THE UNIVERSITY *of* EDINBURGH

Multicentre structural and functional MRI

Viktoria-Eleni Gountouna

Doctor of Philosophy

University of Edinburgh

2013

Abstract

Neuroimaging techniques are likely to continue to improve our understanding of the brain in health and disease, but studies tend to be small, based in one imaging centre and of uncertain generalisability. Multicentre imaging studies therefore have great appeal but it is not yet clear under which circumstances data from different scanners can be combined. The successful harmonisation of multiple Magnetic Resonance Imaging (MRI) machines will increase study power, flexibility and generalisability. I have conducted a detailed study of the performance of three research MRI scanners in Scotland under the name CaliBrain, with the aims of developing reliable, valid image acquisition and analysis techniques that will facilitate multicentre MRI studies in Scotland and beyond. Fourteen healthy volunteers had two brain scans on each of three 1.5T MRI research machines in Aberdeen, Edinburgh and Glasgow. The scans usually took place 2-3 weeks apart. Each scan was performed using an identical scanning protocol consisting of a detailed structural MRI (sMRI) and a range of functional MRI (fMRI) paradigms. The quality assurance (QA) of scanner performance was monitored in all three sites over the duration of the study using a three-part protocol comprising a baseline assessment, regular measures and session specific measures. The analyses have demonstrated that the data are comparable but also that within- and between-scanner variances are evident and that harmonisation work could enhance the level of agreement. The QA data suggest that scanner performance was similar between and within machines over the course of the study. For the structural MRI scans an optimised methodology was utilised to minimise variation in brain geometry between scanners and fit all the scanned brains into a common stereotactic space, such that repeated measures analyses yielded no significant differences over time for any of the three scanners. I examined the reproducibility of the fMRI motor task within and between the three sites. Similar results were obtained in all analyses; areas consistently activated by the task include the premotor, primary motor and supplementary motor areas, the striatum and the cerebellum. Reproducibility of statistical parametric maps was evaluated within and between sites comparing the activation extent and spatial agreement of maps at both the subject and the group level. The results were within the range reported by studies examining the reproducibility of similar tasks on one scanner and reproducibility was found to be comparable within and between sites, with between site comparisons often exceeding the within site measures. A components of variance analysis showed a relatively small contribution of the factor site with subject being the main source of variation. Similar results were obtained for the working memory

task. The analysis of the emotional face processing task showed poor reproducibility both within and between sites. These findings suggest that multicentre structural and functional MRI studies are feasible, at least on similar machines, when a consistent protocol is followed in all participating scanning sites, a suitable fMRI task is employed and appropriate analysis methods are used.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified.

(*Viktoria-Eleni Gountouna*)

Acknowledgements

Many thanks to my supervisors, all the collaborators and volunteers in the CaliBrain project, everyone in the Psychiatry department who provided support and everyone else who was there for me when I needed them.

Viktoria-Eleni Gountouna was supported by an MRC studentship. The CaliBrain study was funded by a Chief Scientist Office (Scotland) Project Grant (CZB/4/427), Chief Investigator Stephen Lawrie.

To Ari.

Table of Contents

List of Figures

List of Tables

Glossary

Consistent : without contradiction, semanticly or syntacticly.

Reliable : the ability of a person or system to perform and maintain its functions in routine circumstances, as well as hostile or unexpected circumstances.

Reproducible : The ability of a single component, or an entire experiment, or study to be reproduced, or by someone else working independently.

Valid : That data are collected according to the criteria specified, and that the resulting data are a consequence of the specified criteria.

Chapter 1

Introduction

1.1 Overview

In this thesis I present work done as part of the CaliBrain project, a multicentre structural and functional Magnetic Resonance Imaging (MRI) study conducted across Edinburgh, Glasgow and Aberdeen. In chapter 1 the motivation for multicentre MRI is presented, along with a brief description of the study and some background information on the methods and issues relevant to multicentre MRI. In chapter 2 the Calibrain study design, protocol and data acquisition are presented in detail. Quality Assurance (QA) analysis methods and results are described in chapter 3. In Chapter 4 the methods and analysis results for structural MRI component of the CaliBrain study are presented. Chapters 5 and 6 deal with the methods and results of the analysis of the CaliBrain fMRI motor and face processing tasks respectively. Finally, chapter 7 presents a discussion of all the findings of the CaliBrain study so far, including Quality Assurance, structural MRI harmonisation methods and the assessment of the reproducibility of the CaliBrain motor, working memory and face processing fMRI tasks within and between scanners.

1.2 Motivation for multicentre MRI

A relatively large proportion of the worldwide population suffers from some form of neurological or psychiatric disorder and these conditions cause more disability than any other category of disease. The biological basis of many of these, schizophrenia being a prominent example, has for years been elusive, making diagnosis, outcome prediction and treatment choice really challenging for clinicians. However over the last 25 years imaging studies have revealed both structural and functional differences between those affected with psychiatric disorders and those who are well. The current state of understanding and treatment of these conditions could benefit greatly from advancements in the area of brain imaging. Early diagnostic tests would allow earlier, more effective and possibly preventative therapeutic measures.

Magnetic Resonance Imaging (MRI) is a particularly flexible technique for examining brain structure and function, and is now widely available. Its rapid and wide adoption can be attributed to the high quality of images it produces, offering excellent resolution and tissue differentiation in conjunction with its non-invasive character, as it can provide clear and detailed representations of internal organs with a minimal risk for the patient. It is not associated with any known hazard and this allows multiple examinations, so as to follow subjects' progress over time.

Structural and functional Magnetic Resonance Imaging (sMRI & fMRI) in particular are increasingly popular research tools because they are non-invasive and can provide valuable information about the brain's anatomy and functional organisation both in healthy and clinical populations, such as neuropsychiatric patients. Although these techniques are swiftly gaining recognition as useful tools yielding meaningful results, there is vast potential and indeed a need for further refinement and standardisation of the methods currently used, as well as for a better understanding of the nature of the acquired data.

This becomes evident if one considers the possibility of large clinical studies, which are so common in other areas of medical research. Due to the wide diversity of available hardware, scanning sequences, image processing techniques and data analysis methods, actual implementations can and often do vary dramatically between different institutions. Therefore, neuroimaging studies have so far typically been restricted to small samples and carried out locally. Multicentre studies are an obvious next move, but very little work has been done to date to quantify and correct for these differences.

To ensure the successful harmonisation of different scanners, methodological advancement is required before the objective of standardisation can be reached. The first step in this endeavour is to determine factors contributing to uncertainty between scanners and identify or create appropriate metrics to quantify their effect. Because of as yet unresolved technical and methodological differences, comparison and integration of data across different research centres is hampered, which results in suboptimal use of resources. This is not an uncommon issue and most new technologies suffer from lack of standardisation at the early stages of their development, but it presents a problem that needs to be addressed soon, if we are to take full advantage of the possibilities structural and functional MRI can offer. Multicentre trials will promote the collaboration and sharing of resources, data and expertise between groups, while other benefits will include greater flexibility in subject recruitment, as well as the ability to use larger, better matched samples and the associated gains in statistical power. Moreover, this will open up a number of possibilities for the more efficient exploitation of existing data, by combining datasets already available. Large-scale studies at the national, European and even international level will make it possible to tackle questions that have so far been relatively difficult to address, including combinations of neuroimaging and genetics research.

1.3 The Calibrain project

In this thesis I present the CaliBrain project, a study designed with the goals to a) develop and refine reliable, valid image acquisition and analysis techniques that will facilitate multicentre MRI studies, b) to identify or develop appropriate measures for assessing inter-scan agreement for structural and functional MRI and c) to establish which of a battery of fMRI paradigms are suitable for use in a multicentre setting.

We scanned a total of 14 right-handed healthy volunteers twice on each of three MRI machines over a six month period. This data set made it possible to examine the critical methodological and technical issues arising from a prospective study using different scanners. The data collected allowed the evaluation of the suitability of harmonisation methods developed for sMRI data. Furthermore, a variety of different functional tasks were employed, likely to activate a wide range of typical areas of interest and generally known to have good reproducibility, to evaluate their suitability for use in multicentre studies. Finally, scanner performance was monitored throughout the study using 'phantoms' (test objects of known properties). The knowledge and experience gained from Calibrain will be of great value in conducting reliable and valid multicentre structural and functional MR imaging results experiments in the future.

1.4 Magnetic Resonance Imaging

1.4.1 Basic principles

Brain imaging is a fairly new and continuously developing field. The concepts behind the many techniques available are sometimes complex, and their implementation even more so. Magnetic Resonance Imaging (MRI) relies on the phenomenon of nuclear magnetic resonance. All atoms have a nuclear 'spin', and magnetic moment. This magnetization is created by the spin and electrical charges. By applying a strong magnetic field, these atoms tend to align or counter align with the static magnetic field of the scanner, a property which can be manipulated and magnetic imaging systems can be 'tuned' to detect specific types of nuclei (Matthews, 2001).

1.4.2 Structural MRI

Structural Magnetic Resonance Imaging (sMRI) of the brain is employed in the assessment of a wide range of neuropsychiatric disorders. Voxel Based Morphometry (VBM) has been established as a leading method for analysing large sMRI studies using anatomical images of high resolution. VBM is a fully automated process that is used to localise differences in brain parenchyma (Ashburner and Friston, 2000, 2005). The VBM implementation segments T1-weighted MRI scans into voxel-wise maps of grey and white tissue and Cerebrospinal Fluid (CSF) and allows the statistical comparisons of these maps over time or between different populations. VBM requires good quality image co-registration at the voxel level and can be sensitive to even slight differences between MRI scanners. Potential sources of systematic variation are presented in section 1.4.4.

1.4.3 Functional MRI

Blood Oxygen Level Depended Functional Magnetic Resonance Imaging (or BOLD fMRI), measures relative changes in oxygen levels in the brain. Oxygenated blood flow increases in areas that are active and this is what reveals their location in fMRI (Hüttel et al., 2004). What is being measured in BOLD fMRI is the haemodynamic response to neuronal activity and not the activity itself. This has very important consequences on experimental design, since the time course of the haemodynamic response (measured in seconds) does not correspond to that of the underlying activity (measured in milliseconds). The differences that are being measured are very slight.

In functional MRI studies a time series of images is acquired, usually while the subject performs a number of repetitions of a cognitive task in the scanner. Various preprocessing steps are performed on the images, including registration in a common space, and a statistical model is fitted in order to determine in which brain areas signal change is consistently correlated with the task.

1.4.4 Potential sources of variability

Ideally, images produced by MRI systems would be imperfection free and completely reproducible across different machines. In reality this is not the case, as various sources contribute to image noise. This is especially important in fMRI, where the signal-to-noise and contrast-to-noise ratios are particularly low (Hüttel et al., 2004). These imperfections gain even more weight when considered from the perspective of multi-centre imaging, as systematic differences between data collected on different scanners could influence analysis results.

Contributions to noise are made by both imaging hardware and subject physiology (Krüger and Glover, 2001; Hüttel et al., 2004). System noise sources include drift and imperfections in RF, gradient, and shim subsystems. Field inhomogeneities caused by gradient nonlinearities and instabilities and RF coil loading effects cause spatial distortions in the images, as well as intensity variations such as scanner drift. Systematic between scanner differences in the characteristics of all these factors could be a concern in multicentre MR imaging. Physiological noise also has a variety of sources, including fluctuations in the basal cerebral metabolism, blood flow and volume, as well as cardiac and respiratory activity and subject movement, which could cause image distortions. Systematic differences in motion artefacts especially could be an important factor in multicentre functional imaging studies, as different sites use different methods for head stabilisation, some less effective than others.

The reproducibility of fMRI results however can be influenced by other factors as well. The consumption of substances like caffeine, nicotine or alcohol (Stefanovic et al., 2006; Brown and Eyler, 2006), or a high fat meal (Noseworthy et al., 2003), and psychological factors, such as anxiety caused by the scanner environment (Raz et al., 2005), can affect global cerebral blood flow, which in turn can have an effect of the magnitude and dynamics of the BOLD signal. The potential effects of the time of day on physiology and cognition (Carrier and Monk, 2000), especially in combination with the substance effects mentioned above, are another important consideration with

regards to fMRI reproducibility. Therefore, consistent scan scheduling differences between sites could also introduce systematic differences in the data.

Identifying and eliminating or at least reducing the impact of potential sources of systematic variation on imaging data acquired in different machines is a complex, multifaceted problem, compounded by the fact that it is also possible that some of these sources may interact in unknown and unpredictable ways.

1.5 Quality Assurance

It would be ideal if images could be independent of scanner characteristics and reflect only the state of the subject at the time they were scanned (Tofts, 1998). However this is usually not the case, due to the existence of the issues presented above in section 1.4.4 . The ultimate aim of any Quality Assurance (QA) program is to *'detect changes in performance before they can adversely affect clinical images'* (Firbank et al., 2000). In MRI, this is achieved by ensuring that specific image quality measures meet a set standard. For a longitudinal, multi-centre study, QA is particularly important as it ensures that any change in the data of a subject can be attributed to genuine, subject-specific changes, and not changes in the equipment or analysis protocols (Koller et al., 2006).

In a busy clinical setting, a compromise has to be reached between the scanning time dedicated to QA and number and frequency of the tests performed. As the most important measure of image quality, the signal to noise ratio (SNR) is most often cited in QA protocols since it provides a sensitive - albeit non-specific - measure of the performance of an MR system (Lerski et al., 1998). When measured from the same test object, using the same coil and sequence parameters, the SNR should remain stable hence SNR is usually measured routinely as part of local and multi-centre QA programmes (Koller et al., 2006; Colombo et al., 2004). However, there is no consensus on the frequency of SNR measurements in routine or previous multi-centre QA programmes, with daily, weekly and monthly intervals reported, most commonly using the head coil (Firbank et al., 2000; Koller et al., 2006).

1.6 Multicentre Voxel Based Morphometry

The goal of the Alzheimers Disease Neuroimaging Initiative (ADNI) project is a longitudinal analysis of ageing, and to facilitate this within site MRI reproducibility was tested on a range of scanners and sequences. Research for the ADNI project demonstrated that to pool scans from multiple sites, it is important to minimise differences between sites (Hua et al., 2008; Jack et al., 2008; Leow et al., 2006). Based upon this research an MR-RAGE sequence was recommended for multiple site scanning and a scheme of corrections that includes field mapping and geometry correction is applied. ADNI researchers investigated the use of B1 field mapping to correct for within scanner variation in the RF inhomogeneity for phased array head coils (Leow et al., 2006), and indicated that this technique has limitations.

Reports of VBM analyses that combine scans from different sites for analysis have applied validity assessments (Stonnington et al., 2008; Meda et al., 2008). In a validity assessment, a VBM contrast of control subjects between the contributing sites is used to map the regions of significant difference between scanners, which is used to form a masking image that charts these regions. These masked regions are then excluded from VBM reporting as results in these regions could be driven by artifactual scanner differences (Tofts, 1998; Jovicich et al., 2006). Meda et al. (2008) demonstrated in a VBM study of psychosis at four centres that it is possible to limit the effects of scanner differences by validity masking and by ensuring that the patients and controls included in the pooled analysis are drawn equally from all contributing sites. Stonnington et al. (2008) presented a VBM analysis of data acquired using six scanners, where the extent of validation masking required could be limited to a single region in the thalamus through the use of equivalent scan sequences and good quality control. In VBM however, the use of validity masking is undesirable because it limits the analyses to less than whole brain coverage.

In VBM analyses, the harmonisation constraints for the use of multiple scanners are difficult to define as VBM requires the provision of corrections for scanner differences at the voxel level. Previous work has shown that it is possible to pool scans from multiple centres in parcellated volumetric studies. In a volumetric analysis of images from multiple scanners, van Haren et al. (2003) utilised a semi-automated method, where global corrections for the tissue classification were computed separately for each scanner. The methods reported summary volumes for grey and white matter in the cerebrum, and cerebellum and lateral ventricle volumes (Schnack et al., 2004). This

set-level segmentation method employed global estimates of the intensity values that marked the transitions between tissue types and CSF. These globally applied transitions were adjusted for each scanning site.

A methodology that seeks to minimise the differences between scanners through an integration of scan sequence parameters into the segmentation functions was proposed by Fischl et al. (2004), which gives global adjustment in the intensity to tissue mapping. These global corrections are appropriate in studies where the inferences drawn are limited to lobar tissue occupancy. A volumetric method that addresses the localised intensity to tissue mappings has been proposed by Han and Fischl (2007) that recognises that localised adjustments for the intensity to tissue mapping within the brain are necessary for scan pooling to be valid for parcellation studies.

1.7 Functional MRI reproducibility

Comparatively little work has been done on the test-retest reliability of fMRI within a single scanner, and even less on the reproducibility of findings between different systems. A variety of methods have been used to address this, focusing on different aspects of the issue. Most studies have concentrated on the visual and motor systems, but higher order cognitive systems also have been targeted to a lesser extent.

1.7.1 Assessing the reproducibility of fMRI activation

The most widely used measures for assessing the reproducibility of fMRI activation are the size and overlap ratios, examining the stability of activation extent and spatial agreement of statistical parametric maps. These are usually applied to thresholded maps so that only statistically significant voxels are considered (Ramsey et al., 1996; Yetkin et al., 1996; Rombouts et al., 1998; Tegeler et al., 1999; Wagner et al., 2005; Rutten et al., 2002; Raemaekers et al., 2007; Miki et al., 2000; Rau et al., 2007; Maldjian et al., 2002; Machielsen et al., 2000; Havel et al., 2006; Feredoes and Postle, 2007), however region of interest approaches have also been used (Duncan et al., 2009; Yoo et al., 2005; Swallow et al., 2003; Machielsen et al., 2000; Harrington et al., 2006b,a).

Various implementations of the Intraclass Correlation Coefficients (ICCs) are also quite popular, usually calculated for a region of interest (Manoach et al., 2001; Schunck et al., 2008; Wei et al., 2004; Aron et al., 2006; Kong et al., 2007; Johnstone et al., 2005; Zou et al., 2005; Friedman et al., 2008; Bosnell et al., 2008). Voxel-wise approaches have also been applied, either for the whole brain (Raemaekers et al., 2007; Freyer et al., 2009) or for statistically significant voxels only (Specht et al., 2003; Caceres et al., 2009). These are usually calculated on the basis of percent signal change (Manoach et al., 2001; Specht et al., 2003; Friedman et al., 2008; Kong et al., 2007; Bosnell et al., 2008; Schunck et al., 2008) or contrast values (Johnstone et al., 2005; Aron et al., 2006; Caceres et al., 2009; Freyer et al., 2009), but statistic values have been also been used (Wei et al., 2004; Raemaekers et al., 2007).

Other approaches to reproducibility have been applied, highlighting different aspects of the issue, such as the Pearson correlation (Tegeler et al., 1999; Miller et al., 2002, 2009; Wagner et al., 2005), the Coefficient of Variation (Loubinoux et al., 2001; Marshall et al., 2004; Leontiev and Buxton, 2007; Magon et al., 2009), kappa (Le and Hu, 1997; Thirion et al., 2007; Liou et al., 2009) and ROC curves (Le and Hu, 1997; Manoach et al., 2001), among others.

1.7.2 Reproducibility within and between sites

Regardless of the method used, it is generally found that regional patterns of activation are qualitatively repeatable but are quantitatively of high variability, both within and between individual subjects (e.g. McGonigle et al., 2000; Marshall et al., 2004), while within session reproducibility of results tends to be better than that of sessions performed on different days (Yoo et al., 2005). Amplitude of brain activation is more likely to be reliable than extent (Cohen and DuBois, 1999; Marshall et al., 2004), and additional benefits can come with optimization of analysis methodology e.g. with respect to spatial alignment processing and time-series statistics (Smith et al., 2005; Thirion et al., 2007). Finally, and importantly for clinical multicentre studies, reproducibility of activation maps are shown to be better at the group level than at the subject level (Yoo et al., 2005; Chee et al., 2003). However, clinical populations may exhibit greater variability than healthy subjects (Manoach et al., 2001), so potential population differences in reproducibility should be taken into account.

Some studies have looked at inter-site reproducibility. Casey et al. (1998) qualitatively assessed the reproducibility of a working memory task across four different 1.5T scanners and found good agreement in the patterns of activation. Vlieger et al. (2003) examined the reproducibility of a visual task within and between two similar 1.5T scanners and found inter-scanner reproducibility ratios to be comparable to those within-scanner.

Krasnow et al. (2003) compared activation maps for various tasks between 1.5T and 3T. They observed substantial increases in activation volume in the 3T data and also found that the higher strength scanner offered greater sensitivity for detecting activation in a number or areas. Fera et al. (2004) however, investigated the effect of echo times and bandwidth on differences between 1.5T and 3.0T using a motor task and found that the noise increase in higher strengths attenuated to some extent the potential benefits in terms of increases in statistical values. Voyvodic (2006) investigated the reproducibility of a hand motor task examining the effect of scanning sequence (gradient echo versus spiral) and field strength (1.5T and 4T) and found that while activation level and spatial extent varied, location was found to be stable.

Zou et al. (2005) examined the effect of many factors, including field strength, manufacturer, subject and visit, on the reproducibility of activation extent in a sensorimotor task and found subject, field strength and k-space differences to have a significant impact on reproducibility. Studies employing a variance components analysis to examine

the relative contributions of site and other factors to the variance in multicentre data sets found the effect of site to be small compared to that of subject and residual unexplained variance (Friedman et al., 2008; Costafreda et al., 2007; Suckling et al., 2008; Sutton et al., 2008).

In summary, activation patterns are generally consistent within and between subjects, visits and sites. Nevertheless, the actual level of the reproducibility of the results varies considerably and depends on many factors, including the task employed, the analysis methods used, the scanner field strength and choice of scanning parameters among others.

1.7.3 Reproducibility of visual perception activation

A few studies have examined the reproducibility of visual activation within and and between sites. Rombouts et al. (1998) scanned ten healthy subjects twice within one visit and once more in second visit while passively viewing red flickering lights in a block design paradigm. They report consistent task related activation in the primary visual cortex, the cuneus and the precuneus. The reproducibility of the size and overlap of activation clusters was better within than between subjects, with the size of clusters being generally more reproducible than the overlap. Specht et al. (2003) examined the reproducibility of visual activation in five healthy subjects in two separate visits using an event-related paradigm. Subjects were asked to vary their attentional effort while watching a flickering checkerboard pattern with varying letters in the centre. They report consistent activations across all conditions in primary visual areas, the inferior and medial occipital gyrus and the lingual and fusiform gyrus. Increased attentional load was associated with increased and more reliable activation in primary visual areas, the prefrontal cortex and the middle temporal gyrus.

The reproducibility of visual activation has also been studied across different sites. Krasnow et al. (2003) scanned fourteen subjects in machines of different strengths with a paradigm involving the passive viewing of a flickering checkerboard pattern. They report significant activation in the striate, extrastriate, and posterior parietal cortices for both scanners, while increased volume of activation was observed in the striate and extrastriate areas for the higher strength. Sutton et al. (2008) employed a similar paradigm and scanned four subjects fifteen times in each of two identical systems. They found good reproducibility of activation in visual areas across the two sites. In the Friedman et al. (2008) multicentre study five subjects were scanned twice on each

of ten scanners with a sensorimotor task which included a flickering checkerboard pattern. They reported consistent activation in the primary visual cortex across all sessions and reliability estimates for this region were amongst the highest of all the areas studied within and between scanners.

Functional activation in the primary visual cortex induced by a simple perception paradigm appears to be rather robust across different subjects, visits and sites. Increased attention can increase the level and reliability of activation, while scanners of higher strength are associated with increased volume, but not necessarily higher reproducibility of activation.

1.7.4 Reproducibility of motor activation

Motor systems are well studied and therefore popular for testing new methods and techniques. A few studies have addressed the reproducibility issue employing some kind of motor task. Mattay et al. (1998) for example examined the reproducibility of a sequential and random finger tapping task in eight subjects. They found consistent activation in typical motor regions, including the primary sensorimotor cortex, the premotor and supplementary motor area (PMA and SMA) and the cerebellum. Scholz et al. (2000) employed a variety of motor tasks including finger tapping and scanned twenty two healthy subjects two to eight times on different days to study reproducibility in motor areas. Consistent activations were found in the primary motor cortex, while activation of the SMA and the basal ganglia was less robust. Primary motor cortex activation also showed less variability than the SMA and basal ganglia within subjects. Between subjects the same pattern was observed, with between subject variability being generally larger than within subject.

Yoo et al. (2005) utilised a sequential finger tapping task to examine the long term reproducibility of motor activation. Eight subjects took part and they had nine scans eight weeks apart. In the first session they repeated the task with a 30min delay while remaining in the scanner. Task related activations were observed in the bilateral precentral, superior frontal and postcentral gyri, the inferior parietal lobule and transverse temporal gyrus, the thalamus, the left putamen and bilateral cerebellum. The researchers examined the reproducibility of activation in regions of interest only, namely the left primary and premotor area, supplementary motor area and the ipsilateral cerebellum. They found that group results were more reproducible than single subjects. Intra-session reproducibility appeared to be better than inter-session, however on the

whole ratios were comparable to other studies examining reproducibility over a shorter period of time.

Ramsey et al. (1996) examined the reproducibility of a finger opposition task in eleven healthy volunteers. Two series were acquired within one scanning sesssion, while nine of them had another scan on a different day. Consistent overlap of activation patterns was observed in the primary sensorimotor area within and between sessions with less reliable activation found in the PMA and SMA. While the location of activation in the primary sensorimotor area was consistent, the extent varied between subjects and sessions. Yetkin et al. (1996) scanned four subjects using a finger opposition task, with the task being repeated within the same scanning session, and also observed stable activation patterns in typical motor cortical regions. Tegeler et al. (1999) also employed a finger opposition task. Six subjects performed the task three times within one session. Activation in the primary sensorimotor area was found to be most reproducible followed by the cerebellum, while the supplementary motor area had poorer reproducibility.

The reproducibility of motor activation has also been studied across different scanners. Costafreda et al. (2007) scanned five volunteers twice on each of five identical systems with a finger tapping task. Significant activation clusters were observed in all occasions in the primary motor cortex. While some variability in activation amplitude and extent was evident, activation patterns were very similar for single subjects across visits and sites and variance between subjects accounted for the largest part of the variance. Friedman et al. (2008) scanned five subjects twice in each of ten scanners using a sensorimotor task which included finger tapping. They found significant clusters of activation in the primary motor cortex and the SMA across all subjects and sites. Region of interest reproducibility analyses in these areas yielded results similar to those reported for a single site.

On the whole, finger tapping and similar tasks appear to be fairly reproducible across different visits and sites both in the short and in the long term. The area showing the highest reproducibility is the primary motor cortex, followed by the cerebellum, while activation in the SMA, PMA and basal ganglia is less reliable.

1.7.5 Reproducibility of working memory activation

Most studies investigating the reproducibility of working memory related activation have used some version of the n-back task, which involves the recollection of infor-

mation presented some elements back, usually up to three. Casey et al. (1998) for example investigated the reproducibility of a spatial n-back task in eight subjects across four sites and reported areas exhibiting significant activation across all or most subjects and sites. The most reliable activation was observed in the right dorsolateral prefrontal cortex (DLPFC) and the right superior parietal lobule, with significant clusters being present in all analyses across subjects and sites. They also report relatively consistent activation in the left supplementary motor area (SMA) and premotor area (PMA) and the left insula. Marshall et al. (2004) presented reproducibility data for nine older healthy subjects performing a spatial n-back task in three sessions separated by weekly intervals. They report qualitatively similar patterns of activation, but the amplitude and extent of activation varied considerably. A coefficients of variation analysis showed similar results within and between sessions for activation amplitude, but much greater variation between sessions than within for activation extent.

Wei et al. (2004) investigated the long term reproducibility of an auditory n-back task. They scanned eight subjects twice within one session and once in each of seven follow up sessions conducted over six months. They report that a qualitative assessment of group activation maps for each session revealed consistent activation patterns. The results of a coefficients of variance analysis of an activation index combining amplitude and extent information in task related regions of interest suggested small longitudinal variability of activation. Furthermore, variation between subjects was found to be larger than variation within subjects. In terms of individual regions, relatively stable activations were observed bilaterally in the DLPFC, the SMA and PMA, Broca's area and the parietal lobe. The right DLPFC exhibited larger within-subject variability than the other areas and a significant session effect was observed in this region. The authors attribute this finding to the fact that this area has no distinct boundaries and is defined functionally rather than anatomically.

Caceres et al. (2009) scanned ten subjects twice separated by three months employing a verbal n-back task. Using an Intraclass Correlation Coefficients (ICCs) analysis they showed that voxels showing high activation at the group level were more reliable than voxels across the rest of the brain, but also pointed out that the voxels showing low group activation but high ICC can be explained by stable signals across sessions that are not well explained by the task model. They report that the areas with the highest reliability were the right frontal pole (rFPC) and the parietal cortex bilaterally, followed by the DLPFC bilaterally, the right ventrolateral prefrontal cortex (VLPFC), the PMA bilaterally and the right SMA. The lowest reliability was found in the left VLPFC.

Overall, different variations of the n-back task appear to reliably activate a number of regions. Most reliable activations have been consistently reported for the DLPFC and the parietal cortex, followed by the SMA and the PMA.

1.7.6 Reproducibility of affect processing activation

Relatively little is known about the reproducibility of fMRI activation in the affect processing network. Stark et al. (2004) investigated the reproducibility of brain activation patterns in response to fear-inducing, disgust-inducing and neutral pictures by scanning twenty four healthy subjects twice within one week. They report a significant activation decrease for the second visit in many brain regions including frontal, temporal and subcortical structures, which they attribute to novelty effects. Furthermore, they found activation patterns to be more stable in response to fear than to disgust. Suckling et al. (2008) scanned twelve healthy volunteers twice on two identical 1.5T systems employing both a block and an event-related version of a face processing task using sad facial expressions of varying emotional intensity. They conducted a components of variance analysis and examined the contributions of subject, site, visit and paradigm. They found the proportional variance attributed to site, visit and paradigm to be much less than that for subjects. They report significant effects of the paradigm factor in two regions of the left orbitofrontal cortex and in the right putamen and no significant effects of centre or visit, but a trend effect of site was present across the entire network.

Some of the literature has specifically focussed on the amygdala. Johnstone et al. (2005) examined the long term reproducibility of a face processing task. They scanned fifteen volunteers in three separate occasions over two months and used Intraclass Correlation Coefficients (ICCs) to assess the reproducibility of the task focussing on the activation of the amygdala in response to fearful, neutral and happy facial expressions. Fearful expressions were found to produce more reproducible activation than neutral or happy expressions, while a lateralisation effect was also observed with activation in the right amygdala being less reliable than the left. Krasnow et al. (2003) investigated the reproducibility of activation in the amygdala between scanners of different field strength, 1.5T and 3T. They scanned fourteen healthy volunteers with a face processing task using fearful, angry and neutral expressions. They detected no significant differences in height or extent of activation between the two scanners in this region and suggest that greater susceptibility artifacts at higher strengths may attenuate potential

increases in task-related activation.

Overall, fear seems to be the emotion inducing the most reproducible activation in the affect processing network in general and in the amygdala specifically. Moreover, there is no clear evidence for an effect of field strength or type of paradigm design on the reproducibility of activation in this region.

Chapter 2

The Calibrain study

2.1 Study overview

The CaliBrain project is a multicentre initiative with three participating research centres in Scotland aiming to assess the feasibility of multicentre structural and functional MRI and to highlight the critical methodological and technical issues arising from a prospective study using different scanners. Fourteen healthy volunteers had two brain scans on each of three 1.5T MRI research machines in Aberdeen, Edinburgh and Glasgow. The scans usually took place 2-3 weeks apart. Each scan was performed using an identical scanning protocol consisting of a detailed structural MRI (sMRI) and a range of functional MRI (fMRI) paradigms. The Quality Assurance (QA) of scanner performance was monitored in all three sites over the duration of the study using a three-part protocol comprising a baseline assessment, regular measures and session specific measures.

This project was conceived by myself as an idea for my doctorate degree. The main precept was to try and minimise potential sources of variation as much as possible, rather than explore the effects of using a wide variate of equipment, which was the fBIRN approach. The study outline was devised by myself, with the input of other members of the department's imaging team and our collaborators from Aberdeen and Glasgow. I also contributed to the writing of the grant application to the Chief Scientist Office, which was successful and enabled us to fund the study.

I devised the detailed study protocol, seeking specialist advice as appropriate, e.g. I relied on the advice of the participating medical physicists regarding the specifics of imaging sequences and their potential suitability for multicentre imaging and discussed paradigm design with fMRI specialists. Furthermore, I trained staff at the three sites to the study protocol, conducted pilot studies, recruited and trained participants and travelled to the sites as needed to facilitate data acquisition. I also devised a data management and organisation scheme appropriate for handling the study dataset, organised data transfer and generally liaised with the study collaborators as needed to overcome any arising difficulties in the management of the project. Finally, I devised and executed the data analysis protocol for the fMRI motor task, which was also applied on the working memory dataset by Victoria Gradin and Gordon Waiter in Aberdeen and later by myself on the face processing dataset. I also collaborated with T. William Moorhead on the development of the absolute distance metric and some of the structural data analysis and with Katherine Lymer on the implementation of the Signal to Noise metric.

2.1.1 Roles in the Calibrain study

People:

Edinburgh

VEG: Elvina Gountouna

KL: Katherine Lymer

DEJ: Dominic Job

SML: Stephen Lawrie

BM: Bill Moorhead

DMcG: David McGonigle

JH: Jeremy Hall

Glasgow

DB: David Brennan

Aberdeen

GW: Gordon Waiter

VG: Victoria Gradin

Project management:

Edinburgh

VEG: project management, project liaison, coordination of centers, meetings

KL: QA logistics

DEJ: IT, data transfer, project liaison

SML: project finances.

Recruitment:

Edinburgh: VEG

Aberdeen: GW

Glasgow: DB & VEG.

QA: Edinburgh:

Design: VEG, KL

Coding: VEG, KL

Testing: KL

Data acquisition: KL & VEG in Edinburgh, KL, DB & VEG in Glasgow and KL, GW & VEG in Aberdeen.

Structural MR:

Design, Protocol: VEG, DEJ, GW, BM

Testing: VEG, DEJ, GW, BM

Analysis: VEG, DEJ, BM

Functional MR:

Design, Protocol: VEG, DB, GW, DMcG, JH

Visual ER: VEG

Motor: VEG

N-back: GW

Faces: JH

Breath Hold: VEG

Testing VEG
Analysis VEG, DEJ, VG

Visual ER: McG (analysis not completed due to illness and staff leaving)

Motor: VEG

N-back: VG

Faces: VEG

Breath Hold (analysis not completed due DBs student taking different project)

2.2 Quality Assurance

As a first step towards multi-centre imaging in Scotland, we sought to assess the performance of the existing MRI research systems with the aim of using this information to optimise and develop future acquisition and post-processing structural and functional MRI protocols. The development of a Quality Assurance (QA) protocol suitable for multi-centre imaging was an important component of the study design.

2.2.1 Design

The quality assurance (QA) of scanner performance was monitored in all three sites over the duration of the study using a three-part protocol comprising a baseline assessment, regular measures and session specific measures. Scanning parameters were kept constant whenever possible to allow for direct comparisons of the datasets to be made.

A detailed assessment of scanner performance was conducted at baseline using a commercially available phantom set to obtain a series of baseline measures. Measures included signal to noise ratio (SNR)/contrast to noise (CNR), image uniformity, resolution, slice positioning, ghosting, geometric distortion and temporal stability. Data acquisition was carried out by Katherine Lymer and myself in Edinburgh, K. Lymer, David Brennan and myself in Glasgow and K. Lymer, Gordon Waiter and myself in Aberdeen.

To facilitate direct comparison between scanners and allow assessment of scanner stability, each site performed regular weekly and monthly QA using identical test phantoms and an identical protocol, which included measures of SNR, geometric distortion and temporal stability. Data was acquired by local staff in the three sites.

Finally, SNR measurements were also performed immediately before and after each human scanning session. To allow comparison of the SNR measurements from the test objects to those taken from the healthy volunteers, as well as the weekly and baseline QA data, all scans were acquired using the same image parameters. A schematic representation of the design is presented in Figure 2.1.

2.2.2 Test objects

Two sets of test objects (phantoms) were purchased to facilitate this approach. Three site-specific identically manufactured 3-D geometry MRI phantoms (Data Spectrum) were purchased to be used in the regular and session-specific QA (Figure 2.2). Before

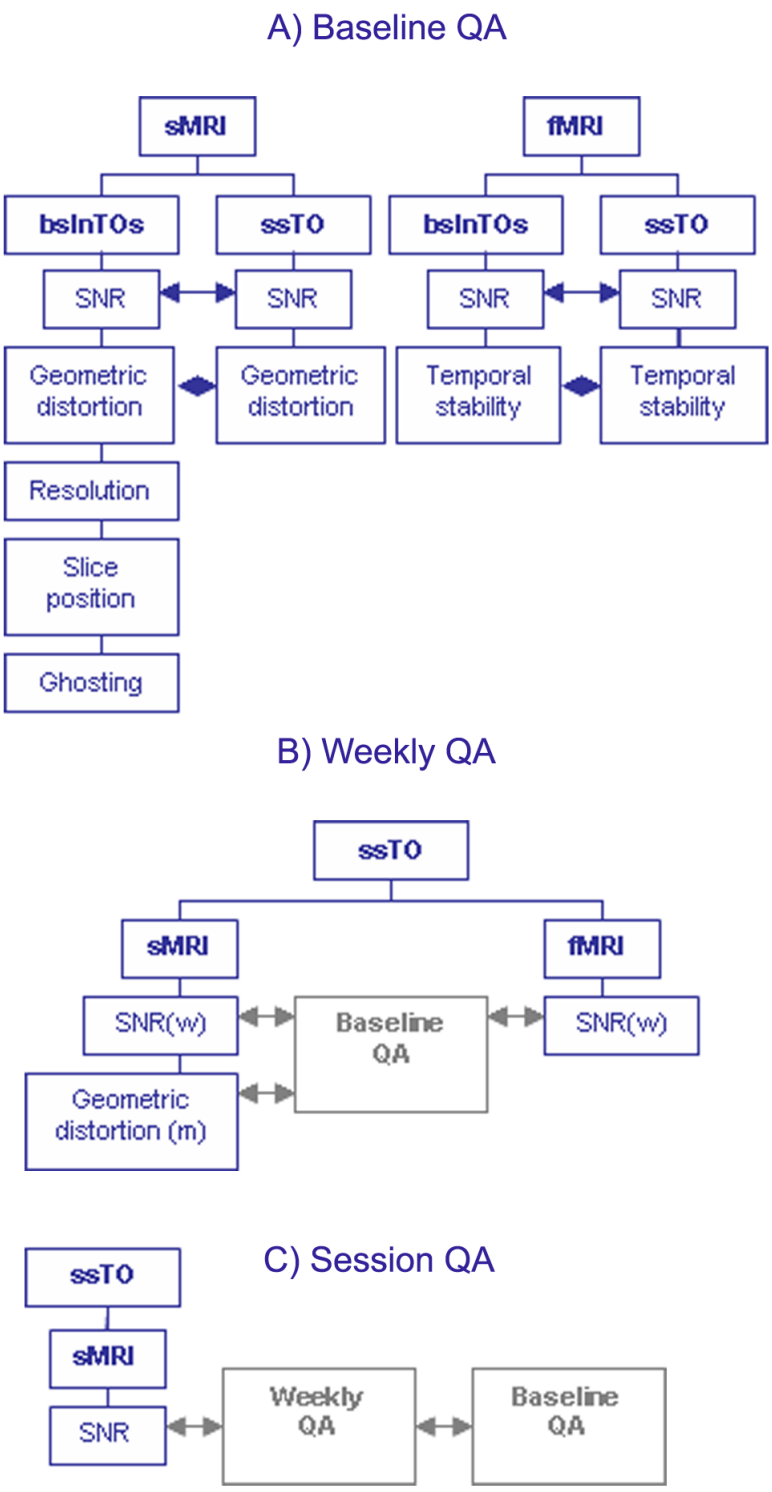

Figure 2.1: Quality Assurance Protocol. A: Baseline QA images were acquired using standard flood field / geometry phantoms and protocols. B: Weekly and Monthly QA images were acquired using the same procedure as those at baseline to allow comparison between the datasets. C: Time of Scan QA. A single image was acquired at the time of subject scanning to allow comparison to the baseline and routine QA.

Figure 2.2: Data Spectrum 3-D geometry MRI phantoms used in routine QA.

distributing the geometry phantoms to the three Calibrain sites, each object was CT-scanned to provide reference data sets for the geometric distortion measurements. A set of standard flood field and geometry phantoms (MagNet) were also purchased to allow detailed assessment of scanner performance at baseline (Figure 2.3). The MagNet set includes a flood field SNR/Uniformity phantom, a Geometric Linearity/Distortion and Slice Width phantom, a Spatial Resolution phantom, and a Slice Position phantom and a Copper Sulphate filled bottle phantom.

2.2.3 Scanning protocol

For the baseline assessment the MagNet set was transported to the three sites the previous day to allow the phantoms to settle. The following measurements were taken at each site. Scanner characteristics are presented in detail in section 2.4.1. For the SNR/Uniformity measurements data was collected using the Flood Field phantom with a T2-weighted fast spin echo sequence (T2FSE) in three planes (transverse, sagittal, coronal) and using the following parameters: TE 102ms, TR 6300ms, NEX 2, FOV 240mm, Matrix 256x256, Slice width 5mm, 1 slice. SNR measurements were also taken with the Data Spectrum 3D Geometry phantoms using the same parameters.

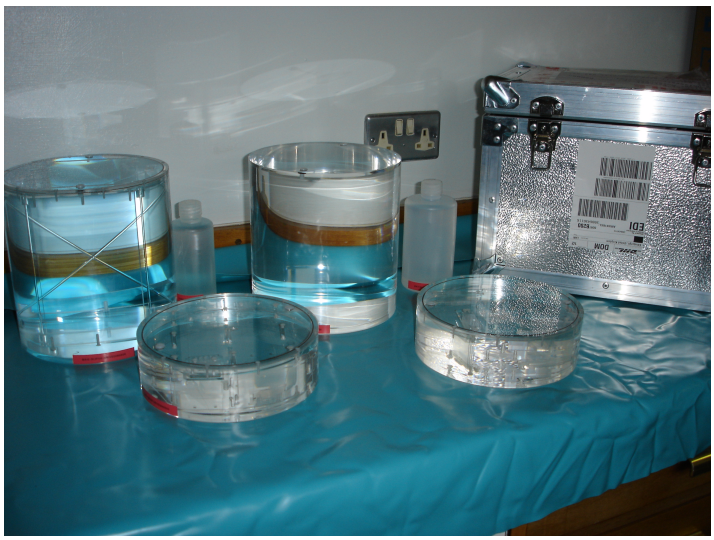

Figure 2.3: Standard Magnet Phantom set for baseline QA including a ood eld SNR/Uniformity phantom, a Geometric Linearity/Distortion and Slice Width phantom, a Spatial Resolution phantom, and a Slice Position phantom and a Copper Sulphate lled bottle phantom

For the Geometric Distortion measurement the Geometric Linearity/Distortion and Slice Width phantom was scanned in three planes (transverse, sagittal, coronal) using a T2-weighted spin echo sequence (SE) with the following parameters: TE 30ms, TR 1000ms, NEX 1, FOV 250mm, Matrix 256x256, Slice width 5mm, 1 slice. Geometric Distortion measurements were also taken with the Data Spectrum 3D Geometry phantoms using a T1-weighted fast spoiled 3D gradient echo sequence at 6 positions after performing 15^o rotations with the following parameters: TE min., TR min., NEX 1, FOV 240mm, Matrix 256x256, Slice width 2mm, 116 slices.

For the Resolution measurement the Resolution phantom was scanned in three planes (transverse, sagittal, coronal) using a T2-weighted spin echo sequence (SE) with the following parameters: TE 30ms, TR 1000ms, NEX 1, FOV 250mm, Matrix 256x256 and 512x512, Slice width 5mm, 1 slice.

For the Slice position measurement the Slice position phantom was scanned in the transverse plane using a T2-weighted spin echo sequence (SE) with the following parameters: TE 30ms, TR 1785ms, NEX 1, FOV 250mm, Matrix 256x256, Slice width 5mm, 45 slices.

For the Ghosting measurement the Copper Sulphate filled bottle was scanned in three planes (transverse, sagittal, coronal) using a T2-weighted multiple spin echo sequence (MSE) with the following parameters: TE 30/60/90/120ms, TR 1000ms, NEX 1, FOV 250mm, Matrix 256x256, Slice width 5mm, 1 slice.

