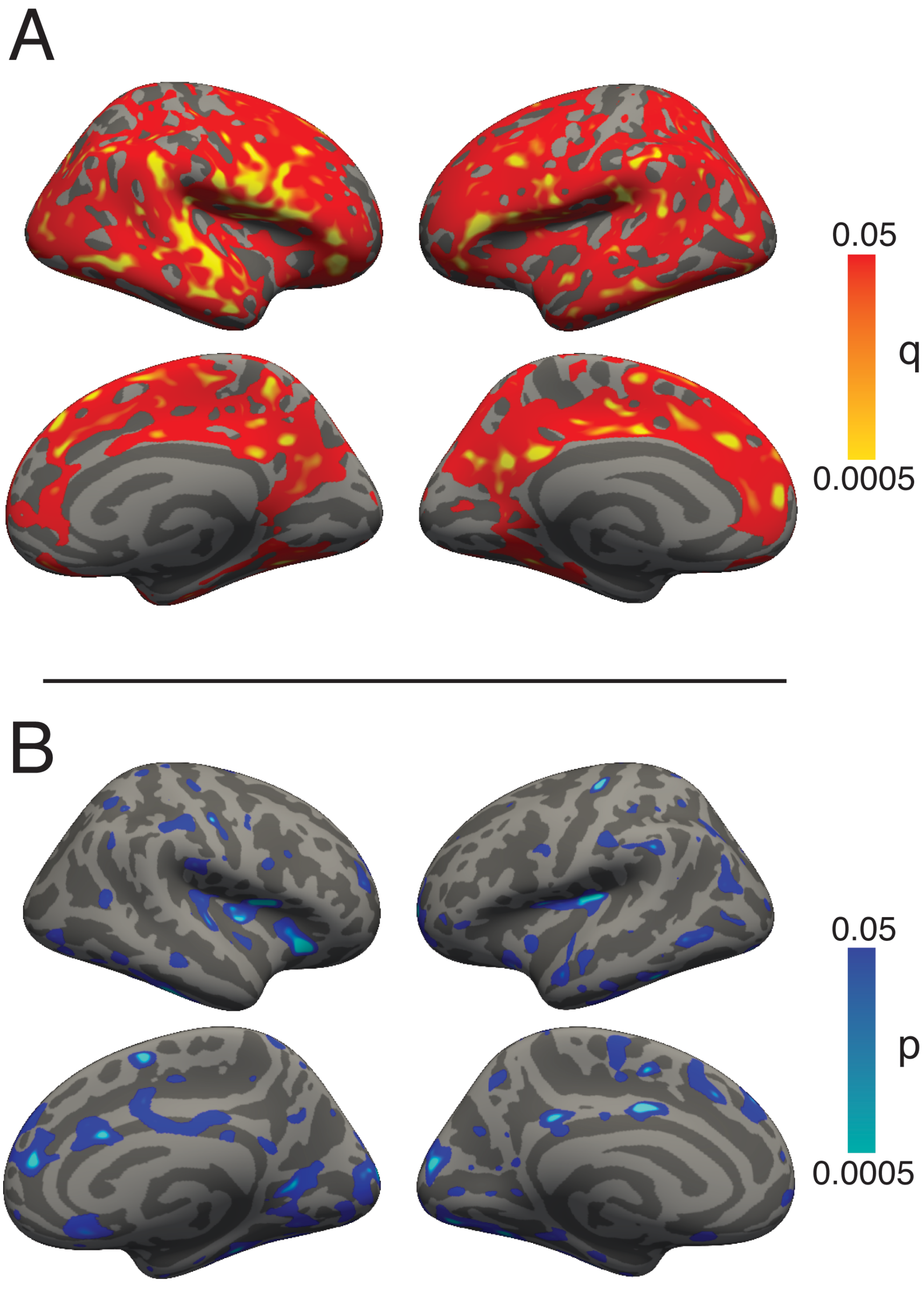


Figure 2. A. Cortical thickness measurements obtained using MRI scans from a 20 channel and 32 channel coils are equivalent over 61% of the cortex (60.4% left hemisphere, 62.4% right hemisphere, TOST q < 0.05 (FDR), median equivalence margin 0.13 mm). B. Differences between two coils assessed using paired Student's T-test. Note there were no suprathreshold thickness differences when using FDR correction; the figure shows uncorrected p-values (alpha = 0.05).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.808v2 | CC-BY 4.0 Open Access | rec: 29 Jul 2015, publ: 29 Jul 2015

The ABIDE study University of Michigan sample 1 consisted of 110 structural MRI scans (55 autism cases and 55 controls). Eleven corpus callosum segmentations required manual edits. Paired TOST analysis of the dataset indicated that corpus callosum areas obtained using both readers were equivalent (equivalence margin = 28.3 mm^2, TOST p = 1.28 × 10^{-15}, 90% confidence interval = [-6.3,4]). No significant differences were observed between estimates from both readers (area difference = -1.2 mm^2, paired T-test t(117) = -0.37, p = 0.71).