For the fMRI SNR and temporal stability measurement the Data Spectrum 3D Geometry phantoms were scanned in the transverse plane using a gradient-echo echo-planar imaging sequence (EPI) with the following parameters: TE 40ms, TR 2500ms, NEX 1, FOV 240mm, Matrix 64x64, Slice width 5mm, 30 slices. Three series were acquired, one with 100 volumes and two with 300 volumes.

For the regular measurements the identical Data Spectrum 3D Geometry phantoms were always used. SNR T2 FSE and EPI measurements were collected weekly and geometric distortion measurements were collected monthly, all using the same parameters described above for the baseline scans. T2 FSE SNR measurements were collected at the start and end of each human scanning session again using the same parameters.

2.3 Participants

Fifteen healthy participants were initially recruited, out of which fourteen completed the study. One participant was not able to complete the study due to time constraints and another was scanned but subsequently excluded from some of the analyses due to an artefact in the interhemispheric fissure, caused by physiological calcification in the falx. Thirteen (nine male and four female; mean age 35.2 years, SD 7.3) were included in the motor task analysis. All participants were native English speakers, right-handed (self reported) and met the inclusion criteria for the study: no cardiac pacemakers, metallic implants and prostheses, aneurysm clips, metal teeth braces, pregnancy or potential pregnancy, history of diagnosed neurological disorder or major psychiatric disorder or treatment with psychotropic medication, including treatment for substance misuse. The participants were not paid, but they were reimbursed for travel and sustenance. On one occasion a participant was also reimbursed for loss of pay. Two participants were also reimbursed for an overnight stay away from home.

2.3.1 Recruitment

Due to the nature of the study, the recruitment of participants presented a challenge. Participation required multiple scans and extensive travel within the space of a few months. The schedule was not particularly flexible because of the demands placed by the counterbalancing of the order of scans and the fact that all scans were to be conducted before noon on working days. Furthermore, each scanning session lasted more than an hour and included five functional runs, which the pilot scans showed to be rather demanding, due to the fatigue caused by the long total duration of the functional component ($>$45min). Recruiting additional participants to replace people unable to continue would have been difficult, so it was important to find reliable volunteers. Recruitment was therefore actively done by researchers in all three centres, through posters and word of mouth. Some of the participants were staff or postgraduate students based in one of the three participating universities and were able to take time off in order to take part in the study. During the briefing it was explained to all volunteers what participation entailed and it was made clear that they could withdraw from the study at any time. All subjects provided written informed consent and the study was approved by the Grampian Research Ethics committee.

2.3.2 Training

All participants received a detailed briefing and training. Participants were informed of the purpose of the study and they were made aware of the effect of certain substances, like alcohol, caffeine and nicotine on the reproducibility of results. They were not asked to refrain from consumption, but to be consistent on the days of scanning and the night before. They were also trained on all fMRI tasks, by practising shorter versions of all tasks on a laptop on a different day from that of their first scan, and they were given the opportunity to ask questions.

Because training was conducted locally at the three sites, it was not possible for all participants to be trained by the same person. To ensure consistency, a training protocol was devised (Appendix B). David Brennan and Gordon Waiter received training by myself and were responsible for training in Glasgow and Aberdeen respectively. In Edinburgh training was done by myself.

2.4 Design

The main aim of the study was to test a range of tasks that cover the span of tasks currently used in psychiatric fMRI research studies. This is limited by the amount of time a subject can stay in a scanner for health reasons, and may also be limited by the subject's ability to attend to multiple lengthy tasks, and endure the physical aspects such as noise and the cramped conditions of the scanner enclosure. Also limited by the cost of MR scanning - approximately £550/hour. These factors, namely cost, health and safety, and attention span drove the selection of tasks. Priority was given to tasks that would produce the most useful results with respect to potential future multicentre studies, between the same centres in the study and in a more general case.

The minimum number of scanners required to allow for reproducibility testing is three. As a rule-of-thumb, the minimum number of subjects in fMRI in general is approximately fourteen, although more recent publications have suggested sixteen subjects would be a safer minimum (Friston, 2012).

All participants had six scanning sessions, two at each of the three sites.The initial design included two more scans in the Glasgow 3T scanner, but this was dropped in favour of conducting a separate 3T study at a later date. The order of visits was counterbalanced to avoid practice effects, for more details see section 2.7. Each session was identical and consisted of a localiser scan, two field mapping scans, six functional scans, a clinical T2-weighted scan and a high resolution anatomical T1-weighted scan. A phantom was scanned at the beginning and end of each session using the T2-weighted sequence. Scanning parameters were kept constant across scanners, with some minor variations arising due to hardware and software differences. The technical characteristics of the equipment used are presented in section 2.4.1 and the full scanning protocol is presented in section 2.4.2. The functional component included five tasks. The task design is presented in section 2.5. Pilot sessions were run at each site before the start of the data acquisition phase, these are briefly presented in section 2.6.

2.4.1 Equipment

Three General Electrics 1.5T scanners were used in this study, with some differences in hardware and software versions. A projector and Presentation software were used for the delivery of stimuli in Aberdeen and Glasgow and an LCD screen and IFIS/E-Prime in Edinburgh.

In Aberdeen scanning was conducted with a General Electrics (GE) 1.5T Signa

NVi/CVi scanner (software version 9.1; Echospeed gradients with max. amplitude 40mT/m and max. slew rate 150T/m/s; standard quadrature head coil). The ADW console (version 4.1) was used for the acquisition and reconstruction of fMRI data. A projector was used for the presentation of stimuli. Presentation v9.9 (Neurobehavioural Systems) was used for the delivery of stimuli and recording of behavioural responses, which were collected using an MR-compatible 4-button response unit.

In Edinburgh scanning was conducted with a General Electrics (GE) 1.5T Signa LX scanner (software version 9.1M4; Echospeed gradients with max. amplitude 22mT/m and max. slew rate 120T/m/s; standard quadrature head coil). The ADW console (custom installation) was used for the acquisition and reconstruction of fMRI data. The delivery of stimuli and the recording of behavioural responses was handled by IFIS-SA (Invivo), IFIS software vR14 with E-Prime v1.1SP3 (Psychology Software Tools). Stimuli were presented using an LCD screen, part of the IFIS system, mounted on the head coil and responses were collected with an MR-compatible 5-button response unit.

In Glasgow scanning was conducted with a General Electrics (GE) 1.5T Signa scanner (software version 11M3/11M4SP[1]; Echospeed gradients with max. amplitude 40mT/m and max. slew rate 150T/m/s; standard quadrature head coil). The main console was used for the acquisition and reconstruction of fMRI data. A projector was used for the presentation of stimuli. Presentation v9.9 (Neurobehavioural Systems) was used for the delivery of stimuli and recording of behavioural responses, which were collected using an MR-compatible 2-button response unit. Detailed scanner and fMRI equipment characteristics are presented in Table 2.1.

2.4.2 Scanning protocol

The scanning parameters were kept constant across the three scanners, allowing for some minor deviations arising from differences in scanner hardware and software. Tables with detailed scanning parameters for the three sites are presented in Appendix A.

A localiser scan in three planes was followed by two scans using an echoplanar imaging (EPI) sequence to be used for field mapping[2]. These were acquired with the following parameters: orientation parallel to the AC/PC plane; repetition time (TR) 2500ms; echo time (TE) 30/40ms; slice thickness 5mm without a gap; matrix 64x64;

[1] The software was upgraded on 4/8/2006.

[2] It was at a later date discovered that the parameters for the field mapping sequence were incorrect, rendering them unusable.

Table 2.1: Equipment

Site	*Aberdeen*	*Edinburgh*	*Glasgow*
Manufacturer	General Electrics	General Electrics	General Electrics
Field Strength	1.5T	1.5T	1.5T
Platform	Signa NVi/CVi	Signa LX	Signa
Software Version	9.1	9.1M4	11M3/11M4 SP[a]
fMRI Control	ADW 4.1	ADW custom installation	Main Console
fMRI Reconstruction	Auto (ADW)	Auto (ADW)	Auto (Main Console)
Gradients Max. Amplitude (mT/m)	40	22	40
Gradients Max. Slew Rate (T/m/s)	150	120	150
Head Coil	Quadrature	Quadrature	Quadrature
Stimulus Presentation Hardware	Projector	IFIS LCD screen	Projector
Stimulus Delivery Software	Presentation v9.9	E-Prime v1.1SP3 / IFIS vR14?	Presentation v9.9
Response Recording	4-Button Response Unit	5-Button Response Unit	2-Button Response Unit

Scanner and fMRI hardware and software characteristics across the three sites. [a] The software was upgraded on 4/8/2006.

field of view (FOV) 240mm^2; flip angle 90^o; 30 slices; 1 volume.

Six functional runs were performed with the same sequence. These were acquired with the following parameters: orientation parallel to the AC/PC plane; repetition time (TR) 2500ms; echo time (TE) 40ms; slice thickness 5mm without a gap; matrix 64x64; field of view (FOV) 240mm^2; flip angle 90^o; 30 slices; number of volumes varied in the different tasks (finger tapping 160; n-back 260; face processing 129; visual perception 100; breath holding 124). The first four images of each functional run were discarded to allow for stabilisation of the signal.

Two anatomical scans were also performed. A T2-weighted clinical scan was acquired using a fast spin echo (FSE) sequence with the following parameters: orientation parallel to the AC/PC plane; repetition time (TR) 6300ms; echo time (TE) 102ms; slice thickness 5mm with a 1.5mm gap; matrix 256x256; field of view (FOV) 240mm^2; flip angle 90^o; 19 (Aberdeen; Glasgow) or 20 (Edinburgh) slices. A high resolution T1-weighted scan was acquired using a 3D inversion recovery-prepared fast gradient echo volume sequence with the following parameters: orientation coronal; repetition time (TR) 5.9ms (Aberdeen; Glasgow) or 8.2ms (Edinburgh); echo time (TE) 1.9ms (Aberdeen) or 3.3ms (Edinburgh) or 1.4ms (Glasgow); slice thickness 1.7mm without a gap; inversion time (TI) 600ms; matrix 256x192; field of view (FOV) 220mm^2; flip angle 15^o; 128 slices.

At the beginning and end of each scanning session one volume was acquired using the 3D geometry phantoms (see section 2.2.2) and the same sequence as that of the human T2-weighted clinical scan.

2.5 fMRI tasks

In each visit participants performed a finger tapping task (section 2.5.1), a N-Back task (section 2.5.2), a face processing task (section 2.5.3), a simple visual perception task (section 2.5.4) and a breath holding task (section 2.5.5). Each task consisted of one run, except for the emotional task, which consisted of two runs. Details of the task properties are summarised in Table 2.2.

2.5.1 Finger tapping

The finger tapping task employed a block design with three conditions, 'sequential tapping', 'random tapping' and 'rest'. For the sequential condition participants were instructed to tap the fingers of their right hand sequentially in time with a flashing '#' symbol, starting with the thumb and finishing with the little finger. For the random condition participants were asked to tap their fingers in a random way in time with a flashing '?' symbol. During the rest condition they were asked to just fix their gaze on a flashing '+' symbol.

In all conditions the symbol was flashing with a frequency of 1Hz. Each block had a duration of 30s and included 28 trials and a 2s verbal prompt at the beginning with the words 'sequence', 'random' and 'rest' respectively. Each run included four repetitions of each tapping condition and five repetitions of the rest condition.

2.5.2 N-Back

The N-back task employed a parametric block design with four conditions of increasing difficulty, '0-Back', '1-Back', '2-Back' and '3-Back'. Consonants of the Latin alphabet were used as stimuli. In the control condition ('0-Back') participants were asked to press the middle finger button whenever the letter 'X' appeared and the index finger button for any other letter. In the '1-Back' condition participants were asked to press the middle finger button if the letter on the screen was the same as the immediately previous one and the index finger button if not. In the '2-Back' and '3-Back' conditions participants were asked to press the middle finger button if the letter on the screen was the same as the one two or three letters back and the index finger button if not. Each block had a duration of 40s which included a 5s verbal prompt and comprised 14 trials. Each trial had a duration of 2.5s, stimuli were presented for 1.5s followed by a blank screen. Blocks were always presented in a fixed order, '0-Back' –

Table 2.2: fMRI task properties

Task	*Nr. of Runs*	*Run Duration*	*Nr. of Volumes*	*Design*	*Conditions (Repetitions)*	*Block/Event Duration*	*Block/Event Structure*
Finger Tapping	1	390s	156	Block	Rest (5), Sequential Tapping (4), Random Tapping (4)	30s	2s verbal prompt, 28 1s trials (0.5s fix 0.5s blank)
N-Back	1	640s	256	Block	4 x (0-Back, 1-Back, 2-Back, 3-Back)	40s	5s verbal prompt, 14 2.5s trials (1s fix 1.5s blank)
Face Processing	2	312.5s	125	Block	Rest (9), Neutral (4), Fear (4)	12.5s rest, 25s face	1s verbal prompt, rest 11.5 fix, face 6x4s trials (3.5s face 0.5s fix)
Visual Perception	1	240s	96	Event-Related	Visual Events (24)	1s	checkerboard flickering at 8Hz
Breath Holding	1	300s	120	Block	Rest (7), Breath Holding (6)	30s rest, 15s breath	1s verbal prompt, rest 29s fix, breath 14s fix

Design and characteristics of the fMRI tasks employed in the Calibrain study, presented in the order they were emplyed in the scanner.

'1-Back' – '2-Back' – '3-Back', and each run comprised four complete cycles.

2.5.3 Face processing

The face processing task employed a block design with three conditions, 'fear', 'neutral' and 'rest'. In the fear condition faces expressing the emotion of fear were presented and in the neutral condition faces with a neutral emotional expression were presented (see 2.4, bottom). During rest periods subjects were instructed to look at a fixation cross. In each face block six greyscale pictures of faces from the Ekman and Friesen series (Ekman and Friesen, 1976) were presented in a random order, three male and three female. Participants were asked to press the index finger button if they thought the face presented was male and the middle finger button if they thought the face was female. The two response choices ('Male' and 'Female') were displayed in the bottom part of the screen during the presentation of each face. The same individuals were shown in both the fear blocks and neutral blocks which differed only in the emotion shown on the faces.

Fear blocks and neutral blocks were alternated and had a duration of 25s, which included a 1s visual prompt with the word 'gender?'. Rest blocks had a duration of 12.5s and included a 1s visual prompt with the word 'rest'. Each trial had a duration of 4s, the face was presented for 3.5s followed by a blank screen. Each run consisted of 4 fear blocks, 4 neutral blocks and 9 interleaved rest blocks. The participants completed two runs of the task in each visit.

2.5.4 Visual perception

For the visual perception task an event-related design was used. The stimulus was a black and white checkerboard (see 2.4, top) flickering at 8Hz and each trial had a duration of 1s. The remainder of the time a simple fixation cross appeared in the centre of the screen. Participants were asked to focus on the cross and just look at the image whenever it appeared. Each run contained 24 events with a pseudo-randomised inter-stimulus interval (ISI) varying between 4 and 14s.

2.5.5 Breath holding

A block design was used for the breath holding challenge with conditions, 'rest' and 'breath holding'. During rest periods a fixation cross appeared on the screen and par-

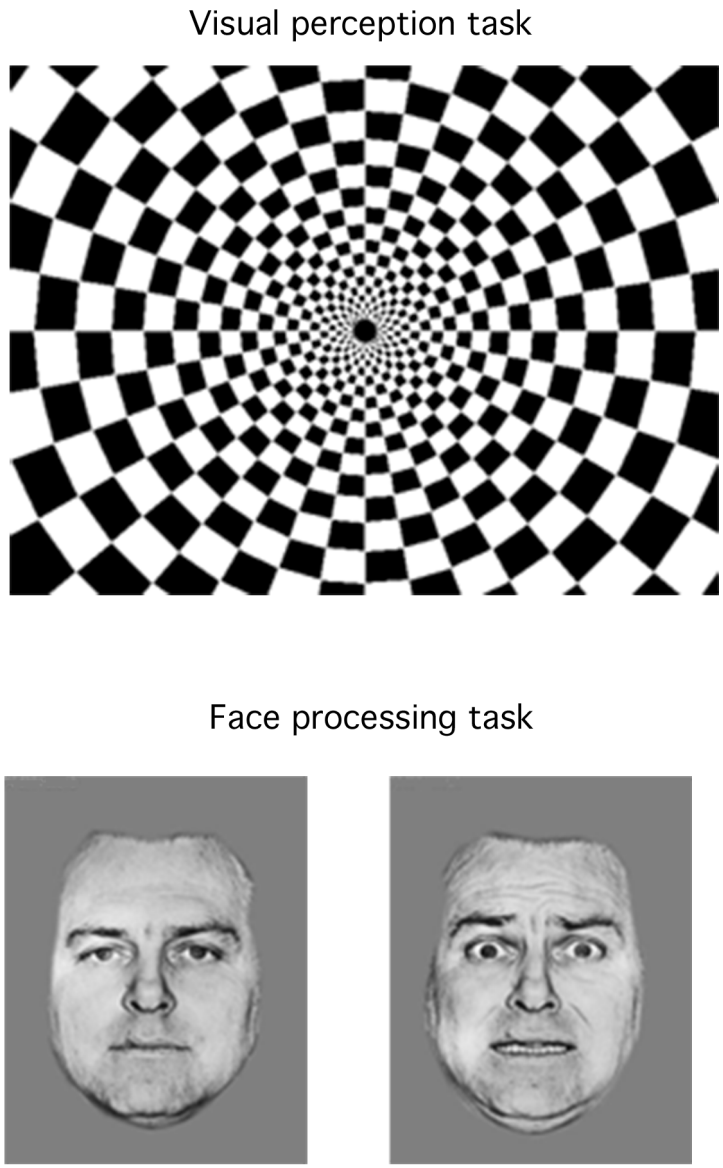

Figure 2.4: Functional task stimuli. Top: Black and white checkerboard flickering at 8Hz. Bottom: Pictures of faces with a neutral (left) or fearful (right) expression.

ticipants were asked to focus on the cross and ‘breath normally’. When the ’breath out and hold’ prompt appeared they were asked to breath out naturally and hold their breath for the whole duration of the block, while focussing on a ‘#’ symbol.

Rest blocks had a duration of 30s and breath holding blocks had a duration of 15s. Both types of blocks included a 1s verbal prompt. Each run included 7 rest and 6 breath holding blocks.

2.6 Pilots

Pilot scans were conducted before the onset of the data acquisition phase. The same subject (myself) was scanned at all sites with the full scanning protocol to check for inconsistencies and to qualitatively assess the differences in the subjective scanning experience at the three sites. Some minor revisions of scanning parameters and debugging of the stimulus presentation scripts followed the pilot scans.

2.7 Data acquisition

Data were acquired using three GE 1.5T Signa scanners (GE Medical, Milwaukee, Wisconsin), in Edinburgh (SFC Brain Imaging Research Centre, Division of Clinical Neurosciences, University of Edinburgh), Glasgow (Department of Clinical Physics, Institute of Neurological Sciences, Southern General Hospital) and Aberdeen (Department of Radiology, University of Aberdeen). Technical characteristics of the three scanners and other equipment used are presented in section 2.4.1, scanning parameters in section 2.4.2. The order of scans was counterbalanced to avoid practice effects, with five participants having their first scan in Edinburgh, four in Aberdeen and four in Glasgow. Some scans were done out of order due to scheduling restrictions. While an effort was made to keep the interval between scans constant, this was not possible in practice due to the difficulties presented by the necessity to reconcile the busy schedules of scanning sites and participants. The mean interval between visits was 25.2 days with an SD of 17.8 days. Details of the data acquisition schedule are presented in Table 2.3.

Table 2.3: Data acquisition schedule

				Group A			
Subject	*Gender*	*Aberdeen 1*	*Glasgow 1*	*Edinburgh 1*	*Aberdeen 2*	*Glasgow 2*	*Edinburgh 2*
S007	M	21/06/2006	12/07/2006	08/08/2006	14/08/2006	30/08/2006	15/09/2006
S005	M	14/06/2006	19/07/2006	15/08/2006	28/08/2006	13/09/2006	19/09/2006
S012	F	19/06/2006	05/07/2006	11/08/2006	28/08/2006	02/09/2006	07/09/2006
S011	F	19/06/2006	05/07/2006	11/08/2006	21/08/2006	02/09/2006	07/09/2006
				Group B			
Subject	*Gender*	*Edinburgh 1*	*Aberdeen 1*	*Glasgow 1*	*Edinburgh 2*	*Aberdeen 2*	*Glasgow 2*
S002	M	19/06/2006	05/07/2006	11/08/2006	28/08/2006	02/09/2006	07/09/2006
S010	F	06/06/2006	24/07/2006	30/08/2006	14/09/2006	23/10/2006	22/11/2006
S004	M	08/06/2006	03/07/2006	16/08/2006	12/09/2006	06/11/2006	15/11/2006
S003	M	08/06/2006	30/06/2006	18/10/2006[a]	21/09/2006	23/10/2006	01/11/2006
S001	M	01/06/2006	24/08/2006	30/11/2006[a]	08/09/2006	26/10/2006	21/12/2006[a]
				Group C			
Subject	*Gender*	*Glasgow 1*	*Edinburgh 1*	*Aberdeen 1*	*Glasgow 2*	*Edinburgh 2*	*Aberdeen 2*
S008	M	01/07/2006	20/07/2006	07/08/2006	23/08/2006	12/09/2006	20/11/2006
S013	F	28/06/2006	03/08/2006	10/08/2006	23/08/2006	23/09/2006	28/10/2006
S014	F	02/08/2006	14/08/2006	21/08/2006	13/09/2006	20/09/2006	12/10/2006
S009	M	20/09/2006[a]	05/07/2006	11/09/2006	22/11/2006[a]	21/09/2006	30/10/2006

[a] These scans were performed out of order.

Data acquisition schedule. The order of scans was counterbalanced across sites to avoid practice effects. The mean interval between visits was 25.2(SD 17.8) days.

2.8 Data management

The study required a significant amount of work in terms of electronic data management. This includes the issues of data transfer, presented in section 2.8.1, as well as the naming and physical organisation of data, presented in section 2.8.2.

2.8.1 Data transfer

The transfer of scanning data from the scanning centres to the UoE DoP data server was accomplished in different ways. Data acquired in Edinburgh was transferred immediately after the scan over the university computer network, using an established secure and automatic procedure. This was not possible at the other centres, due to security restrictions and policies of hospitals and universities. From Aberdeen all data was transferred using optical disks (DVD). In Glasgow data had to be manually copied to a different computer and transferred over the internet to the data server using the secure SSH protocol.

2.8.2 Data organisation

The large quantity of data dictates that most operations on the data are accomplished by batch scripting. To facilitate ease of data manipulation and minimise error, the consistent organisation and naming of data was deemed necessary. A method was devised that is independent of scanning site and clearly denotes the contents of each directory and file, which are uniquely named. It is also appropriate for use with regular expressions and filters to select directories and files for processing and analysis.

A simple structure was created for the hierarchy of directories on the file system, illustrated in Figure 2.5. A root directory was created for the study containing one directory per exam (scanning visit). Each exam directory contains all image and analysis files for that exam in clearly named directories. All directory and file names are coded by subject, site, visit and exam (e.g. *S001a1_10071_motor_img*). Some files created by the analysis software are not assigned unique names. The ones used in further analyses, like contrast images and statistical parametric maps, are renamed using a shell script (e.g. *S001a1_10071_motor_1stlevel_pstcon_0005.img*).

Each exam directory name consists of the subject code, the site code and the corresponding visit number, followed by the date of the exam (e.g. *S001a1_240806*). This directory contains all image and analysis data for this exam, coded by subject

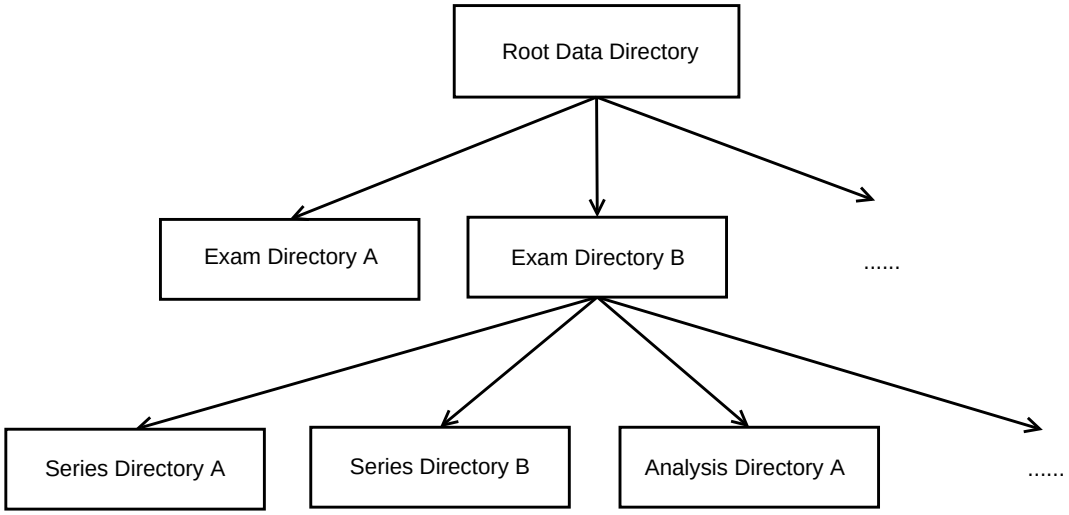

Figure 2.5: Data Structure. Diagram showing the organisation of image data on the le system.

code, site code and visit number, followed by the exam number assigned by the scanner and a descriptive name (e.g. *S001a1_10071_t1_img*, *S001a1_10071_motor_img*, *S001a1_10071_motor_1stlevel_pst*). The file name for each raw image consists of the parent directory name and volume number (e.g. *S001a1_10071_motor_img0001.img*). A short prefix is then added to the name of processed images by the batch scripts. Some files created by the analysis software are not assigned unique names. The ones used in further analyses, like contrast images and statistical parametric maps, are renamed using a shell script (e.g. *S001a1_10071_motor_1stlevel_pstcon_0005.img*).

2.8.3 File format conversion

Due to differences in the way EPI data acquisition was handled at the three sites, the raw data came in different formats. The raw data from Edinburgh and Aberdeen were acquired using the ADW console and came in a proprietary GE format. They were converted to SPM compatible ANALYZE format using the GE2SPM SPM extension, version 3.1 (Souheil Inati, http://www.fil.ion.ucl.ac.uk/spm/ext/). In Glasgow EPI data were acquired using the main console and the raw images came in DICOM format. They were converted to SPM compatible ANALYZE format using the SPM2 DICOM toolbox.

Chapter 3

Quality Assurance

3.1 Overview

Ensuring that the reproducibility of data collected at different sites and over time is an important component of multicentre imaging. In order to address this issue, Quality Assurance (QA) data was collected using both test object and human images to assess the stability of the three scanners used in the CaliBrain study.

The signal to noise ratio (SNR) is the most often cited measure in QA protocols, since it is easy to calculate and provides a sensitive if non-specific measure of the performance of an MR system (Lerski et al., 1998). However, it is not clear what the optimal frequency of SNR measurements should be. The frequency of SNR measurements was varied during Calibrain, in order to determine the most appropriate sampling interval. In addition, the relationship between the variability of in-vitro and in-vivo measurements was investigated in order to determine which measures would most accurately reflect data quality.

The Quality assurance (QA) analysis presented here was performed by Katherine Lymer using methods developed in collaboration by her and myself.

3.2 Methods: Signal to Noise

Fourteen healthy volounteers were scanned twice on each of the three scanners, at approximately three monthly intervals. At each site, in-vitro SNR measurements were taken at three discrete frequencies: (1) Once at study baseline; (2) once a week as part of the longitudinal QA; (3) immediately before and after each subject was scanned. Identical test objects were scanned at each site to allow direct comparison of the resulting data using a standard T2-weighted sequence with identical acquisition parameters for both the in-vivo and in-vitro scanning except for the number of slices. Details of the protocol are presented in 2.2.

Regions of interest (ROI) were placed within the high-signal region of the test-object image and away from the internal structural boundaries. ROIs of the same size were placed bilaterally in the white matter (WM) of the longitudinal fasiculi in each subject. This homogenous WM region was chosen to approximate position of the single slice from the test object. Individual ROIs were drawn for each subject / test object image to ensure correct placement. The mean and SD of the signal intensities were computed in each of these regions and the SNR was calculated using SNR= (signal/SDsignal)(Firbank et al., 2000). The variance was analysed to identify any significant differences between data series.

The temperature dependence of the T1 in paramagnetic salts has been well documented (Lerski et al., 1998) and so to allow the effects of temperature to be separated from real system changes, each site was asked to record the magnet room temperature at the time of scanning. Since the test-objects were permanently stored in this room, the room temperature was believed to be an accurate reflection of that in the test object.

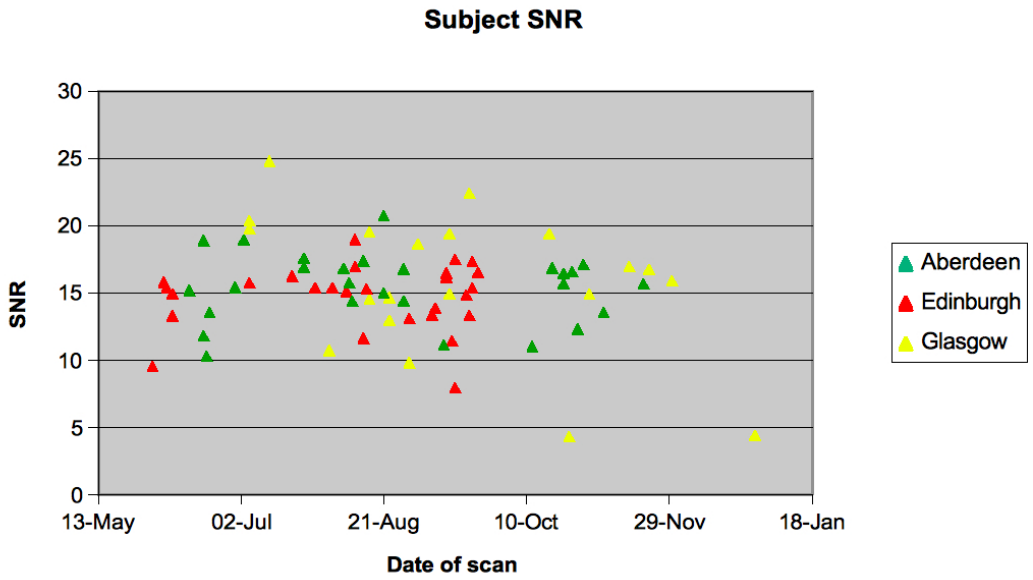

Figure 3.1: Signal to Noise (SNR) data for all subjects at each of the three sites.

3.3 Results: Signal to Noise

3.3.1 Variance of in-vivo SNR measurements

The SNR measurements acquired from the healthy subjects in each of the three sites are shown in Figure 3.1. Comparison of the variance between these measurements showed no significant differences between those obtained in Aberdeen and Edinburgh but significant differences between the SNR measures from Glasgow and the both Aberdeen and Edinburgh. This is due to the much larger variance of the Glasgow data (var = 28.80 compared to var = 6.38 in Aberdeen and var = 5.77 in Edinburgh).

3.3.2 Variance of test-object SNR measurements

The weekly SNR measurements for all three sites are plotted in Figure 3.2. There are significant differences between the variance of the Aberdeen SNR (var = 3.55) compared to both that in Edinburgh (var = 28.40) and Glasgow (var = 13.50). Despite these large variances, the coefficients of variation (CV) for these measurements are all $\leq$10%, namely 5% for Aberdeen, 10% for Edinburgh and 9% for Glasgow.

Full data sets from the in-vitro SNR measurements were only available for two of the three centres (Aberdeen and Edinburgh) as a considerable proportion of the 'time of scan' data is missing from Glasgow. A summary of these results are presented in Figure 3.3 and Figure 3.4. The scanning room temperatures in both Edinburgh and

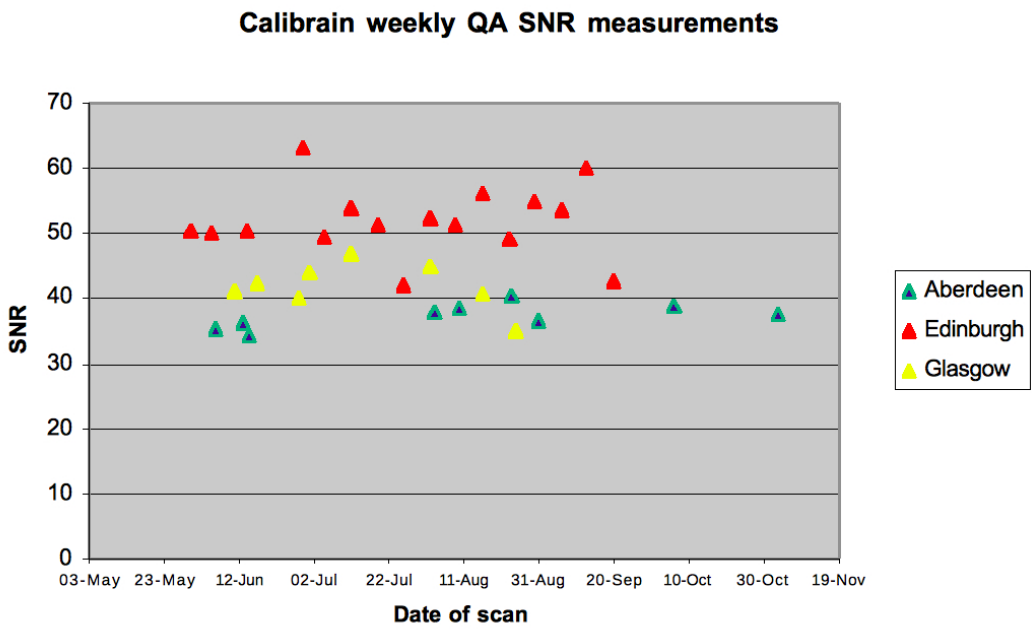

Figure 3.2: Weekly SNR measurements in all three Calibrain sites.

Aberdeen we stable for the duration of the study, with both sites recording temperature changes of $\pm 1^{o}C$.

There was no significant difference between the before and after 'time of scan' in-vitro variances from either Aberdeen or Edinburgh. This was also true of the comparison of the variance between the weekly QA and both the time of scan measurements in the Edinburgh data set. There were, however, significant differences between the weekly QA and the both the time of scan SNR variances in Aberdeen, with the 'before scan' and 'after scan' SNRs showing much higher levels of variance (25.92 and 36.26 respectively compared to the weekly QA variance of 3.55).

Comparison of the variance of the in-vivo SNR results to all of the in-vitro measurements showed no significant differences in Edinburgh dataset. However, in Aberdeen there were significant differences between the variance of the in-vivo SNR measures and those recorded immediately before and after scanning with the in-vitro measurements showing much larger degrees of variance (before scanning = 25.9, after scanning = 36.2, in-vivo = 6.38). Comparison of the weekly SNR and in-vivo variances showed no significant differences.

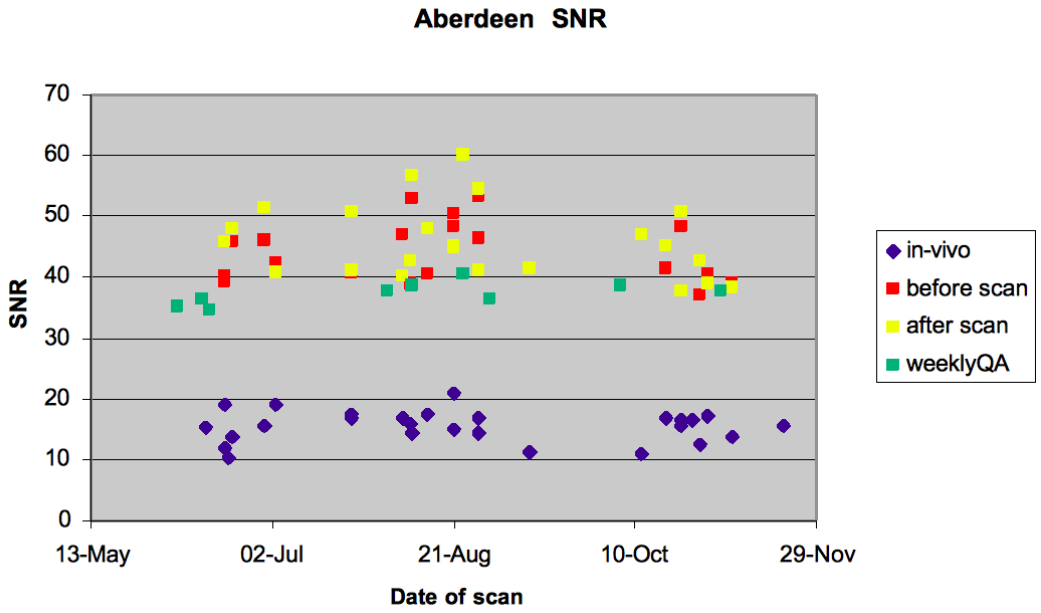

Figure 3.3: SNR results for the Aberdeen scanner, showing in vivo and in vitro data acquired during subject scanning and in vitro data acquired during routine weekly QA.

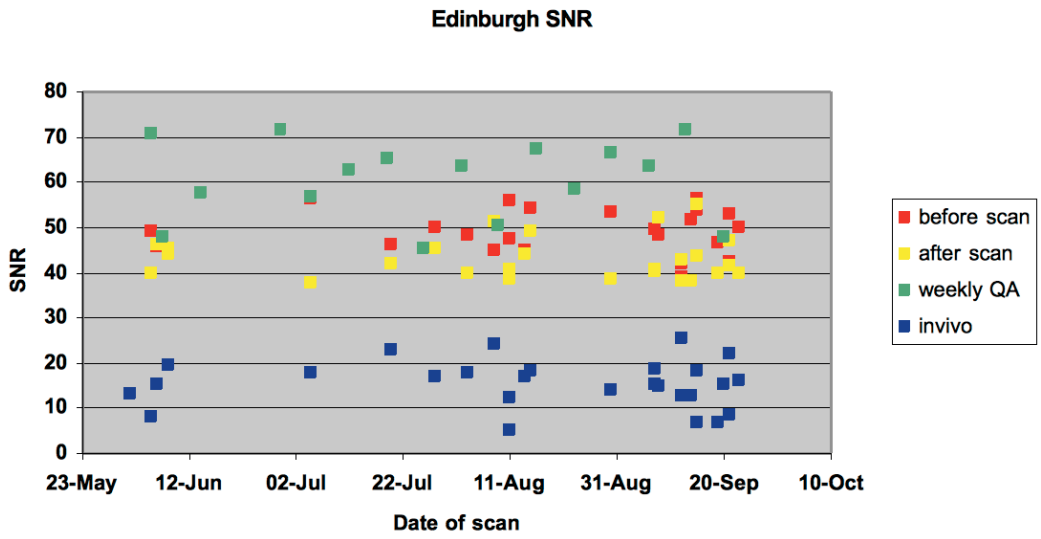

Figure 3.4: SNR results for the Edinburgh scanner, showing in vivo and in vitro data acquired during subject scanning and in vitro data acquired during routine weekly QA.

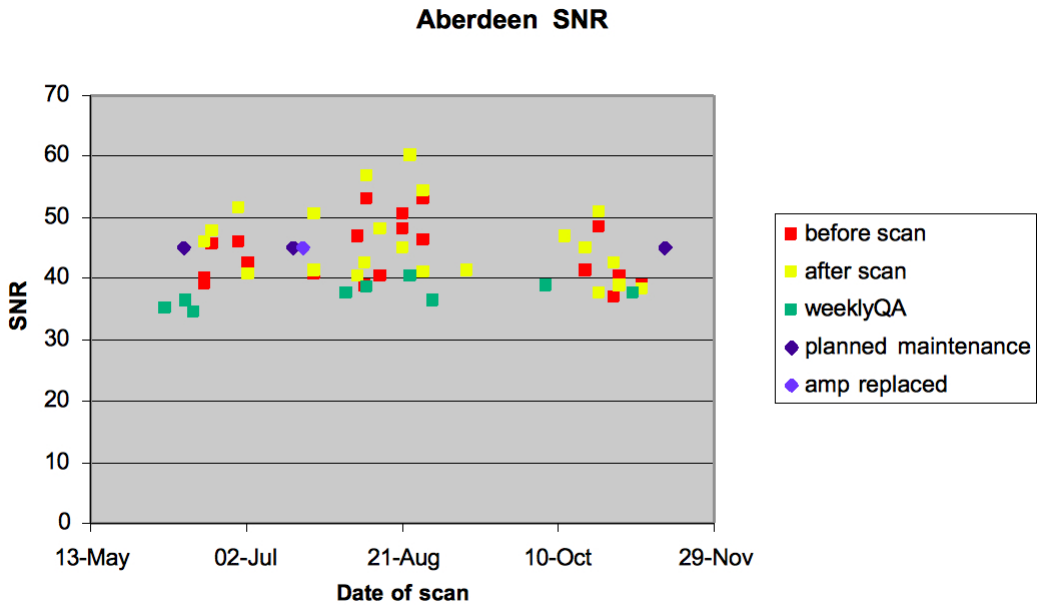

Figure 3.5: SNR measures in Aberdeen in reference to planned maintenance and repair.

3.3.3 Effect of scanner service

To date, only Aberdeen has provided data on the planned maintenance of the MRI scanner during the scanner period for Calibrain. Preliminarily results from the SNR measurement are shown in Figure 3.5 and suggest no observable change even following hardware replacement.

Chapter 4

Structural MRI

4.1 Overview

In this chapter I present the methods and analysis results for structural Magnetic Resonance Imaging (sMRI). Structural imaging data acquired as part of the CaliBrain project was used to examine scanner differences at the three sites and to assess the practicality of pooling scans for multicentre VBM studies. T1-weighted sequences from all sites were analysed to assess within scanner variability and between scanner differences in structure. A metric was developed to correct for between scanner differences by adjusting the probability mappings of tissue priors, used in SPM5 for tissue classification, resulting in scanner specific tissue priors.

Voxel Based Morphometry (VBM) analyses and our metric tests were used to assess how scanner specific priors reduced tissue classification differences between scanners and we compare those remaining differences to within scanner variability, the ideal for scanner harmonisation.

I reviewed existing methods, composed the acquisition protocol, and prepared the raw data for processing. The sMRI data analysis and harmonisation method development was done by T. William Moorhead, the absolute distance metric and sMRI data analysis were developed by both T. William Moorhead and myself. For completeness the full details of the metrics, analysis and methods are given here. This work was published in Moorhead et al. (2009), included in Appendix E.

4.2 Methods: Multicentre Voxel Based Morphometry

As VBM draws its inferences from voxel-wise comparisons it is necessary to apply fine grain corrections of the sMRI tissue classification to avoid any need for validity masking. To compensate for scanner differences separate sets of segmentation priors were developed for each CaliBrain scanner using proportional feedback. These scanner specific priors were created from the first visit scans of six subjects and were tested using our metrics and VBM analyses on the seven subjects who were excluded from the priors adjustment protocol. Metric tests were applied to quantify within scanner variability and between scanner differences. The absolute distance metric assesses the absolute distance between segmentations. The metrics are applied at the voxel level and are averaged to report an overall distance inclusive of noise and systematic differences. These metrics were applied at baseline and on the adjusted segmentations. The method presented here is in keeping with this existing work as corrections were implemented at a scale that is close to the analysis scale for VBM.

4.2.1 VBM Preprocessing and Segmentation

The sMRI data, (T1-weighted images), were pre processed using SPM5 (Wellcome Department of Cognitive Neurology and collaborators, Institute of Neurology, London, UK; http://www.fil.ion.ucl.ac.uk/spm/software/spm5/) running on Matlab Version XXXXXX (Mathworks, Natick, MA, USA).

Prior to segmentation all the scans were co-registered and resliced to ensure that they were aligned to the anterior-posterior commissure axis (ACPC) in the standard MNI template space. As part of this process the scans were re-sampled to a resolution of 1 1 1 mm.

The T1 images were segmented to produce baseline grey and white tissue maps and CSF maps. Creating study specific tissue priors specifically for a study cohort acquired at one scanning centre is an established practice (Wilke et al., 2003; Moorhead et al., 2004; Job et al., 2006; Spencer et al., 2008; Wilke et al., 2008). Here we extend this approach of adjusting the priors so that they compensate for multiple scanners differences. To achieve this goal, an iterative adjustment method that employed proportional feedback to develop scanner specific priors for each scanner, was applied.

The SPM priors used in the baseline tissue classifications were taken from a study of psychosis which employed a scan sequence that was equivalent to that used for the CaliBrain acquisitions (McIntosh et al., 2007). These scanner specific priors in

the psychosis study were drawn from scans of young adults with a family history of schizophrenia and control subjects with no family history of psychosis. All 93 subjects used to create these priors were well at time of scanning.

The final adjusted segmentations were obtained using the scanner specific priors, derived in our priors adjustment procedure, details in the next section. The SPM5 segmentations were run using the default settings, the 'Number of Gaussians per class' was set to [2 2 2] and the 'Bias regularization' was set to 'medium'. In keeping with the established practice in research in normal controls, the segmented results were output as unmodulated and normalized to the MNI template. The normalization employed the SPM5 default normalization with the 'Nonlinear Frequency Cutoff = 25'. Also, in keeping with established VBM practice in tissue density analyses the SPM5 segmentations were smoothed using an isotropic 12 mm Full Width Half Maximum (FWHM) kernel.

4.2.2 Procedure for creating Scanner Specific Priors

Our procedure compensates for scanner differences using proportional feedback to iteratively create sets of scanner specific priors where each set is a small adjustment to the previous segmentation to partially reduce scanner differences.

The process flow diagram in Figure 4.1 gives an overview of this procedure. One scanner is designated as the target scanner and a second as the object scanner. The scans from the target scanner are segmented and for this segmentation the priors were taken from our psychosis study (McIntosh et al., 2007). The priors applied to the target scanner remain unchanged throughout the iterative procedure. The object scanner segmentation is also initialized with the priors taken from our psychosis study (McIntosh et al., 2007). During our iterative procedure these object scanner priors are incrementally adjusted. These adjustments are set to compensate for the segmentation differences between the target and object scanners.

The process illustrated in Figure 4.1 adjusts the object priors through comparisons based upon grey and white segmentations. This dependence of the adjusted priors on both grey and white tissue is implemented by alternating the prime comparison between the grey and white tissue types. When grey is the prime comparison segment we adjust the grey prior to correct for the voxel level differences found between the target and object scanners. Also, at the voxel level, we apply a balancing adjustment to the white or CSF prior to ensure that the sum of the priors at the voxel level is

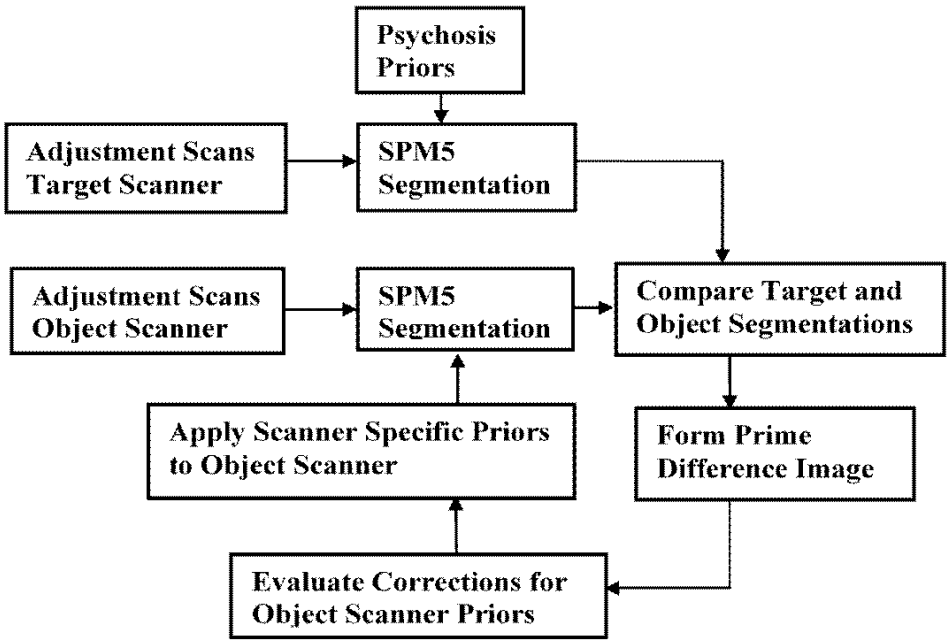

Figure 4.1: Process Flow diagram. Process Flow for procedure that develops scanner specific priors to correct for segmentation differences between the object and target scanners. Adjustment of the object scanner priors is used to minimise the difference between the scanners. The final adjusted object scanner priors are output as the scanner specific priors.

maintained at its nominal sum of unity. Similarly when the prime comparison is made upon the white segment we adjust the white prior to correct for the differences between the target and object scanner and we apply a balancing adjustment to the grey or CSF prior.

On a subject by subject basis the prime comparison segmentations obtained from the target and object scanners are subtracted at the voxel-level. These subtractions were averaged across the subjects included in the priors adjustment process. The averaged voxel-level differences were used to form a difference image that was then smoothed to reduce sampling noise and subject bias. Next a proportion of the smoothed difference image is used to adjust the grey, white and CSF priors applied to the object scanner. These adjusted priors are then used in the next iteration of the procedure.

This process is repeated until the segmentations given by the object scanner converge with those given by the target scanner. We assess this convergence through the use of metrics described below.

The evaluation of the prime difference image, *Pgdiff* for the grey segment G is given in equation (4.1), and *Pwdiff* for the white segment W, is given by equation (4.2). In these calculations of the prime difference images, the processed scans are designated by subscripts (*subject, visit, scanner*), with $N = 6$, the number of compared subjects. These comparisons were limited to the first round scans. The averaging across the adjustment subjects reduces individual differences.

$$Pgdiff = \frac{1}{N}\sum_{n=1}^{N}(G_{(n,visit,Object)}) - (G_{(n,visit,Target)}) \tag{4.1}$$

$$Pwdiff = \frac{1}{N}\sum_{n=1}^{N}(W_{(n,visit,Object)} - (W_{(n,visit,Target)} \tag{4.2}$$

$$adjGPrior = curGprior - sPgdiff * beta \tag{4.3}$$

$$adjWPrior = \begin{Bmatrix} curWPrior + (sPgdiff * beta) & if(curWPrior > curCPrior \\ 0 & if(curCPrior >= curWPrior \end{Bmatrix} \tag{4.4}$$

$$adjCPrior = \begin{Bmatrix} curCPrior + (sPgdiff * beta) & if(curCPrior >= curWPrior \\ 0 & if(curWPrior > curCPrior \end{Bmatrix} \tag{4.5}$$

When the primary segment is grey the adjusted prior *adjGprior* is given by equation (4.3). In this voxel-level process the current grey prior *curGprior* has a proportion

beta of the smoothed prime difference *sPgdiff* image subtracted. When the prime comparison segment is grey the evaluations of the adjusted white *adjWprior* and CSF priors *adjCprior* are given by equations (4.4) and (4.5). In these equations, the changes applied to the grey prior are balanced by equivalent additions to the white or CSF priors. At the voxel level we test the relative occupancy of the white and CSF priors and assign the balancing adjustment to which ever prior exhibits greater occupancy.

When the prime comparison segment is white, the adjusted priors evaluations are equivalent to those given in equations (4.3), (4.4) and (4.5) with the exceptions that the prime difference image is given by *Pwdiff* and the grey and white priors are interchanged. This averaged difference images *Pgdiff* and *Pwdiff* are smoothed using an isotropic kernel, with a FWHM of 10 mm.

Smoothing at this level reduces sampling noise and limits the subject bias that results from the relatively small number of subjects that we have used to create the scanner specific priors.

The value of beta determines the proportion of the difference image that is used to adjust the priors for each iteration of the protocol, and therefore has an important bearing on this protocol. A high setting for beta could lead to instability whilst using value that is too low could result in slow convergence. As part of the development of this method, we experimented with the beta setting and found that setting beta to 0.33 or greater could lead to instability in the priors adjustment process. We found that a beta setting of 0.15 allowed for stable convergence of the segmentations from different scanners. We found that further reductions of the beta value did not improve the degree of convergence that was obtained from the adjustment process. The reductions in beta did increase the number iterations required to attain convergence.

Key to this process is our between scanner distance metric, which assesses the degree of convergence between the target and object scanners. We terminated the adjustment procedure when the incremental change in the between scanner distance was less than 0.1% and remained at this level in subsequent iterations.

4.2.3 Testing Scanner Specific Priors Procedure

We tested the operation of this method, (our priors adjustment procedure), by randomly selecting six subjects from the available data. We applied the scanner specific priors adjustment method to the 1st round scans of these subjects. These priors were then used to segment all the T1 scans for our final analyses. Our metrics and VBM contrast

analyses were used to assess the scanner differences at baseline and for subsequent iterations of partially adjusted segmentations. The seven subjects that were not included in the six subjects from the scanner specific priors adjustment process, formed a test group upon which we could assess the viability of our protocol.

We designated the scanners in the CaliBrain project as scanners A, B and C. Scanner A was set as the target scanner and scanner B as the object scanner and we developed a set of scanner specific priors for scanner B. We also developed scanner specific priors for scanner C with scanner A set as the target scanner.

Throughout these adjustment procedures the priors set used for scanner A was fixed as the priors drawn from our study of psychosis (McIntosh et al., 2007). The scanner specific priors developed for scanners B and C were initialised with the priors from our psychosis study. The choice of scanner A as the target scanner was based upon the baseline metric results that indicated that scanner A has a low within scanner Variability and that it exhibited the lowest overall between scanner differences.

4.2.4 VBM Statistical Analysis

Using SPM5 we implemented VBM statistical analyses of the grey and white segmentations at baseline and for our adjusted segmentations. In these we treated the visits and scanners as separate grouping components and thus formed a factorial analysis matrix that was composed of six groups. We designated the Independence variable as 'NO' to account for the fact that we have repeated measures on the same subjects. We reported the overall F-test for main effect of scanner in the CaliBrain study. Also, we used this design matrix to report t-test contrast results for within scanner variability and between scanner differences.

The t-tests for between scanner differences were made by combining the two visits at each scanner. All t-tests and the F-test were carried out with an uncorrected threshold of 0.001, and we reported Family wise error (FWE) correction for multiple comparisons. All groups were composed of the same subjects and as the scans were all acquired within a six month period there was no requirement to covary for age or gender.

4.2.5 Voxel-wise distance metrics

We employed a percentage distance metric to quantify the within scanner variability and between scanner differences. In SPM, tissue occupancy is assigned at the voxel

level for grey, white and CSF as full occupancy or as partial volumes. At full occupancy the voxel is assigned as either grey or white or CSF with occupancy of 1.0. At the interface between tissue types partial occupancy is assigned on a continuous scale from 0.0 to 1.0 and the sum of the assigned occupancy for each voxel does not exceed 1.0. In order to evaluate the distance between two tissue classifications we computed as a percentage the absolute distance.

The general form of the absolute percentage distance computation is illustrated in equation (4.6) where we compare two voxels V1 and V2. This reports the percentage absolute difference with respect to the average value of the compared voxels. We chose this metric because it accentuates the differences in the compared segmentations.

$$AbsolutePercentageDistance = \frac{200 * |V1 - V2|}{(V1 + V2)} \tag{4.6}$$

In keeping with established VBM analyses the metrics were applied to the smoothed segmentations and limited to valid-voxels where the compared segments had occupancy of greater than 0.05. The summary value reported by the metric is an average of the absolute percentage difference found at the valid voxels in the normalised and smoothed segmentations. The metrics are applied on a subject basis and for each subject we evaluate the within scanner variability for scanners A, B and C and we evaluate the between scanner differences for the scanner pairs AB, AC and BC. A paired sample t-test is used to compare the baseline and adjusted metric results and to report the mean difference and its significance.

Table 4.1: Grey matter metric results for the subjects used in priors generation

Scanner Comparison	Baseline*	Adjusted*	MD** (paired sample significance)
AA	2.1 (0.5)	2.1 (0.5)	0.00 ($p < 1.00$)
BB	3.0 (0.5)	3.0 (0.4)	0.02 ($p < 0.79$)
CC	2.1 (0.6)	2.1 (0.5)	0.02 ($p < 0.36$)
AB	7.2 (0.8)	3.7 (0.4)	3.5 ($p < 0.001$)
BC	8.0 (0.8)	4.0 (0.6)	4.0 ($p < 0.001$)
AC	3.1 (0.4)	2.3 (0.3)	0.8 ($p < 0.001$)

Grey matter metric results averaged over the six subjects used to generate the scanner specific priors. For the within and between scanner comparison both baseline and adjusted absolute percentage distances are recorded. *Absolute Percentage distance % (std dev) **Mean Difference

4.3 Results: Multicentre Voxel Based Morphometry

4.3.1 Metric Results

We applied the percentage distance metric to the grey matter segmentations to obtain measures of within scanner variability and between scanner differences. The metric was applied at baseline and after adjustment using the scanner specific priors. Table 4.1 gives the grey matter metric results averaged across the six subjects used to generate the scanner specific priors. Table 4.2 gives the grey matter metric results averaged across the seven subjects who were excluded from the process that developed the scanner specific priors. In Tables 4.1 and 4.2 we note the mean difference between baseline and adjusted analyses and we give the p value for the paired sample t-test as measure of significance in the adjustment process.

4.3.2 Baseline VBM Results

The F-test for the baseline (unadjusted) grey matter analysis, uncorrected threshold of $p < 0.001$, main effect Maximum Intensity Projection (MIP), is shown in Figure 4.2. This reveals significant scanner effects in the frontal lobes, temporal poles, the thalamus, brain-stem, parietal lobes and occipital lobes. The results of the baseline grey matter F-test are given in Table 4.3. In this we report the significant maximal voxels, their MNI coordinates, and the anatomical location. We also report the Family

Table 4.2: Grey matter metric results

Scanner Comparison	Baseline*	Adjusted*	MD** (paired sample significance)
AA	2.8 (0.6)	2.8 (0.6)	0.00 ($p < 1.00$)
BB	3.4 (0.6)	3.4 (0.6)	0.00 ($p < 1.00$)
CC	2.6 (0.7)	2.5 (0.7)	0.05 ($p < 0.078$)
AB	7.3 (0.6)	5.1 (0.7)	2.2 ($p < 0.001$)
BC	8.2 (0.8)	5.5 (0.8)	2.7 ($p < 0.001$)
AC	4.0 (0.6)	3.6 (0.5)	0.36 ($p < 0.003$)

Grey matter metric results averaged over the seven subjects excluded from the scanner specific priors adjustment procedure. For the within and between scanner comparisons both baseline and adjusted absolute percentage distances are recorded. *Absolute Percentage distance % (std dev) **Mean Difference

Wise Error (FWE) corrected p-value and the extent of the cluster associated with the maximal voxel.

These baseline grey matter differences were investigated by applying t-test contrasts. T-test comparisons of 1st and 2nd round scans of each scanner revealed that there were no significant within scanner differences. T-test comparisons between the scanners revealed that there were no significant differences between scanners A and C. T-test comparisons between scanners B and C were shown to correspond to the differences reported in the F-test for main effect of scanner. T-test comparisons between scanners B and A also corresponded to the differences reported in the F-test for main effect of scanner. The F-test results for the baseline white matter VBM analysis are given in Figure 4.3 and Table 4.4. This illustrates the spatial distribution of the between scanner differences with significant differences in the right middle frontal gyrus and the thalamus. We investigated sources of these baseline white matter differences by applying t-test contrasts. Comparing 1st and 2nd round scans demonstrated that there were no within scanner differences. In the between scanner tests we found no significant differences for the A-C contrasts and we found significant differences in the A-B and B-C contrasts. The B-C white matter differences were more extensive that those found in the A-B contrasts.

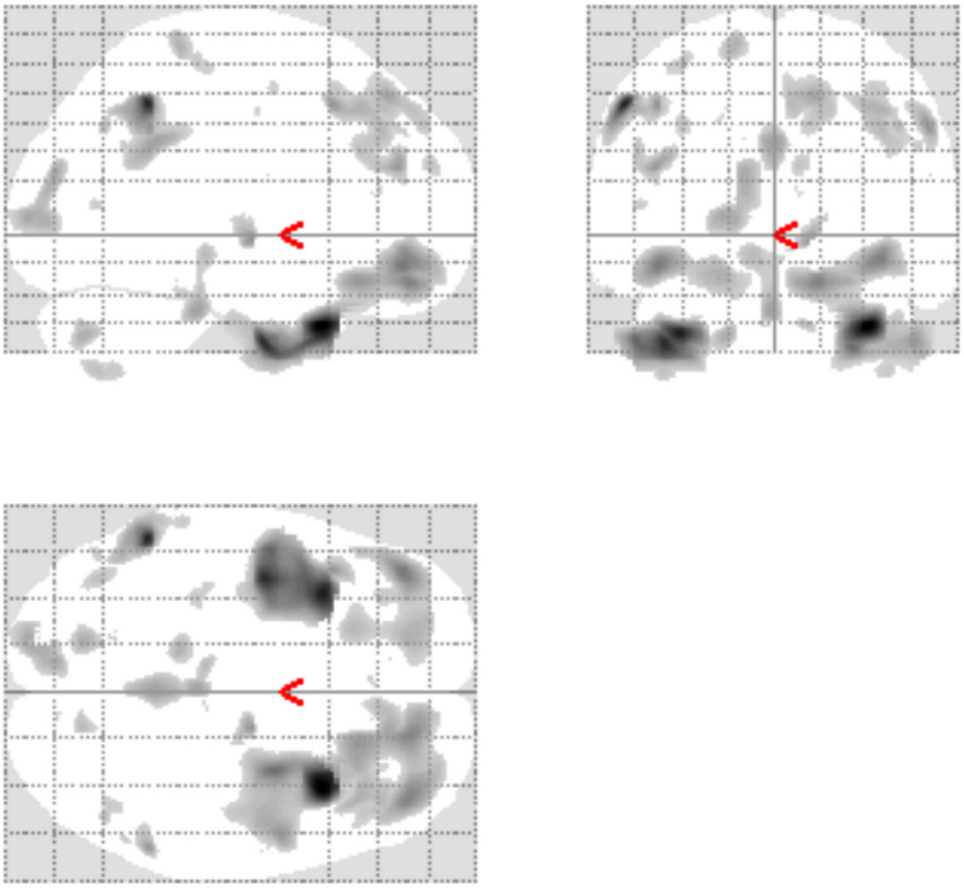

Figure 4.2: Grey Matter Baseline Results. Grey Matter Baseline Maximum Intensity Projection for the CaliBrain Project. Illustrates the regions where the scanners differ when the uncorrected threshold is $p<0.001$.

Table 4.3: VBM Grey matter baseline tests for the effect of scanner

F-test Cluster Anatomical location	Maximal voxel MNI coordinate	FWE p-corrected
Right Temporal Pole	34, 15, -33	0.001
Left Temporal Pole	-35, 16, -36	0.001
Left Inferior Parietal lobule	-55, -48, 47	0.001
Left Inferior frontal gyrus	-42, 45, -11	0.001
Thalamus	13, -10, -1	0.006
Right Inferior Parietal lobule	56, -48, 40	0.009
Left Middle frontal gyrus	-43, 20, 46	0.014
Left Middle frontal gyrus	-44, 44, 24	0.023
Cingulate gyrus	0, 46, 32	0.04
Brain stem	-1, -30, -28	0.045
Right Superior frontal gyrus	15, 40, 49	0.052

VBM Grey matter baseline tests for the effect of scanner. Reporting the extent of the F-test cluster for an uncorrected threshold of $p < 0.001$. Giving the anatomical location of the maximal voxel, the MNI coordinate of this maximal voxel and the p-corrected significance.

Table 4.4: VBM White matter baseline tests for the effect of scanner

F-test Cluster Anatomical location	Maximal voxel MNI coordinate	FWE p-corrected
Right Middle frontal gyrus	18, 45, -19	0.021
Thalamus	15, -10, 0	0.053

VBM White matter baseline tests for the effect of scanner. Reporting the extent of the F-test cluster for an uncorrected threshold of $p < 0.001$. Giving the anatomical location of the maximal voxel, the MNI coordinate of this maximal voxel and the p-corrected significance.

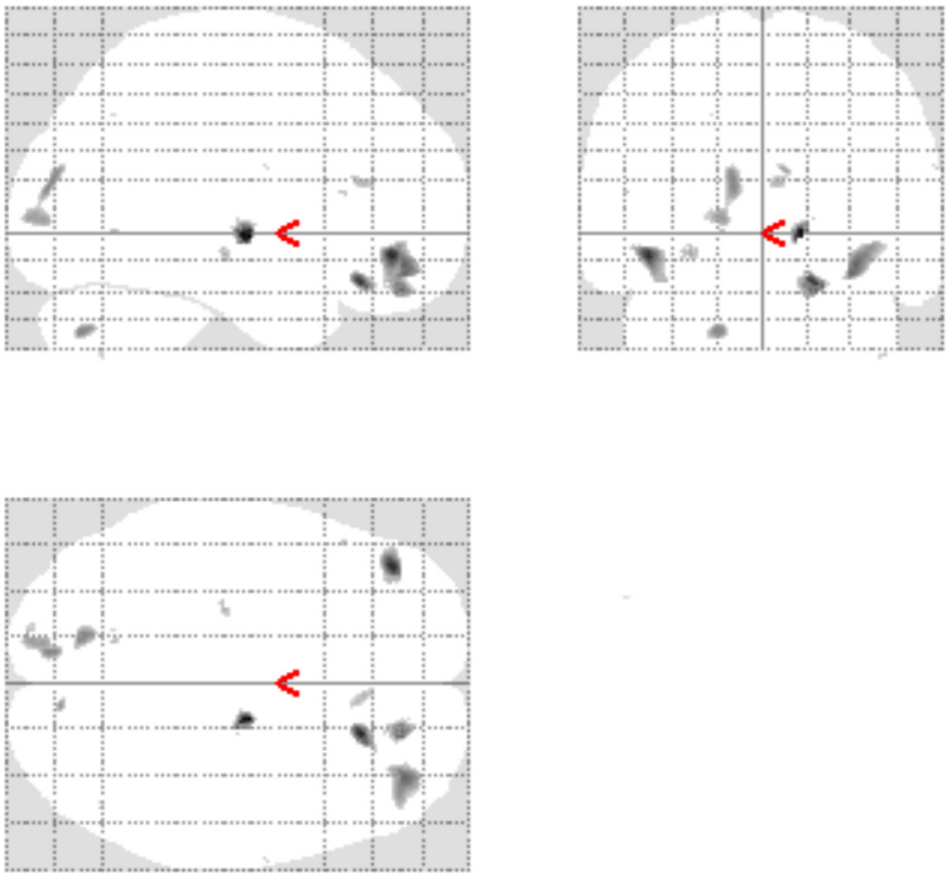

Figure 4.3: White Matter baseline Results. White matter baseline Maximum Intensity Projection for the CaliBrain project, Illustrates the regions where the scanners differ when the uncorrected threshold is $p<0.001$.

4.3.3 Adjusted VBM Results

We applied VBM analyses to the adjusted segmentations obtained from the seven subjects who were excluded from the scanner specific priors development process. We applied the same tests as applied in the baseline tests. In these F-tests for the main effect of scanner we found that there were no significant differences for either the grey or white matter analyses. We repeated the adjusted VBM analyses with all 13 CaliBrain subjects for whom we had complete records and these analyses confirmed that no significant differences remained between the pooled scanners.

Chapter 5

Functional MRI: Motor task

5.1 Overview

In this chapter I present the methods and analysis results for the fMRI motor task. Data was analysed using standard methods for each site and visit separately and then the reproducibility of the resulting activation maps was assessed within and between subjects, visits and sites. A components of variance analysis was performed to determine the contributions to the variance of the factors subject, visit and site.

5.2 Methods: Data analysis

5.2.1 Preprocessing

The fMRI data for the motor task was preprocessed using SPM2 (Wellcome Department of Cognitive Neurology and collaborators, Institute of Neurology, London, UK; http://www.fil.ion.ucl.ac.uk/spm/software/spm2/) running on Matlab Version 6.5 R13 SP1 (Mathworks, Natick, MA, USA).

EPI volumes were realigned to the mean image in the series using a rigid body transformation. This step estimates and attempts to correct subject movement throughout the time-series. The graphical output of the realignment procedure was visually inspected for large spikes or periodic changes. No subjects had to be excluded due to excessive movement ($>$3mm in less than 20 volumes) or large correlations (> 0.5) between movement parameters and task regressors. All subjects exhibited considerably less movement than the pre-established exclusion criteria.

The images were then normalized to the standard SPM2 MNI EPI template. This step is necessary for performing group analyses and reporting coordinates of activation foci in a standardised space. Normalisation parameters were estimated using the mean image for each run and these were applied to all volumes of that run. A linear affine transformation was applied followed by non-linear deformations using the SPM2 default parameters and a custom bounding box for writing the output images (-95:95 -112:105 -80:95 in mm, relative to the anterior commisure) .

Normalized images were spatially smoothed with a 8mm^3 FWHM (full width half maximum) Gaussian kernel. This step helps reduce the effect of noise and alleviate differences arising due to functional and anatomical variations between subjects.

5.2.2 Subject-level statistics

Statistical analysis at the subject level was performed using the General Linear Model as implemented in SPM2, the default settings were used unless otherwise specified. The design matrix included three conditions, random tapping, sequential tapping and rest. The six movement parameters estimated in the realignment step were also included as covariates of no interest. The canonical heamodynamic response function (hrf) was convolved with the regressors to model the data. A high-pass filter with a 180s cut-off was applied to remove low-frequency components. Serial autocorrelations were modelled using AR(1).

A mask image containing only voxels with an intensity at least 80% of the global mean is created by default in SPM to constrain the analysis space. Regions of signal drop-out like the orbitofrontal cortex are thus sometimes excluded from the analysis. Taking into account the added issue of potential differences between sites, it was decided to use an explicit mask. A generous analysis mask was created by thresholding the default SPM EPI template at 0.30. This mask was used in all subject-level analyses and is presented in Figure 5.1.

5.2.3 Data quality assessment

To assess data quality the raw and preprocessed images and the 1st-level maps were inspected visually. Sessions with visible problems or unusual maps were explored further using the ArtRepair tools version 2.1 (http://spnl.stanford.edu/tools/ArtRepair/ArtRepair.htm). All sessions of one subject were excluded due to signal drop-out in the interhemispheric fissure caused by a physiological artefact. One session of another subject was excluded due to an irreparable artefact caused by a scanner instability, which caused an excessive signal increase a few volumes after the start of the scan. Two further sessions contained volumes with less severe artefacts and were repaired using the ArtRepair software. The ArtRepair tool box was used to correct slices or images by averaging the slices/images either side of the problematic slice/image.

5.2.4 Group-level statistics

Contrast images for the sequential tapping versus rest and random tapping versus rest conditions were taken forward to random effects group analyses. The first and second visits to the three scanners were treated separately and six one-sample t-tests were performed using the default SPM settings. A one-sample t-test is used to test whether a population mean is significantly different from some hypothesized value. A tighter mask was created for these analyses by creating a mean image of the smoothed normalised mean images of all sessions and thresholding it by eye to exclude non-brain voxels. To assess within-scanner inter-session differences repeated measures t-tests were performed. All statistical maps were thresholded at a level of $P<0.05$ corrected for multiple comparisons using the False Discovery Rate method (FDR) and using a 20 voxel cluster size threshold.

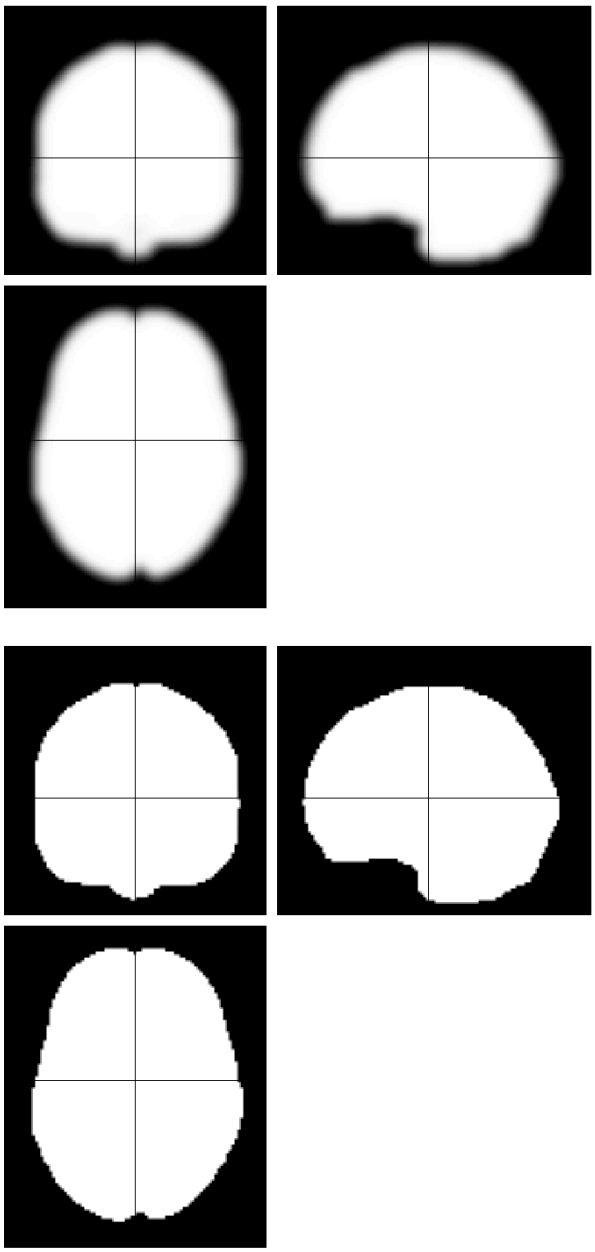

Figure 5.1: Mask comparison. Top: Default SMP mask. Bottom: Custom mask created by thresholding the default SPM EPI template at 0.30.

5.3 Methods: Assessment of scanner effects

5.3.1 Registration evaluation

One issue that was explored was whether systematic differences in registration existed after normalisation between images acquired in the three scanners. In order to investigate this a 'grand mean' image per visit was created from the normalised mean images of all subjects using the SPM ImCalc function. Six difference images were then created using ImCalc by subtracting one image from the other. This was done for each scanner separately comparing the first and second visit and the three paired comparisons between the first visits in Aberdeen and Edinburgh, Edinburgh and Glasgow, Glasgow and Aberdeen. These difference images were inspected visually to locate any obvious areas of misregistration qualitatively.

To assess the degree of overall differences in registration within and between scanner a sum of absolute difference metric was employed. To obtain a summary value the absolute values of all voxels in the difference images were summed and then divided by the total number of voxels. Custom matlab scripts were written to implement the analysis.

This procedure was initially carried out on the original unscaled data to assess global intensity differences. Then images were scaled to the global mean and normalised to 100 using the spm_global function and the procedure was repeated. This was done to reflect the process followed by SPM2 during statistical analysis, where all volumes are scaled to the session mean intensity and normalised to a value of 100.

5.3.2 Reproducibility

To assess the reproducibility of statistical parametric maps an overlap and a size measure were used (Rombouts et al., 1998). The overlap ratio:

$$R^{ij}_{overlap} = 2 * V^{ij}_{overlap} / (V_i + V_j) \tag{5.1}$$

where V_i represents the voxels with t-values exceeding the defined threshold in the statistical map ıi and $V^{ij}_{overlap}$ the intersection of both maps; and the size ratio:

$$R^{ij}_{size} = 2 * V^{ij}_{smallest} / (V_i + V_j) \tag{5.2}$$

where $V^{ij}_{smallest}$ is the smallest of the two volumes compared and V^{ij}_{size} the intersection of both maps. Values for both ratios range between 0 and 1 with 1 indicating

perfect agreement. Statistical parametric maps were thresholded at 0.001 uncorrected and reproducibility was assessed at both the subject and the group level, within and between scanner. Custom matlab scripts were written to implement the analysis.

5.3.3 Voxel-wise Intraclass Correlation Coefficients

Reliability was assessed using approaches similar to those described previously (Specht et al., 2003; Aron et al., 2006). Measures were computed for all within and between scanner comparisons, comparing either the two visits on the same scanner or the first visit on a pair of different scanners. ICC values, representing the ratio of between subject variance to the total variance, were calculated on a voxel-by-voxel basis on contrast images representing sequential versus rest and random tapping versus rest. The resulting 3D ICC maps of ICC values were then masked with a binary image representing the main network of regions robustly activated by the task. The mask was generated by performing a one sample t-test on subjects at visit 1 and at visit 2 on each scanner as described in section 5.2.4. The resulting group activation maps were thresholded at 0.05 (p corrected FDR, cluster extent 20 voxels), and a binary mask was created. ICC maps therefore only present voxels that were activated at all six visits.

5.3.4 Components of variance

To examine the impact of systematic variability contributed by different subjects, visits and scanning sites on the data, a components of variance analysis was performed. The model fits each parameter as a random effect (with a mean and normal distribution) and implicitly assumes that the patients, scanners and visits are taken randomly from a larger population. Although the number of potential scanners is finite in Scotland, whether this effect was modelled as a fixed or random effect had little influence on the solution obtained.

Data for this analysis was extracted from both contrast images and statistical parametric maps in three regions of interest, the primary motor area, the supplementary motor area and the basal ganglia. The images were masked so that only voxels in these areas that were significant in all visits across scanners at $p < 0.001$ uncorrected were included.

Variance components were estimated using the MIVQUE0 method (Hartley et al., 1978) as implemented in SAS PROC VARCOMP (SAS version 9, Cary, NC). The following model was employed:

$$Y_{ijk} = mean + site_j + visit_i + subject_k + subject_k * site_j + unexplained_{ijk} \quad (5.3)$$

with Y_{ijk} denoting the dependent measure for visit i, site j, and subject k. All factors were treated as random effects. More complex models including other interactions between scanners, visits and and subjects were also examined, but their inclusion had very little effect, so the simpler model was preferred. An alternative method of estimation employing a Restricted Maximum Likelihood algorithm was also explored, with similar results.

Reproducibility of the measurements within and between sites was calculated using Intraclass Correlation Coefcients (ICCs):

$$ICC_{within} = (VD_{subject} + VD_{site} + VD_{subjectbysite})/Total\ Variance \quad (5.4)$$

$$ICC_{between} = VD_{subject}/Total\ Variance \quad (5.5)$$

where '$VD_{subject}$' sigifies 'variance due to subject'. While the size and overlap ratios provide useful information on the reproducibility of activation patterns at the whole brain level, this analysis allows us to examine testretest and between site reproducibility and to get more detailed information about the effect of including data from multiple sites.

5.4 Results: Single centre analysis

The sequential tapping versus rest and random tapping versus rest contrasts were examined separately. Data were grouped by scanning site and visit in that site, resulting in six one-sample T-tests per contrast, presented below. The inverse contrasts, rest versus sequential tapping and rest versus random tapping are not presented in detail here. The statistical parametric maps were visually inspected and the results were found to conform to what is predicted by the relevant literature. The within scanner repeated measures analyses yielded no significant results for any of the three scanners.

Thirteen datasets were included in the Aberdeen and Edinburgh analyses and twelve datasets were included in the Glasgow analyses. One Glasgow dataset was not available for the reasons stated in section 5.2.3.

5.4.1 Sequential tapping versus Rest

The sequential tapping versus rest contrast demonstrated activations in areas commonly associated with the task in all analyses, including peaks in the primary and premotor cortices, the superior and inferior parietal lobules, the basal ganglia, and the cerebellum. Details of the results are presented in Tables 5.3 to 5.8.

In the analysis examining the first visit in Aberdeen several activation peaks were noted in a large cluster covering parts of the left precentral, inferior frontal and postcentral gyri and the inferior parietal lobule, including peaks in the primary motor, premotor and supplementary motor areas. Smaller foci of activation were observed in the right precentral and postcentral gyri. Bilateral activations were observed on the superior and inferior banks of the anterior sylvian fissure and the insula. Activation was also observed in the left striatum and thalamus, the right inferior occipital gyrus and the cerebellum bilaterally.

Similar results were obtained in the analysis examining the second visit in Aberdeen. Several activation peaks in the left primary premotor and supplementary motor areas, the inferior frontal gyrus, middle sylvian fissure, the insula, the superior and inferior parietal lobules. Weaker activations were observed in some homologous right hemisphere areas, including the right precentral gyrus, the anterior sylvian fissure, insula and inferior parietal lobule, as well as a right inferior occipital gyrus region. Bilateral activations were also observed in the striatum, the thalamus and the cerebellum.

The analysis of the first Edinburgh visit yielded a large cluster covering parts of the left frontal and parietal lobes, the insula, the basal ganglia and extended to cover

part of the right medial precentral gyrus. Distinct activation peaks in this cluster were observed in the left primary motor and premotor areas, the supplementary motor area bilaterally, the left inferior frontal gyrus, middle sylvian fissure and insula, the left postcentral gyrus and inferior parietal lobule, the left inferior and middle occipital gyri, the left basal ganglia and the left thalamus. In the right hemisphere, weaker activations were observed in the premotor area, the postcentral gyrus and inferior parietal lobule, the sylvian fissure, insula and basal ganglia, and a stronger activation in the inferior and middle occipital gyri. Bilateral activations were also evident in the cerebellum.

The analysis of the second Edinburgh visit demonstrated smaller and more focussed clusters. In the left hemisphere activation peaks were observed in the premotor, primary motor and supplementary motor areas, the inferior parietal lobule and the sylvian fissure. In the right hemisphere peaks were observed in the precentral, inferior frontal, superior temporal and inferior occipital gyri. Clusters were also noted in the left thalamus and bilateral basal ganglia and cerebellum.

In the analysis results of the first Glasgow visit activation peaks were noted in the left premotor, primary motor and supplementary motor areas, the anterior inferior bank of the sylvian fissure, the superior and inferior parietal lobules, the thalamus and the basal ganglia. In the right hemisphere activations were observed in the precentral and postcentral gyri and the inferior parietal lobule. Bilateral activations were also observed in the cerebellum.

No significant voxels were found in the analysis of the second Glasgow visit at the 0.005 FDR corrected threshold. When the threshold was lowered to 0.01 FDR corrected a similar pattern to the other analyses emerged. At the lower threshold activation left-hemispheric peaks were observed in the premotor, primary motor and supplementary motor areas, the anterior cingulate, the postcentral gyrus, the insula and the posterior part of the superior temporal gyrus. In the right hemisphere peaks were observed in the inferior frontal, anterior and middle cingulate, precentral, superior and middle temporal, inferior and middle occipital and lingual gyri. Activations were also observed in the left thalamus, bilateral basal ganglia and bilateral cerebellum.

In summary, the sequential tapping vs. rest contrast demonstrated activations in areas commonly associated with the task in all analyses, including peaks in the precentral and postcentral gyri, the superior and inferior parietal lobules, the basal ganglia and the cerebellum.

5.4.2 Random tapping versus Rest

The random versus rest contrast demonstrated activations in regions commonly associated with this finger tapping task including clusters in the frontal and parietal lobes, the basal ganglia and the cerebellum. Some additional areas in the frontal cortex were observed, not seen in the sequential tapping condition contrast. SPMs of the six group analyses are presented in Figure 5.2. Detailed listings of results are presented in Tables 5.9 to 5.14.

In the analysis examining the first visit in Aberdeen activation peaks were noted in the left middle and inferior frontal, precentral, and postcentral gyri and the inferior parietal lobule, including the primary motor, premotor and supplementary motor areas. Foci of activation were also observed in the right middle frontal, precentral and postcentral gyri. Bilateral activations were observed on the superior and inferior banks of the anterior sylvian fissure and the insula. Activation was also observed in the basal ganglia and the cerebellum bilaterally.

In the analysis of the second visit in Aberdeen several activation peaks were noted in the left primary motor, premotor and supplementary motor areas, the inferior frontal gyrus, middle sylvian fissure and the superior parietal lobules. In the right hemisphere peaks were oserved in the right middle and inferior frontal gyrus and the anterior cingulate, the anterior sylvian fissure, insula, the inferior parietal lobule, as well as the right inferior temporal sulcus. Bilateral activations were also observed in the basal ganglia and the cerebellum.

In the first Edinburgh visit bilateral activation peaks were observed in the middle and inferior frontal gyri, precentral gyrus, superior parietal lobule and the banks of the sylvian fissure. Activation was also noted in the right anterior cingulate cortex. Bilateral activations were evident in the basal ganglia and the cerebellum.

In the analysis of the second Edinburgh visit the clusters were small compared to the other analyses for this contrast, but in the same regions. In the left hemisphere activation peaks were observed in the middle frontal gyrus, the premotor, primary motor and supplementary motor areas, the insula and the sylvian fissure. No significant voxels were observed in the right hemisphere. Bilateral activations were observed in the cerebellum.

In the analysis results of the first Glasgow visit activation peaks were noted in the left middle and inferior frontal gyri, the premotor, primary motor and supplementary motor areas, the banks of the sylvian fissure, the superior and inferior parietal lobules,

the precuneus the thalamus and the basal ganglia. In the right hemisphere activations were observed in the middle and inferior frontal gyri, the precentral and postcentral gyri, the banks of the sylvian fissure, the superior parietal lobule and the basal ganglia. Bilateral activations were also observed in the cerebellum.

In the second Glasgow visit activation left-hemispheric peaks were observed in the premotor, primary motor and supplementary motor areas, the postcentral gyrus, the middle cingulate gyrus, the superior parietal lobule and the left inferior occipital gyrus. In the right hemisphere peaks were observed in the inferior and middle frontal gyri, anterior and middle cingulate, the insula, and the inferior temporal gyrus. Activations were also observed in the left thalamus, bilateral basal ganglia and bilateral cerebellum.

The random vs. rest contrast demonstrated activations in regions associated with this finger tapping task including clusters in the frontal and parietal lobes, the basal ganglia and the cerebellum, demonstrating a very similar pattern to the sequential tapping contrast. Overall, stronger activations were observed in this condition. Some additional areas were also observed, bilaterally in the dorsolateral prefrontal cortex.

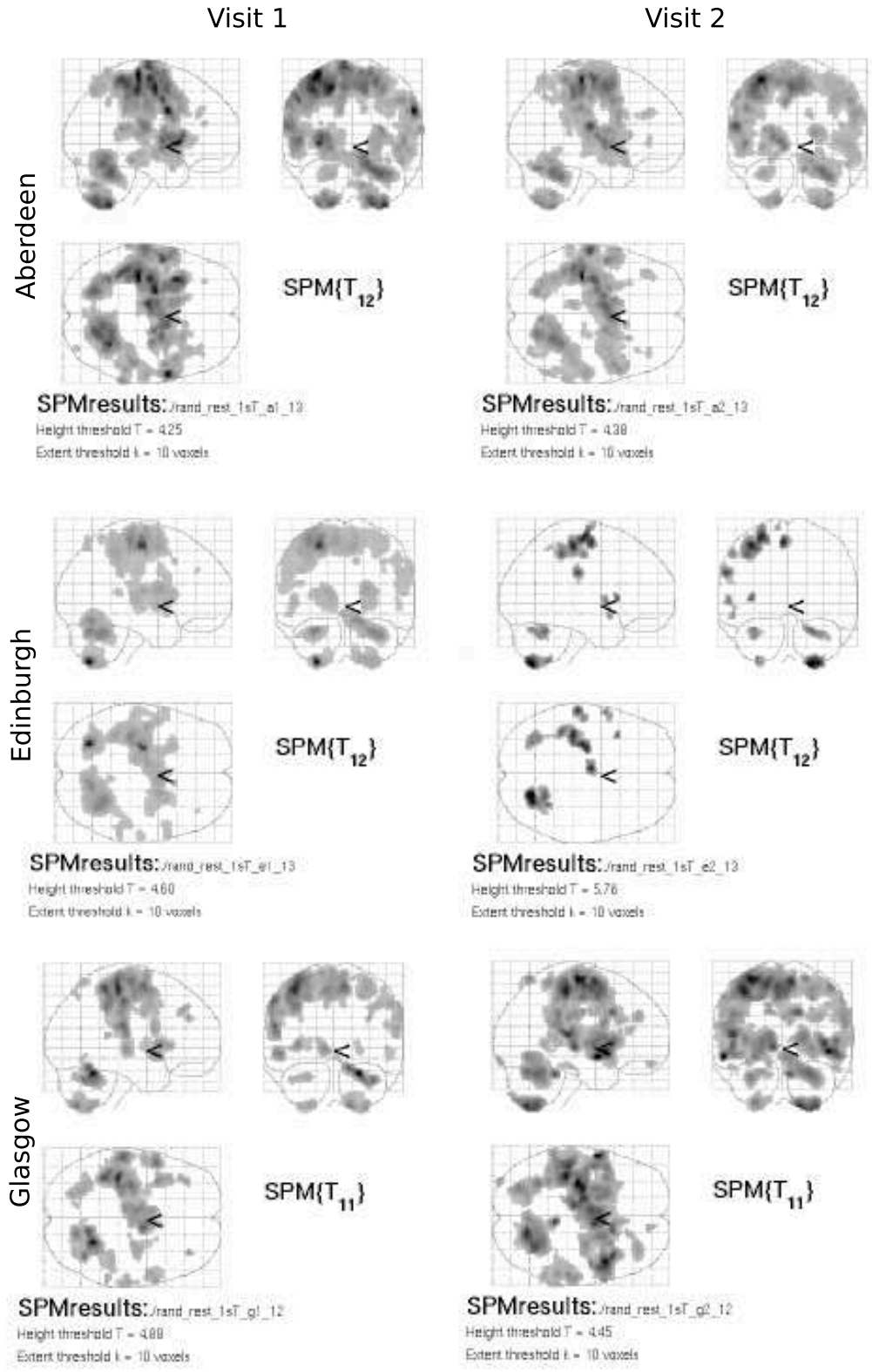

Figure 5.2: Group maps using neurological convention for the random tapping vs. rest contrast, with a threshold at a voxel level of $p<0.0001$ uncorrected and using a 20 voxel cluster extent threshold for illustration purposes.