Sample size calculations were obtained using simulated data and a retrospective analysis of automated hippocampal volume estimates obtained from different head coils. The simulation-based approach demonstrated that a larger number of participants are required to conduct a well-powered equivalence analysis if the mean difference between the paired measurements is close to the equivalence margin (Figure 3). Increased variance in the difference between the two measures required a larger number of subjects for a well-powered equivalence analysis, as is the case for a traditional sample size calculation using tests for differences between groups.

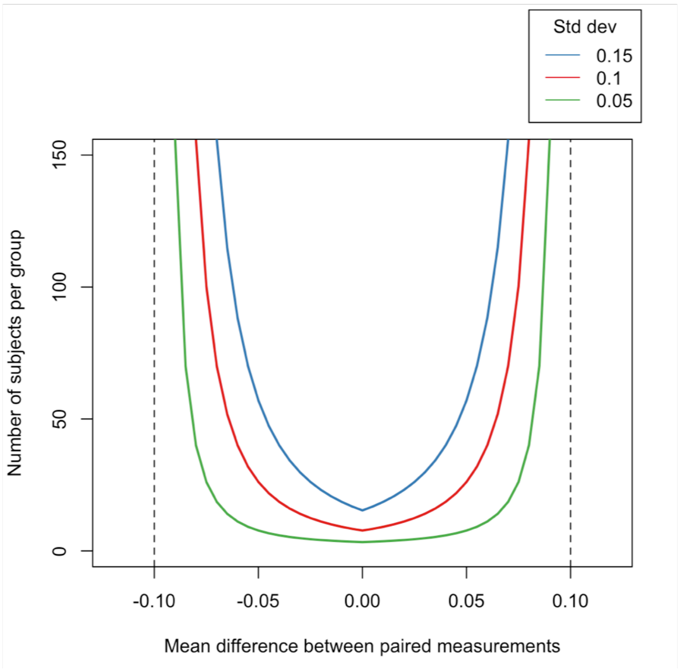


Figure 3. The distance between the equivalence margin, indicated by the dashed lines, and the mean difference in paired measurements is an additional consideration when estimating required sample sizes for determining if morphometric measurements obtained under variable conditions are equivalent. Here the equivalence margin is [-0.1,0.1]. The sample size for a well-powered equivalence analysis (power = 0.8) also depends on the standard deviation of the paired measurements, indicated by the blue (standard deviation = 0.15), red (standard deviation = 0.1) and green (standard deviation = 0.05) lines.

The automated hippocampal volume estimates obtained using different head coils were used to demonstrate how sample size varies with the equivalence margin. The mean difference in paired hippocampal volume measurements was 8.4 mm^3 and the standard deviation of the difference in paired measurements was 90.9 mm^3. If a smaller equivalence margin is used, then more subjects are required for a well-powered analysis (Figure 4). Note that our prespecified equivalence margin of 5% of the mean automated hippocampal volume = 230 mm^3, which would require only a handful of subjects according to Figure 5. This indicates that the use of 40 healthy controls for an equivalence margin of 230 mm^3 provided ample power for demonstrating equivalence. The authors wish to caution that an equivalence margin should always be defined before carrying out a prospective analysis and requires familiarity with the desired application to decide on an appropriate margin.