5.5 Results: Harmonisation evaluation

5.5.1 Registration

Image registration was evaluated qualitatively by visually inspecting mean visit difference images within and between scanner and numerically by a distance metric. The original intensity images were examined first.

Within scanner, in Edinburgh and Glasgow some small differences were observed in areas prone to signal drop-out and distortions like the orbitofrontal cortex. In Aberdeen, a slight displacement was apparent between the first and second visit which manifested as an edge effect around the left and right side of the image.

Between scanner, the most striking difference is between Glasgow and both Edinburgh and Aberdeen, caused by an overall large difference in intensity, masking any potential more subtle effects. Between Aberdeen and Edinburgh, the same displacement is observed as the Aberdeen within scanner comparison.

The comparison between scaled images displays a very similar pattern within scanner. Between scanner, the alleviation of the large difference in signal intensity between the Glasgow scans and the others allows more information to be viewed. First visit Aberdeen images seem to have the same displacement in comparison to Edinburgh and Glasgow, similar to the within scanner result but more pronounced. Differences between Edinburgh and Glasgow are local, most notably in the orbitofrontal cortex and the cerebellum.

Image registration was then evaluated numerically using an absolute difference metric within and between scanner. Comparing the original mean images difference values within scanner were 292.61 for Aberdeen, 360.09 for Edinburgh and 288.8 for Glasgow, with Edinburgh showing the greatest difference and Glasgow the smallest. For scaled data difference values within scanner were 2.41 for Aberdeen, 1.74 for Edinburgh and 3.08 for Glasgow, with Edinburgh showing the smallest difference and Glasgow the greatest.

Between scanner comparisons using the original data yielded difference values of 1360 between Aberdeen and Edinburgh, 7410 between Aberdeen and Glasgow and 6300 between Glasgow and Edinburgh. Glasgow showed a big difference in overall intensity compared to Edinburgh and Aberdeen. For scaled data the difference between Aberdeen and Edinburgh was 7.02, between Aberdeen and Glasgow 6.69 and between Glasgow and Edinburgh 3.68. When global intensities are taken into account Aberdeen shows the greatest difference.

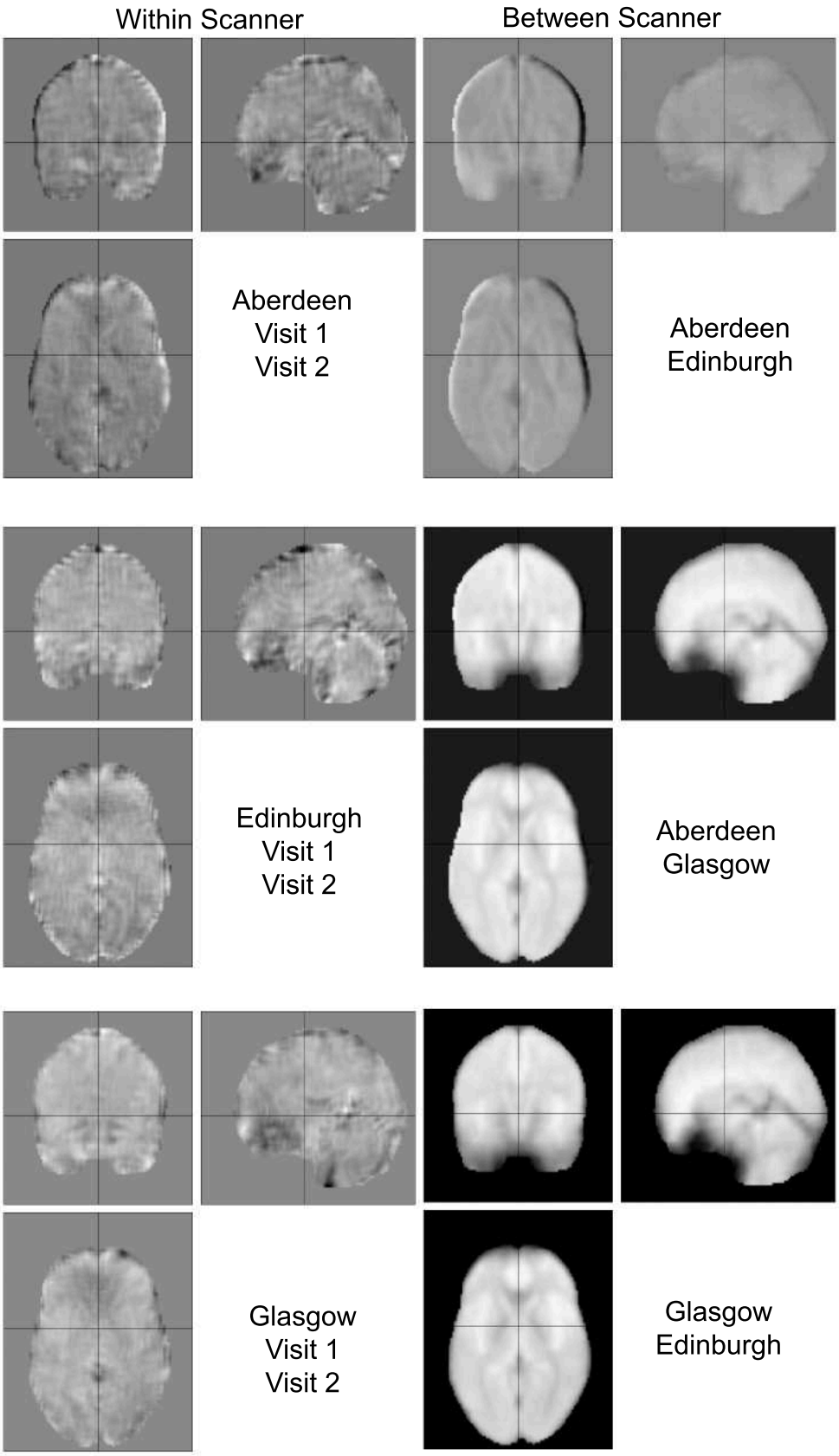

Figure 5.3: Difference images - Original Intensity. Image registration comparison using mean visit difference images within and between scanner.

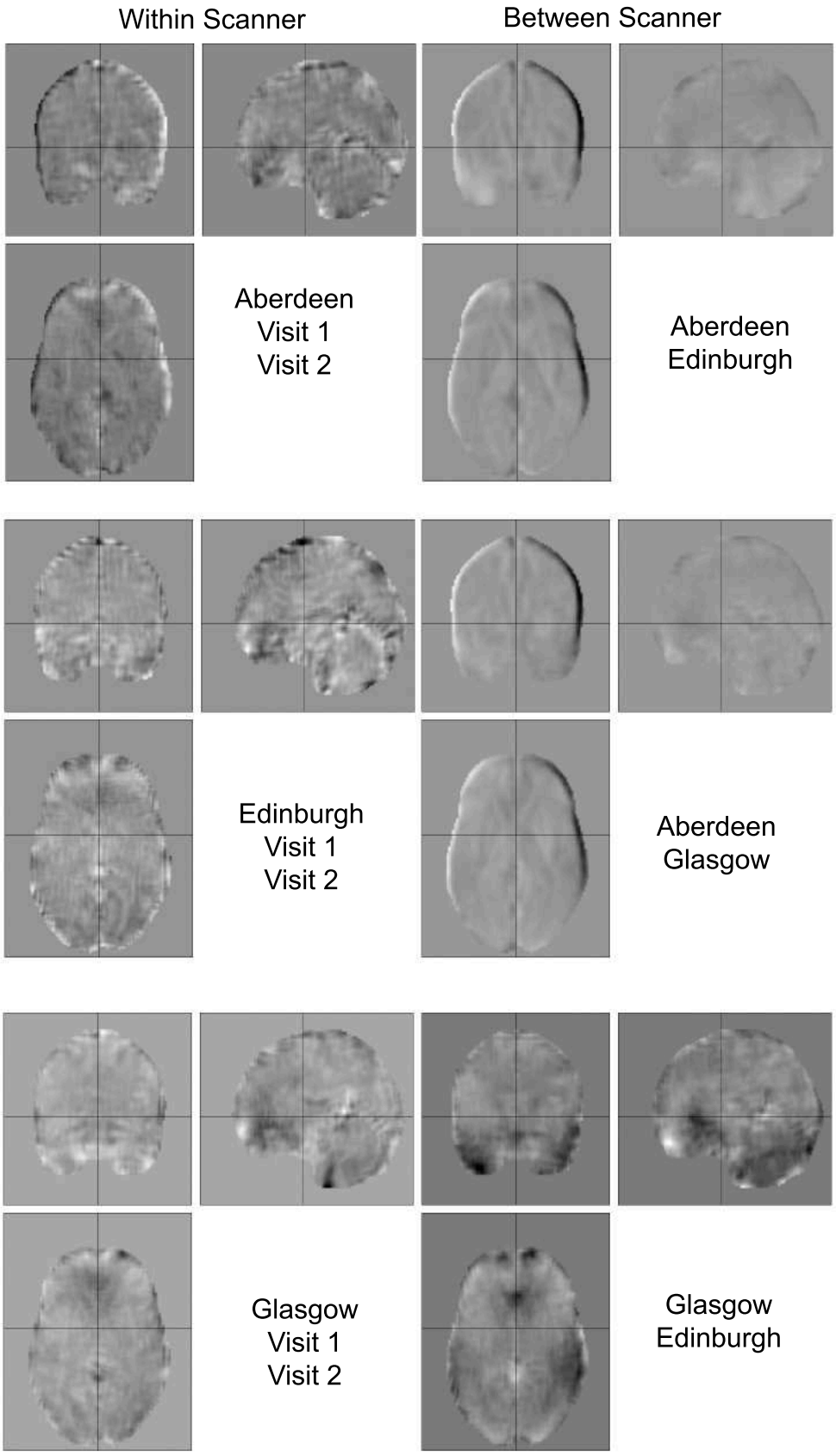

Figure 5.4: Difference images - Scaled Intensity. Image registration comparison using mean visit difference images within and between scanner after scaling.

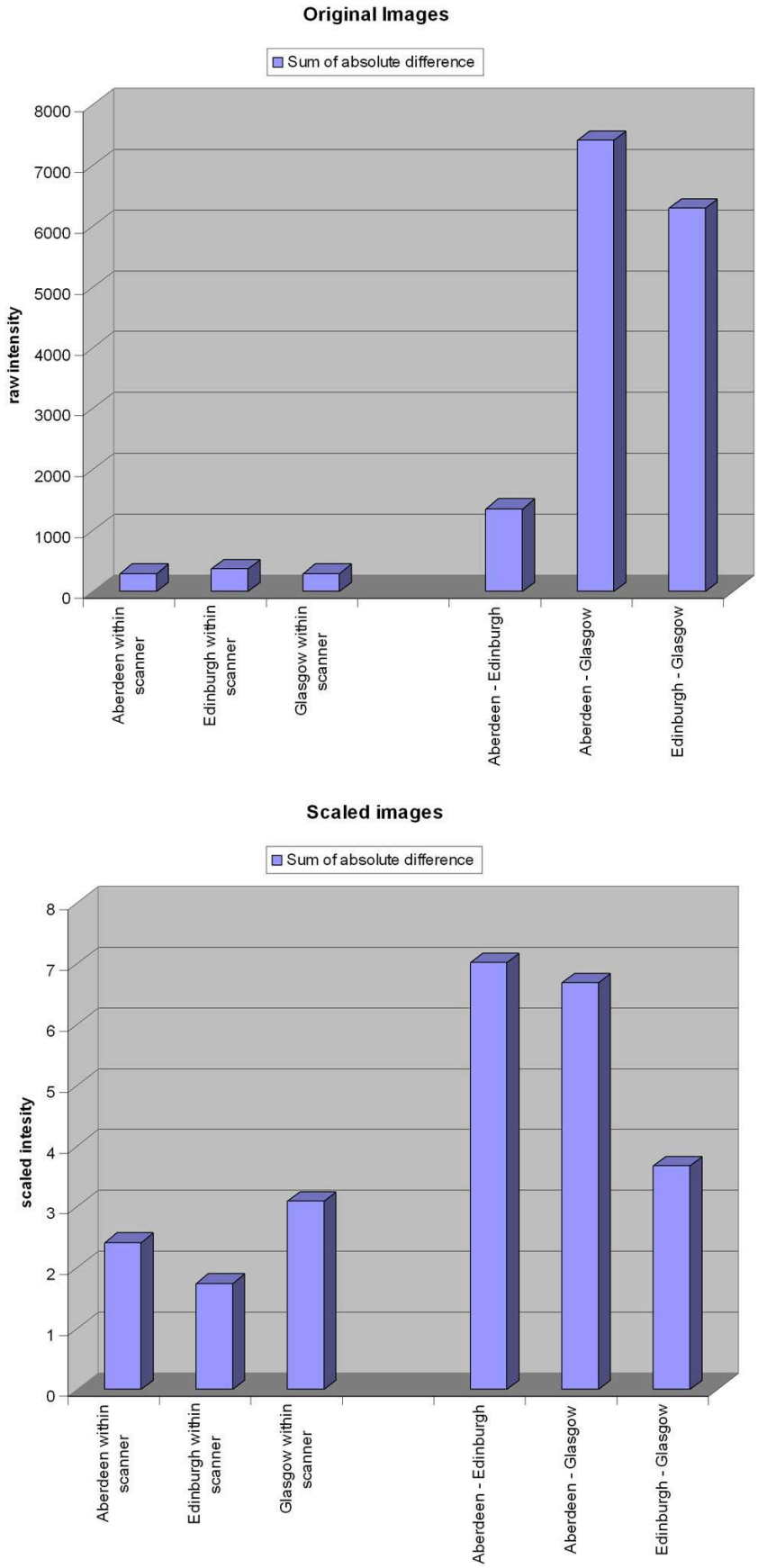

Figure 5.5: Sum of absolute difference. Graphical representation of the absolute difference metric results evaluatiing image registration within and between scanners. Top: original intensity mean images. Bottom: Scaled intensity mean images.

Table 5.1: Sum of absolute difference results

Comparison	Raw Intensity	Scaled Intensity
Aberdeen within scanner	292.61	2.41
Edinburgh within scanner	360.09	1.74
Glasgow within scanner	288.80	3.08
Aberdeen - Edinburgh	1360	7.02
Aberdeen - Glasgow	7410	6.69
Edinburgh - Glasgow	6300	3.68

Image registration was evaluated using an absolute difference metric within and between scanner comparing mean images of both original and scaled intensities.

5.6 Results: Reproducibility

5.6.1 Size and Overlap ratios

Overlap and size ratios within and between sites were similar for all analyses. Mean size ratios ranged from 0.60 to 0.75 within sites and 0.65 to 0.73 between sites for the single subject analyses. Mean overlap ratios for the single subject analyses ranged from 0.41 to 0.50 within sites and from 0.44 to 0.48 between sites. The reproducibility of group maps was generally higher. Size ratios for group maps ranged from 0.73 to 0.97 within sites and from 0.71 to 0.95 between sites. Overlap ratios for group maps ranged from 0.55 to 0.67 within sites and from 0.58 to 0.67 between sites. A summary of this data, with means across subjects, is presented graphically in Figure 5.6.

5.6.2 Voxel-wise Intraclass Correlation Coefficients

ICC maps for the regions activated for the sequential tapping versus rest and random tapping versus rest contrasts within and between scanner are presented in Figure 5.7 for groups of voxels activated at all visits. ICC values for the sensorimotor and supplementary motor cortices were medium to high, while values in the thalamus and basal ganglia ranged from medium-low to negative. Negative ICCs can occur when the between-subject variance is relatively small compared to the within-subject variance.

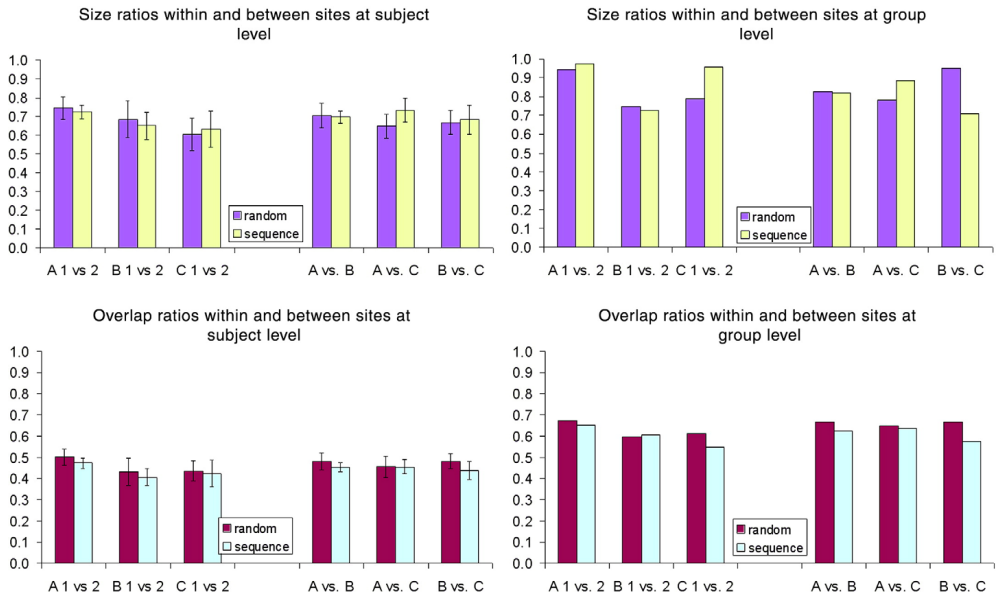

Figure 5.6: Reproducibility of statistical parametric maps within and between sites at subject and group levels, for the random tapping vs. rest and sequential tapping vs. rest contrasts. Left: mean ratios for subjects level analyses with standard error bars. Right: ratios for the group level analyses.

5.6.3 Components of variance

The percentage of total variance in our three regions (significant in all group analyses) contributed by the site component varied between 0% and 13.3% across all analyses, while that of visit varied between 0% and 6.3%. The contribution of the subject by site interaction ranged from 0% to 14.3% while that of subject alone was larger, varying between 15.6% and 61.0% with unexplained variance from 23.0% to 81.0%. The analysis of contrast images yielded ICCs within sites from 0.23 for the striatum to 0.72 for the precentral ROI, and ICCs between sites from 0.23 for the striatum to 0.61 for the precentral ROI. The equivalent analysis of T-statistic images gave lower reproducibility values, within sites from 0.17 for the striatum, to 0.55 for the precentral ROI, and between sites from 0.16 for the striatum, to 0.54 for the precentral ROI. These results are presented in Table 5.2.

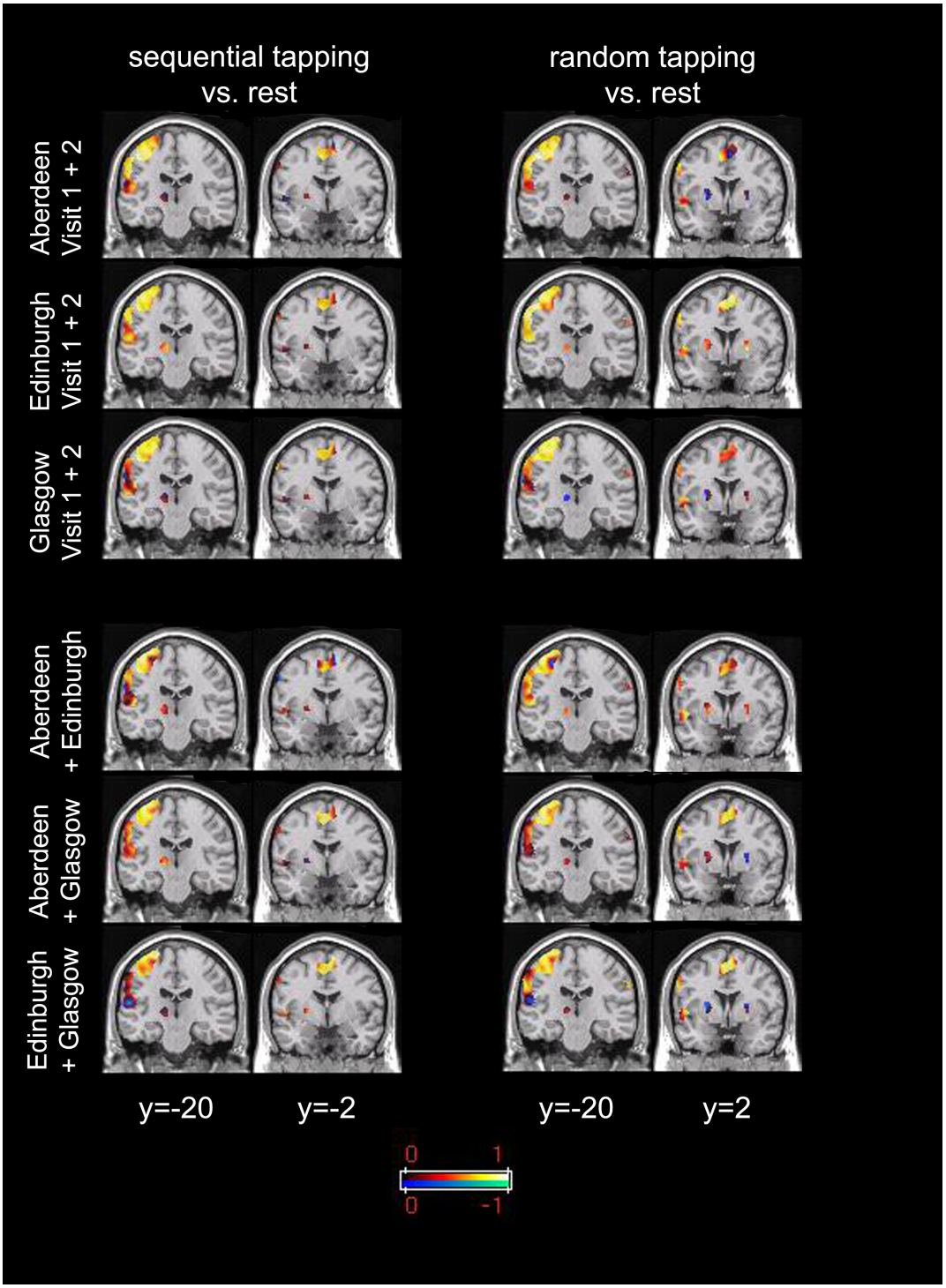

Figure 5.7: Intraclass Correlation Coefficient maps. ICC maps for the regions activated for the sequential tapping versus rest and random tapping versus rest contrasts within and between scanner.

Table 5.2: Percentage of total variance and reproducibility estimates

Anat. area	Site	Visit	Subject	Subject by site	Error	ICC_{within}	$ICC_{between}$
Random vs. rest							
Contrast mean							
Precentral	0.00	4.97	57.76	14.29	22.98	0.72 (g)	0.58 (f)
SMA	13.31	0.00	49.43	0.76	36.50	0.63 (g)	0.49 (f)
Striatum	0.00	3.28	31.15	0.00	65.57	0.31 (p)	0.31 (p)
T-statistic mean							
Precentral	0.00	6.35	50.62	0.00	43.04	0.51 (f)	0.51 (f)
SMA	8.00	0.00	39.12	0.00	52.88	0.47 (f)	0.39 (p)
Striatum	2.71	2.46	15.57	0.00	79.26	0.18 (p)	0.16 (p)
Sequence vs. rest							
Contrast mean							
Precentral	1.39	4.53	60.98	10.11	23.00	0.72 (g)	0.61 (g)
SMA	3.61	3.61	54.22	0.00	38.55	0.58 (f)	0.54 (f)
Striatum	0.00	3.57	23.21	0.00	73.21	0.23 (p)	0.23 (p)
T-statistic mean							
Precentral	0.69	5.89	54.04	0.00	39.38	0.55 (f)	0.54 (f)
SMA	0.00	2.68	41.31	0.00	56.01	0.41 (f)	0.41 (f)
Striatum	0.00	1.81	17.28	0.00	80.91	0.17 (p)	0.17 (p)

Variance components as percentage of total variance contributed by site, visit, subject by site interaction and unexplained variance. Intraclass Correlation Coefficients (ICCs) within and between site. The analysis was run in a region of interest (ROI) in the fusiform gyrus bilaterally in the fearful faces vs. rest contrast images and statistical parametric maps. According to a priori criteria we use ‘p’, ‘f’ and ‘g’ to refer to ‘good’, ‘fair’ and ‘poor’ for the ICCs (Cicchetti, 2001).

Table 5.3: Sequential tapping versus rest Aberdeen 1st visit

Anatomical Area	*Extent*	*Voxel p*	*Voxel T*	*x*	*y*	*z*
frontal lobe:						
L precentral gyrus[b]			12.21	-54	-2	44
R precentral gyrus	222	<0.001	8.61	58	2	34
R precentral gyrus			5.64	60	0	20
R precentral gyrus[c]			5.92	24	0	42
R rolandic operculum[d]			6.62	62	6	8
parietal lobe:						
L postcentral gyrus[b]	6886	<0.001	16.88	-36	-36	46
L postcentral gyrus			15.40	-38	-30	52
R postcentral gyrus	241	<0.001	7.27	60	-20	36
R postcentral gyrus			5.70	52	-20	28
temporal and limbic lobe:						
L temporal operculum	191	<0.001	11.77	-52	-4	6
L insula	31	0.001	6.16	-36	0	10
L insula	34	0.001	5.97	-28	-16	20
L insula			5.27	-34	-18	10
R temporal operculum[d]	295	0.001	6.67	50	0	-2
R temporal operculum			6.02	50	8	-6
R insula	29	0.001	6.38	38	16	6
R insula[c]	44	0.001	6.16	24	-8	34
R superior temporal gyrus	129	0.002	5.74	52	-44	10
R superior temporal gyrus			5.19	60	-42	10
R inferior temporal sulcus	41	0.002	5.78	56	-36	-18
occipital lobe:						
R inferior occipital gyrus	115	<0.001	7.20	26	-102	-8
subcortical structures:						
L globus pallidus	408	<0.001	9.20	-24	-8	-6
L putamen			6.34	-32	-16	-4
L putamen			5.18	-24	4	10
L thalamus	176	0.001	6.19	-16	-26	2
Non-brain voxel			5.31	-12	-20	-8
cerebellum:						
L cerebellum	223	<0.001	10.01	-26	-66	-60
L cerebellum			5.69	-18	-76	-54
L cerebellum	31	0.003	5.29	-24	-60	-24
R cerebelllum	1562	<0.001	14.99	24	-52	-26
R cerebellum			10.37	12	-60	-16
R cerebellum	925	<0.001	13.69	18	-66	-58
R cerebellum			10.70	32	-56	-60
R cerebellum			7.06	26	-48	-56

[a] Subjects included = 13.
[b-d] Part of the same clusters.

Brain activations for the sequential tapping vs. rest contrast: Aberdeen 1st visit. P-values FDR whole brain corrected. MNI space.

Table 5.4: Sequential tapping versus rest Aberdeen 2nd visit[a]

Anatomical Area	*Extent*	*Voxel p*	*Voxel T*	*x*	*y*	*z*
frontal lobe:						
L precentral gyrus	99	0.001	7.14	-60	6	20
L precentral gyrus		0.002	5.64	-60	2	34
L precentral gyrus[b]		<0.001	11.95	-38	-28	62
L precentral gyrus	21	0.002	6.18	-4	-36	78
R rolandic operculum	564	<0.001	10.45	44	2	4
R rolandic operculum		<0.001	8.07	58	8	12
R rolandic operculum		0.001	7.74	34	2	8
parietal lobe:						
L postcentral gyrus[b]	4788	<0.001	34.14	-60	-20	18
L postcentral gyrus		<0.001	11.33	-40	-28	52
L superior parietal lobule	22	0.003	5.3	-36	-60	-58
R inferior parietal lobule	215	0.001	6.82	56	-34	30
R central sulcus		0.002	5.69	56	-20	36
R inferior parietal lobule		0.003	5.47	62	-40	24
R postcentral gyrus	64	0.002	5.93	62	-20	20
temporal lobe:						
L temporal operculum[c]	1299	<0.001	12.32	-50	-4	2
occipital lobe:						
R inferior occipital gyrus	201	<0.001	8.75	32	-98	-10
R inferior occipital gyrus		0.001	7.43	22	-100	-10
R inferior occipital gyrus		0.001	6.52	40	-92	-8
subcortical structures:						
L putamen[c]		<0.001	8.19	-26	-8	12
L putamen[c]		0.001	7.47	-28	-4	-2
cerebellum:						
L cerebellum	269	<0.001	8.94	-36	-54	-30
L cerebellum		0.003	5.27	-28	-66	-56
L cerebellum	28	0.001	6.26	-4	-80	-16
R cerebelum	1237	<0.001	11.92	26	-54	-24
R cerebellum		<0.001	9.27	2	-60	-8
R cerebellum		0.001	7.72	12	-58	-18
R cerebellum	215	0.001	7.08	26	-60	-62
R cerebellum		0.004	5.12	12	-68	-44

[a] Subjects included = 13.
[b,c] Part of the same clusters.

Brain activations for the sequential tapping vs. rest contrast: Aberdeen 2nd visit. P-values FDR whole brain corrected. MNI space.

Table 5.5: Sequential tapping versus rest Edinburgh 1st visit[a]

Anatomical Area	*Extent*	*Voxel p*	*Voxel T*	*x*	*y*	*z*
frontal lobe:						
L precentral gyrus	10780	<0.001	13.91	-34	-28	58
L precentral gyrus		<0.001	11.24	-28	-26	68
L precentral gyrus		<0.001	11.11	-26	-18	54
L rolandic operculum[b]		<0.001	7.16	62	6	14
R precentral gyrus	330	<0.001	8.18	42	-6	56
R precentral gyrus	21	0.002	5.33	58	2	44
parietal lobe:						
R postcentral gyrus	1062	<0.001	6.73	64	-18	28
R inferior parietal lobule[c]		0.001	6.58	48	-36	36
R superior parietal lobule	26	0.002	5.24	36	-54	62
temporal and limbic lobe:						
L temporal operculum	383	<0.001	7.82	-48	-4	2
L temporal operculum[b]	507	<0.001	7.48	62	4	26
L temporal operculum		0.001	6.02	46	2	18
L insula	22	0.001	5.44	-40	6	12
L superior temporal sulcus	25	0.004	4.58	-50	-56	10
R superior temporal gyrus[c]		0.001	6.38	60	-36	12
occipital lobe:						
L middle occipital gyrus	31	0.001	5.6	-24	-104	0
L calcarine sulcus	189	<0.001	7.22	-8	-100	-12
L inferior occipital gyrus		0.002	5.15	-36	-92	-14
subcortical structures:						
R putamen	452	<0.001	7.48	22	4	16
R globus pallidus		<0.001	7.19	18	-2	6
cerebellum:						
L cerebellum	503	<0.001	16.24	-30	-64	-58
L cerebellum		<0.001	6.7	-18	-94	-20
L cerebellum	180	0.001	6.44	-16	-26	-22
L cerebellum		0.001	6.14	-10	-22	-18
L cerebellum	906	<0.001	13.21	-30	-68	-24
R cerebellum	5541	<0.001	12.95	10	-62	-14
R cerebellum		<0.001	11.52	22	-62	-58
R cerebellum		<0.001	10.97	16	-66	-20

[a] Subjects included = 13.
[b,c] Part of the same clusters.

Brain activations for the sequential tapping vs. rest contrast: Edinburgh 1st visit. P-values FDR whole brain corrected. MNI space.

Table 5.6: Sequential tapping versus rest Edinburgh 2nd visit[a]

Anatomical Area	*Extent*	*Voxel p*	*Voxel T*	*x*	*y*	*z*
frontal lobe:						
L central sulcus[b]	3582	0.002	9.27	-52	-14	44
L medial frontal gyrus		0.002	8.98	-8	-8	58
L rolandic operculum[c]		0.003	5.58	-58	4	6
R inferior frontal gyrus	52	0.002	5.98	58	6	34
parietal lobe:						
R postcentral gyrus	99	0.002	8.08	66	-18	22
temporal lobe:						
L temporal operculum[b]		0.002	8.55	-56	-22	8
L temporal operculum[c]	127	0.002	6.27	-50	2	-2
R superior temporal gyrus	57	0.003	5.78	60	-38	12
R superior temporal gyrus		0.004	5.2	68	-38	18
occipital lobe:						
R inferior occipital gyrus	20	0.002	6.16	24	-96	4
subcortical structures:						
L putamen	176	0.002	11.49	-26	2	-8
L thalamus	187	0.002	7.36	-10	-22	2
L thalamus		0.002	5.97	-6	-10	2
L thalamus		0.004	5.25	-16	-26	-6
R putamen	39	0.002	6.12	26	0	6
cerebellum:						
L cerebellum	274	0.002	7.12	-42	-66	-26
L cerebellum		0.002	6.7	-28	-60	-30
L cerebellum		0.003	5.63	-16	-64	-28
L cerebellum	95	0.002	6.39	-30	-64	-56
R cerebellum	1802	0.002	9.84	36	-64	-30
R cerebellum		0.002	9.73	2	-54	-22
R cerebellum		0.002	8.36	24	-52	-28
R cerebellum	506	0.002	8.67	22	-66	-58
R cerebellum		0.003	5.38	30	-52	-56
R cerebellum		0.004	5.12	34	-48	-50

[a] Subjects included = 13.
[b,c] Part of the same clusters.

Brain activations for the sequential tapping vs. rest contrast: Edinburgh 2nd visit. P-values FDR whole brain corrected. MNI space.

Table 5.7: Sequential tapping versus rest Glasgow 1st visit[a]

Anatomical Area	*Extent*	*Voxel p*	*Voxel T*	*x*	*y*	*z*
frontal lobe:						
L postcentral gyrus	3921	0.001	11.93	-38	-30	52
L medial frontal gyrus		0.001	11.44	-10	-16	60
L precentral gyrus		0.001	10.90	-54	-2	32
R precentral gyrus	152	0.001	8.79	58	6	42
R precentral gyrus		0.002	6.45	62	6	30
parietal lobe:						
L superior parietal lobule	50	0.001	7.60	-16	-68	48
L superior parietal lobule		0.002	6.75	-18	-62	42
R postcentral sulcus	32	0.002	6.47	34	-36	36
R postcentral gyrus	78	0.002	6.45	60	-18	32
R inferior parietal lobule	24	0.003	5.92	52	-32	46
subcortical structures:						
L thalamus	49	0.002	6.80	-22	-4	10
L thalamus	339	$<$0.001	16.62	-8	-18	-8
L thalamus		0.002	6.94	-12	-22	6
L putamen		0.002	6.49	-12	-26	18
cerebellum:						
L cerebellum	38	0.002	6.25	-34	-60	-28
L cerebellum	25	0.003	5.94	-24	-62	-56
R cerebellum	1692	0.001	12.09	24	-52	-24
R cerebellum		0.001	12.03	34	-56	-34
R cerebellum		0.001	11.24	14	-64	-26

[a] Subjects included = 12.

Brain activations for the sequential tapping vs. rest contrast: Glasgow 2nd visit. P-values FDR whole brain corrected. MNI space.

Table 5.8: Sequential tapping versus rest Glasgow 2nd visit[a]

Anatomical Area	*Extent*	*Voxel p*	*Voxel T*	*x*	*y*	*z*
frontal lobe:						
L precentral gyrus[b]	4168	0.006	9.73	-30	-16	64
L anterior cingulate gyrus	37	0.006	6.31	-12	34	2
R inferior frontal gyrus	41	0.006	6.12	40	20	4
R precentral gyrus	44	0.006	5.87	58	4	44
R precentral gyrus		0.006	5.65	54	-2	52
R middle cingulate gyrus	69	0.006	7.94	16	16	34
R anterior cingulate gyrus		0.006	5.85	8	14	20
parietal lobe:						
L postcentral gyrus[b]		0.006	8.99	-38	-30	54
L superior parietal lobule	56	0.006	5.72	-36	-48	62
temporal lobe:						
L temporal operculum[b]		0.006	8.34	-42	-38	12
L temporal operculum	102	0.006	5.63	-48	0	2
L temporal operculum		0.009	4.9	-48	-10	0
R superior temporal gyrus	21	0.007	5.47	52	12	-10
R middle temporal gyrus	43	0.006	6.21	68	-40	6
R middle temporal gyrus		0.010	4.71	60	-44	2
occipital lobe:						
R inferior occipital gyrus	329	0.006	8.55	32	-98	-8
R lingual gyrus		0.006	7.94	18	-102	-10
R middle occipital gyrus		0.006	5.96	24	-100	6
subcortical structures:						
L putamen	624	0.006	6.72	-20	-4	8
L putamen		0.006	6.36	-22	2	-10
L putamen		0.006	6.13	-30	10	4
L thalamus	287	0.006	8.24	-14	-24	4
R putamen	133	0.006	5.62	20	-6	12
cerebellum:						
L cerebellum	80	0.006	6.43	-30	-58	-28
L cerebellum	33	0.006	5.51	-26	-68	-56
R cerebellum	296	0.006	7.17	30	-60	-56
R cerebellum		0.006	6.47	16	-66	-58
R cerebellum	856	0.006	7.15	22	-66	-24
R cerebellum		0.006	7.07	22	-52	-24
R cerebellum		0.006	6.76	6	-62	-14

[a] Subjects included = 12.
[b] Part of the same cluster.

Brain activations for the sequential tapping vs. rest contrast: Glasgow 1st visit. P-values FDR whole brain corrected. MNI space.

Table 5.9: Random tapping versus rest Aberdeen 1st visit[a]

Anatomical Area	*Extent*	*Voxel p*	*Voxel T*	*x*	*y*	*z*
frontal lobe:						
L precentral gyrus	14233	<0.001	14.96	-34	-26	68
L superior frontal sulcus		<0.001	13.65	-22	-10	54
L precentral gyrus		<0.001	13.03	-34	-16	64
L superior frontal sulcus	13	0.002	5.06	-30	38	20
R middle frontal gyrus	68	0.002	5.11	36	34	32
R precentral gyrus[b]	3872	<0.001	14.07	60	2	32
parietal lobe:						
R intraparietal sulcus	437	0.001	6.31	20	-68	52
R superior parietal lobule		0.001	6.29	32	-62	56
R inferior occipital gyrus	19	0.002	5.12	34	-86	-16
temporal lobe:						
R temporal operculum		<0.001	8.36	50	0	2
R superior temporal gyrus	329	<0.001	7.97	66	-42	14
R superior temporal gyrus		<0.001	6.60	54	-38	14
R superior temporal sulcus		0.001	5.94	50	-46	8
R middle temporal gyrus	107	<0.001	7.23	48	-34	-16
subcortical structures:						
L caudate nucleus	18	0.003	4.84	-16	22	-2
R putamen[b]		<0.001	8.19	28	-2	14
cerebellum:						
L cerebellum	711	<0.001	10.47	-22	-68	-58
L cerebellum		<0.001	9.42	-34	-56	-58
R cerebellum	3724	<0.001	12.60	30	-56	-60
R cerebellum		<0.001	11.08	18	-66	-58
R cerebellum		<0.001	10.74	26	-54	-28

[a] Subjects included = 13.
[b] Part of the same cluster.

Brain activations for the random tapping vs. rest contrast: Aberdeen 1st visit. P-values FDR whole brain corrected. MNI space.

Table 5.10: Random tapping versus rest Aberdeen 2nd visit[a]

Anatomical Area	*Extent*	*Voxel p*	*Voxel T*	*x*	*y*	*z*
frontal lobe:						
L middle frontal gyrus	102	0.001	6.16	-40	34	30
L central sulcus[b]	11285	<0.001	16.50	-36	-28	62
L precentral gyrus		<0.001	12.10	-32	-16	62
R inferior frontal gyrus	48	0.002	5.13	38	36	2
R inferior frontal sulcus	11	0.003	4.86	36	26	28
R anterior cingulate gyrus	35	0.001	5.88	10	14	30
parietal lobe:						
L postcentral gyrus[b]		<0.001	12.68	-60	-20	16
L superior parietal lobule	25	0.001	6.00	-18	-62	44
R postcentral sulcus	1780	<0.001	7.49	58	-24	32
R postcentral sulcus		0.001	6.90	40	-40	46
R postcentral sulcus		0.001	6.87	46	-36	50
temporal lobe:						
R inferior temporal sulcus	125	0.001	5.79	58	-64	-12
R inferior temporal sulcus		0.002	5.55	56	-52	-8
subcortical structures:						
L globus pallidus	1645	<0.001	10.65	-16	-6	2
L putamen		<0.001	10.29	-24	-12	8
L putamen		<0.001	8.60	-24	0	14
R putamen	553	<0.001	8.37	16	-4	10
R putamen		0.001	6.18	24	16	8
R globus pallidus	56	0.002	5.39	20	-20	-6
R globus pallidus		0.003	4.84	10	-26	-8
R amygdala	188	0.001	6.07	32	0	-18
R amygdala		0.001	6.02	24	-8	-14
R amygdala		0.002	5.66	12	-6	-12
cerebellum						
L cerebellum	342	<0.001	8.83	-20	-72	-56
L cerebellum		<0.001	7.32	-28	-66	-56
L cerebellum	403	0.001	7.17	-34	-54	-28
L cerebellum		0.001	6.22	-30	-62	-24
R cerebellum	1745	<0.001	10.56	26	-56	-28
R cerebellum		<0.001	10.37	30	-48	-32
R cerebellum		<0.001	9.58	18	-72	-24
R cerebellum	645	<0.001	9.22	26	-58	-62
R cerebellum		<0.001	9.00	20	-66	-60
R cerebellum		<0.001	8.10	32	-56	-48

[a] Subjects included = 13.

[b] Part of the same cluster.

Brain activations for the random tapping vs. rest contrast: Aberdeen 2nd visit. P-values FDR whole brain corrected. MNI space.

Table 5.11: Random tapping versus rest Edinburgh 1st visit[a]

Anatomical Area	*Extent*	*Voxel p*	*Voxel T*	*x*	*y*	*z*
frontal lobe:						
L precentral gyrus	6250	<0.001	17.84	-26	-18	56
L precentral gyrus		<0.001	11.79	-34	-28	62
L superior frontal sulcus		<0.001	8.83	-30	-12	72
L central sulcus	84	0.001	5.95	-60	8	26
R anterior cingulate	22	0.001	6.14	10	16	40
R inferior frontal gyrus	201	0.002	5.90	58	8	24
R frontal operculum		0.002	5.46	52	6	16
R middle frontal gyrus	10	0.004	4.91	38	38	32
R central sulcus	85	0.001	5.92	62	-16	32
parietal lobe:						
L superior parietal lobule	49	0.001	6.10	-14	-68	54
L superior parietal lobule		0.003	5.30	-20	-70	64
R supramarginal gyrus	287	0.001	6.33	54	-38	30
R supramarginal gyrus		0.002	5.88	62	-38	32
R postcentral sulcus		0.002	5.64	50	-30	42
R superior parietal lobule	41	0.002	5.63	36	-54	64
temporal lobe:						
L temporal operculum	171	0.001	7.17	-50	-2	2
L superior temporal gyrus		0.002	5.46	-54	8	-10
R superior temporal gyrus	36	0.002	5.61	58	-38	12
R temporal operculum	45	0.003	5.23	52	8	-2
subcortical structures:						
L thalamus	1230	<0.001	8.49	-18	-14	12
L thalamus		0.001	7.64	-14	-22	6
L putamen		0.001	7.09	-26	12	10
R putamen	784	<0.001	8.99	20	2	14
R putamen		<0.001	7.81	28	6	12
R globus pallidus		0.001	7.11	18	-2	-2
cerebellum:						
L cerebellum	422	<0.001	22.58	-28	-68	-56
L cerebellum		0.002	5.48	-16	-68	-46
L cerebellum	623	<0.001	12.91	-30	-66	-26
R cerebellum	1840	<0.001	11.66	28	-52	-28
R cerebellum		<0.001	11.23	12	-66	-16
R cerebellum		0.001	7.53	26	-70	-22
R cerebellum	633	<0.001	8.31	22	-64	-54
R cerebellum		<0.001	7.90	14	-68	-54
R cerebellum		<0.001	7.88	30	-54	-60

[a] Subjects included = 13.

Brain activations for the random tapping vs. rest contrast: Edinburgh 1st visit. P-values FDR whole brain corrected. MNI space.

Table 5.12: Random tapping versus rest Edinburgh 2nd visit[a]

Anatomical Area	*Extent*	*Voxel p*	*Voxel T*	*x*	*y*	*z*
frontal lobe:						
L precentral gyrus[b]	798	0.005	8.06	-26	-16	64
L precentral gyrus		0.005	8.02	-36	-28	56
L middle frontal gyrus	55	0.005	6.81	-54	-18	48
L suppl. motor area	109	0.005	8.05	-4	-10	60
L suppl. motor area	22	0.005	6.63	-10	-10	74
parietal lobe:						
L supramarginal gyrus[b]		0.005	7.41	-40	-38	48
L postcentral gyrus	104	0.005	7.72	-60	-22	30
temporal and limbic lobe:						
L insula	24	0.005	7.45	-36	16	6
L temporal operculum	58	0.005	6.93	-58	4	4
L superior temporal sulcus		0.005	6.58	-54	8	-12
cerebellum:						
L cerebellum	106	0.005	6.85	-34	-66	-26
L cerebellum	68	0.005	6.97	-30	-66	-56
R cerebellum	229	0.005	7.29	22	-54	-26
R cerebellum		0.005	6.66	32	-60	-30
R cerebellum		0.005	6.23	26	-68	-26
R cerebellum	292	0.005	8.56	22	-68	-58
R cerebellum		0.005	7.30	32	-52	-56

[a] Subjects included = 13.
[b] Part of the same cluster.

Brain activations for the random tapping vs. rest contrast: Edinburgh 2nd visit. P-values FDR whole brain corrected. MNI space.

Table 5.13: Random tapping versus rest Glasgow 1st visit[a]

Anatomical Area	*Extent*	*Voxel p*	*Voxel T*	*x*	*y*	*z*
frontal lobe:						
L superior frontal sulcus	86	0.001	7.52	-34	38	40
L middle frontal gyrus		0.003	5.45	-32	40	28
L inferior frontal gyrus		0.001	7.63	-38	22	2
L central sulcus	334	0.001	7.11	-58	6	22
L precentral gyrus		0.001	7.03	-56	-4	34
L precentral gyrus		0.002	6.32	-56	6	38
R superior frontal sulcus	25	0.001	8.06	34	40	30
R superior frontal gyrus	10	0.002	5.87	18	-16	68
R precentral gyrus	393	0.001	7.45	30	-10	60
R precentral gyrus		0.001	7.14	26	-12	52
R central sulcus		0.001	6.64	36	-12	52
R inferior frontal gyrus	345	0.001	7.38	60	8	18
R precentral gyrus		0.001	7.33	54	4	44
R precentral gyrus		0.003	5.49	60	10	36
parietal lobe:						
L postcentral gyrus	5725	<0.001	13.73	-38	-28	58
L postcentral gyrus		<0.001	12.11	-38	-42	58
L postcentral gyrus		0.001	10.88	-58	-22	22
L precuneus	32	0.002	6.06	-10	-66	50
L superior parietal lobule		0.003	5.40	-14	-72	54
L superior parietal lobule		0.004	5.00	-20	-68	58
R postcentral gyrus	177	0.001	7.73	64	-18	42
R postcentral sulcus		0.001	6.97	66	-22	30
R postcentral gyrus		0.002	6.41	58	-20	28
R postcentral sulcus	267	0.001	6.59	44	-34	42
R supramarginal gyrus		0.003	5.42	48	-36	56
R superior parietal lobule	61	0.002	6.08	28	-60	64
R superior parietal lobule		0.004	5.06	22	-68	60
temporal and limbic lobe:						
L temporal operculum	571	0.001	9.62	-52	8	-6
L insula		0.001	8.86	-36	12	4
R superior temporal gyrus	26	0.001	6.74	60	-38	12
R temporal operculum	54	0.003	5.67	52	12	-2
subcortical structures:						
L thalamus	219	0.001	8.31	-10	-20	0
R putamen	54	0.003	5.69	26	2	-2
R putamen		0.004	5.05	22	6	6
cerebellum:						
L cerebellum	176	0.001	7.45	-26	-64	-28
L cerebellum		0.002	6.19	-38	-58	-30
L cerebellum	269	0.001	7.42	-22	-74	-56
L cerebellum		0.001	6.72	-32	-58	-56
R cerebellum	842	<0.001	15.49	24	-52	-24
R cerebellum		0.001	10.94	16	-56	-20
R cerebellum		0.001	10.39	38	-60	-32
R cerebellum	596	0.001	9.49	16	-68	-56
R cerebellum		0.001	7.66	30	-54	-58

[a] Subjects included = 12.

Brain activations for the random tapping vs. rest contrast: Glasgow 1st visit. P-values FDR

Table 5.14: Random tapping versus rest Glasgow 2nd visit[a]

Anatomical Area	*Extent*	*Voxel p*	*Voxel T*	*x*	*y*	*z*
frontal lobe:						
L precentral gyrus[b]		0.001	11.93	-32	-18	66
R superior frontal gyrus	23	0.002	5.82	12	52	-18
R middle frontal gyrus	217	0.001	6.27	36	42	30
R inferior frontal gyrus	21	0.001	6.22	28	14	36
L middle cingulate gyrus	23	0.001	6.61	-14	-30	38
L middle cingulate gyrus	89	0.001	6.69	-16	-14	34
R middle cingulate gyrus		0.001	6.37	2	-16	30
R anterior cingulate gyrus	35	0.002	5.60	18	32	16
R middle cingulate gyrus	31	0.002	5.75	10	-4	34
temporal and limbic lobe:						
L inferior temporal gyrus	27	0.002	5.41	-50	-22	-20
L inferior temporal gyrus		0.003	4.88	-44	-18	-24
L superior temporal sulcus[b]	15536	0.001	12.22	-58	2	-10
R insula		0.001	11.43	44	10	-6
R inferior temporal gyrus	50	0.002	5.93	52	-24	-20
R inferior temporal gyrus		0.003	5.12	50	-34	-20
parietal lobe:						
L superior parietal lobule	118	0.001	7.53	22	-64	66
R supramarginal gyrus	1109	0.001	7.39	58	-26	28
R central sulcus		0.001	6.90	62	-20	32
R supramarginal gyrus		0.001	6.58	56	-36	34
L superior parietal lobule	34	0.003	5.14	-22	-66	66
L superior parietal lobule		0.004	4.75	-18	-72	62
occipital lobe:						
L inferior occipital gyrus	28	0.001	6.26	-34	-94	-14
cerebellum:						
L cerebellum	462	0.001	7.62	-32	-56	-30
L cerebellum		0.001	6.69	-22	-68	-26
L cerebellum		0.001	6.02	-14	-66	-28
L cerebellum	373	0.001	7.59	-28	-76	-54
L cerebellum		0.002	5.95	-26	-82	-48
L cerebellum		0.002	5.31	-14	-72	-50
R cerebellum	673	0.001	10.93	16	-66	-58
R cerebellum		0.001	9.52	30	-60	-60
R cerebellum		0.002	5.28	10	-74	-52
R cerebellum	1850	0.001	9.90	6	-40	-26
R cerebellum		0.001	8.77	26	-56	-26
R cerebellum		0.001	8.55	22	-66	-26

[a] Subjects included = 12.
[b] Part of the same cluster.

Brain activations for the random tapping vs. rest contrast: Glasgow 2nd visit. P-values FDR whole brain corrected. MNI space.

Chapter 6

Functional MRI: Face processing

6.1 Overview

In this chapter the data for the face processing fMRI task is presented. Data was analysed using the methods given in chapter 5 for the motor task. Each visit was examined separately and reproducibility was assessed within and between subjects, visits and sites. In addition, an analysis was performed to examine habituation effects in the amygdala.

6.2 Methods: Data analysis

6.2.1 Preprocessing

The fMRI data for the faces task was preprocessed using SPM8 running on Matlab Version 7. EPI volumes were realigned to the mean image in the series using a rigid body transformation. The graphical output of the realignment procedure was visually inspected for large spikes or periodic changes. No subjects had to be excluded due to excessive movement ($> 3mm$ in less than 20 volumes). All subjects exhibited considerably less movement than the pre-established exclusion criteria. The images were then normalised using the coregistered anatomical image to determine transformation parameters. Normalised images were spatially smoothed with a $8mm^3$ FWHM (full width half maximum) Gaussian kernel.

6.2.2 Subject-level statistics

Statistical analysis at the subject level was performed using the General Linear Model as implemented in SPM8, the default settings were used unless otherwise specified. The design matrix included three conditions, fearful faces, neutral faces and rest. The six movement parameters estimated in the realignment step were also included as covariates of no interest. The canonical haemodynamic response function (hrf) was convolved with the regressors to model the data. A high-pass filter with a 180s cut-off was applied to remove low-frequency components. Serial autocorrelations were modelled using AR(1).

6.2.3 Data Quality Assessment

Data for four sessions of different subjects were not available due to scanner problems. Data for eighty sessions total were included in the faces task analysis. To assess data quality the raw and preprocessed images and the 1st-level maps were inspected visually. Sessions with visible problems or unusual maps were explored further using the ArtRepair tools. Three sessions of different subjects were excluded due to large scanner artefacts.

6.2.4 Group-level analyses

Data for 77 sessions total were included in the group level analyses, 14 in both visits in Aberdeen, 13 in each Edinburgh visit, 12 repeated, 11 in the first Glasgow visit and 12 in the second, 9 repeated.

Contrast images for the fearful faces versus rest condition were taken forward to random effects group analyses. The first and second visits to the three scanners were treated separately and six one-sample t-tests were performed using the default SPM settings. A one-sample t-test is used to test whether a population mean is significantly different from some hypothesised value. All statistical maps were thresholded at a level of $p < 0.001$ uncorrected for multiple comparisons and using a 20 voxel cluster size threshold. Clusters were deemed significant at a level of $p < 0.05$ corrected for multiple comparisons using the False Discovery Rate method (FDR). Small volume corrections were applied in three a priori areas, reported by Fusar-Poli et al. (2009) to be activated in this contrast, bilateral amygdala, bilateral fusiform gyrus and right medial frontal cortex. Masks for these analyses were created using the WFU PickAtlas software and results were deemed significant at $p < 0.05$ corrected for multiple comparisons using the Family-Wise Error method (FWE). Analyses were also performed the same way on the data from the first visit in each scanner examining habituation effects in the amygdala between runs in the fearful faces condition with the method used in Hall et al. (2008) (Figure 6.1).

6.2.5 Reproducibility

To assess the reproducibility of statistical parametric maps an overlap and a size measure were used (Rombouts et al., 1998). The overlap ratio:

$$R^{ij}_{overlap} = 2 * V^{ij}_{overlap} / (V_i + V_j) \tag{6.1}$$

where V_i represents the voxels with t-values exceeding the defined threshold in the statistical map ıi and $V^{ij}_{overlap}$ the intersection of both maps; and the size ratio:

$$R^{ij}_{size} = 2 * V^{ij}_{smallest} / (V_i + V_j) \tag{6.2}$$

where $V^{ij}_{smallest}$ is the smallest of the two volumes compared and V^{ij}_{size} the intersection of both maps. Values for both ratios range between 0 and 1 with 1 indicating perfect agreement. Statistical parametric maps were thresholded at 0.001 uncorrected

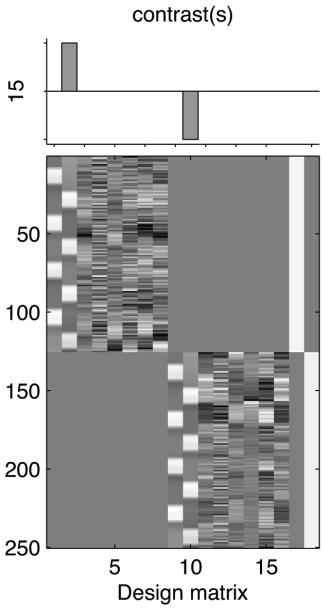

Figure 6.1: Design matrix for the habituation analysis, contrasting the first versus the second run for the fear versus rest contrast.

and reproducibility was assessed at both the subject and the group level, within and between scanner. Custom matlab scripts were written to implement the analysis.

6.2.6 Analysis of Variance

A components of variance analysis was also performed, as exemplified in section 5.3.4. The model fits each parameter as a random effect (with a mean and normal distribution) and implicitly assumes that the patients, scanners and visits are taken randomly from a larger population.

Data for this analysis was extracted from both contrast images and statistical parametric maps in the only a priori area (Fusar-Poli et al., 2009) where significant activation was observed consistently, bilateral fusiform gyrus. The images were masked so that only voxels in these areas that were significant in any visits across scanners at $p < 0.001$ uncorrected were included.

Variance components were estimated using the Restricted Maximum Likelihood method as implemented in SAS PROC VARCOMP (SAS version 9, Cary, NC). The following model was employed:

$$Y_{ijk} = mean + site_j + visit_i + subject_k + subject_k * site_j + unexplained_{ijk} \quad (6.3)$$

with Y_{ijkl} denoting the dependent measure for visit i, site j, and subject k. All

factors were treated as random effects.

Reproducibility of the measurements within and between sites was calculated using Intraclass Correlation Coefcients (ICCs):

$$ICC_{within} = (VD_{subject} + VD_{site} + VD_{subjectbysite})/Total\ Variance \tag{6.4}$$

$$ICC_{between} = VD_{subject}/Total\ Variance \tag{6.5}$$

where '$VD_{subject}$' sigifies 'variance due to subject'.

6.3 Results

6.3.1 Single visit group analysis

Activations were detected in expected areas in most groups in response to the presentation of fearful faces. In the first visit across scanners extensive activations were detected in the occipital lobe bilaterally, also the fusiform gyrus bilaterally, the striatum bilaterally, left hemisphere motor areas, and the medial frontal gyrus among others.

Reduced activation was observed in the second visits across scanners, especially in the fontal region, possibly suggesting learning effects. Consistent significant activations were still present in the occipital lobe and the fusiform gyrus bilaterally and in the left hemisphere motor areas. Statistical maps of the analyses are presented in Figure 6.2, detailed listings of results are presented in Tables 6.2 to 6.7.

After small volume corrections were applied, significant clusters were detected for the amygdala bilaterally in all sites and visits except for the Edinburgh second visit results, where the right amygdala cluster did not reach significance. In the second visit significant activation was detected only in the right amygdala in the Aberdeen analysis. For the fusiform gyrus, significant clusters were detected bilaterally in both visits across all scanners. For the right medial frontal gyrus, significant clusters were detected in the first visit only across all scanners.

None of the habituation analyses in the first visit groups for the three scanners yielded significant results.

6.3.2 Reproducibility

Mean size ratios for the fearful faces contrast were at similar levels within and between sites. The mean size ratios for the single subject analyses ranged from 0.63 to 0.79 within sites and from 0.6 to 0.8 between sites. Mean overlap ratios were lower for the within site comparisons, ranging from 0.52 to 0.61, while between sites they were in the range of 0.58 to 0.73. For the group analyses size ratios ranged within sites from 0.5 to 0.68 and between sites from 0.68 to 0.89. Group level overlap ratios were similar, with a range of 0.45 to 0.61 within sites and 0.6 to 0.62 between sites. The data is presented graphically in Figure 6.3.

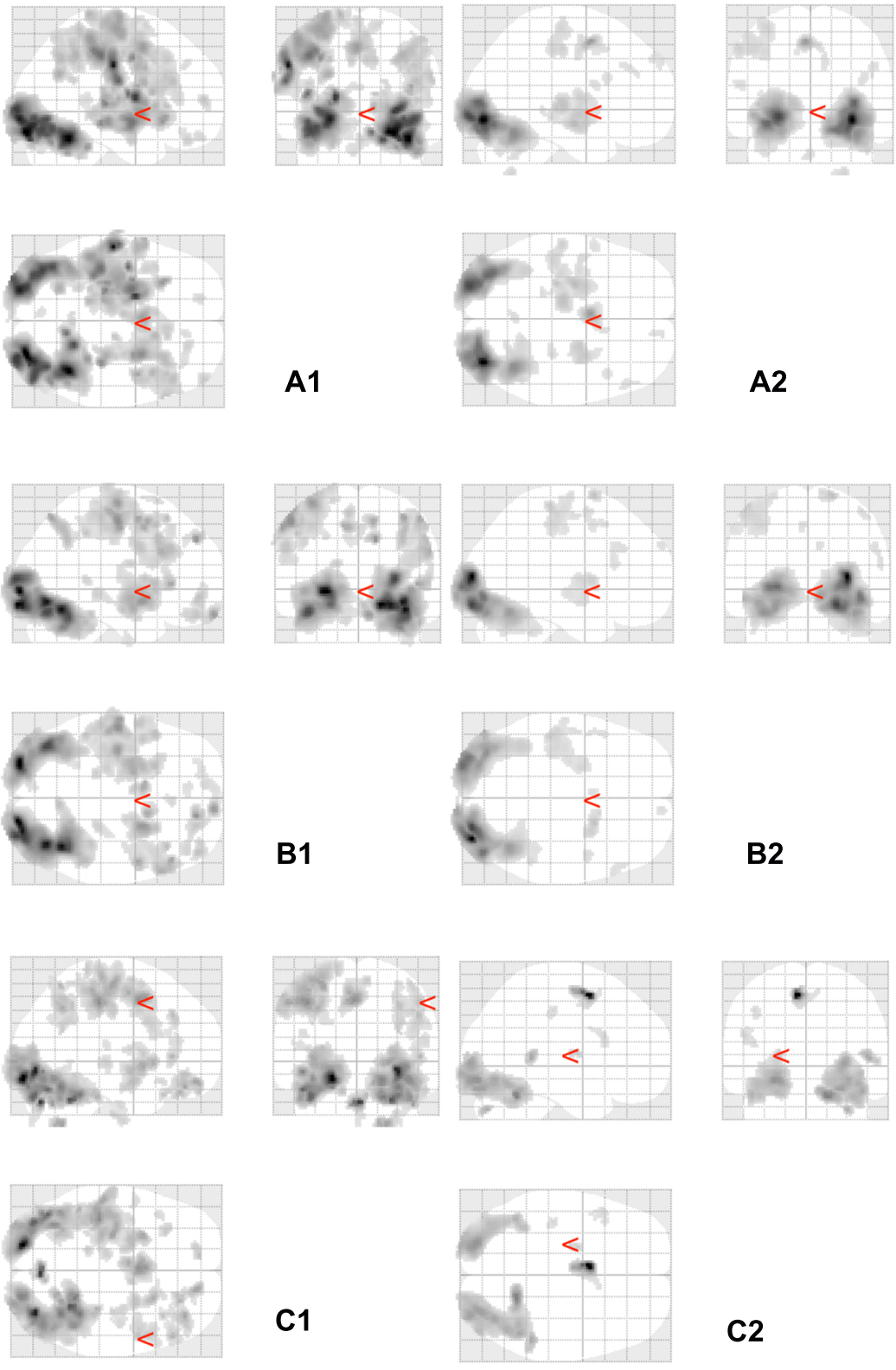

Figure 6.2: Group maps using neurological convention for the fearful faces vs. rest contrast. Threshold at a voxel level $p < 0.001$ uncorrected and 20 voxel cluster extent. From top to bottom: sites A=Aberdeen, B=Edinburgh and C=Glasgow. Left: visit 1. Right: visit 2.

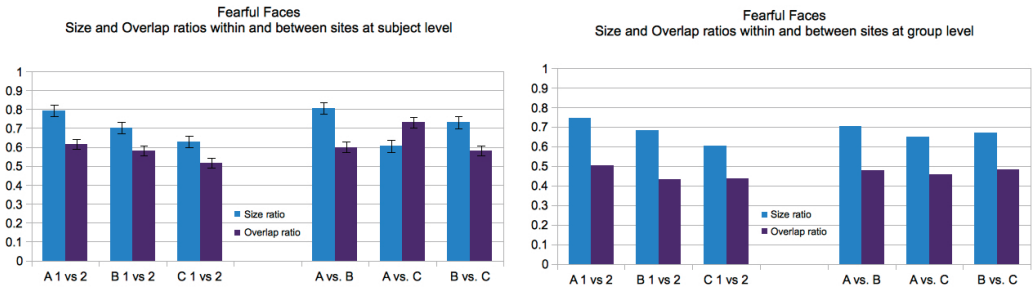

Figure 6.3: Reproducibility of statistical parametric maps within and between sites at subject and group levels, for fearful faces vs. rest contrast. Left: mean ratios for subjects level analyses with standard error bars. Right: ratios for the group level analyses.

6.3.3 Components of variance

The percentage of total variance contributed by the site component in the region of interest analysis was minor, 0.04% for the left and 0.19% for the right fusiform gyrus in the contrast image analysis, while that of visit was 1.65% and 3.31% respectively. A larger part of the variance was explained by the subject factor, 23.9% and 40.89%, with the largest part remaining unexplained (74.4% and 55.61%). In the analysis of statistical maps, no contribution was found by the site factor. Visit (7.9% and 15.54%) and subject (24.14% and 9.49% explained some of the variance, with the largest part again remaining unexplained, at 67.95% and 74.98% respectively. ICCs were poor in most cases, though better or worse within and between sites. ICCs were 0.24 in the left fusiform for contrast and statistical map analyses, while in the right they were 0.41 for the contrast images but 0.09 for the statistical maps.

Table 6.1: Variance Components

Anat. area	Site	Visit	Subject	Subject by site	Error	ICC_{within}	$ICC_{between}$
Contrast mean							
Fusiform L	0.04	1.65	23.90	0.00	74.40	0.24 (p)	0.24 (p)
Fusiform R	0.19	3.31	40.89	0.00	55.61	0.41 (f)	0.41 (f)
T statistic mean							
Fusiform L	0.00	7.90	24.14	0.00	67.95	0.24 (p)	0.24 (p)
Fusiform R	0.00	15.54	9.47	0.00	74.98	0.09 (p)	0.09 (p)

Variance components as percentage of total variance contributed by site, visit, subject by site interaction and unexplained variance. Intraclass Correlation Coefficients (ICCs) within and between site. The analysis was run in a region of interest (ROI) in the fusiform gyrus bilaterally in the fearful faces vs. rest contrast images and statistical parametric maps. According to a priori criteria we use 'p', 'f' and 'g' to refer to 'good', 'fair' and 'poor' for the ICCs (Cicchetti, 2001).

Table 6.2: Fearful faces vs. rest Aberdeen 1st visit

Anatomical Area	*Coordinates*	*Cluster p*	*Extent*	*Voxel T*
Right Occipital Inferior Gyrus	[28;-88;-16]	< 0.001	5999	14.88
Right Fusiform Gyrus	[40;-54;-22]			14.31
Right Occipital Middle Gyrus	[28;-96;0]			12.33
Left Postcentral Gyrus	[-58;-18;36]	< 0.001	9480	13.53
Left Putamen	[-20;0;12]			12.35
Left Postcentral Gyrus	[-60;-14;24]			10.48
Left Inferior Occipital Gyrus	[-24;-94;-6]	< 0.001	3961	13.51
Left Fusiform Gyrus	[-40;-76;-14]			11.80
Left Fusiform Gyrus	[-42;-68;-14]			11.52
Right Hippocampus	[34;-28;-10]	< 0.001	3266	8.62
Right Putamen	[26;4;-2]			8.52
Right Amygdala	[26;-2;-10]			7.92
Left Supplementary Motor Area	[-2;6;50]	< 0.001	1184	7.75
Left Middle Cingulum	[-6;-10;46]			5.76
Right Supplementary Motor Area	[4;4;60]			5.53
Right Parietal Inferior Gyrus	[34;-48;44]	< 0.001	899	6.66
Right Angular Gyrus	[36;-60;42]			6.25
Right Angular Gyrus	[40;-58;32]			5.45

Brain activations for the fearful faces vs. rest contrast: Aberdeen 1st visit. P-values FDR 'whole brain' corrected. MNI space. Subjects included = 14.

Table 6.3: Fearful faces vs. rest Aberdeen 2nd visit

Anatomical Area	*Coordinates*	*Cluster p*	*Extent*	*Voxel T*
Right Occipital Inferior Gyrus	[34;-84;-8]	< 0.001	4781	18.95
Right Middle Occipital Gyrus	[36;-88;6]			12.85
Right Lingual Gyrus	[24;-86;-16]			12.30
Left Occipital Inferior Gyrus	[-26;-94;-6]	< 0.001	3802	14.64
Left Occipital Inferior Gyrus	[-34;-80;-10]			12.63
Left Occipital Middle Gyrus	[-46;-84;-4]			8.03
Left Hippocampus	[-28;-10;-10]	< 0.001	1285	7.15
Left Thalamus	[-14;-18;4]			6.69
Left Putamen	[-28;-24;2]			6.29
Right Amygdala	[22;-4;-12]	0.030	329	6.07
Right Putamen	[26;8;-2]			4.24
Left Parietal Inferior Gyrus	[-58;-20;48]	0.005	539	5.51
Left Parietal Inferior Gyrus	[-46;-26;48]			5.33
Left Precentral Gyrus	[-36;-20;64]			5.29

Brain activations for the fearful faces vs. rest contrast: Aberdeen 2nd visit. P-values FDR 'whole brain' corrected. MNI space. Subjects included = 14.

Table 6.4: Fearful faces vs. rest Edinburgh 1st visit

Anatomical Area	*Coordinates*	*Cluster p*	*Extent*	*Voxel T*
Right Putamen	[22;6;-10]	0.000	700	8.74
Right Putamen	[24;16;0]			6.48
Right ParaHippocampal Gyrus	[18;2;-22]			4.84
Right Frontal_Superior_Orbital Gyrus	[8;64;-20]	0.002	443	8.60
Right Frontal Midddle Orbital Gyrus	[22;50;-20]			5.96
Right Frontal Inferir Orbital Gyrus	[42;42;-20]			5.27
Right Inferior Parietal Gyrus	[38;-68;56]	0.011	290	8.00
Right Angular Gyrus	[38;-60;48]			6.85
Right Angular Gyrus	[36;-56;40]			6.70
Right Inferior Frontal Gyrus	[48;30;4]	0.000	1480	7.93
Middle Frontal Gyrus	[50;2;56]			7.78
Middle Frontal Gyrus	[58;22;30]			7.36
Left Amygdala	[-22;-2;-14]	0.000	1371	7.73
Left Putamen	[-26;4;-4]			6.93
Left Insula	[-26;16;4]			6.54
Right Supplementary Motor Area	[12;10;54]	0.106	129	7.65
Left Supplementary Motor Area	[-10;8;50]	0.028	221	7.46
Left Supplementary Motor Area	[-8;0;60]			6.15
Middle Frontal Gyrus	[42;52;-6]	0.039	193	6.29
Middle Frontal Gyrus	[36;52;-12]			5.34

Brain activations for the fearful faces vs. rest contrast: Edinburgh 1st visit. P-values FDR 'whole brain' corrected. MNI space. Subjects included = 13.

Table 6.5: Fearful faces vs. rest Edinburgh 2nd visit

Anatomical Area	*Coordinates*	*Cluster p*	*Extent*	*Voxel T*
Right Middle Occipital Gyrus	[32;-92;10]	0.000	8925	22.43
Right Lingual Gyrus	[24;-92;-12]			17.48
Left Middle Occipital Gyrus	[-26;-100;2]			14.50
Left Postcentral Gyrus	[-48;-28;66]	0.004	804	6.18
Left Postcentral Gyrus	[-52;-24;60]			5.76
Left Postcentral Gyrus	[-40;-24;52]			5.41

Brain activations for the fearful faces vs. rest contrast: Edinburgh 2nd visit. P-values FDR 'whole brain' corrected. MNI space. Subjects included = 13.

Table 6.6: Fearful faces vs. rest Glasgow 1st visit

Anatomical Area	*Coordinates*	*Cluster p*	*Extent*	*Voxel T*
Left Lingual Gyrus	[-20;-90;-14]	0.000	4249	18.88
Left Cerebelum	[-38;-82;-28]			13.00
Left Cerebelum	[-44;-70;-22]			12.90
Right Inferior Occipital Gyrus	[34;-80;-8]	0.000	4936	16.80
Right Cerebelum	[28;-52;-30]			13.94
Right Cerebelum	[34;-44;-26]			11.66
Vermis	[2;-76;-32]	0.002	262	15.41
Right Cerebelum	[8;-72;-26]			7.23
Left Cerebelum	[0;-82;-46]			4.69
Left Postcentral Gyrus	[-40;-28;46]	0.000	3254	10.38
Left Precentral Gyrus	[-30;-16;62]			9.03
Left Postcentral Gyrus	[-48;-30;48]			9.00
Left Supplementary Motor Area	[-2;4;48]	0.000	805	10.28
Left Supplementary Motor Area	[-6;-4;54]			8.27
Right Supplementary Motor Area	[8;-2;64]			6.40
Right Cerebelum	[24;-58;-50]	0.059	106	9.02
Right Superior Frontal Orbital Gyrus	[14;52;-16]	0.000	460	8.75
Right Inferior Frontal Orbital Gyrus	[36;42;-20]			8.70
Right Superior Frontal Orbital Gyrus	[24;38;-24]			7.92
Left Putamen	[-28;8;-4]	0.000	624	8.61
Left Putamen	[-26;-2;-6]			7.40
Left Amygdala	[-20;-2;-16]			7.10
Right Putamen	[22;16;0]	0.001	319	8.14
Right Amygdala	[26;2;-14]			6.94

Brain activations for the fearful faces vs. rest contrast: Glasgow 1st visit. P-values FDR 'whole brain' corrected. MNI space. Subjects included = 11.

Table 6.7: Fearful faces vs. rest Glasgow 2nd visit

Anatomical Area	*Coordinates*	*Cluster p*	*Extent*	*Voxel T*
Left Supplementary Motor Area	[-6;6;56]	0.002	303	14.18
Right Supplementary Motor Area	[8;14;56]			4.36
Right Cerebelum	[18;-54;-24]	0.000	3025	8.86
Right Cerebelum	[28;-84;-18]			7.57
Right Middle Occipital Gyrus	[32;-92;2]			7.40
Left Inferior Occipital Gyrus	[-24;-88;-10]	0.000	1933	8.09
Left Inferior Occipital Gyrus	[-24;-98;-6]			6.91
Left Fusiform Gyrus	[-44;-54;-22]			6.64

Brain activations for the fearful faces vs. rest contrast: Glasgow 2nd visit. P-values FDR 'whole brain' corrected. MNI space. Subjects included = 12.

Chapter 7

Discussion

7.1 Project management

The CaliBrain study is a multicentre effort, involving sites in three different cities and a large amount of collected data. Planning and implementing an imaging study of this nature presented some challenges not typically relevant in a single site study.

7.1.1 Challenges of a geographically distributed project

The three sites participating in the project were located in Edinburgh, Aberdeen and Glasgow. In order to organise meetings we had to arrange travel for the researchers involved, which proved to be a challenge given everyone's commitments, therefore most of the communication took place by email. There were few meetings where all participating researchers were present, but in many cases I travelled to the different sites as needed to liaise with local research staff.

7.1.2 Design stage

In the initial design stages of the study there were various concepts and ideas explored which had to be weighed against practical constraints such as financial costs, time allocated for the project, staff experience with procedures involved and equipment available. In order to cover expenses such as scanning costs, participant and researcher travel expenses and occasionally overnight stays, a Project Grant was secured from the Chief Scientist Office (CSO Scotland). In comparison to running a single site project every stage of the study took longer to complete, on one hand because of the challenges posed by the distributed nature of the project and on the other hand because some of the work, like the development and implementation of scanning protocols and the programming of the fMRI tasks, had to be done in duplicate or triplicate.

Staff at the different sites had various degrees of experience with running an fMRI study. It also became evident that in each site specific procedures were followed which could result in undesirable differences in the data collected. In order to resolve these issues, a detailed site specific protocol was produced and extensive staff training was conducted by Katherine Lymer for the QA component and by myself for the human imagining in all sites.

The main factor influencing the specifics of the study was the equipment available at each site. In all sites we used 1.5 T GE scanners which eliminated some potential concerns, as using machines by different manufactures and of different field strength

could introduce systematic differences in the data collected. However, some variations were unavoidable, such as differences in gradient strength and scanner software versions, as well as differences in stimulus presentation and behavioural response collection software and hardware available. One of the main study goals was to minimise differences in the equipment used as much as possible. However, due to financial constraints and in view of realistic considerations in future projects, it was decided to use already existing equipment in each of the sites.

7.1.3 Data acquisition stage

Due to the complexity of the study design it was not easy to find fifteen volunteers who fulfilled the requirements and were able to commit to having six scans each, four of which would involve travel to a different city in Scotland. Fifteen participants were initially recruited, out of which fourteen completed the study, a satisfactory rate for a project of this scale.

Another challenge was adhering to the scan schedule. Each participant had to have the scans in a specific order within a certain time interval, which in many cases conflicted with their personal time available and the busy scanner schedules. The study protocol was long and demanding, so staff and participant training was conducted prior to scanning in order to ensure good data quality. The process of collecting all the data in a centralised location was also a consideration, due to large file sizes and the confidential nature of the data, which was accomplished in different ways from each site.

7.1.4 Analysis stage

A file naming convention and data filing system was developed to ensure anonymity and facilitate automated data processing and analysis. The quality assurance data was analysed in Edinburgh by Katherine Lymer and myself, the structural MRI data was analysed in Edinburgh by T. William Moorhead, the motor fMRI data was analysed in Edinburgh by myself, all using both standard software and locally developed methods. The memory fMRI data was analysed in Aberdeen by Victoria Gradin and Gordon Waiter using the same procedures followed for the motor task.

7.1.5 General comments

For a project of this nature it is important to strike a balance between distributed and centralised management with a clear allocation of tasks and responsibilities. Furthermore, it became evident that there is a need for detailed documentation, both at each stage of the project and for the final data set in order to facilitate future uses of the data.

7.2 Quality Assurance

The Quality assurance (QA) analysis discussed here was performed by Katherine Lymer using methods developed in collaboration by her and myself. QA is an essential part of multicentre studies as it provides an indication of the precision of the measurements made in the absence of any systematic bias (Tofts, 2003). Since control subjects are not always scanned as part of a study, test objects (phantoms), with their known configuration and composition, are convenient for repeated scanning and accurate determination of scanner performance (Tofts, 2003). However, QA is of practical use if it can accurately reflect changes that may be observed in-vivo and for this reason it was examined whether signal-to-noise (SNR) measures obtained from the human data were reflected in any of the in-vitro SNR measurements.

7.2.1 Signal-to-noise measurements

The SNR is a non-specific measure of image quality and can be affected by a number of parameters but should remain stable when measured using the same coil and sequence parameters (Koller et al., 2006), as used in Calibrain. The SNR measured in the phantoms was consistently higher than that measured in the human data. Ideally, the QA phantoms would have been scanned within an annuls to ensure that the resistive loading of the test object match those of a human subject (Firbank et al., 2000). However, this was not possible due to financial constraints. Although the in-vivo and in-vitro SNR measurements cannot be compared directly, comparisons of the variance of these measures are valid, since the scanning protocols were identical except for the number of slices acquired.

More specifically, the variance in the weekly QA SNR and 'time of scan' SNR data was compared to that of the human data. Overall, the results suggest that in the CaliBrain data set weekly QA provides a more accurate indication of scanner performance despite the shorter time interval between the human and 'time of scan' phantom scans. In the absence of any significant system changes, as indicated by the absence of substantial change in the weekly QA, or changes in the scan room temperature, it is likely that the large variance observed in the 'time of scan' measurements is due to operator set-up of the phantom.

7.2.2 Sources of variation

Accurate positioning of test objects is a known source of variance in QA and, as such, preventative measures were taken in the form of explicitly marking the test object and training the staff in the specific of the scanning protocol. However, it is possible that, because of the time pressure when scanning human volunteers, phantom set-up in these sessions was less reliable than in a dedicated QA session. In addition, staffing arrangements varied across sites with some employing radiographers for all human scanning and QA, physicists for all research human scanning and QA and radiographers for human scanning and physicists for QA. Since the before and after scans of the test objects were acquired as part of the control subject scanning protocol, it is possible that this change in staff with their potential inexperience of QA also contributed to the increase in variance in the 'time of scan' SNR measurements observed in one of the sites.

Variations in the level of experience of the staff acquiring the data may also have contributed to the differences in the variance observed in the in-vivo measurements. For example, the variance in one of the sites was significantly higher and this may reflect a relative inexperience of the scanning staff with this specific procedure. If true, this re-enforces the importance of thorough staff training in both the human and phantom components of the scanning protocol.

7.2.3 Conclusions

Finally, QA results also made it possible to confirm that routine maintenance in one of the sites had no observable effect on scanner performance. Conducting regular QA measurements will be useful for assessing the impact of planned and unplanned hardware changes in future studies.

Overall, the weekly QA data suggests stable operation of all three Calibrain scanners, with the SNR fluctuations falling within the expected range of 5 -10 % (Lerski et al., 1998). Furthermore, these analyses suggest that the exact timing of data acquisition is not critical for the purposes of scanner calibration. Practically, this provides greater latitude to perform QA on a regular basis when convenient, in terms of scanner load, rather than at a proscribed time.

7.3 Structural MRI

Structural imaging data acquired as part of the CaliBrain project was used to examine scanner differences at the three sites and to assess the practicality of pooling scans for multicentre VBM studies. The sMRI data analysis and harmonisation method development was done by T. William Moorhead, the absolute distance metric was developed by T. William Moorhead and myself. Detailed methods and results are presented in Chapter 4 and were published in Moorhead et al. (2009), included in Appendix E.

7.3.1 Scanner Harmonisation

Previous research (ADNI; Hua et al., 2008; Jack et al., 2008; Leow et al., 2006; Han and Fischl, 2007), recommend that to pool scans from multiple sites it is important to:

- minimise hardware differences in
  - vendor
  - field strength
  - head coil
  - sequences
- minimise
  - software differences between sites
  - protocol differences between sites
  - by e.g. using an MR-RAGE sequence, or scanner 'invariant' sequences
- make global or regional corrections (limited to region of interest studies)
- use validity masking (limits the analyses to less than whole brain coverage)

For within scanner variation:

- correct RF inhomogeneity using B1 field mapping (requires extra scanning time)

In the CaliBrain project within scanner variability and between scanner differences were examined, aiming to reduce the between scanner differences to the level of within scanner variability. In keeping with the ADNI recommendations it was sought

to minimise the scanner differences in terms of vendor, field strength, head coil and sequences.

We investigated this issue in the CaliBrain data set using T1-weighted images for thirteen subjects scanned twice at each of three sites. Baseline analyses of the CaliBrain T1 segmentations indeed revealed significant differences between the scanners and these differences were of an order that would require validity masking. However, one of the scanners in the CaliBrain project does differ from the other two in terms of maximum gradient amplitude and maximum slew rate. Therefore, while two of the scanners are well matched and data from these two sites could be pooled without further adjustment or compensation, the third scanner exhibits significant differences with respect to the other two.

The method that we have proposed is in keeping with this existing work as we have implemented corrections at a scale that is close to the analysis scale for VBM. However, B1 field mapping can be applied as an addition to the priors adjustments protocol that we have developed. It is possible that the inclusion of field mapping would further reduce the between scanner differences in the CaliBrain project. However, the scan time acquisitions necessary for correction of the B1 field are not available in the CaliBrain project.

Our findings indicate that the development of scanner specific adjusted priors for use in VBM analyses can assist in the pooling of structural imaging data from different sites. Six subjects were found to be adequate for the purpose of matching the scanners in the CaliBrain project. In the typical clinical study the range of tissue presentations would be expected to be greater than that seen here. Thus it is likely that in a clinical study a larger number of travelling subjects would be required, with the exact number depending on the diversity of tissue presentation in the study and the nature of the differences in the scanners employed.

The method presented here may be limited to multi-site studies following a design similar to CaliBrain, which provides an optimal environment for multiple site scan pooling. Different field strengths and image acquisition protocols could have very different tissue contrasts that would lead to marked differences in segmentation results. In such cases the differences in tissue classification may well be beyond the scope of this compensatory method. However, this development can facilitate data pooling and allow for improvements in the statistical power of multicentre brain imaging studies where there are no major hardware and acquisition protocol differences across sites.

The metric results demonstrated that the use of scanner specific priors can reduce the tissue classification differences between scanners. These reductions were not sufficient to bring the between scanner differences down to the level of within scanner variability. However, in VBM analyses of the segmentations based upon scanner specific priors it was found that the baseline differences which would previously have required validity masking, were removed.

Combining structural MRI scans from different scanners could increase the statistical power in VBM analyses. Our aim was to extend SPMs segmentation processes to reduce the effects of scanner differences which currently limit multi-centre MRI pooling (Stonnington et al., 2008; Meda et al., 2008). Although these scanners are well matched we found significant between scanner differences in tissue segmentations, using standard methods. We have demonstrated that scanner specific priors can reduce between scanner differences.

In the CaliBrain project we consider within scanner variability and between scanner differences and our aim was to reduce the between scanner differences to the level of within scanner variability. In keeping with the ADNI recommendations we have sought to minimise the scanner differences in terms of vendor, field strength, head coil and sequences. However, scanner B in the CaliBrain project does differ from scanners A and C in terms of maximum gradient amplitude and maximum slew rate. Our baseline results indicate that scanners A and C are well matched and scans from these two sites could be pooled without further adjustment or compensation. However, our baseline results also demonstrate that scanner B exhibits significant differences with respect to both scanners A and C.

In order to reduce the differences between the scanners in the CaliBrain project we have developed a procedure that employs proportional feedback to adjust the priors for each of the scanners. We have scan records for 13 healthy subjects who were scanned twice at three scanners within a six month period. We demonstrate our protocol for creating scanner specific priors using the 1st round scans of six subjects. We test the adequacy of these scanner specific priors through metric and VBM analyses. The tests for adequacy are applied to the seven subjects who were excluded from the priors adjustment protocol. These tests are limited by the number of subject scans available and we are unable to evaluate the full effects of subject variation expected in a multi-centre clinical study. Clinical studies that could benefit from the scanner specific priors method are expected to have subject numbers considerably greater than those available for the CaliBrain project. In a multi-centre clinical study, with the exception of the

travelling subjects used to develop the scanner specific priors, the subjects would be recruited and scanned independently at the contributing centres. In such a clinical study a test for adequacy of scanner harmonisation could be implemented through comparisons of the healthy control scans recruited from the contributing centres (Stonnington et al., 2008; Meda et al., 2008).

The metric that we report assesses the absolute distance between segmentations. The metrics are applied at the voxel level and are averaged to report an overall distance inclusive of noise and systematic differences. Metic results on the scans that were used to implement the scanner specific priors procedure indicate that the within scanner variability ranges from 3.0% in scanner B to 2.1% in scanners A and C. The adjustment process gives rise to a reduction in the within scanner variability. However, the paired-t tests reveal that these within scanner adjustments do not represent a significant change. The baseline between scanner differences are at a maximum for the B and C comparison. Here the adjustment procedure gave rise to significant reductions in all three scanner comparison metrics.

We then consider the effects of the scanner specific priors on the scans of seven subjects who were excluded from the priors adjustment process. At baseline the within scanner variability and between scanner distances were equivalent to the baseline results in the first dataset. Consequently, the use of the scanner specific priors resulted in significant reductions in all three scanner comparisons. However, for the comparisons that include scanner B, the reductions are not sufficient to bring the between scanner difference down to the level of within scanner variability.

The VBM analyses that we applied demonstrated that at baseline there are no significant differences between scanners A and C, However, we found that comparisons of scanners A and C with scanner B gave rise to differences that would require validity mapping such as that employed in VBM analyses by (Stonnington et al., 2008; Meda et al., 2008). After developing scanner specific priors for scanners B and C and re-segmenting the scans we found that the requirement for validity mapping was removed, because we recorded no significant differences in the grey and white matter F-tests for scanner effect.

7.3.2 Conclusions

Our results indicate the development of scanner specific priors for the SPM application can assist in the pooling of scan resources from different research centres. This

development can facilitate scan pooling and allow for improvements in the statistical power of multi-centre brain imaging studies. Our results indicate that six subjects were adequate for the purpose of matching the scanners in the CaliBrain project.