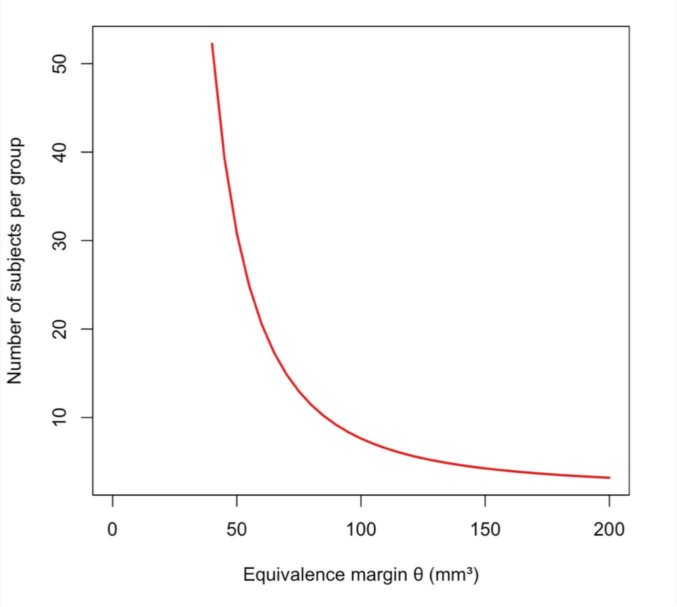


Figure 4. Sample size estimates for well-powered equivalence analyses using mean difference and standard deviation of paired measurements obtained from automated hippocampal volumes measured using whole brain MRI scans obtained on 20-channel and 32-channel receive coils. For this dataset the mean difference in paired measurements = 8.3 mm^3 and the standard deviation = 90.9 mm^3. Approximately 30 subjects per group would be required to demonstrate equivalence using an equivalence margin = [-50,50]; less than 10 subjects would be required is using a more liberal equivalence margin = [-100,100].

Discussion

In this study we have applied a statistical inference method to morphometric estimates obtained under variable conditions to determine if these measures can be considered equivalent within a prespecified interval, known as an equivalence margin. We recommend the use of equivalence testing for determining if MRI-based morphometric estimates obtained under variable conditions can be pooled. If measures are demonstrated to be equivalent under these variable conditions, they may be pooled with a reasonable level of certainty that no systematic bias has been introduced. The specific outcomes of our experiments indicated that cortical thickness measurements obtained from MRI scans acquired using different coils are statistically equivalent to within 5% of mean vertex-wise cortical thickness over approximately 61% of the cortex. Similarly hippocampal volumes obtained using the Freesurfer automated subcortical segmentation algorithms are equivalent to within an equivalence margin of 230 mm^3 (5% of the average overall hippocampal volume) when obtained from a 20 channel and 32 channel head coils. Based on these analyses, we would infer that studies may reasonably pool MRI data acquired from these two coils for morphometric analyses of hippocampal volume, as long as researchers are hoping to detect an effect size significantly larger than 230 mm^3. If researchers are hoping to detect effects of the order of 230 mm^3 or less, scans should be limited to a single coil or approximately equal numbers of controls and subjects of interest should be imaged using both coils, with the knowledge that the coil may be a confounding factor or potentially modify the estimated effect size in their study.

With regard to vertex- or voxel-wise measures, our results indicate that measures may be equivalent only across regions of the cortex; in the case of the cortical thickness estimates obtained from different head coils, approximately 61% of vertices were equivalent (see Figure 2 and Table 1). The finding indicates an important distinction between (i) the use of an equivalence test (TOST $p < 0.05$) and (ii) the erroneous interpretation of a T-test or similar with p > 0.05 as evidence in support of equivalence. Our vertex-wise analysis indicated that there were no brain regions with significant differences in cortical thickness (when using FDR correction for multiplicity). Using the incorrect interpretation a researcher could then erroneously infer that vertices were equivalent over the whole cortex. Our equivalence analysis indicates that we can only demonstrate equivalence over 61% of the cortex using the dataset of 16 healthy controls.

Method	Variable Condition	TOST p-value	T-test p-value	Mean difference in paired estimates
1 a. Cortical thickness	Coils	0.024*	0.33*	0.12 mm*
1 b. Hippocampal volume (automated)	Coils	4.28×10^{-15}	0.61	8.31 mm^3
2. Hippocampal volume (manual)	Readers	0.97	3.42×10^{-8}	214.9 mm^3
3. Corpus callosum area	Readers	1.28×10^{-14}	0.71	1.2 mm^2

Table 1. Equivalence testing p-values and comparative paired T-test values for morphometric parameters derived under variable experimental conditions. In all analyses the equivalence margin was set to 5% of mean estimate of interest. *Median value of p-values/thickness difference measured over the cortical surface. See Figure 1 for images showing distribution of TOST p values over the cortical surface.

An important limitation of the vertex-wise equivalence analyses in this study are that they are only valid for univariate or mass univariate analyses. Although this covers a large number of neuroimaging analyses, there are a growing number of studies that employ multivariate methods for detecting differences between groups, including machine learning approaches for subject classification. These allow for the detection of patterns of morphometric differences across brain regions that may not reach the threshold for statistical significance, when considered at a univariate level (vertex- or voxel-wise). Multivariate approaches may provide an increase in sensitivity for the detection of differences between

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.808v2 | CC-BY 4.0 Open Access | rec: 29 Jul 2015, publ: 29 Jul 2015

groups; however this increased sensitivity may also mean that multivariate techniques are more sensitive to changes in experimental conditions. The equivalence testing approach described in this study does not provide complete protection against potential systematic errors introduced by using data acquired under variable conditions if multivariate statistical methods are used.