In the typical clinical study the range of tissue presentations could be greater than that seen in our study of healthy controls. Thus it is possible that in a clinical study that more than six travelling subjects would be required. The number of travelling subjects required would depend upon the diversity of tissue presentation in the study and upon the nature of the differences in the scanners pooled.

The method that we have suggested may also be limited to multi-site studies in which there are :

- Minor hardware or acquisition protocol differences across sites
- scanners from the same vendor
- the same field strengths
- the same head coils
- and the same or matched sequences

This provides an optimal environment for multiple site scan pooling. Different field strengths and image acquisition protocols could have very different tissue contrasts that would lead to marked differences in segmentation results. In such cases the differences in tissue classification may well be beyond the scope of our compensatory method.

7.4 fMRI motor task

We investigated the reproducibility of a sequential and a random motor tapping task across multiple sites and visits, examining it both in terms of single subjects and group analyses. A detailed description of the methods and results can be found in Chapter 5 and in Gountouna et al. (2010), included in Appendix C. We found mostly consistent activations in expected areas, including cortical and subcortical motor areas and the cerebellum. As anticipated, comparison ratios were overall higher and more stable in the group-level analyses than at the subject level. Reproducibility between sites was similar to that of different visits within the same site. Reproducibility in overlap and size estimates were similar between and within sites and acceptable at both group and subject levels.

7.4.1 Reproducibility: Activation locations

Robust activations were observed in the left premotor, primary motor and supplementary motor areas. These were present across all sites and visits. Right-hemispheric counterparts were also activated but are weaker and less consistent. The opposite pattern was observed in the cerebellum, with generally consistent ipsilateral but weaker contralateral activations. Left thalamus and basal ganglia were also detected in most cases. Mattay et al. (1998) employed a sequential and random tapping task and reported activations in the primary motor, somatosensory and premotor areas, the SMA, parietal cortex, putamen and cerebellum. They also reported a prefrontal cluster (BA 9) in the random tapping condition, which was detected in some of our analyses. Yoo et al. (2005) investigated the reproducibility of a sequential finger tapping task over a longer period of time, also employing size and overlap ratios in selected regions of interest. They found consistent activations in the primary motor, premotor, SMA and cerebellum; activations were less consistent in the basal ganglia and thalamus. Scholz et al. (2000) examined the reproducibility of functional activations in motor areas using a variety of motor tasks including finger tapping and also found relatively reduced signal change and decreased reproducibility in the basal ganglia. Casey et al. (1998) employed a task including a motor condition similar to this paradigm in a multicentre study and reported reliable activations in the left premotor, primary motor and supplementary motor areas and the right cerebellum and less reliable activations in the thalamus and basal ganglia. These findings are compatible with our results and suggest that in these smaller regions increased noise possibly has a negative effect on the

reproducibility of activation. Alternatively, this could also be due to the signal being less reliable in these regions.

7.4.2 Reproducibility: Overlap and size ratios

A qualitative evaluation of the results indicates that at the group level, location and extent of activation is robust, both within and between sites. Quantitatively the results were within the range reported by studies examining the reproducibility of similar tasks on a single scanner (e.g. Ramsey et al., 1996). Furthermore both size and overlap ratios, and size by volume and size by volume and intensity, were found to be comparable within and between sites. Other studies investigating between scanner reproducibility also did not find large differences in comparisons within and between sites. Voyvodic (2006) reported stable spatial patterns even between scanners of different strengths. Vlieger et al. (2003) found that the average inter-scanner agreement did not differ significantly from the average within scanner reproducibility of the site with the worst reproducibility. Sutton et al. (2008) found similar spatial extent of activation across scanners of the same strength.

The reproducibility in overlap and size ratios for both mean and maximum measures were comparable across all analyses with no clear advantage of the one over the other. A similar pattern was found by Friedman et al. (2008) who compared median and maximum values. A clear difference however emerges when comparing the reproducibility in overlap and size ratios in contrast and T-statistic images, with the reproducibility in contrast values being consistently higher. This was hypothesised by Friedman et al. (2008) who used a finite impulse response (FIR) method to get a surrogate value for signal change and suggested the method employed here as a possible alternative.

Much of the reproducibility literature to date has been focussed on the subject level, but in many cases it is the group-level results that are given weight in the interpretation of MRI results. Our analyses indicate that group-level maps for this task are overall more reproducible, and the range of obtained values is much narrower. This is in agreement with what has been reported in a single centre analysis (Seghier et al., 2004). Critically, subject is a much greater source of variability than scanner, however subject by scanner variance is low, and within subject variance (visit) is also low and of an acceptable level. Merging of data across sites is therefore possible in studies where the reproducibility of individual subject measurements is essential, e.g. in treatment

response studies.

7.4.3 Variance components and reliability estimates

This outcome is supported by the results of the variance components analysis. The contribution of the factors site and visit to the total variance was relatively small regardless of measure used, and subject was the largest component. However a large part of the variance remained unexplained in some of the analyses, notably in the striatum, presumably due to the small volume of the region, although this variability could also reflect a difference in the striatum's more complex neurovascular dynamics and the task used. ICCs were very similar within and between sites, compatible with the results of the whole brain size and overlap ratios.

Friedman et al. (2008), reported a variance components analysis examining the effects of site, subject and visit in a dataset of five subjects who conducted repeated visits in ten sites. Costafreda et al. (2007) also employed a variance components analysis to examine the relative contributions of site and subject to the variance in a motor task in five subjects across five scanners. Both found the effect of site to be small compared to that of subject and residual unexplained variance. Suckling et al. (2008) performed a similar study employing two versions of an affect processing task in twelve subjects and two sites. They examined the contributions of site, visit, task and subject and also found the contribution of site to the variance to be relatively small. Sutton et al. (2008) scanned four subjects fifteen times in each of two sites employing a motor and visual task. They found the contribution of site to the variance to be small compared to that of subject, with most of the variance remaining unexplained. Bosnell et al. (2008) conducted a clinical multicentre study across five sites using a hand tapping task comparing variability across seventeen multiple sclerosis patients and twenty two healthy controls. They also scanned five healthy volunteers at each of four of the sites to assess the contributions of site to the variance and found that variation between subjects greatly dominated over variation between visits or sites.

It is of note that little site effect was observed in the striatal ROI in any of the analyses performed. It is possible that the location of this structure near the middle of the brain renders it less vulnerable to scanner specific distortions and therefore makes it a suitable target area for future multicentre studies of clinical interest. However, its small size means it is very susceptible to noise. A lot of the variance remained unexplained in these analyses and reproducibility in overlap and size of activation was

found to be poor in comparison to the larger neocortical ROIs. ICC values for these analyses are similar or higher than those reported elsewhere (Friedman et al., 2008; Kong et al., 2007).

Brown et al. (2011) performed a multicentre study employing a working memory task in four sites and eighteen subjects. They found the contribution of site to the variance to be small in most anatomical regions, however highlighted the potential problems introduced by large site-by-subjects interactions even if the contribution of the site factor is small. The trade-off between number of subjects and number of averaged runs also is discussed, along with the possible effect of the nature of the contrast used, low vs high level cognitive control, with the low level control possibly offering better reliability. However in this case this coincided with a greater number of runs. Also, the authors point out that effect size and ICCs don't always go together and stress the importance of examining both when planning multicentre studies. Finally they discuss the possible effects of poor reliability on studies attempting to correlate with an external variable, such as genotype.

7.4.4 Conclusions

In summary, the reproducibility of a finger tapping task was evaluated across three sites in thirteen subjects on two visits each. Robust activations were detected in typical motor areas. Reproducibility of activation location and extent were similar for the sequential and random tapping conditions. Reproducibility was comparable within and between sites, and critically was acceptable within single subjects, the majority of variance being between subjects and in unexplained variance. The contributions of site and visit to the variance were low and reproducibility in overlap and size was similar between and within sites. However, reproducibility was poor in smaller anatomical areas and mostly fair to good in the larger areas, representative of the difficulties that face fMRI in general. This indicates that we can have confidence in the results produced by multicentre functional MRI when a consistent scanning and analysis protocol is followed, but that more work needs to be done and care taken in selecting homogeneous subject groups and suitable tasks. Possible strategies to combine data in one analysis include the use of one large mixed model or a more traditional meta-analysis approach.

7.5 fMRI Working memory task

The n-back working memory data set was analysed by Victoria Gradin and Gordon Waiter in Aberdeen using methods similar to those presented in 5 and in Gountouna et al. (2010), included in Appendix C. A detailed discussion of methods and results can be found in Gradin et al. (2010), included in Appendix D. The n-back working memory task was implemented as a parametric design with different n-back levels occurring in blocks. Subjects were presented with letters and had to respond by indicating whether the letter matched the target or not up to three letters back ('n'=number of letters back). The difficulty of the task increased with 'n'. Recorded behavioural measures were letter choice, allowing determination of accuracy and reaction time. Data for fourteen subjects across sites and visits was included in the analysis. One subject was found to have an abnormality resulting in a loss of BOLD signal in a small region of the dorsal anterior cingulate (dAC), however while data for this subject was excluded from the analysis of the motor task, its inclusion here was not expected to affect the planned tests. Behavioural responses and functional imaging data across sites were comparable to findings of similar studies conducted in a single site.

7.5.1 Reproducibility: Behavioural responses

The behavioural responses which consisted of reaction times and accuracy data were analysed using within-subject analyses of variance. Accuracy was compared between two sites only as data was not recorded at one of the sites. As expected, mean reaction times increased significantly with the n-back level of difficulty. Additionally though, mean reaction times were significantly slower in one site. This is unlikely to be due to practice effects as the order of visits was counterbalanced. Instead, this result could be attributed to the fact that different software was used for the presentation of stimuli and recording of responses in this site. Reaction times were not significantly affected by whether a subject was visiting a scanner for the first or second time. Accuracy was affected by both site, with data from one site having higher accuracy than the other, and visit, with higher accuracy on a subject's second visit than on the first visit, potentially indicating a practice effect. Finally, as expected, accuracy tended to decrease with increasing difficulty.

7.5.2 Reproducibility: Activation locations

Changes in brain (de)activation with increasing n-back task difficulty were examined at the group level for each scanner on each visit. Significant activations across groups were found in the dorsal anterior cingulate (dAC), lateral anterior prefrontal cortex (laPFC), dorsolateral prefrontal cortex (dlPFC), right posterior parietal lobe (PPL), insula and cerebellum. Additionally, significant deactivations were found in the dorsal posterior cingulate (dPC), retrosplenial cortex (RCS), medial anterior prefrontal cortex (maPFC), hippocampal-amygdala complex (HAC) and auditory cortex. In summary, significant and qualitatively similar patterns of (de)activation were found for each scanner and visit, which are in agreement with the existing literature (Owen et al., 2005). Using a within-subject analysis of variance, no significant effect of site, visit or their interaction was found.

7.5.3 Reproducibility: Overlap and size ratios

Reproducibility measures yielded similar results to those obtained for the motor task and an examination of the contribution of site to variance in functionally defined regions of interest showed site to have a very small effect, which is in agreement with the motor task data. For the brain activations, the best reliabilities were found in the dlPFC and the insula bilaterally, the worst in the dAC and left laPFC. For the deactivations, the RSC showed the best reliability, the dPC showed poor reliability and the auditory cortex worst reliability. These results are in agreement with the findings of Caceres et al. (2009), who examined the reproducibility of a very similar version of n-back and found the dlPFC amongst the most reliable regions and the ventrolateral PFC the least reliable. Wei et al. (2004) examined the reproducibility of an auditory n-back task in a single site and found consistent activations in the dlPFC, anterior cingulate and the insula. Casey et al. (1998) investigated the reproducibility of the n-back task across four sites and also reported reliable activations in the right dlPFC.

7.5.4 Variance components and reliability estimates

For most brain regions examined by Gradin et al. (2010), a large part of the variance was attributed to the subject factor or remained unexplained. Notably, the within-scanner ICCs reported were similar to the between-scanner ICCs, reflecting the fact that the scanner factor accounted for a small percentage of the total variance for all

brain regions studied. Brown et al. (2011) employed an emotional working memory task in a multicentre study across four sites and examined ICCs both voxel-wise and in functionally defined ROIs. They found that between site reliability in areas with large activations was comparable to within site, while poorer reliability was evident in regions nearer the brain edges. They also report that averaging across runs increased reliability substantially. The variance components reported in their ROI analysis are comparable to those reported by Gradin et al. (2010), however the contributions of site-by-subject interactions seem to be more pronounced in their results. This could be explained by the fact that scan order was not counterbalanced in the Brown et al. (2011) study, in contrast to the CaliBrain study.

7.5.5 Motor vs. working memory

In comparison to the motor task, the working memory task involves higher level cognitive processes and could therefore be expected to show more variability due to individual differences between subjects, e.g. strategy, psychological state at the time of the scan, or physiology (Miller et al., 2009). However, reproducibility measures were very similar at the group level, with mean values for size and overlap ratios being comparable within and between scanners across both tasks. Reliability estimates were also very similar within and between sites for both tasks. This result is encouraging in view of future multicentre clinical studies, as the n-back task engages higher cognitive of relevance to psychiatric research (e.g. Rose et al., 2006).

7.5.6 Other considerations

Manoach et al. (2001) however, compared the reproducibility of results between schizophrenic patients and healthy controls using a different working memory task and found that patients consistently showed less reliable activation than controls in regions associated with cognition including the dorsolateral prefrontal cortex and the insula. This suggests that potential population differences in reproducibility should be investigated when planning clinical studies. Furthermore, Krasnow et al. (2003) compared activation maps for a spatial n-back task between scanners of different strengths and found substantial differences in activation volume for the higher strength scanner in a number of areas, including the frontal and parietal lobes. Similar findings were reported by Zou et al. (2005) who employed a sensorimotor task across ten scanners of varying strengths. While this was not a concern in the current study, it should be taken into

account in future studies including scanners of different strengths.

7.6 fMRI face processing task

The final Calibrain fMRI task to be examined here is a face processing task involving the presentation of fearful and neutral faces. The task was analysed following the same procedures as the other two tasks. A detailed presentation of methods and results can be found in Chapter 6.

7.6.1 Reproducibility: Activation locations

The analysis of this dataset was focused on the fear versus rest contrast because it has been reported to produce stronger and more reliable activation in the affect processing network in general and the amygdala in particular (e.g. Fusar-Poli et al., 2009; Stark et al., 2004; Johnstone et al., 2005). However amygdala activation did not prove to be robust across sites and visits, with significant activations appearing in only one of the second visit analyses. This was true also of frontal activations, which appeared markedely reduced in the later visits. This result is in accord with previously reported findings by Stark et al. (2004). In the first visit analyses activations were also observed in both hemispheres across sites in other areas associated with this task, like the putamen and the hippocampus.

A number of areas commonly associated with face processing (Fusar-Poli et al., 2009) were consistently activated in the second visits as well; these include bilaterally visual areas in the occipital lobe, the fusiform and the lingual gyri. Persistent activations were also observed in lefthemispheric motor areas around the central sulcus and the cerebelum.

7.6.2 Reproducibility: Overlap and size ratios

Size ratios are similar across all analyses, whether within or between sites. Overlap ratios appear overall lower in the group analyses, a reverse trend to the one observed in the motor task results. However this is a higher level task and it is possible that individual differences contribute to this.

Ratios of both size and overlap are lower than those observed in the motor and n-back data, showing the poorest reproducibility of the tasks examined.

7.6.3 Variance components and reliability estimates

Subject and visit were the factors found to contribute most of the explained variance in the region examined, namely the fusiform gyrus bilaterally, as would be expected. Reliability estimates were poor to fair. Plichta et al. (2012) also report poor reliability estimates for an emotional face processing task, though in the amygdala rather than the fusiform. They do report relatively robust activations in the amygdala, however this is not directly comparable to this study, due to design differences. One thing to consider is that while the data are grouped into first and second visit, this refers to the local visit to the specific site, while in reality these are grouppings of visits 1-3 compared against 4-6, so that could also account for the more pronounced effects of visit than those found by Suckling et al. (2008). Brown et al. (2011) report fair to good reliability estimates for an emotional working memory task. However, this was only achieved after averaging over a large number of runs within one scanning session. Also, the task they employed was quite different and presumably more engaging than the one presented here.

While the areas commonly asocciated with face processing were reproducible across centres, those areas specific to fear processing didn't show reliable activations over repeat sessions. This is an issue with the design of this task within the study as it is not possible to differentiate within first to third actual visit or within fourth to sixth actual visit and still enough subjects to do group analyses, i.e. it would be five subjects per group. In order to definitevly assess this, it would be required to have a much larger group of subjects, e.g. at least three times the current sample. One possible explantion for these results could be fatigue, not just habituation to the task, but fatigue due to travelling a distance and undergoing the same tedious and quickly learned task six times.

7.6.4 Overview of the fMRI results across tasks

Because all other conditions are identical, any differences in reproducibility seen cannot be attributed to any factor other than the nature of the task itself. All data was acquired in the same scanning session on the same equipment with the same subjects. Data was analysed using the same methods, allowing for minor differences in analysis protocol dictated by logisitcs, like the version of available software.

The motor task appears very stable but is not as interesting from a clinical perspective as the other two. The face processing task, at least as implemented here, is

not showing as much promise for use in a clinical multicentre trial, as reproducibility appears to be the poorest among the tasks examined. The n-back task is the most promising, with possible clinical applications and a wide literature base. Despite it being a higher level cognitive task, the reproducibility is reasonable for 'fMRI standards' and between sites reproducibility is comparable to that within site.

7.7 Strengths and limitations

7.7.1 Study design

This study allows us to examine the feasibility of multicentre MRI taking into account certain limitations, some by design and others by necessity. In terms of equipment, it was decided to use scanners of the same manufacturer and field strength and identical head coils in all sites and any differences between scanning sequences were negligible. Although this made it possible to combine data with little difficulty, it does not allow us to examine the possible effects of these factors. We made an effort to control for systematic differences between sites and reduce variability; however, some differences in hardware and software, as well as more subtle sources of variation, e.g. site-specific practices of the scanning staff, and differences in the displays used, could not be avoided. These are inevitable and realistic in relation to future multicentre studies. The strengths of the present study also include the use of a consistent staff and participant briefing protocol and extensive staff and participant training. Practice effects in the fMRI tasks were distributed evenly by the counterbalancing of visits.

While it was attempted to utilise a wide variety of tasks of possible clinical interest, a language task was not included nor were any tasks employed using auditory stimuli. It was decided during the design stage that while it would be interesting to incorporate these, it was not feasible due to time constraints and technical considerations. Furthermore, the possibility was considered to record respiratory data during functional imaging, which could be potentially be used during analysis to aid harmonisation, but it was not technically possible to employ this method in all sites, therefore it was not included in the protocol. Finally, participants were not restricted from consuming substances such as alcohol, caffeine and nicotine on the night before and the morning of the scan, but were instead instructed to be consistent in their consumption. While this could be a possible additional source of variation in the data, it was more in keeping with what could be expected in a clinical study.

7.7.2 Data analysis

The studies that have systematically examined the reproducibility of fMRI within and between sites have typically employed small numbers of subjects and therefore do not readily lend themselves to the examination of the reproducibility of results at the group level. Here we obtained a data set with larger numbers of subjects in each group

in order to better focus on group level analyses, which are of particular interest to psychiatric studies.

A standard analysis protocol was used in the analysis of the fMRI data, with no steps added to account for intersite differences. Future developments in this area are likely to improve reproducibility further. One limitation of the reproducibility measures employed here (overlap and size rations) is their sensitivity to the threshold used and the fact that they only give information about extent, not peak or amplitude. However, some indication of the effect of threshold and amplitude can be seen in the differences between the size by volume and the size by volume and intensity measures. It can be seen that the volume by intensity measures are significantly more reproducible than the size by volume measure alone. An analysis of all thresholds would be required to confirm this apparent effect. However, this still leaves the possibility that between centre factors, such as peak and amplitude, could increase the total variance sufficiently to obscure any effects, e.g. type II errors. Further to this, it is possible that one site could contribute much more to the total signal so as to bias the results, leading to type I errors. However, activation location and extent are important considerations in multicentre studies and it is possible to interpret the findings in relation to what is known about the reproducibility of activation maps within a single centre.

Despite the coarse level of information, evaluation of these measures at the group level provided useful insights regarding the level of agreement that can be expected between maps obtained from the same group of individuals on different occasions, both in single and multiple sites. The whole brain approach adopted here does not give information on specific regions; this was addressed in the subsequent variance components analysis which allowed us to examine in detail the partition of the variance and the reproducibility in overlap and size ratios of contrast images and statistical parametric maps. It would be advisable to perform pilot studies and employ these or similar measures to examine potential site effects and reproducibility in overlap and size ratios in targeted brain areas when planning clinical multicentre studies.

7.8 Future work

This project produced a large data set which was not possible to be fully utilised within the given timeframe of this degree. The further exploration of this dataset will enable us to identify the key characteristics in the acquisition and analysis of fMRI paradigms which influence their suitability for multicentre MRI studies.

7.8.1 Analysis of the other functional tasks

In this study we have presented the analysis results of the motor and face processing tasks and discussed the memory task. However, there was also data collected using an event-related visual perception task and a breath holding task. These data sets can be examined in the future to provide further insights. Analysis of the data collected using the other tasks will allow the evaluation of reproducibility across different paradigm designs and cognitive domains. The visual perception data set will allow us to examine the feasibility of employing event related tasks in a multicentre setting. Finally, the breath holding data can be used for the normalisation of the BOLD signal of other functional tasks (Cohen et al., 2004; Friedman et al., 2006a; Chiarelli et al., 2007; Thomason et al., 2007), which could further assist in harmonising functional imaging data from different scanners.

7.8.2 fMRI image registration

Image registration was evaluated qualitatively and quantitatively in the motor data set within and between scanner. An edge effect could be observed when comparing one of the sites to the other two, which could be attributed to a slight displacement or distortion of the data. The absolute difference metric values also reflected a lesser degree of agreement in comparison to the other results. More work needs to be done in order to better understand this issue and further improve registration between sites.

7.8.3 Smoothness equalisation

Friedman et al. (2006a) highlighted the potential importance of between scanner differences in image smoothness and its relationship to activation effect size and proposed a method for smoothness equalisation, demonstrated to reduce inter-scanner variation in processed images. Most marked differences in the Friedman study were found between scanners of different manufacturer and field strength, none of which apply here.

Costafreda et al. (2007) however used five 1.5T GE scanners and identical scanning sequences and found that reducing the size of the smoothing filter reduced the amount of variance explained by both the subject and the site factors in a random effects analysis. A comparison of image smoothness in the CaliBrain n-back data set revealed significant differences between scanners (Gradin et al., 2010). It would therefore be of interest to investigate whether applying the smoothness equalisation method proposed by Friedman et al. (2006a) would further reduce between scanner variance in the data analysis of the functional imaging data.

7.8.4 Noise estimation and reduction

The functional imaging data presented here was analysed using the most popular statistical methods as implemented in the SPM software package. However, a variety of additional methods exist, some of which could potentially be of particular relevance in analysing a multicentre data set. One potential method worth exploring was proposed by Diedrichsen and Shadmehr (2005), which aims to detect and adjust for noise and artifacts in functional imaging data using a weighted least squares method. While they suggest that this method could be of use in data sets where a lot of movement artifacts are likely, this way of estimating and accounting for noise could also be of special interest in multicentre data sets.

7.9 Conclusions

7.9.1 CaliBrain outcomes

The CaliBrain project was initiated with the goal to assess the feasibility of multicentre structural and functional MRI and to highlight the critical methodological and technical issues arising from a prospective study using different scanners. This is not a simple question of whether it is possible or not, but a complex and multifaceted issue dependent on a number factors. A major design decision was whether to use scanners of the same strength or not. It was decided to use identical equipment and scanning protocols in order to minimise systematic differences, which meant that these issues could not be investigated within the scope of the study but went a long way towards eliminating many potential problems from the start. A rich data set, which included phantom and human imaging data, was collected in order to investigate other aspects of the multicentre imaging challenge.

More specifically, Quality Assurance (QA) was identified as an area of special importance in multicentre imaging and relevant practices were investigated. One main point addressed was whether employing identical equipment and protocols and conducting extensive staff training to avoid differences due to site specific practices would be an adequate answer to the multicentre question, both for structural and functional imaging. Another issue that was specifically targeted was whether functional activation by a variety of tasks in different anatomical areas would be affected by differences between sites to the extent that it would invalidate analyses pooling data from these sites.

A QA protocol appropriate for use in multicentre studies was created, which provides information on scanner performance over time, while being realistic for use in busy research centres with limited scanning time available. High resolution anatomical imaging data was collected for a group of healthy volunteers which can be used to assess scanner harmonisation methods. One such method was developed using this data set involving the creation of scanner specific segmentation priors for use in Voxel Based Morphometry (VBM) analyses, which improved agreement between scanners. An appropriate measure was created in order to quantify differences in images acquired in different visits and sites. The absolute difference metric was used to assess the effectiveness of this harmonisation method and can also be used in the future to evaluate other harmonisation methods and techniques.

Despite efforts to minimise differences between sites, some variations were un-

avoidable, both in terms of scanner hardware and software and in terms of staff experience with the required procedures. For example, marked differences in the level of variance in the QA data may reflect a different degrees of experience for the staff performing the various procedures. In terms of hardware, differences in gradient strength in one scanner could account for some differences observed in the analyses of the structural data. It would be unrealistic to expect the complete elimination of even minor differences in any multicentre study. Instead these should be compensated for. This could be achieved by conducting more extensive staff training to address the former issue and by applying harmonisation methods, such as the one described above, to deal with hardware related differences.

Finally, data was acquired using a variety of different functional tasks which were likely to activate a wide range of typical areas of interest and were generally known to have good reproducibility, to evaluate their suitability for use in a multicentre setting. Two of these tasks have been analysed using standard methods and the results support data pooling if scanners of the same strength and manufacturer are used.

The motor paradigm employed has been widely used and is well understood, while the working memory paradigm is a task engaging higher cognitive process of clinical interest in neuropsychiatric research. Appropriate measures for assessing fMRI reproducibility in a multicentre setting were identified and implemented and an assessment of the suitability of these tasks showed them to be appropriate for multicentre imaging studies using scanners of the same strength and manufacturer. Reproducibility was comparable within and between sites, and critically was acceptable within single subjects, the majority of variance being between subjects and in unexplained variance. However, reproducibility was poor in smaller anatomical areas and mostly fair to good in the larger areas, representative of the difficulties that face fMRI in general. This indicates that we can have confidence in the results produced by multicentre fMRI when a consistent scanning and analysis protocol is followed, but that more work needs to be done and care taken in selecting homogeneous subject groups and suitable tasks.

Moreover, further improvements can be achieved even in this optimised setting. For example, differences in image smoothness were observed in the functional imaging data and correcting for these differences would further improve the agreement between sites. Furthermore, systematic differences were also observed in the behavioural data collected, which could be attributed to hardware and software differences between sites. Determining the precise source of the problem and, if no other solution is possible, replacing the relevant component would address this issue.

The analysis of the rest of the tasks will allow the assessment of the reproducibility of activation in more anatomical areas of interest between sessions and sites. fMRI in general is a rapidly evolving field and multicentre efforts are now becoming more common, so advances in relevant preprocessing techniques and statistical models are continuously being made. This data set can be used to test new methods suitable for attenuating between scanner differences as they become available. Furthermore, it can be used in the development and evaluation of a wider range neuroimaging analysis methods and has been utilised in this capacity to assess the reproducibility of a structural MRI analysis method (Tijms et al., 2011).

The multicentre nature of the project also meant the collaboration and sharing of resources, data and expertise between the participating researchers, who gained a better understanding of the relevant issues. The centres involved gained valuable experience and some of the infrastructure created as part of CaliBrain was useful for their participation in other multicentre imaging initiatives, such as Neuro/PsyGrid (Suckling et al., 2011).

7.9.2 Designing prospective multicentre MRI studies

According to the existing literature, most notably the work done by the Biomedical Informatics Research Network (BIRN) (see Zou et al., 2005; Magnotta et al., 2006; Jovicich et al., 2006; Friedman et al., 2006b, 2008), employing a wide variety of scanner strength and/or manufacturer, head coils, scanning sequences and sequence parameters in a study substantially increases the obstacles that need to be overcome post-acquisition in order to be able to pool the data in one analysis. It would therefore seem advisable to avoid such complications whenever possible. While realistically this may not always be feasible, adopting an 'aggressive' prospective harmonisation philosophy and striving for consistency across sites in these areas from the beginning should decrease unwanted influences.

It would also greatly benefit any multicentre endeavour to conduct extensive staff training at each site to reduce variations in local practices. In addition, ensuring that detailed and up to date documentation of protocols is available to all staff participating in data collection and analysis would further aid in improving data quality and avoiding errors and omissions. To facilitate this, it would be advisable to have a clearly delineated organisational hierarchy, ensuring that staff exists at each site who is well informed regarding the details of the study and is responsible for ensuring consistency.

Before embarking on a large multicentre study involving a lot of time, effort and financial cost, it would be advisable to test the protocol using a 'human phantom' data set, such as the one presented here (see also Zou et al., 2005; Costafreda et al., 2007; Suckling et al., 2008; Sutton et al., 2008; Bosnell et al., 2008). This can be used to verify scanning sequences, test the functional paradigms and also perform power calculations to determine the number of subjects needed at each site (e.g. Suckling et al., 2008). In the undesirable case that considerable differences are observed in this data set and steps cannot be taken to improve agreement to a satisfactory degree, ensuring that similar numbers are scanned at each site should control for any differences.

Experience gained from the Calibrain project indicated that the automation of data collection and quality assurance procedures would be essential in any large multicentre effort. It was very time consuming to gather all the data in one central location, check it and convert it to the same file formats, orientations etc. This work is usually only necessary for one site and done using one procedure, however this was not the case here, as each site had their own idiosyncrasies. This also increased the potential for error. This issue would be of even more critical importance in a clinical study involving large numbers, as despite the large number of data collected for CaliBrain, this is only a fraction of what can be expected in a clinical study examining differences between patients and controls over time for example. Automating the transfer of data, establishing clearly defined QA procedures, checking the data for compliance with the protocol and using a consistent study specific file naming convention and file system structure would greatly facilitate the manipulation and processing of data in large multicentre studies.

7.9.3 5 simple rules for running a multicentre study

1) Ensure that hardware and software differences between sites are minimised, that staff under go detailed training and have clearly defined roles. Automate file transfer and checking.

Example from Calibrain: Roles were not defined well enough to ensure that all members of staff have the time and the expertise to complete their roles.

2) All tasks and QA should be piloted at each site before the study begins, to confirm that there are no significant issues that would prevent the study from working.

Example from Calibrain: B0 field mapping was collected, but not tested, and due to a design error in the sequence used, the data was unusable.

3) QA is required primarily to ensure that all scanners used in the study are performing as expected and that there are no unexpected changes during the study. B0 field mapping could have been used to correct for scanner inhomogeneities.

Example from Calibrain: The Aberdeen scanner underwent maintenance during the study, and the QA showed that there was no discernible impact.

4) Tasks should be selected that are known to be reproducible, including longitudinally if that is part of the study design.

Example from Calibrain: The faces task showed poor reproducibility.

5) The smaller the number of staff or subjects, the greater the impact of absence.

Example from Calibrain: Multiple instances of staff changes resulted in data sets not being analysed.

7.9.4 Practical considerations

People and reliability:

In a small study like ours the availability of resources, e.g. people with the necessary expertise or available time on the scanner, has a critical impact upon the project. Often with only one person per centre, usually in a speciality that does not cover all areas of expertise necessary to solve all of the potential problems involved in a multicentre study, a small problem can cause significant errors or delays in data collection and transfer. Additionally, if an analysis task assigned to an expert is not completed as that expert becomes unavailable, it can be difficult to cover the task. For example, does your physicist at site 'B' know how to transfer data through the firewall to the central database?

Scale and complexity of study:

Over and above the many considerations and recommendations that should be taken into account, given in this study and according to those previously published by similar studies, we would add these additional considerations, or perspectives.

The first is the question of scale. As the study changes, by numbers of staff, complexity of the experiments, or the number of volunteers or patients, so the problems change and the impact of various aspects of the study change. For example, our study was relatively small in the number of staff and volunteers. This creates the potential

for significant issues relating to staff or volunteer illness, or the ability to attend for scans or meetings and the ability to cross multidisciplinary boundaries.

Consider the impact on the study as a whole of a volunteer or member of staff being unavailable and cannot make a scan - how much redundancy do you have? Do you have any contingency? If a member of staff is overloaded with work, how well might they perform their part of the study or scan procedure if they have to, for example, leave their family out of hours, or are called upon to attend a clinical emergency.

What contingency plans do you have if a member of staff has to leave the project? Will there be someone who can cover their role? During the Calibrain multicentre study, two members of staff left their posts, and were not replaced within the time frame of the study. Additionally another member of the team had a long term absence, and one member of the project that was to be recruited was never recruited. Fortunately some of these tasks were completed by the remaining staff, and by greatly extending the time frame of the original study (by 4 years), but additional funds were needed. Other components, primarily analyses, were not completed to their planned conclusion.

As a study grows in the number of staff involved, it would be expected that the project would be more resilient to loss of staff or expertise. As the study tasks or logistics grow in complexity, it would be expected that the numbers of problems in generating and gathering valid data would increase.

7.9.5 Potential applications

Structural and functional MRI have an important role to play in the diagnosis of neuropsychiatric disorders, while objective brain imaging biomarkers of disease and trial outcomes are likely to prove useful in the development and evaluation of new treatments. Large multicentre clinical studies are common in other areas of medical research, but most neuroimaging studies so far have been restricted to using small numbers of participants and are based in one imaging centre. Sufficient numbers of clinical populations are difficult to recruit in one location, an issue that becomes even more important in studies seeking to combine neuroimaging with genetics research. The ability to conduct multicentre studies would increase flexibility in recruitment and the potential increase in statistical power offered by multicentre studies (Suckling et al., 2008) would allow the detection of more subtle differences between groups. Multicentre imaging studies have great appeal as a means of generating large datasets relatively

quickly and are complementary to initiatives that seek to develop clinical research networks.

Developments in this area would benefit other kinds of clinical research employing neuroimaging as well. Bosnell et al. (2008) conducted a multicentre study investigating the reproducibility of fMRI activation in Multiple Sclerosis patients and healthy controls, with a view towards using fMRI in a clinical therapeutic trial. Other areas of research using neuroimaging could also benefit from advances in this field, namely any research where recruitment from geographically distributed locations is desirable or even necessary. One such field is cultural neuroscience, the study of cultural differences in neurocognitive processes, where the ability to employ more than one site would greatly facilitate recruitment (Sutton et al., 2008).

7.9.6 Summary

The ability to perform multicentre neuroimaging studies is an important and exciting development which will expand the range of research possibilities in the field of psychiatry. While we are still in the early days of multicentre neuroimaging, initial findings indicate that running such studies is feasible if care is taken to employ appropriate methods and ensure consistency in data collection and analysis techniques.

A more thorough understanding of the issues involved will allow the structural and functional MRI research community to better plan future studies. These developments would put researchers in the field in a stronger position to plan and conduct novel multicentre MRI studies to improve our understanding and management of major neuropsychiatric disorders.

Appendix A

Scanning Protocol

CaliBrain Scanning Protocol Aberdeen

Scan	*Seq.*	*Orient.*	*TE*	*TR*	*TI*	*Phase FOV*	*Slice Thick.*	*Slice Gap*	*Matrix*	*FOV*	*Flip Angle*	*Options*	*NEX*	*Freq Dir*	*Band-width*	*ETL*	*No. Slices*	*EPI vols*	*Dura-tion*
phantom1	T2 FSE	Ax	102	6300	--	1	5	1.5	256x256	240	90	FC,VB,TRF, Z512	2	A/P	15.63	24	1	--	2.37
Loc	GRE	3 Plane	min	min	--	1	5	5	256x128	240	30	--	1		31.25		9	--	0.23
Field map1	EPIRT	AC/PC	30	2500	--	1	5	0	64x64	240	90	CV -> rhrcctrl=3	1	R/L	62.5	--	30	5	0.13
Field map2	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	CV -> rhrcctrl=3	1	R/L	62.5	--	30	5	0.13
Motor	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	160	6.40
Nback	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	260	10.50
Face1	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	129	5.23
Face2	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	129	5.23
Visual	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	100	4.10
Breath	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	124	5.10
T2	T2 FSE	AC/PC	102	6300	--	1	5	1.5	256x256	240	90	FC,VB,TRFZ512	2	A/P	15.63	24	19	--	2.37
T1	3DIR PREP	Cor	1.9	7.2	600	22	1.7	0	256x192	220	15	IrP, Fast	1	S/I	31.25	--	128	--	6.48
phantom2	T2 FSE	Ax	102	6300	--	1	5	1.5	256x256	240	90	FC,VB,TRFZ512	2	A/P	15.63	24	1	--	2.37

EPI Nr. of volumes and duration include 4 discarded.

CaliBrain Scanning Protocol Edinburgh

Scan	*Seq.*	*Orient.*	*TE*	*TR*	*TI*	*Phase FOV*	*Slice Thick.*	*Slice Gap*	*Matrix*	*FOV*	*Flip Angle*	*Options*	*NEX*	*Freq Dir*	*Band-width*	*ETL*	*No. Slices*	*EPI vols*	*Dura-tion*
phantom1	T2 FSE	Ax	102	6300	--	1	5	1.5	256x256	240	90	FC,VB,TRF, Z512	2	A/P	15.63	24	1	--	2.37
Loc	GRE	3 Plane	min	min	--	1	5	5	256x128	240	30	--	1		31.25		9	--	0.23
Field map1	EPIRT	AC/PC	30	2500	--	1	5	0	64x64	240	90	CV -> rhrcctrl=3	1	R/L	62.5	--	30	5	0.13
Field map2	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	CV -> rhrcctrl=3	1	R/L	62.5	--	30	5	0.13
Motor	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	160	6.40
Nback	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	260	10.50
Face1	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	129	5.23
Face2	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	129	5.23
Visual	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	100	4.10
Breath	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	124	5.10
T2	T2 FSE	AC/PC	102	6300	--	1	5	1.5	256x256	240	90	FC,VB,TRFZ512	2	A/P	15.63	24	20	--	2.37
T1	3DIR PREP	Cor	3.3	8.2	600	22	1.7	0	256x192	220	15	IrP, Fast	1	S/I	31.25	--	128	--	7.12
phantom2	T2 FSE	Ax	102	6300	--	1	5	1.5	256x256	240	90	FC,VB,TRFZ512	2	A/P	15.63	24	1	--	2.37

EPI Nr. of volumes and duration include 4 discarded.

CaliBrain Scanning Protocol Glasgow

Scan	*Seq.*	*Orient.*	*TE*	*TR*	*TI*	*Phase FOV*	*Slice Thick.*	*Slice Gap*	*Matrix*	*FOV*	*Flip Angle*	*Options*	*NEX*	*Freq Dir*	*Band-width*	*ETL*	*No. Slices*	*EPI vols*	*Dura-tion*
phantom1	T2 FSE-XL	Ax	102	6300	--	1	5	1.5	256x256	240	90	FC,VB,TRF, Z512, fast	2	A/P	15.63	24	1	--	2.44
Loc	GRE	3 Plane	min	min	--	1	5	5	256x128	240	30	--	1		31.25		9	--	0.23
Field map1	EPIRT	AC/PC	30	2500	--	1	5	0	64x64	240	90	CV -> rhrcctrl=3	1	R/L	62.5	--	30	5	0.13
Field map2	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	CV -> rhrcctrl=3	1	R/L	62.5	--	30	5	0.13
Motor	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	160	6.40
Nback	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	260	10.50
Face1	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	129	5.23
Face2	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	129	5.23
Visual	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	100	4.10
Breath	EPIRT	AC/PC	40	2500	--	1	5	0	64x64	240	90	EPI	1	R/L	62.5	--	30	124	5.10
T2	T2 FSE-XL	AC/PC	102	6300	--	1	5	1.5	256x256	240	90	FC,VB,TRFZ512	2	A/P	15.63	24	19	--	2.44
T1	3DIR PREP	Cor	1.3	5.9	600	22	1.7	0	256x192	220	15	IrP, Fast	1	S/I	31.25	--	128	--	5.59
phantom2	T2 FSE-XL	Ax	102	6300	--	1	5	1.5	256x256	240	90	FC,VB,TRFZ512, fast	2	A/P	15.63	24	1	--	2.44

EPI Nr. of volumes and duration include 4 discarded.

Appendix B

Participant Briefing/Training Procedure

1. Make sure participant fits selection criteria:
    - Right-handed
    - No history of neurological or psychiatric disorders
    - MRI safety criteria (pacemakers etc.)
2. Briefly explain study purpose and design. Participants will have 6 scans with an approx. interval of three weeks, 2 each in Edinburgh, Glasgow and Aberdeen, plus 2 more scans in Glasgow for which scheduling is more flexible.
3. Inform them that their anatomical scan will be checked by a radiologist and sent to their GP.
4. Explain the effect of coffee, alcohol and other substances on the reproducibility of the BOLD response and make sure they understand the importance of consistency for this study. Do not ask them to refrain from coffee etc., but do ask them to make sure they are consistent about it!
5. Give them the relevant documents and ask them to read them carefully. They need to sign and return the consent form before their first scan. Screening forms are included for information only. Document pack contains:
    - Information sheet
    - Volunteer consent form
    - Edinburgh screening form
    - Aberdeen screening form
    - Glasgow screening form
    - GP letter

6. Get the following information:
    - Name
    - Date of Birth
    - Address
    - Telephone/Email
7. Describe what will happen during the scan and in what order and verbally explain the functional tasks.
8. Do a run through the tasks using the ‘CaliBrain Practice’ version. Save the logs using participant initials.

Appendix C

Motor task publication

NeuroImage 49 (2010) 552–560

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Functional Magnetic Resonance Imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task

Viktoria-Eleni Gountouna [a,1], Dominic E. Job [a,*,1], Andrew M. McIntosh [a], T. William J. Moorhead [a], G. Katherine L. Lymer [a], Heather C. Whalley [a], Jeremy Hall [a], Gordon D. Waiter [e], David Brennan [f], David J. McGonigle [i], Trevor S. Ahearn [e], Jonathan Cavanagh [g], Barrie Condon [f], Donald. M. Hadley [h], Ian Marshall [c], Alison D. Murray [e], J. Douglas Steele [d], Joanna M. Wardlaw [b], Stephen M. Lawrie [a]

[a] Division of Psychiatry, The University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, EH10 5HF, UK
[b] Division of Clinical Neurosciences, University of Edinburgh, UK
[c] Department of Medical Physics, University of Edinburgh, UK
[d] Centre for Neuroscience, Division of Medical Sciences, University of Dundee, UK
[e] Department Radiology, University of Aberdeen, UK
[f] Department of Clinical Physics, Division of Community Based Sciences, Faculty of Medicine, University of Glasgow, UK
[g] Sackler Institute of Psychobiological Research, Division of Community Based Sciences, Faculty of Medicine, University of Glasgow, UK
[h] Department of Neurosciences and Clinical Radiology, University of Glasgow, UK
[i] Schools of Psychology and Biosciences, University of Cardiff, UK (previously Centre for Functional Imaging Studies, Western General Hospital, Edinburgh, UK)

ARTICLE INFO

Article history:
Received 22 September 2008
Revised 10 July 2009
Accepted 13 July 2009
Available online 23 July 2009

Keywords:
fMRI
CaliBrain
Multicentre
Scanner harmonisation
Motor
Reproducibility

ABSTRACT

Multicentre MRI studies offer great potential to increase study power and flexibility, but it is not yet clear how reproducible the results from multiple centres may be. Here we present results from the multicentre study 'CaliBrain', examining the reproducibility of fMRI data within and between three sites. Fourteen subjects were scanned twice on three 1.5 T GE scanners using an identical scanning protocol. We present data from a motor task with three conditions, sequential and random finger tapping and rest. Similar activation maps were obtained for each site and visit; brain areas consistently activated during the task included the premotor, primary motor and supplementary motor areas, the striatum and cerebellum. Reproducibility was evaluated within and between sites by comparing the extent and spatial agreement of activation maps at both the subject and group levels. The results were within the range previously reported for similar tasks on single scanners and both measures were found to be comparable within and between sites, with between site reproducibility similar to the within site measures. A variance components analysis was used to examine the effects of site, subject and visit. The contributions of site and visit were small and reproducibility was similar between and within sites, whereas the variance between subjects, and unexplained variance was large. These findings suggest that we can have confidence in combined results from multicentre fMRI studies, at least when a consistent protocol is followed on similar machines in all participating scanning sites and care is taken to select homogeneous subject groups.

Introduction

Magnetic Resonance Imaging (MRI) techniques have made major contributions to the understanding of the brain in health and disease. MRI is non-invasive, allows the examination of brain structure and function, and is now widely available. Structural and functional MRI have an important role to play in the diagnosis of neuropsychiatric disorders, while objective brain imaging biomarkers of disease and trial outcomes are likely to prove useful in the development and evaluation of new treatments. Most existing studies however tend use small numbers of participants and are based in one imaging centre. Multicentre imaging studies have great appeal as a means of generating large datasets relatively quickly and are complementary to initiatives that seek to develop clinical research networks. Nevertheless, several technical and methodological issues need to be addressed before data from different scanners can be confidently combined to increase study power. In order to understand the potential impact of the contribution of the inclusion of different scanning systems, assessments need to be made in the context of what is currently known about reproducibility of MRI at a single site.

Comparatively little work has been done on the test–retest reproducibility of fMRI within a single scanner, and even less on the

* Corresponding author. Fax: +44 131 537 6531.
E-mail address: djob@staffmail.ed.ac.uk (D.E. Job).
[1] Shared first authorship.

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.07.026

reproducibility of findings between different systems. The most widely used measures are those introduced by Rombouts (1998), examining the stability of activation extent and spatial agreement of statistical parametric maps (e.g. Yoo, 2005; Vlieger et al., 2003, Harrington et al., 2006; Fernandez et al., 2003; Brannen et al., 2001; Machielsen et al., 2000; Nybakken et al., 2002). Other approaches to reproducibility have also been used, highlighting different aspects of the issue, such as Intraclass Correlation Coefficients (Specht et al., 2003; Aron et al. 2006; Zou et al., 2005), ROC curves (Le and Hu, 1997; Manoach et al., 2001), kappa (Le and Hu, 1997; Thirion et al., 2007), the Coefficient of Variation (Marshall et al., 2004), and the extra-sum-of-squares *F* statistic (McGonigle et al., 2000), amongst others.

Regardless of the method used, it is consistently found that regional patterns of activation are qualitatively repeatable within a single scanner but are quantitatively of high variability, both within (McGonigle et al., 2000) and between (Marshall et al., 2004) individual subjects. Within session reproducibility of results tends to be better than that of sessions performed on different days. For example, Yoo et al. (2005) examined the long-term reproducibility of motor activation over more than one year and found it to be comparable to that of shorter intervals, but within-session agreement was considerably higher. As predicted by Bernoulli's Theorem (the law of large numbers), the reproducibility of activation maps is shown to be better at the group level than at the subject level (Yoo et al., 2005; Chee et al., 2003).

Several studies have looked at inter-site reproducibility (Casey et al., 1998; Vlieger et al. 2003; Voyvodic, 2003; Zou et al., 2005; Costafreda et al., 2007; Friedman et al., 2008, Suckling et al., 2007). Casey et al. (1998) qualitatively assessed the reproducibility of a working memory task across four different 1.5 T scanners in different subject groups and found good agreement in the patterns of activation. Vlieger et al. (2003) examined the reproducibility of a visual task within and between two scanners of the same field strength and found inter-scanner and within-scanner reproducibility ratios to be comparable. Voyvodic (2006) investigated the reproducibility of a hand motor task examining the effect of scanning sequence (gradient echo vs. spiral) and field strength (1.5 T and 4 T) and found that while activation level and spatial extent varied, location was found to be consistent. Zou et al. (2005) examined the effect of many factors, including field strength, manufacturer, subject and visit, on the reproducibility of activation extent in a sensorimotor task and found subject, field strength and *k*-space differences to have a significant impact on reproducibility. Friedman et al. (2008), Costafreda et al. (2007) and Suckling et al. (2007), employed a variance components analysis to examine the relative contributions of site and other factors to the variance in multicentre datasets. They found the effect of site to be small compared to that of subject and residual unexplained variance.

The studies that have systematically examined the reproducibility of fMRI within and between sites have typically employed small numbers of subjects and therefore do not readily lend themselves to the examination of the reproducibility of results at the group level. We have recently completed data collection to assess the feasibility of multicentre fMRI in a detailed study of the performance of MR scanners in three research centres in Scotland, under the name 'CaliBrain'. The participating institutions are the Universities of Aberdeen, Edinburgh and Glasgow. The dataset consists of a high resolution brain volume and fMRI scans during five separate tasks in fourteen subjects, each of whom had scans on two occasions on each of the three 1.5 T MRI scanners at roughly fortnightly intervals. The primary aim of the current investigation was to assess the reproducibility of whole brain activation maps at the subject and group levels, within and between sites. In selected regions of interest we also examined the respective contributions of the factors *site*, *visit* and *subject* to the variance and assessed reproducibility using Intraclass Correlation Coefficients (ICC).

Methods

Participants

Fourteen healthy participants (ten male, mean age 36.3 years, age range 25–51 years) participated in the study. All participants were native English speakers, right-handed (self reported), met the standard MRI safety criteria and had no history of diagnosed neurological disorder, major psychiatric disorder or treatment with psychotropic medication, including substance misuse. All participants provided written informed consent and the study was approved by the appropriate research ethics committee.

All participants were made aware of the effect of certain substances, like alcohol, caffeine and nicotine on the reproducibility of results. They were not asked to refrain from consumption, but to be consistent on the days of scanning and the night before, and all scanning was performed in the morning. All subjects were trained on all fMRI tasks, by practicing shorter versions of the tasks on a laptop on a different day from that of their first scan. Training was conducted locally at the three sites, so to ensure consistency a detailed briefing and training protocol was devised.

Design

As a first step towards multicentre imaging in Scotland, we sought to assess the performance of the existing MRI research systems with the aim of using this information to optimise and develop future acquisition and post-processing structural and functional MRI protocols.

All participants had six scanning sessions in total, two at each of the three sites. The order of visits was counterbalanced to evenly distribute practice effects. Each session was identical and consisted of a localiser, six functional runs, a T_2-weighted structural, and a high resolution anatomical T_1-weighted volume. All scanning parameters were kept constant across scanners, with some minor variations arising due to hardware and software differences.

The finger tapping task employed a block design with three conditions, 'sequential tapping', 'random tapping' and 'rest'. For the sequential condition, participants were instructed to tap the fingers of their right hand sequentially in time with a flashing '#' symbol, starting with the thumb and finishing with the little finger. For the random condition participants were asked to tap their fingers in a random way in time with a flashing '?' symbol. During the rest condition they were asked to fixate on a flashing '+' symbol. In all conditions the symbol was flashing with a frequency of 1 Hz. Each block had a duration of 30 s and included 28 trials (a trial being one flash of a '#', or '?') and a 2 s verbal prompt at the beginning with the words 'sequence', 'random' and 'rest' respectively. Each run included four repetitions of each tapping condition and five repetitions of the rest condition.

Data acquisition

Three General Electric (GE) 1.5 T scanners were used in this study, with differences in hardware and software versions reflecting scanner age. At site A scanning was conducted with a GE 1.5 T Signa NVi/CVi scanner (software version 9.1; gradients with max. amplitude 40 mT/m and max. slew rate 150 T/m/s; standard quadrature head coil). The ADW console (version 4.1) was used for the acquisition and reconstruction of fMRI data. A projector and Presentation v9.9 (Neurobehavioural Systems) were used for the presentation of stimuli. At site B scanning was conducted with a GE 1.5 T Signa LX scanner (software version 9. 1M4; Echospeed gradients with max. amplitude 22 mT/m and max. slew rate 120T/m/s; standard quadrature head coil). The ADW console (custom installation) was used for the acquisition and reconstruction of fMRI data. The delivery

of stimuli was handled by IFIS-SA (Invivo), IFIS software vR14 with E-Prime v1.1SP3 (Psychology Software Tools). Stimuli were presented using an LCD screen, part of the IFIS system, mounted on the head coil. At site C scanning was conducted with a GE 1.5 T Signa scanner (software version 11M3/11M4SP1; gradients with max. amplitude 40 mT/m and max. slew rate 150 T/m/s; standard quadrature head coil). The main console was used for the acquisition and reconstruction of fMRI data. A projector and Presentation v9.9 (Neurobehavioural Systems) were used for the presentation of stimuli.

A localiser scan in three planes was followed by six functional runs, performed using an echo planar imaging (EPI) sequence. These were acquired with the following parameters: orientation parallel to the AC/PC plane; repetition time (TR) 2500 ms; echo time (TE) 40 ms; slice thickness 5 mm without a gap; matrix 64×64; field of view (FOV) 240 mm^2; flip angle 90°; 30 slices. 160 volumes were acquired for the finger tapping task. The first four volumes of each functional run were discarded to allow for the equilibration of T1 signal effects.

Data analysis

Motor fMRI

Due to differences in the way EPI data acquisition was handled at the three sites, the raw data came in different formats. The raw data from sites A and B were acquired using the ADW console and were converted to SPM compatible ANALYZE format using the GE2SPM SPM extension, version 3.1 (Souheil Inati, http://www.fil.ion.ucl.ac.uk/spm/ext/). At site C EPI data was acquired using the main console and was converted to SPM compatible ANALYZE format using the SPM2 DICOM toolbox.

The fMRI data for the motor task was preprocessed using SPM2 (Wellcome Department of Cognitive Neurology and collaborators, Institute of Neurology, London, UK; http://www.fil.ion.ucl.ac.uk/spm/software/spm2/) using Matlab Version 6.5 R13 SP1 (Mathworks, Natick, MA, USA), on a Dell Precision 690 workstation with RedHat Enterprise Linux WS v. 4. EPI volumes were realigned to the mean image in the series using a rigid body transformation. Our pre-established exclusion criteria were a correlation between movement parameters and task regressors greater than 0.5, or excessive movement of greater than 3 mm in less than 20 volumes. These criteria represent our 'best practice', based on many previous fMRI studies (e.g. Whalley et al., 2007), as it is much easier to correct for slow drift in movement than a large single change. No subjects had to be excluded by these criteria. The images were then normalised to the standard SPM2 MNI EPI template. Normalisation parameters were estimated using the mean image for each run and these were applied to all volumes of that run. A linear affine transformation was applied followed by non-linear deformations using the SPM2 default parameters. Normalised images were spatially smoothed with an 8 mm^3 full width half maximum (FWHM) Gaussian kernel, resulting in a final estimated residual smoothness of approximately 12 mm FWHM. No substantial differences in residual smoothness that could be due to scanner differences, such as the *k*-space filters used, were seen.

Statistical analysis at the subject level was performed using the General Linear Model as implemented in SPM2; the default settings were used unless otherwise specified. The design matrix included three conditions, random tapping, sequential tapping and rest. The six movement parameters estimated in the realignment step were also included as covariates of no interest. The canonical haemodynamic response function (hrf) was convolved with the regressors to model the data. A high-pass filter with a 180 s cut-off was applied to remove low-frequency components. Serial autocorrelations were modelled using AR(1). A custom brain mask was used in all subject-level analyses to exclude areas of non-brain tissue.

To assess data quality the raw and preprocessed images and the 1st-level *T*-maps were checked visually and quantitatively using the ArtRepair toolbox version 2.1 (http://spnl.stanford.edu/tools/ArtRepair/ArtRepair.htm). All sessions of one subject were excluded due to a signal dropout artefact in the interhemispheric fissure, caused by calcification in the falx cerebri. Part of the data for a session of another subject was lost due to hardware problems, so the entire session was excluded. One session of a third subject was excluded due to a large and abrupt increase in mean signal several volumes into the scan (total number of useable sessions, $n=81$).

Contrast images for the sequential tapping vs. rest and random tapping vs. rest conditions were taken forward to random-effects group analyses. The first and second visits to the three scanners were treated separately and six one-sample *T*-tests were performed, each representing a single subject's session on one occasion at one of the centres. All statistical maps were thresholded at a voxel level of $p<0.001$ uncorrected using a 20 voxel cluster extent threshold. Clusters were deemed significant at $p<0.05$ corrected for multiple comparisons.

Reproducibility

To assess the reproducibility of statistical parametric maps (SPM *T*-maps) the widely used overlap and size measures (OS) were employed (Rombouts et al., 1998): (1) the *overlap* ratio $R^{ij}{}_{\text{overlap}} = 2*V^{ij}{}_{\text{overlap}} / (V_i \cup V_j)$, where V_i and V_j represent the sets of supra-threshold voxels in *T*-map images *i* and *j* respectively, and $V^{ij}{}_{\text{overlap}} = V_i \cap V_j$, the intersection of both maps; and (2) the *size* ratio $R^{ij}{}_{\text{size}} = 2*V^{ij}{}_{\text{smallest}} / (V_i \cup V_j)$, where V_i and V_j are defined as above and $V^{ij}{}_{\text{smallest}}$ is the smallest of the two volumes compared. Values for both ratios range between 0 and 1, with 1 indicating perfect agreement. The maps were thresholded at a voxel-wise level of $p<0.001$ uncorrected and reproducibility was assessed at both the subject and the group levels, within and between scanners. Custom MATLAB scripts were written to implement the analysis.

Components of variance

To examine the impact of systematic variability contributed by different subjects, visits and scanning sites on the data, a variance component analysis was performed following the approach recommended by Friedman et al. (2008). The model fits each parameter as a random effect (with a mean and normal distribution) and implicitly assumes that the participants, scanners and visits are taken randomly from a larger population. Whether the effect of scanner was modelled as a fixed or random effect had little influence on the solution obtained. However, we adopted a random-effects model as this can afford some protection against the inappropriate generalization of results.

Data for this analysis was extracted from both contrast images (the numerators of the *T*-maps) and SPM t-maps in three regions of interest in the left hemisphere, the primary motor area, the supplementary motor area and the striatum. Firstly, to assess the size and overlap ratios, the images were masked so that only voxels in these areas that were significant in all group analyses, across visits and scanners, ($p<0.001$ uncorrected) were included. Secondly, to assess the size of the activations, the images were masked so that only voxels in these three regions of interest, ($p<0.001$ uncorrected) were included.

Variance components were estimated using Restricted Maximum Likelihood (REML) in SAS PROC VARCOMP (SAS version 9, Cary, NC). The following model was employed with Y_{ijk} denoting the dependent measure for visit *i*, site *j*, and subject *k*:

$$Y_{ijk} = \text{mean} + \text{site}_j + \text{visit}_i + \text{subject}_k + \text{subject}_k{*}\text{site}_j + \text{unexplained}_{ijk}.$$

More complex models including other interactions between site, visit and subject were also examined but these did not differ substantively from the simpler model, so the simpler model was used.

Reproducibility of the measurements within and between sites was calculated using Intraclass Correlation Coefficients (ICCs):

$$ICC_{within} = (VD\_subject + VD\_site + VD\_subject - by - site) / Total\ Variance$$

$$ICC_{between} = VD\_subject / Total\ Variance$$

where 'VD_subject' signifies 'variance due to subject'. While the size and overlap ratios provide useful information on the reproducibility of activation patterns at the whole brain level, this analysis allows us to examine test–retest and between site reproducibility and to get more detailed information about the effect of including data from multiple sites.

A simple General Linear Model (SPM2) voxel-wise analysis, corresponding to a factorial design with three main factors of site, subject and visit and their potential interactions was also implemented to see if there were any significant differences due to any of the above factors or their interaction.

Results

Group-level activations

The sequential tapping vs. rest and random tapping vs. rest contrasts were examined separately. Thirteen datasets were included in the analyses for sites A and B and twelve in the analyses for site C. Data were grouped by scanning site and visit in that site, resulting in six one-sample *T*-tests per contrast, presented below.

The sequential tapping vs. rest contrast demonstrated activations in areas commonly associated with the task in all analyses, including peaks in the precentral and postcentral gyri, the superior and inferior parietal lobules, the basal ganglia, and the cerebellum. The six group analyses for this contrast are presented in Fig. 1. Strong activation in the left precentral and postcentral gyri was present in all analyses, with peaks in the premotor, primary motor and supplementary motor areas. Smaller foci of activation were observed in the right precentral and postcentral gyri. Consistent activations were also noted on the superior and inferior banks of the anterior sylvian fissure and the insula bilaterally. Activation peaks in the left striatum and thalamus were noted in all visits, while activations in their right-hemispheric counterparts were mostly present in the second visits. Distinct peaks were also noted consistently in the left inferior parietal lobule, with the activation in some cases spreading to the superior parietal lobule, while activations in the right inferior parietal lobule were weaker and less consistent. Activation clusters were also present in the right inferior occipital gyrus, sometimes extending into the middle occipital gyrus. Bilateral cerebellar activations were noted in all visits.

The random vs. rest contrast demonstrated activations in regions associated with this finger tapping task including clusters in the frontal and parietal lobes, the basal ganglia and the cerebellum, demonstrating a very similar pattern to the sequential tapping contrast. Overall, stronger activations were observed in this condition. Some additional areas were also observed, bilaterally in the dorsolateral prefrontal cortex. The six group analyses for this contrast are presented in Fig. 2.

Overlap and size ratios

Overlap and size ratios within and between sites were similar for all analyses. Mean size ratios ranged from 0.60 to 0.75 within sites and 0.65 to 0.73 between sites for the single subject analyses. Mean overlap ratios for the single subject analyses ranged from 0.41 to 0.50 within sites and from 0.44 to 0.48 between sites. The reproducibility of group maps was generally higher. Size ratios for group maps ranged from 0.73 to 0.97 within sites and from 0.71 to 0.95 between sites. Overlap ratios for group maps ranged from 0.55 to 0.67 within sites and from 0.58 to 0.67 between sites. A summary of this data, with means across subjects, is presented graphically in Fig. 3. A visual representation of the activations and overlap of all group maps for each scanner in the random tapping vs. rest contrast is shown in Fig. 4.

Components of variance

The percentage of total variance in our three regions (significant in all group analyses) contributed by the site component varied between 0% and 13.3% across all analyses, while that of visit varied between 0% and 6.3%. The contribution of the subject by site interaction ranged from 0% to 14.3% while that of subject alone was larger, varying between 15.6% and 61.0% with unexplained variance from 23.0 to 81.0%. The analysis of contrast images yielded ICCs within sites from 0.23 for the striatum to 0.72 for the precentral ROI, and ICCs between sites from 0.23 for the striatum to 0.61 for the precentral ROI. The equivalent analysis of *T*-statistic images gave lower reproducibility values, within sites from 0.17 for the striatum, to 0.55 for the precentral ROI, and between sites from 0.16 for the striatum, to 0.54 for the precentral ROI. These results are presented in Table 1.

The size of the activations in our three regions of interest, percentage of total variance contributed by the site component varied between 0% and 10.9%, while that of visit varied between 0% and 22.0% across all analyses. The contribution of the subject by site interaction ranged from 0% to 21.5% while that of subject alone was larger, varying between 4.6% and 65.7% with unexplained variance from 14.6% to 77.1%. The analysis of contrast images yielded ICCs within sites from 0.10 for the striatum to 0.83 for the precentral ROI, and ICCs between sites from 0.09 for the striatum to 0.66 for the precentral ROI. The equivalent analysis of *T*-statistic images gave lower reproducibility values, within sites from 0.07 for the striatum, to 0.50 for the precentral ROI, and between sites from 0.05 for the striatum, to 0.50 for the precentral ROI. These results are presented in Table 2.

As expected the results for the size measures, taking only volume without intensity into account, show much lower values. These 'volume of activation' results are given in Supplementary Table 1.

Our General Linear Model (SPM2) voxel-wise analysis, corresponding to a factorial design with three main factors of site, subject and visit and their potential interactions showed no significant results due to site or visit, but as expected, extensive variance between subjects. The interactions between factors did not provide additional results.

Discussion

We investigated the reproducibility of a sequential and a random motor tapping task across multiple sites and visits, examining it both in terms of single subjects and group analyses. We found consistent activations in expected areas, including cortical and subcortical motor areas and the cerebellum. As anticipated, comparison ratios were overall higher and more stable in the group-level analyses than at the subject level. Reproducibility between sites was similar to that of different visits within the same site. Reproducibility in overlap and size estimates were similar between and within sites and acceptable at both group and subject levels.

Robust activations were observed in the left premotor, primary motor and supplementary motor areas. These were present across all sites and visits. Right-hemispheric counterparts were also activated but are weaker and less consistent. The opposite pattern was observed in the cerebellum, with consistent ipsilateral but weaker contralateral activations. Left thalamus and basal ganglia were also detected in most cases. Mattay et al. (1998) employed a sequential and random tapping task and reported activations in the primary motor, somatosensory and premotor areas, the SMA, parietal cortex, putamen and cerebellum. They also reported a prefrontal cluster (BA 9) in

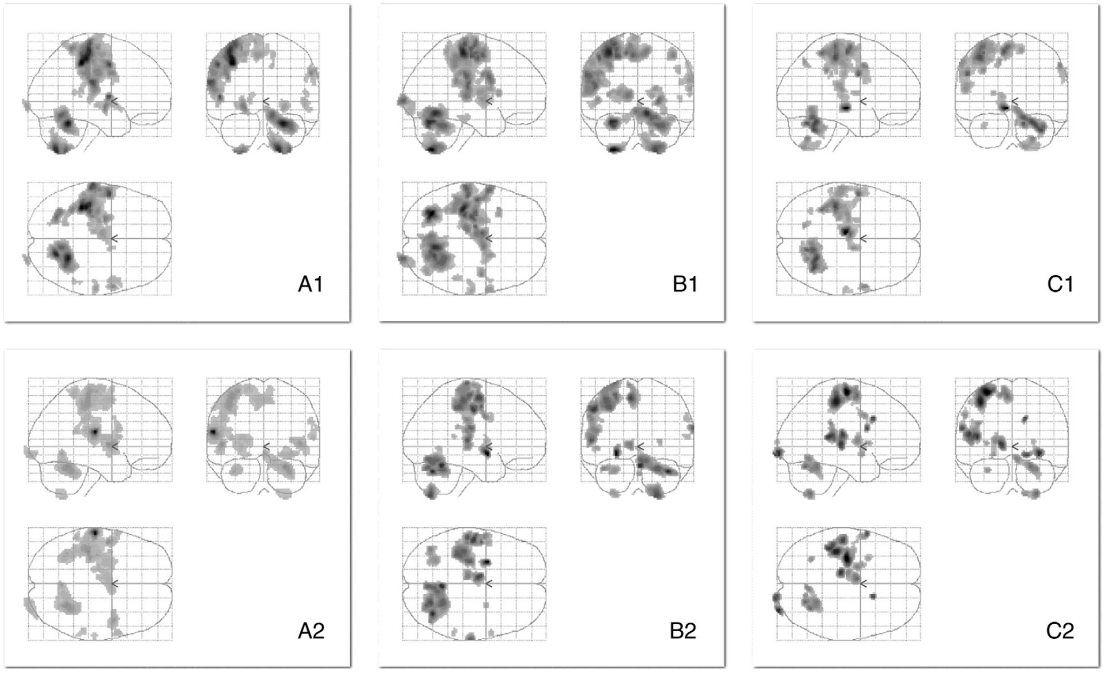

Fig. 1. Group maps, using neurological convention, for the sequential tapping vs. rest contrast, with a threshold at a voxel level of $p<0.0001$ uncorrected and using a 20 voxel cluster extent threshold, for illustration purposes. From left to right: sites A, B and C. Top: visit 1. Bottom: visit 2.

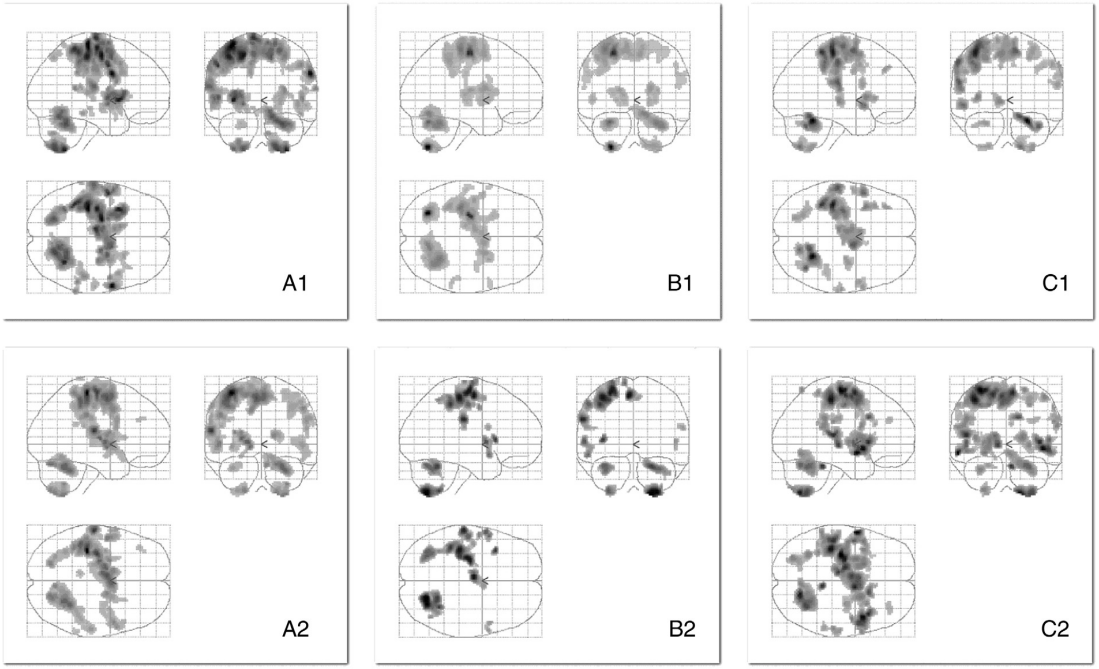

Fig. 2. Group maps, using neurological convention, for the random tapping vs. rest contrast, with a threshold at a voxel level of $p<0.0001$ uncorrected and using a 20 voxel cluster extent threshold, for illustration purposes. From left to right: sites A, B and C. Top: visit 1. Bottom: visit 2.

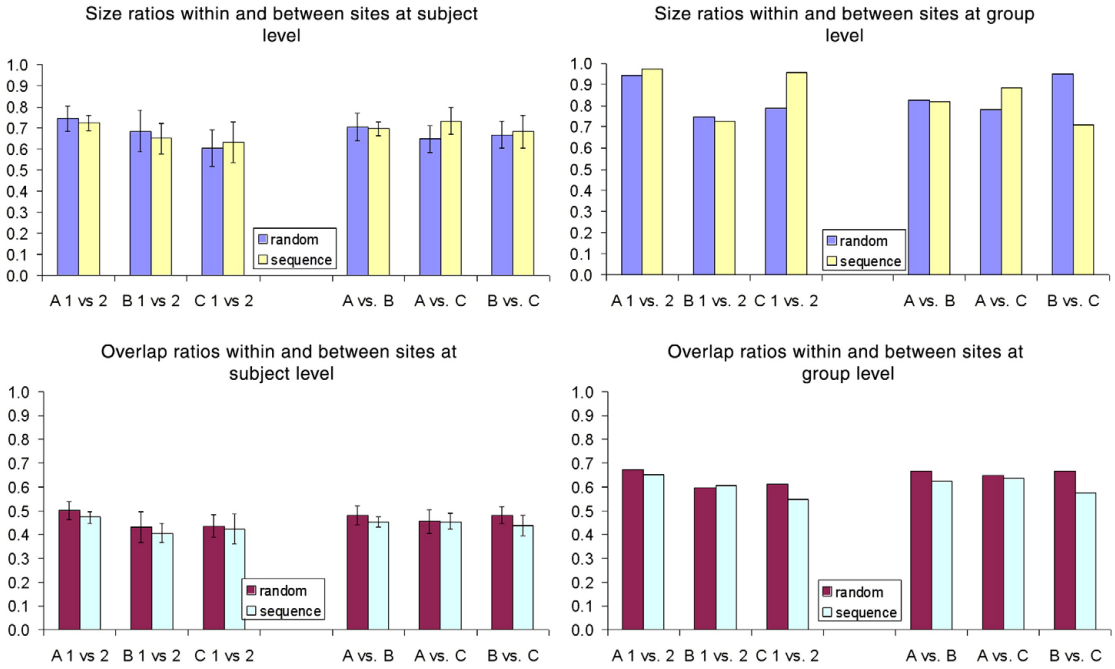

Fig. 3. Reproducibility of statistical parametric maps within and between sites at the subject and group levels for the random vs. rest and sequence vs. rest contrasts. Values for the size and overlap ratio measures range from 0 to 1 and are means across subjects. Columns labelled A–C 1 vs. 2 show within site comparisons. Columns labelled A vs. B, A vs. C and B vs. C show between site comparisons. Top: size ratios. Bottom: overlap ratios. Left: mean ratios for all subjects with standard error bars. Right: ratios for the group-level comparisons.

the random tapping condition, which was detected in some of our analyses. Yoo et al. (2005) investigated the reproducibility of a sequential finger tapping task over a longer period of time, also employing size and overlap ratios in selected regions of interest. They found consistent activations in the primary motor, premotor, SMA and cerebellum; activations were less consistent in the basal ganglia and thalamus. These findings are consistent with our results and suggest that in these smaller regions increased noise has a negative effect on the reproducibility of activation.

A qualitative evaluation of the results indicates that at the group level, location and extent of activation is robust, both within and between sites. Quantitatively the results were within the range reported by studies examining the reproducibility of similar tasks on a single scanner. Furthermore both size and overlap ratios, and size by volume and size by volume and intensity, were found to be comparable within and between sites. Other studies investigating between scanner reproducibility also did not find large differences in comparisons within and between sites. Voyvodic et al. (2006) reported stable spatial patterns even between scanners of different strengths. Vlieger et al. (2003) found that the average inter-scanner agreement did not differ significantly from the average within-scanner reproducibility of the site with the worst reproducibility.

Much of the reproducibility literature to date has been focussed on the subject level, but in many cases it is the group-level results that are given weight in the interpretation of MRI results. Our analyses indicate, in agreement with the law of large numbers, that group-level maps are overall more reproducible, and the range of obtained values is much narrower. This is in agreement with what has been reported in a single centre analysis (Seghier et al., 2004). Critically, subject is a much greater source of variability than scanner, however subject by scanner variance is low, and within subject variance (visit) is also low and of an acceptable level. Merging of data across sites is therefore

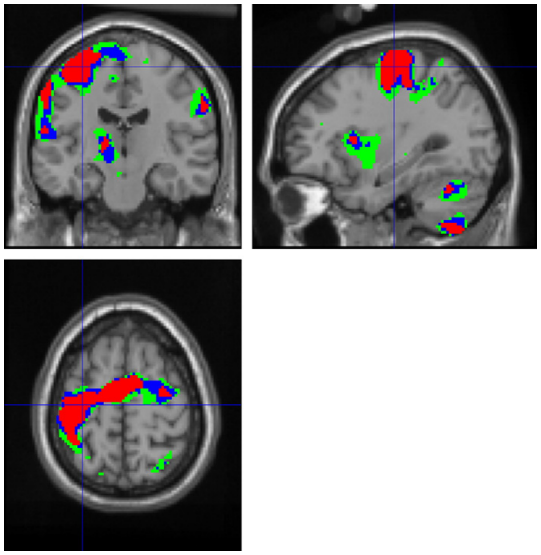

Fig. 4. Orthogonal slices showing the activations and overlap of all group-level maps for each scanner using neurological convention, at $T \geq 5.0$, in the random tapping vs. rest contrast. The cross hairs are at $x = -31$, $y = -18$, $z = 62$ (MNI coordinates, mm). Activations where all 3 scanners overlap are shown in red, where two of the scanners overlap activations are shown in blue, and where only one scanner has activated are shown in green areas.

Table 1
Percentage of total variance and reproducibility estimates.

Anat. area	Site	Visit	Subject	Subject by site	Error	ICC within	ICC between
Random vs. rest							
Contrast mean							
Precentral	0.00	4.97	57.76	14.29	22.98	0.72 (g)	0.58 (f)
SMA	13.31	0.00	49.43	0.76	36.50	0.63 (g)	0.49 (f)
Striatum	0.00	3.28	31.15	0.00	65.57	0.31 (p)	0.31 (p)
T-statistic mean							
Precentral	0.00	6.35	50.62	0.00	43.04	0.51 (f)	0.51 (f)
SMA	8.00	0.00	39.12	0.00	52.88	0.47 (f)	0.39 (p)
Striatum	2.71	2.46	15.57	0.00	79.26	0.18 (p)	0.16 (p)
Sequence vs. rest							
Contrast mean							
Precentral	1.39	4.53	60.98	10.11	23.00	0.72 (g)	0.61 (g)
SMA	3.61	3.61	54.22	0.00	38.55	0.58 (f)	0.54 (f)
Striatum	0.00	3.57	23.21	0.00	73.21	0.23 (p)	0.23 (p)
T-statistic mean							
Precentral	0.69	5.89	54.04	0.00	39.38	0.55 (f)	0.54 (f)
SMA	0.00	2.68	41.31	0.00	56.01	0.41 (f)	0.41 (f)
Striatum	0.00	1.81	17.28	0.00	80.91	0.17 (p)	0.17 (p)

Variance components as percentage of total variance contributed by site, visit, subject, subject by site interaction and unexplained variance. Intraclass Correlation Coefficients (ICCs) within and between sites. According to a priori criteria, (Cicchetti, 2001), we use 'p', 'f', and 'g' to refer to 'poor', 'fair' and 'good', respectively, for the ICC values.

possible in studies where the reproducibility of individual subject measurements is essential, e.g. in treatment response studies.

This outcome is supported by the results of the variance components analysis. The contribution of the factors site and visit to the total variance was relatively small regardless of measure used, and subject was the largest component. However a large part of the variance remained unexplained in some of the analyses, notably in the striatum, presumably due to the small volume of the region, although this variability could also reflect a difference in the striatum's more complex neurovascular dynamics and the task used.

Friedman et al. (2008), reported a variance components analysis examining the effects of site, subject and visit in a dataset of five subjects who conducted repeated visits in ten sites. Costafreda et al. (2007) also employed a variance components analysis to examine the relative contributions of site and subject to the variance in a motor task in five subjects across five scanners. Both found the effect of site to be small compared to that of subject and residual unexplained variance. Suckling et al. (2007) performed a similar study employing two versions of an affect processing task in twelve subjects and two sites. They examined the contributions of site, visit, task and subject and also found the contribution of site to the variance to be relatively small.

It is of note that little site effect was observed in the striatal ROI in any of the analyses performed. It is possible that the location of this structure near the middle of the brain renders it less vulnerable to scanner specific distortions and therefore makes it a suitable target area for future multicentre studies of clinical interest. However, its small size means it is very susceptible to noise. A lot of the variance remained unexplained in these analyses and reproducibility in overlap and size of activation was found to be poor in comparison to the larger neocortical ROIs. ICC values for these analyses are similar or higher than those reported elsewhere (Friedman et al., 2008; Kong et al. 2007).

The reproducibility in overlap and size ratios for both mean and maximum measures were comparable across all analyses with no clear advantage of the one over the other (Supplementary Table 2). A similar pattern was found by Friedman et al. (2008) who compared median and maximum values. A clear difference however emerges when comparing the reproducibility in overlap and size ratios in contrast and *T*-statistic images, with the reproducibility in contrast values being consistently higher. This was hypothesised by Friedman et al., (2008) who used a finite impulse response (FIR) method to get a surrogate value for signal change and suggested the method employed here as a possible alternative. ICCs were very similar within and between sites, consistent with the results of the whole brain size and overlap ratios.

Strengths and limitations

The strengths of the present study include the use of a consistent briefing protocol. The scanners used were of the same field strength and by the same manufacturer. We were able to employ identical head coils and any differences between scanning sequences were negligible. We made an effort to control for systematic differences between sites and reduce variability; however, some differences in hardware and software, as well as more subtle sources of variation, e.g. site-specific practices of the scanning staff, and differences in the displays used, could not be avoided. These are inevitable and realistic in relation to future multicentre studies. Practice effects were distributed evenly by the counterbalancing of visits. A standard analysis protocol was used, with no steps added to account for inter-site differences. Future developments in this area are likely to improve reproducibility further.

One limitation of the reproducibility measures employed is their sensitivity to the threshold used and the fact that they only give information about extent, not peak or amplitude. However, some indication of the effect of threshold and amplitude can be seen in the differences between the size by volume and the size by volume and intensity measures, in Supplementary Tables 1 and 2 respectively. It can be seen that the volume by intensity measures are significantly more reproducible than the size by volume measure alone. An analysis of all thresholds would be required to confirm this apparent effect. However, this still leaves the possibility that between centre factors, such as peak and amplitude, could increase the total variance sufficiently to obscure any effects, e.g. type II errors. Further to this, it is possible that one site could contribute much more to the total signal so as to bias the results, leading to type I errors. However, activation location and extent are important considerations in multicentre studies and it is possible to interpret the findings in relation to what is known about the reproducibility of activation maps within a single centre. Despite the coarse level of information, evaluation of these

Table 2
Size by volume and intensity, reproducibility estimates.

Anat. area	Site	Visit	Subject	Subject by site	Error	ICC within	ICC between
Random vs. rest							
Contrast mean							
Precentral	0.00	2.65	61.68	21.05	14.61	0.83 (e)	0.62 (g)
SMA	10.94	0.00	62.34	0.00	26.72	0.73 (g)	0.62 (g)
Striatum	1.11	19.19	8.73	0.00	70.98	0.10 (p)	0.09 (p)
T-statistic mean							
Precentral	0.00	6.38	49.55	0.00	44.08	0.50 (f)	0.50 (f)
SMA	7.39	0.00	40.62	0.20	51.79	0.48 (f)	0.41 (f)
Striatum	2.81	21.96	4.57	0.00	70.67	0.07 (p)	0.05 (p)
Sequence vs. rest							
Contrast mean							
Precentral	0.00	2.85	65.71	11.58	19.85	0.77 (e)	0.66 (g)
SMA	4.80	3.17	60.97	0.00	31.06	0.66 (g)	0.61 (g)
Striatum	0.00	9.56	15.25	0.00	75.19	0.15 (p)	0.15 (p)
T-statistic mean							
Precentral	0.00	3.47	47.21	0.00	49.33	0.47 (f)	0.47 (f)
SMA	0.17	1.99	38.51	0.59	58.74	0.39 (p)	0.39 (p)
Striatum	0.00	5.06	17.85	0.00	77.09	0.18 (p)	0.18 (p)

Size by volume and intensity, as percentage of total variance contributed by site, visit, subject, subject by site interaction and unexplained variance. Intraclass Correlation Coefficients (ICCs) within and between sites. According to a priori criteria, (Cicchetti, 2001), we use 'p', 'f', 'g' and 'e', to refer to 'poor', 'fair', 'good' and 'excellent', respectively, for the ICC values.

measures at the group level provided useful insights regarding the level of agreement that can be expected between maps obtained from the same group of individuals on different occasions, both in single and multiple sites. The whole brain approach adopted here does not give information on specific regions; this was addressed in the subsequent variance components analysis which allowed us to examine in detail the partition of the variance and the reproducibility in overlap and size ratios of contrast images and statistical parametric maps (*T*-maps). It would be advisable to perform pilot studies and employ these or similar measures to examine potential site effects and reproducibility in overlap and size ratios in targeted brain areas when planning clinical multicentre studies.

Future work

Analysis of the data collected using other tasks will allow the evaluation of reproducibility across different cognitive domains and tasks. The further exploration of this dataset will enable us to identify the key characteristics in the acquisition and analysis of fMRI paradigms which influence their suitability for multicentre MRI studies. A more thorough understanding of the issues involved will allow the fMRI research community to better plan future studies. These developments would put researchers in the field in a stronger position to plan and conduct novel multicentre MRI studies to improve our understanding and management of major neuropsychiatric disorders.

Conclusion

The reproducibility of a finger tapping task was evaluated across three sites in 14 subjects on two visits each. Robust activations were detected in typical motor areas. Reproducibility of activation location and extent were similar for the sequential and random tapping conditions. Reproducibility was comparable within and between sites, and critically was acceptable within single subjects, the majority of variance being between subjects and in unexplained variance. The contributions of site and visit to the variance were low and reproducibility in overlap and size was similar between and within sites. However, reproducibility was poor in smaller anatomical areas and mostly fair to good in the larger areas, representative of the difficulties that face fMRI in general. This indicates that we can have confidence in the results produced by multicentre functional MRI when a consistent scanning and analysis protocol is followed, but that more work needs to be done and care taken in selecting homogeneous subject groups and suitable tasks.

Acknowledgments

VE Gountouna was supported by an MRC studentship. The CaliBrain study was funded by a Chief Scientist Office (Scotland) Project Grant (CZB/4/427), Chief Investigator Prof. S. Lawrie. One of the three participating imaging centres was the SFC Brain Imaging Research Centre (www.sbirc.ed.ac.uk).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2009.07.026.

References

Aron, A.R., Gluck, M.A., Poldrack, R.A., 2006. Long-term test–retest reliability of functional MRI in a classification learning task. NeuroImage 29 (3), 1000–1006.

Brannen, J.H., Badie, B., Moritz, C.H., Quigley, M., Meyerand, E.M., Haughton, V.M., 2001. Reliability of functional MR imaging with word-generation tasks for mapping Broca's area. Am. J. Neuroradiol. 22 (9), 1711–1718.

Casey, B.J., Cohen, J.D., O'Craven, K., Davidson, R.J., Irwin, W., Nelson, C.A., Noll, D.C., Hu, X., Lowe, M.J., Rosen, B.R., Truwitt, C.L., Turski, P.A., 1998. Reproducibility of fMRI results across four institutions using a spatial working memory task. NeuroImage 8 (3), 249–261.

Chee, M.W., Lee, H.L., Soon, C.S., Westphal, C., Venkatraman, V., 2003. Reproducibility of the word frequency effect: comparison of signal change and voxel counting. NeuroImage 18 (2), 468–482.

Cicchetti, D.V., 2001. The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. J. Clin. Exp. Neuropsychol. 23, 695–700.

Costafreda, S.G., Brammer, M.J., Vêncio, R.Z., Mourão, M.L., Portela, L.A., de Castro, C.C., Giampietro, V.P., Amaro, E., 2007. Multisite fMRI reproducibility of a motor task using identical MR systems. J. Magn. Reson. Imaging 26 (4), 1122–1126.

Fernandez, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., Klaver, P., Ruhlmann, J., Reul, J., Elger, C.E., 2003. Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. Neurology 60 (6), 969–975.

Friedman, L., Stern, H., Brown, G.G., Mathalon, D.H., Turner, J., Glover, G.H., Gollub, R.L., Lauriello, J., Lim, K.O., Cannon, T., Greve, D.N., Bockholt, H.J., Belger, A., Mueller, B., Doty, M.J., He, J., Wells, W., Smyth, P., Pieper, S., Kim, S., Kubicki, M., Vangel, M., Potkin, S.G., 2008. Test–retest and between-site reliability in a multicenter fMRI study. Hum. Brain Mapp. 29 (8), 958–972.

Harrington, G.S., Buonocore, M.H., Farias, S.T., 2006. Intrasubject reproducibility of functional mr imaging activation in language tasks. AJNR Am. J. Neuroradiol. 27 (4), 938–944.

Kong, J., Gollub, R.L., Webb, M.J., Kong, J.-T., Vangel, M.G., Kwong, K., 2007. Test–retest study of fMRI signal change evoked by electroacupuncture stimulation. NeuroImage 34 (3), 1171–1181.

Le, T.H., Hu, X., 1997. Methods for assessing accuracy and reliability in functional MRI. NMR Biomed. 10 (4–5), 160–164.

Machielsen, W.C.M., Rombouts, S.A.R.B., Barkhof, F., Scheltens, P., Witter, M.P., 2000. fMRI of visual encoding: reproducibility of activation. Hum. Brain Mapp. 9 (3), 156–164.

Manoach, D.S., Halpern, E.F., Kramer, T.S., Chang, Y., Goff, D.C., Rauch, S.L., Kennedy, D.N., Gollub, R.L., 2001. Test–retest reliability of a functional mri working memory paradigm in normal and schizophrenic subjects. Am. J. Psychiatry 158 (6), 955–958.

Marshall, I., Simonotto, E., Deary, I.J., Maclullich, A., Ebmeier, K.P., Rose, E.J., Wardlaw, J.M., Goddard, N., Chappell, F.M., 2004. Repeatability of motor and working-memory tasks in healthy older volunteers: assessment at functional MR imaging. Radiology 233 (3), 868–877.

Mattay, V.S., Callicott, J.H., Bertolino, A., Santha, A.K., Van Horn, J.D., Tallent, K.A., Frank, J.A., Weinberger, D.R., 1998. Hemispheric control of motor function: a whole brain echo planar fMRI study. Psychiatry Res.: Neuroimaging 83 (1), 7–22.

McGonigle, D.J., Howseman, A.M., Athwal, B.S., Friston, K.J., Frackowiak, R.S.J., Holmes, A.P., 2000. Variability in fMRI: an examination of intersession differences. NeuroImage 11 (6), 708–734.