Applying equivalence analysis to manual hippocampal volume estimates indicates that measures obtained using two different readers are not equivalent. Therefore these measurements should not be pooled without including readers as a factor, as well as requiring each reader to segment a balanced number of cases and controls. The presence of systematic reader-specific differences in manual hippocampal volumes is well known [8]. However, it is important to note that equivalence analyses do not investigate the sensitivity of a particular method to detecting effects of interest. For example, although hippocampal volume measurements obtained manually are not equivalent between readers, they may still be a preferable approach compared with automated measurements because they are more sensitive to detecting disease related hippocampal volume changes. We recently demonstrated the improved sensitivity and specificity of manual hippocampal volumetry over automated methods in individuals with temporal lobe epilepsy [18].

Finally we demonstrated that corpus callosum area estimates, obtained with occasional manual input to correct minor segmentation errors, may be considered statistically equivalent with a 5% equivalence margin. This allows us to be confident that the use of different readers for the error correction process does not introduce systematic differences in the measured corpus callosum area.

The size of imaging studies is increasing, as is the use of multi-site designs. Typically differences between sites are measured empirically [2-4]. The method presented in this paper is a useful inference-based technique for determining if subtle changes in experimental conditions in a study can influence quantitative measurements derived from MRI, and will allow for improved design of large neuroimaging studies. The statistical equivalence testing technique can be adapted for use in other MRI modalities such as diffusion imaging and for metrics derived from functional MRI analyses. There may be additional useful advantages for the approach described in this paper; for example these methods might be adapted to form more homogeneous groupings of individuals based on having “equivalent” brain parameters.

Acknowledgements

Supported by Amazon Web Services Education in Research grants.

References

1. Dewey, J., et al., *Reliability and validity of MRI-based automated volumetry software relative to auto-assisted manual measurement of subcortical structures in HIV-infected patients from a multisite study.* Neuroimage, 2010. **51**(4): p. 1334-44.
2. Han, X., et al., *Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer.* Neuroimage, 2006. **32**(1): p. 180-94.
3. Jovicich, J., et al., *MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths.* Neuroimage, 2009. **46**(1): p. 177-92.
4. Jovicich, J., et al., *Brain morphometry reproducibility in multi-center 3T MRI studies: a comparison of cross-sectional and longitudinal segmentations.* Neuroimage, 2013. **83**: p. 472-84.
5. Nugent, A.C., et al., *Automated subcortical segmentation using FIRST: test-retest reliability, interscanner reliability, and comparison to manual segmentation.* Hum Brain Mapp, 2013. **34**(9): p. 2313-29.
6. Schuff, N., et al., *MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers.* Brain, 2009. **132**(Pt 4): p. 1067-77.
7. Walker, E. and A.S. Nowacki, *Understanding equivalence and noninferiority testing.* J Gen Intern Med, 2011. **26**(2): p. 192-6.
8. Geuze, E., E. Vermetten, and J.D. Bremner, *MR-based in vivo hippocampal volumetrics: 1. Review of methodologies currently employed.* Mol Psychiatry, 2005. **10**(2): p. 147-59.
9. Ardekani, B.A., *yuki module of the Automatic Registration Toolbox (ART) for corpus callosum segmentation*. 2013.
10. Schuirmann, D.J., *A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability.* J Pharmacokinet Biopharm, 1987. **15**(6): p. 657-80.
11. Robinson, A., *equivalence: provides tests and graphics for assessing tests of equivalence*. 2010. p. R package version 0.5.6.
12. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.

13. Fischl, B. and A.M. Dale, *Measuring the thickness of the human cerebral cortex from magnetic resonance images.* Proc Natl Acad Sci U S A, 2000. **97**(20): p. 11050-5.

14. Pardoe, H., *cortex: Sample size estimates for well-powered cortical thickness studies*. 2012.

15. Team, R.C., *R: A Language and Environment for Statistical Computing*. 2013, R Foundation for Statistical Computing.

16. Monk, C.S., et al., *Abnormalities of intrinsic functional connectivity in autism spectrum disorders.* Neuroimage, 2009. **47**(2): p. 764-72.

17. Di Martino, A., et al., *The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism.* Mol Psychiatry, 2013.

18. Pardoe, H.R. and G.D. Jackson, *Manual hippocampal volumetry is a better detector of hippocampal sclerosis than current automated hippocampal volumetric methods.* AJNR Am J Neuroradiol, 2013. **34**(10): p. E114-5.