Nybakken, G., Quigley, M., Moritz, C., Cordes, D., Haughton, V., Meyerand, M., 2002. Test–retest precision of functional magnetic resonance imaging processed with independent component analysis. Neuroradiology 44 (5), 403–406.

Rombouts, S.A., Barkhof, F., Hoogenraad, F.G., Sprenger, M., Scheltens, P., 1998. Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. Magn. Reson. Imaging 16 (2), 105–113.

Specht, K., Willmes, K., Shah, J.N., Jäncke, L., 2003. Assessment of reliability in functional imaging studies. J. Magn. Reson. Imaging 17 (4), 463–471.

Suckling, J., Ohlssen, D., Andrew, C., Johnson, G., Williams, S.C.R., Graves, M., Chen, C.H., Spiegelhalter, D., Bullmore, E., 2007. Components of variance in a multicentre functional MRI study and implications for calculation of statistical power. Hum. Brain Mapp. 1111–1122 Published Online: Aug 6 2007.

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., Poline, J.B., 2007. Analysis of a large fMRI cohort: statistical and methodological issues for group analyses. NeuroImage 35 (1), 105–120.

Vlieger, E.J., Lavini, C., Majoie, C.B., den Heeten, G.J., 2003. Reproducibility of functional MR imaging results using two different MR systems. Am. J. Neuroradiol. 24 (4), 652–657.

Voyvodic, J.T., 2006. Activation mapping as a percentage of local excitation: fMRI stability within scans, between scans and across field strengths. Magn. Reson. Imaging 24 (9), 1249–1261.

Whalley, H.C., Gountouna, V.E., Hall, J., McIntosh, A., Whyte, M.C., Simonotto, E., Job, D.E., Owens, D.G., Johnstone, E.C., Lawrie, S.M., 2007. Correlations between fMRI activation and individual psychotic symptoms in un-medicated subjects at high genetic risk of schizophrenia. BMC Psychiatry 29 (7), 61.

Yoo, S.S., Wei, X., Dickey, C.C., Guttmann, C.R., Panych, L.P., 2005. Long-term reproducibility analysis of fMRI using hand motor task. Int. J. Neurosci. 115 (1), 55–77.

Zou, K.H., Greve, D.N., Wang, M., Pieper, S.D., Warfield, S.K., White, N.S., Manandhar, S., Brown, G.G., Vangel, M.G., Kikinis, R., Wells, W.M.a., 2005. Reproducibility of functional MR imaging: preliminary results of prospective multi-institutional study performed by biomedical informatics research network. Radiology 237 (3), 781–789.

Appendix D

Working memory task publication

Psychiatry Research: Neuroimaging 184 (2010) 86–95

Contents lists available at ScienceDirect

Psychiatry Research: Neuroimaging

journal homepage: www.elsevier.com/locate/psychresns

Between- and within-scanner variability in the CaliBrain study *n*-back cognitive task

Victoria Gradin[a,b], Viktoria-Eleni Gountouna[c], Gordon Waiter[d], Trevor S. Ahearn[d], David Brennan[e], Barrie Condon[e], Ian Marshall[f], David J. McGonigle[g], Alison D. Murray[a], Heather Whalley[c], Jonathan Cavanagh[h], Donald Hadley[i], Katherine Lymer[c], Andrew McIntosh[c], Thomas William Moorhead[c], Dominic Job[c], Joanna Wardlaw[j], Stephen M. Lawrie[c], John Douglas Steele[b,*]

[a] *Department of Mental Health, University of Aberdeen, UK*
[b] *Centre for Neuroscience, Division of Medical Sciences, University of Dundee, UK*
[c] *Division of Psychiatry, School of Molecular and Clinical Medicine, University of Edinburgh, UK*
[d] *Aberdeen Biomedical Imaging Centre, University of Aberdeen, UK*
[e] *Department of Clinical Physics, NHS Glasgow, UK*
[f] *Department of Medical Physics, School of Molecular and Clinical Medicine, University of Edinburgh, UK*
[g] *Schools of Psychology and Biosciences, University of Cardiff, UK*
[h] *Department of Sackler Institute of Psychobiological Research, University of Glasgow, UK*
[i] *Department of Neurosciences and Clinical Radiology, University of Glasgow, UK*
[j] *Division of Clinical Neurosciences, School of Molecular and Clinical Medicine, University of Edinburgh, UK*

ARTICLE INFO

Article history:
Received 29 August 2009
Received in revised form 15 August 2010
Accepted 19 August 2010

Keywords:
Multi-centre fMRI
Reproducibility
Reliability

ABSTRACT

Psychiatric neuroimaging techniques are likely to improve understanding of the brain in health and disease, but studies tend to be small, based in one imaging centre and of unclear generalisability. Multicentre studies have great appeal but face problems if functional magnetic resonance imaging (fMRI) data from different centres are to be combined. Fourteen healthy volunteers had two brain scans on different days at three scanners. Considerable effort was first made to use similar scanning sequences and standardise task implementation across centres. The *n*-back cognitive task was used to investigate between- and within-scanner reproducibility and reliability. Both the functional imaging and behavioural results were in good accord with the existing literature. We found no significant differences in the activation/deactivation maps between scanners, or between repeat visits to the same scanners. Between- and within-scanner reproducibility and reliability was very similar. However, the smoothness of images from the scanners differed, suggesting that smoothness equalization might further reduce inter-scanner variability. Our results for the *n*-back task suggest it is possible to acquire fMRI data from different scanners which allows pooling across centres, when the same field strength scanners are used and scanning sequences and paradigm implementations are standardised.

1. Introduction

Structural and functional magnetic resonance imaging (sMRI and fMRI) show promise in their ability to facilitate the early diagnosis of disorders such as schizophrenia and dementia (Job et al., 2006; Whalley et al., 2006), and neuroimaging biomarkers of disease and clinical outcome data are increasingly regarded as useful in the development and evaluation of new treatments. However, psychiatric imaging studies tend to be small and based in one imaging centre. Multicentre studies, involving recruitment of subjects at several scanning centres, therefore have great appeal, but data pooling faces potential problems related to differences in scanner field strength and pulse sequences (Zou et al., 2005; Friedman et al., 2008).

Before undertaking large psychiatric multi-centre fMRI studies, it is important to investigate the extent to which it is possible to obtain reproducible and reliable data, taking account of both between-scanner and within-scanner (over time) variability. If most of the variance in images were due to scanner-related differences, rather than to subject- and task-related factors of interest, imaging data would be largely scanner-dependent and of unclear generalisability (Costafreda et al., 2007). Such heterogeneity would raise questions about pooling fMRI data. Whilst a number of previous studies have focused either on within-scanner or between-scanner effects, to our knowledge, only two recent studies have examined both between- and within-scanner fMRI variability in the same subjects (Friedman et al., 2008; Suckling et al., 2008) as here. As will be discussed, a limited amount of work has been done on within-scanner reliability and less on between-scanner reproducibility. Here we report the results of an analysis of the *n*-back cognitive task (Owen et al., 2005) fMRI data, obtained as part of the Scottish 'CaliBrain' multi-centre's MRI and fMRI initiative.

* Corresponding author. Advanced Interventions Service, Area 7, Level 6, South Block, Ninewells Hospital & Medical School, Dundee, DD1 9SY. Tel.: +44 1382 632 121.
E-mail address: d.steele@dundee.ac.uk (J.D. Steele).

0925-4927/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.pscychresns.2010.08.010

Our first hypothesis was that significant patterns of brain activation would be found with the three scanners. This was a prerequisite to testing our second hypothesis, which was no significant within- and between-scanner image differences *with regard to these significant patterns of activity*. Investigation of reproducibility and reliability of brain activation is a central focus of this study. Our third hypothesis was that after attention to 'scanner harmonization' (using the same field strength scanners and implementing the task in as similar a way as possible at each scanner), reproducibility values would be in the range reported for single scanner studies, and within-scanner reliability would be similar to between-scanner reliability.

2. Methods

2.1. Subjects

The study was approved by local ethics committees. Fifteen healthy right-handed volunteers from the participating laboratories gave written informed consent. However, data for only one scanning session were obtained for one subject, so that subject was excluded from analysis. Ten subjects were male. The mean age for the whole group was 36.3, range 25–51 years. Volunteers were not taking medications or drugs that might alter brain activity and did not have a current illness or a history of a serious head injury that might alter brain function. Previous psychiatric disorder was not an exclusion criterion. The order of the two visits to each of the three scanners was counterbalanced across sites. The mean time between scans at different centres was 3 weeks and the mean time between the two visits to the same scanner was 8 weeks.

2.2. Image acquisition

Participants were not asked to be abstinent from smoking or drinking caffeine as this would be difficult to achieve in clinical fMRI studies with psychiatric patients. However, subjects were asked to be consistent in their use before each scan. As is common in psychiatric studies, all scanning was done at a similar time of day (between 10 am and mid-day). The three scanners were as follows:

'Scanner A': GE Medical Systems Signa 1.5 T NVi/CVi (software version 9.1M4; gradients with max. amplitude 40 mT/m and max. slew rate 150 T/m/s; standard quadrature head coil). A GE Medical Systems secondary console (Real Time Image Processing, RTIP) was used for the acquisition and reconstruction of the fMRI data. A projector and Presentation v0.99 (Neurobehavioural Systems) were used for the presentation of stimuli.

'Scanner B': GE Medical Systems 1.5 T Signa LX scanner (software version 9.1M4; Echospeed gradients with max. amplitude 22 mT/m and max. slew rate 120 T/m/s; standard quadrature head coil). The ADW console (custom installation) was used for the acquisition and reconstruction of fMRI data. The delivery of stimuli was handled by IFIS-SA (Invivo), IFIS software vR14 with E-Prime v1.1SP3 (Psychology Software Tools). Stimuli were presented using an LCD screen, part of the IFIS system, mounted on the head coil.

'Scanner C': GE Medical Systems Signa 1.5 T (software version 11M3/11M4SP1; gradients with max. amplitude 40 mT/m and max. slew rate 150 T/m/s; standard quadrature head coil). The main console was used for the acquisition and reconstruction of the fMRI data. A projector and Presentation v0.99 (Neurobehavioural Systems) were used for the presentation of stimuli.

At each site, images were acquired parallel to the AC-PC plane with a repetition time (TR) 2.5 s, echo time (TE) 40 ms, slice thickness 5 mm, each slice acquired contiguously, matrix 64 × 64, field of view (FOV) 240 mm^2, and a flip angle of 90º. Each of the 260 volumes for the *n*-back task was made up of 30 slices. The first four volumes of each functional run were discarded to allow the magnetization to reach a steady state.

2.3. The n-back task and behavioural data acquisition

The *n*-back working memory task was implemented as a parametric design with different *n*-back levels occurring in blocks. Subjects were presented with letters and had to respond by indicating whether the letter matched the target (an 'x' for $n=0$, the immediately previous letter for $n=1$, similarly for two or three letters back; $n=2$ and $n=3$, respectively) or not. The difficulty of the task ("cognitive load") increases with '*n*'. Each block consisted of 14 2.5-s trials (1-s stimulus, 1.5-s blank) with a total duration of 40 s, which included a 5-s prompt. There were no fixation stimuli. Four blocks starting with $n=0$ and ending with $n=3$ made up a cycle; four such cycles were repeated. Recorded behavioural measures were letter choice (allowing determination of accuracy) and reaction time.

2.4. Behavioural analysis

Reaction times were analyzed using a $3\times2\times4$ (scan-centre × visit × *n*-back level) within-subject analysis of variance (ANOVA). Accuracy was analyzed with a $2\times2\times4$ (scan-centre × visit × *n*-back level) design since only two centres (Scanner A and Scanner B) recorded the accuracy of the responses. The threshold of significance was defined as $P<0.05$.

2.5. Image analysis

Image data were converted to Analyze format and visually inspected to detect obvious abnormalities, scanner artefacts, and problems with the conversion process. Statistical Parametric Mapping (SPM2) (Friston et al., 2007) was used to implement a random effects event-related analysis with a repeat measures ANOVA at the second level. For pre-processing, images were slice time corrected, and then each time series was realigned to the first image of each series using a rigid body affine transformation. No subject had more than 5 mm translation during scanning. The average realigned image was used to derive parameters for spatial normalisation to the SPM2 template; then the parameters were applied to each image in the realigned time-series. The resultant time-series realigned and spatially normalised images were smoothed with an 8 mm full width half maximum Gaussian kernel.

For the first level within-subject analysis, the sequence and timing of the picture presentation and behavioural response events, and the *n*-back level of difficulty of each trial, were extracted from each subject's log file. The design matrix was constructed according to a standard linear *parametric modulation* model (Chapter 9, Friston et al., 2007). For each subject, the covariate of interest was the event times multiplied (modulated) by the *n*-back level of task difficulty, with the result convolved with the SPM2 hemodynamic response function (without time or dispersion derivatives). The covariates of no interest were as follows: the event times convolved with the hemodynamic response function, six motion realignment terms to allow for any residual movement artefacts not removed by pre-processing realignment, and a constant term modelling the baseline of unchanged neural activity. The first covariate of no interest was necessary to isolate the effects of stimuli presentation and motor responses (of no interest) from *n*-back level of task difficulty (covariate of interest). For each subject, the resultant covariate image of interest was the SPM2 "beta" image, comprising linear regression coefficients at each voxel, between the *n*-back level of task difficulty and the observed BOLD signal. These beta images were taken to second level analyses (Seymour et al., 2004; Friston et al., 2007).

Two second level between-subjects, random-effects analyses were conducted to test the first two hypotheses. The first consisted of testing,

for each of the six groups, the null hypothesis of no significant (de) activation associated with an increase in *n*-back level of task difficulty. This was done by performing six one-group *t*-tests. The threshold of significance was defined as $P<0.05$ with FDR (Genovese et al., 2002) adjustment to significance values (Yekutieli and Benjamini, 1999) for the whole brain. As is conventional, for display the images were thresholded at $P<0.001$ uncorrected to demonstrate the spatial extent of the signals and the maximum *z* value within a cluster of (de)activations reported. The second level analysis consisted of entering the covariate images of interest for the six data subject groups into a 3×2 (scanner×visit) within-subject ANOVA applying a correction for possible non-sphericity. Again the threshold of significance was defined as $P<0.05$ with FDR correction for the whole brain. Images representing the effect of the scanner, the effect of the visit, and the interaction between both factors are shown thresholded at $P<0.001$ uncorrected.

The third hypothesis was tested by an investigation of reproducibility and reliability. To assess the *reproducibility* of statistical parametric maps, overlap and the size ratio measures of (de) activation clusters were calculated (Rombouts et al., 1998). The overlap ratio was computed as $R^{ij}{}_{\text{overlap}} = 2{*}V^{ij}{}_{\text{overlap}}/(V_i + V_j)$, where V_i and V_j are the number of voxels with *t*-values exceeding a given threshold in statistical maps *i* and *j*. $V^{ij}{}_{\text{overlap}}$ is the number of voxels that are commonly activated in both maps. The size ratio was computed as $R^{ij}{}_{\text{size}} = 2{*}V^{ij}{}_{\text{smallest}}/(V_i + V_j)$, where V_i and V_j are defined as above and $V^{ij}{}_{\text{smallest}}$ is the smallest of the two volumes. Values for both ratios range from 0 (no agreement between the compared maps) to 1 (complete agreement between maps). For analysis, the second level group maps were thresholded at $P<0.001$ uncorrected, and reproducibility was assessed both within and between scanners.

Within- and between-scanner reliability was assessed using ICCs (Friedman et al., 2008). Regions of interest (ROIs) were defined as areas consistently activated in all group analyses across scanners and visits. For each contrast, the second level random effects analysis maps for the six groups were thresholded at $P<0.001$ uncorrected and masked allowing identification of ROIs across groups. From each subject's beta image of regressors (observed BOLD vs. *n*-back level of task difficulty), the mean value of regressor across voxels for each ROI was extracted and used as the dependent variable for the reliability analysis. Variance components were estimated using the Restricted Maximum Likelihood (REML) method implemented in SPSS (Chicago, Illinois) version 16 ('Varcomp' routine). The following model was used:

$$Y_{ijk} = \text{mean} + \text{subject}_i + \text{scanner}_k + \text{subject}_i{*}\text{scanner}_k + \text{unexplained}_{ijk}$$

where Y_{ijk} denotes the dependent measure for subject *i*, visit *j* and scanner *k*. In this model each factor is considered a random effect. Other models, including other interactions between the subject, visit and scanner factors were also examined, but no appreciable changes in the results of the reliability analyses were observed. According to the above model, a total variance term can be written as:

$$\text{Total Variance} = \text{VD\_subject} + \text{VD\_scanner} + \text{VD\_subject}{*}\text{scanner} + \text{VD\_unexplained}$$

where VD means "variance due to". The ICCs for within- and between-scanner reliability were computed as:

$$\text{ICC}_{\text{between}} = \text{VD\_subject}/\text{Total Variance}$$

$$\text{ICC}_{\text{within}} = (\text{VD\_subject} + \text{VD\_scanner} + \text{VD\_subject}{*}\text{scanner}) / \text{Total Variance}$$

ICCs range from 0 to 1 where the latter is best reliability (Friedman et al., 2008).

3. Results

Complete image data were obtained for 82 scanning sessions and complete reaction time data for 84 scanning sessions. Behavioural accuracy data were obtained for all 56 Scanner A and Scanner B sessions. Two complete Scanner C imaging data sets were unavailable and no behavioural accuracy data were available for the Scanner C sessions, although reaction time measures were recorded. Although one subject was found to have an abnormality resulting in a loss of BOLD signal in a small region of the dorsal anterior cingulate (dAC), that subject's data were included as their inclusion was not expected to affect the planned tests. Visual inspection of each subject's first BOLD volume in their time series, SPM2 realignment translation and rotation graphs, and mean realigned and spatially normalised BOLD images, did not indicate substantial scanner or pre-processing problems likely to have a major impact on the analyses. Therefore, all available data were included in the planned analyses.

3.1. Behavioural results

3.1.1. Reaction times

The results of the 3×2×4 within-subject ANOVA are shown in Fig. 1. There was a significant main effect of scanner ($F_{(2,26)} = 16.948$, $P<0.001$) and a significant main effect of *n*-back level ($F_{(1.981,25.759)} = 35.658$, $P<0.001$). There was not a significant main effect of visit ($F_{(1,13)} = 0.245$, $P = 0.629$). No interactions were significant. A significant linear trend was found for the *n*-back level ($F_{(1,13)} = 58.326$, $P<0.001$). *Post-hoc* pairwise comparisons after Bonferroni correction identified significant differences in reaction times between Scanner A and Scanner B ($P = 0.003$) and between Scanner C and Scanner B ($P = 0.001$). Pairwise comparisons identified significant differences in reaction times between difficulty levels of the *n*-back task. Specifically, significant differences were found between 0-back and 1-back ($P = 0.012$), 0-back and 2-back ($P<0.001$), 0-back and 3-back ($P<0.001$), 1-back and 2-back ($P<0.001$) and 1-back and 3-back ($P<0.001$). Other pairwise comparisons were non-significant. In summary, as expected, mean reaction times increased significantly with the *n*-back level of difficulty. Additionally though, mean reaction times were significantly slower in Scanner B than in Scanner A and Scanner C. However, reaction times were not significantly affected by whether a subject was visiting a scanner for the first or second time.

3.1.2. Accuracy

The accuracy of subjects was investigated using a 2×2×4 (scan-centre×visit×*n*-back level of difficulty) within-subject ANOVA. As shown in Fig. 1, there were significant main effects of scanner ($F_{(1,13)} = 5.685$, $P = 0.033$), visit ($F_{(1,13)} = 5.523$, $P = 0.035$) and *n*-back level ($F_{(1.23,15.994)} = 8.625$, $P = 0.007$. None of the interactions were significant. Analysis of the *n*-back level identified a significant linear trend ($F_{(1,13)} = 9.771$, $P = 0.008$). *Post hoc* pairwise comparisons after Bonferroni correction identified significant differences in accuracy between the 0-back and 2-back ($P = 0.012$) and between the 1-back and 2-back ($P = 0.009$). Other pairwise comparisons between *n*-back levels were not significant. In summary, accuracy was affected by the scanner (higher accuracy in Scanner A than in Scanner B), the visit (higher accuracy on a subject's second visit than on the first visit) and, as expected, the *n*-back level of difficulty (accuracy tended to decrease with increasing *n*-back level).

3.2. Imaging results

3.2.1. First hypothesis

Six second level one-group *t*-tests (one test for each scanner on each visit) were used to test the null hypothesis of no change in brain (de)activation with increasing *n*-back task difficulty. Significant

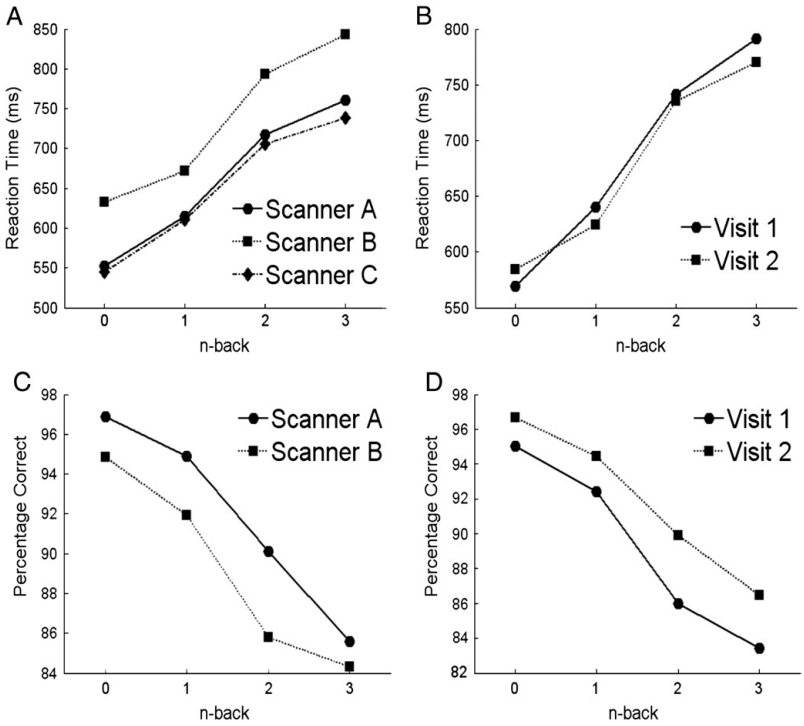

Fig. 1. Behavioural results: (A) Mean reaction time (ms) at each level of the *n*-back task for every scan-centre; (B) mean reaction time (ms) at each level of the *n*-back task at each visit; (C) mean accuracy (percentage correct) at each level of the *n*-back task for every scan-centre; (D) mean accuracy (percentage correct) at each level of the *n*-back task at each visit.

activations across groups were found in the dorsal anterior cingulate (dAC), lateral anterior prefrontal cortex (laPFC), dorsolateral prefrontal cortex (dlPFC), right posterior parietal lobe (PPL), insula and cerebellum (see Fig. 2). Additionally, significant deactivations were found in the dorsal posterior cingulate (dPC), retrosplenial cortex (RSC), medial anterior prefrontal cortex (maPFC), hippocampal–amygdala complex (HAC) and auditory cortex (see Fig. 3). Tables 1 and 2 summarise regions of significant (de)activations. Therefore, consistent with our first hypothesis, significant (and qualitatively similar) patterns of (de)activation were found for each scanner.

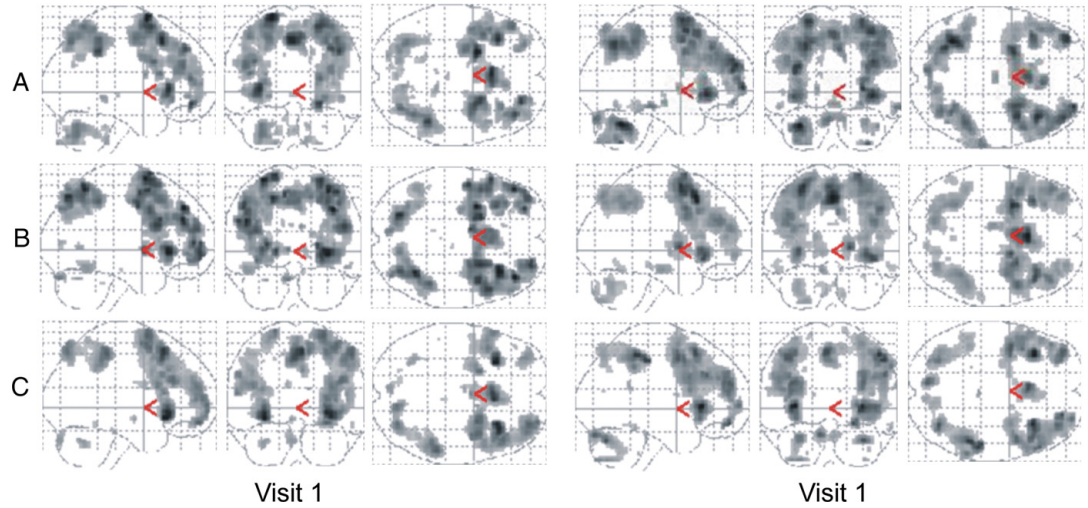

Fig. 2. Group maps for the activations (increase in activation as task difficulty ('n') increases). Images are thresholded at $P<0.001$ uncorrected. Capital letters indicate the scanner.

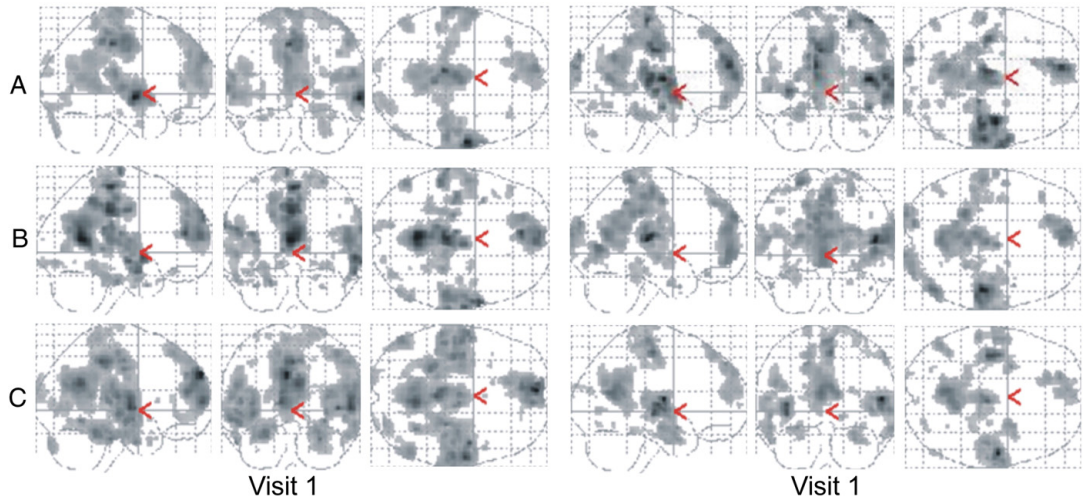

Fig. 3. Group maps for the deactivations (decrease in activation as task difficulty ('n') decreases). Images are thresholded at $P<0.001$ uncorrected. Capital letters indicate the scanner.

3.2.2. Second hypothesis

Using a 3×2 (scan-centre × visit) within-subject ANOVA, no significant effect of the scanner, the visit or their interaction was found (see Fig. 4). As above, whilst the images are thresholded at $P<0.001$ uncorrected (to demonstrate the spatial extent of any significant signal), significance is defined as $P<0.05$ corrected. As none of the regions visible in Fig. 4 reached the threshold of significance, the regions only represent chance findings. This is consistent with our second hypothesis.

3.2.3. Third hypothesis

Reproducibility ratios for within- and between-scanner reproducibility are shown in Fig. 5. Size ratios ranged from 0.58 to 0.98 within scanners and from 0.86 to 1 between scanners. Overlap ratios ranged from 0.45 to 0.61 within scanners and from 0.63 to 0.69 between scanners. Consistent with our third hypothesis, the within-scanner reproducibility was similar to the between-scanner reproducibility. Regarding reliability, Fig. 6 shows 10 of 14 ROIs used in the reliability analysis. The remaining four ROIs are similar to the corresponding

Table 1
Brain *activations* with increasing *n*-back level of difficulty.

		Scanner A			Scanner B			Scanner C		
		Coordinate	*z*	*P*	Coordinate	*z*	*P*	Coordinate	*z*	*P*
dAC	Visit 1	(−2,20,45)	5.27	0.001	(2,20,43)	4.34	0.002	(6,22,49)	5.03	0.002
	Visit 2	(0,18,49)	4.76	0.001	(2,18,45)	5.91	<0.001	(−2,20,49)	5.06	0.001
right laPFC (BA 10–11)	Visit 1	(26,46,−12)	5.01	0.001	(26,54,−4)	4.81	0.002	(30,52,−14)	4.56	0.003
	Visit 2	(30,58,1)	5.04	0.001	(30,54,−6)	4.21	0.002	(28,58,1)	4.58	0.002
left laPFC (BA 10–11)	Visit 1	(−28,49,5)	4.22	0.003	(−26,49,3)	4.77	0.002	(−34,49,14)	4.08	0.005
	Visit 2	(−36,53,19)	5.06	0.001	(−44,42,24)	4.75	0.001	(−36,51,20)	4.87	0.001
right dl*PFC* (BA 6)	Visit 1	(24,11,57)	4.98	0.001	(26,9,60)	4.86	0.002	(26,7,62)	4.58	0.003
	Visit 2	(28,9,60)	4.84	0.001	(26,9,66)	5.18	<0.001	(32,5,57)	4.39	0.003
left dl*PFC* (BA 6)	Visit 1	(−20,7,64)	5.45	0.001	(−30,7,62)	5.17	0.002	(−22,5,62)	4.05	0.005
	Visit 2	(−24,5,61)	5.01	0.001	(−28,5,55)	5.59	<0.001	(−42,4,44)	4.94	0.001
right dl*PFC* (BA 6–8–44)	Visit 1	(48,11,29)	3.71	0.006	(50,15,21)	4.78	0.002	(44,28,23)	4.48	0.004
	Visit 2	(40,6,33)	4.51	0.001	(48,8,36)	4.98	<0.001	(53,17,29)	4.26	0.002
left dl*PFC* (BA 6–8–44)	Visit 1	(−38,11,25)	4.55	0.002	(−46,15,23)	4.55	0.002	(−42,15,20)	4.2	0.005
	Visit 2	(−38,5,29)	3.98	0.003	(−42,9,31)	4.05	<0.001	(−40,13,27)	4.59	0.002
right PPL (BA 7–39–40)	Visit 1	(46,−41,37)	4.97	0.001	(44,−41,44)	4.45	0.002	(53,−32,50)	4.6	0.003
	Visit 2	(44,−42,44)	4.62	0.001	(36,−46,43)	4.53	0.001	(48,−36,50)	5.27	0.001
left PPL (BA 7)	Visit 1	(−22,−68,48)	4.12	0.003	(−20,−67,49)	4.95	0.002	−−−	−−−	−−−
	Visit 2	(−28,−62,44)	5.02	0.001	(−30,−50,41)	4.51	0.001	(−42,−37,37)	3.9	0.004
right Insula	Visit 1	(32,27,−6)	4.26	0.003	(32,25,−3)	5.17	0.002	(30,21,−4)	5.15	0.002
	Visit 2	(38,25,−5)	5.14	0.001	(40,25,−6)	5.28	<0.001	(30,21,1)	5.43	0.001
left Insula	Visit 1	(−34,25,2)	5.16	0.001	(−30,21,1)	4.57	0.002	(−30,20,8)	5.35	0.002
	Visit 2	(−38,19,−4)	4.75	0.001	(−32,21,−1)	5.23	<0.001	(−34,23,1)	5.64	0.001
Cerebellum	Visit 1	(−30,−66,−30)	4.39	0.002	(−6,−81,−20)	4.12	0.003	(−32,−60,−26)	3.84	0.007
	Visit 2	(−34,−59,−24)	5.41	0.001	(−34,−65,−27)	4.01	0.002	(−28,−69,−20)	4.54	0.002
Cerebellum	Visit 1	(24,−74,−38)	3.58	0.008	−−−	−−−	−−−	(−4,−81,−23)	3.84	0.007
	Visit 2	(34,−60,−39)	*4.62*	0.001	−−−	−−−	−−−	(34,−69,−22)	*4.99*	0.001

dAC, dorsal anterior cingulate; laPFC, lateral anterior prefrontal cortex; dlPFC, dorsolateral prefrontal cortex; PPL, posterior parietal lobe. Brodmann areas are indicated between brackets. *P*-values are FDR "whole brain" corrected. Coordinates are in Talairach space.

Table 2
Brain *deactivations* with increasing *n*-back level of difficulty.

		Scanner A			Scanner B			Scanner C		
		Coordinate	*z*	*P*	Coordinate	*z*	*P*	Coordinate	*z*	*P*
dPC	Visit 1	(−6,−27,46)	5.65	<0.001	(0,−27,38)	5.56	<0.001	(2,−33,44)	5.43	<0.001
	Visit 2	(−6,−16,38)	5.15	0.001	(4,−17,54)	4.35	0.006	(4,−17,52)	5.17	0.001
RSC	Visit 1	(−2,−56,16)	4.37	0.002	(−2,−54,14)	5.98	<0.001	(6,−53,27)	5.48	<0.001
	Visit 2	(−10,−53,32)	4.39	0.002	(0,−61,20)	4.41	0.006	(−4,−57,19)	5.05	0.001
maPFC	Visit 1	(−8,60,28)	4.74	0.001	(2,57,12)	5.01	<0.001	(−4,60,28)	6.2	<0.001
	Visit 2	(−10,54,21)	5.31	0.001	(−8,60,28)	4.46	0.006	(−4,51,12)	4.03	0.004
right HAC	Visit 1	−−−			(28,−34,−10)	3.51	0.006	(30,−18,−18)	4.40	0.003
	Visit 2	(32,−22,−19)	3.66	0.006	(34,−24,−16)	3.27	0.013	(26,−1,−−17)	4.37	0.002
left HAC	Visit 1	(−24,−20,18)	3.63	0.005	(−26,−32,−19)	4.05	0.002	(−36,−13,−20)	4.45	0.003
	Visit 2	(−24,−26,−17)	4.01	0.003	(−28,−36,−17)	3.58	0.008	(−24,−3,−22)	3.75	0.007
right Auditory cortex	Visit 1	(61,−8,−3)	6.28	<0.001	(59,4,−40)	5.34	<0.001	(57,−8,4)	5.67	<0.001
	Visit 2	(46,−21,12)	5.24	0.001	(48,−23,14)	5.61	0.002	(53,−9,12)	6.15	<0.001
left Auditory cortex	Visit 1	(−65,−19,6)	4.82	0.001	(−53,−32,22)	4.89	<0.001	(−57,−6,−6)	4.93	<0.001
	Visit 2	(−53,−21,5)	4.23	0.002	(−53,−23,16)	4.12	0.006	(−53,−6,−3)	4.36	0.002

dPC, dorsal posterior cingulate; RSC, retrosplenial cortex; maPFC, medial anterior prefrontal cortex; HAC, hippocampus–amygdala complex. *P* values are FDR "whole brain" corrected. Coordinates are in Talairach space.

contralateral ROIs displayed in Fig. 6. Components of variance and ICCs are shown in Table 3. For most of the brain regions, the total variance was either attributed to the *subject* factor or remained unexplained. In some cases, small positive variances for the *scanner* factor and the *subject*scanner* interaction were obtained; in other cases, the estimates were zero. This was due to the REML method setting to zero negative variance estimates. Negative variance estimates are usually an underestimate of a variance component with a small or zero true value (Brown and Prescott, 2006). When a negative estimate of a variance component occurs, the usual approach is either to remove the random effect from the model (not possible here since assessing the *scanner* effect was the objective of this study) or fix the variance component to zero (Brown and Prescott, 2006). Notably, the probability of obtaining a negative variance component estimate increases if the ratio of the true variance component to the residual variance is small, and if the number of random effects categories is small (low degrees of freedom) (Brown and Prescott, 2006). Both conditions are likely to occur in this study, as the variance due to the *scanner* factor was very small compared with the residual variance (which represents the *within-subject* variability, which is usually high in fMRI studies) and since a small number of scanners (three) were used. For brain regions where the REML method did not report a positive variance estimate, the ICC calculation was affected by the true positive variance's not being included in the ICC calculation (and possibly a negative variance estimate). Since a negative variance estimate typically occurs when the true value of the variance is small, this will not affect ICC estimates (Brown and Prescott, 2006). To test the above, the calculation was repeated setting the *scanner* factor as a fixed effect. This resulted in very similar *subject* factor variance estimates, and unexplained variance estimates (compared with setting the *scanner* factor as a random effect). Specifically, across brain regions with zero value variance estimates (Table 3), the median percentage change in ICC_{bet} (from random to fixed scanner effect models) was 1.34% (25th percentile = 1.03, 75th percentile = 2.90). This indicates that the zero variance estimates for the *scanner* factor are accurate.

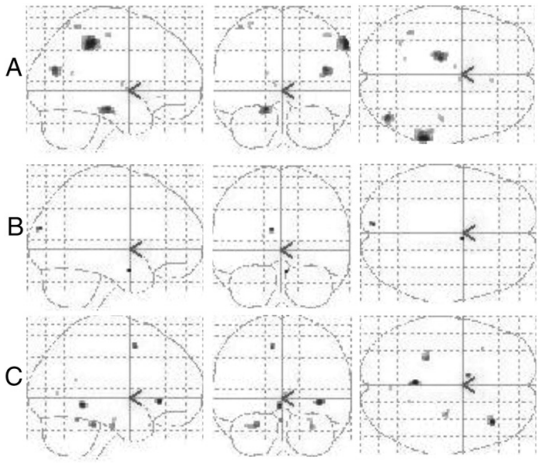

Fig. 4. Lack of a significant difference in brain activity: (A) different scanners; (B) different visits to the same scanners; (C) interaction between scanners and visits. For display, images are thresholded at a conventional level of $P<0.001$ uncorrected. None of the visible regions satisfy criteria for significance ($P<0.05$ corrected).

For the brain activations, the median within-scanner reliability was 0.52 (25th percentile = 0.44, 75th percentile = 0.59) and the median within-scanner reliability was 0.48 (25th percentile = 0.39, 75th percentile = 0.56). The best reliabilities were found in the dlPFC and the insula bilaterally, the worst in the dAC and left laPFC. For the deactivations, the median within-scanner reliability was 0.46 (25th percentile = 0.15, 75th percentile = 0.6) and the median between-scanner reliability was 0.28 (25th percentile = 0.1, 75th percentile = 0.55). The RSC showed the best reliability, the dPC showed poor reliability and the auditory cortex worst reliability. Notably though, the within-scanner ICCs were similar to the between-scanner ICCs, reflecting the fact that the scanner factor accounted for a small percentage of the total variance, for all brain regions studied, as hypothesised.

3.2.4. Post hoc analysis of image smoothness

Differences in image smoothness can contribute to between-scanner variability of activation (Friedman et al., 2006). Residual image smoothness is estimated by SPM using random field theory (Friston, 2004). Mean smoothness values are summarised in Table 4. Three 3×2 (scan-centre × visit) within-subject ANOVAs were used to test the null hypothesis of no difference in smoothness. There was a significant effect of the scanner in all three dimensions: *x*-dimension ($F_{(2,22)}=9.056$, $P=0.001$); *y*-dimension ($F_{(2,22)}=8.583$, $P=0.002$); *z*-dimension ($F_{(2,22)}=5.077$, $P=0.015$). The visit factor and the interactions were not significant. *Post hoc* pairwise comparisons after Bonferroni correction identified significant differences in smoothness for the *x*-dimension between scanners A and B ($P=0.003$), for the *y*-dimension between scanners A and B ($P=0.002$) and for the *z*-dimension between scanners A and B ($P=0.02$) and scanners B and C ($P=0.044$). In summary, the residual image smoothness was highest for scanner B.

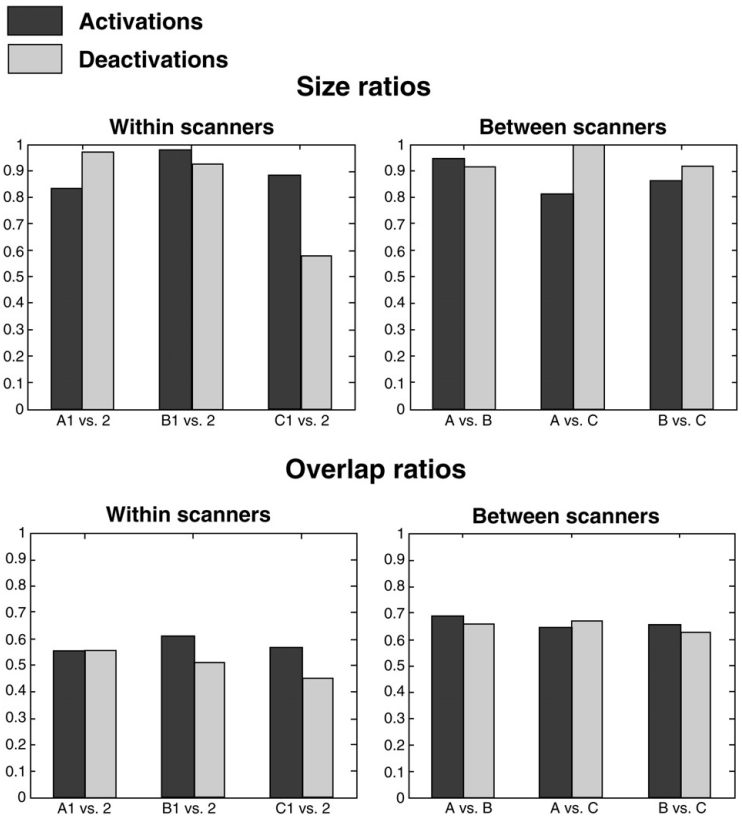

Fig. 5. Reproducibility of statistical parametric maps within and between scanners at the group levels for the activations and deactivations. Columns labelled A–C 1 vs. 2 show within-scanner comparisons. Columns labelled A vs. B, A vs. C and B vs. C show between-scanner comparisons.

4. Discussion

Consistent with our first hypothesis, significant patterns of brain activation and deactivation were identified using each of the three scanners. Qualitatively, the patterns of activation were very similar. A meta-analysis of fMRI results from 24 primary studies of healthy subjects performing the *n*-back task has been done (Owen et al., 2005). Activation loci found at each of the three scanner sites appears in good agreement with this meta-analysis. In many versions of the *n*-back task, as task difficulty increases, reaction times and error rates typically increase (Jansma et al., 2000; Nystrom et al., 2000; Watter et al., 2001; Jensen and Tesche, 2002; Goldberg et al., 2003; McMillan et al., 2007). Our behavioural results are therefore also in good agreement with the literature.

Consistent with our second hypothesis, no significant differences in brain activation were found between or within scanners; hence no evidence was found to reject the null hypotheses. However, significant behavioural differences were found: reaction times were increased in Scanner B. These are unlikely to be due to practice effects as the order of visits was counterbalanced. Instead, the differences in reaction times could be due to differences in the software and hardware used at each site. In site B, the software used to present stimuli and record responses was “IFIS”, whilst in sites A and C “Presentation” was used. We have noted that subtle differences in the way a paradigm is implemented in IFIS may result in consistent small delays in stimuli presentation. Further work is underway to clarify this issue.

Although apparent behavioural differences were found between sites, the image analysis method was not affected as the timing of the responses was extracted from each subject's log file (i.e. no assumptions were made about particular reaction times). Accuracy was highest in Scanner A and subjects performed more accurately on their first visit to a scanner than on their second. Although a weak relationship between *n*-back error rate and brain activation has been reported (Marshall et al., 2004), the average *z*-values from each of the three scanners were very similar (see Tables).

Our third hypothesis was that reproducibility values would be in the range reported for single scanner studies, and within-scanner reliability would be similar to between-scanner reliability. Our results are therefore similar to previous work (Friedman et al., 2008) supporting our hypothesis. Notably, estimates of reproducibility can be affected by the threshold chosen for the images. Rombouts et al. investigated how the overlap ratio was altered by varying the image threshold (Rombouts et al., 1998). Very low and high thresholds resulted in lower estimated values of reproducibility, perhaps due to type 1 and 2 errors in the images, respectively. A similar exploration of the relationship between reproducibility measures and threshold was beyond the scope of this study. Instead, we chose a threshold level of $P<0.001$ uncorrected, which is very common in psychiatric fMRI studies and therefore useful to other researchers.

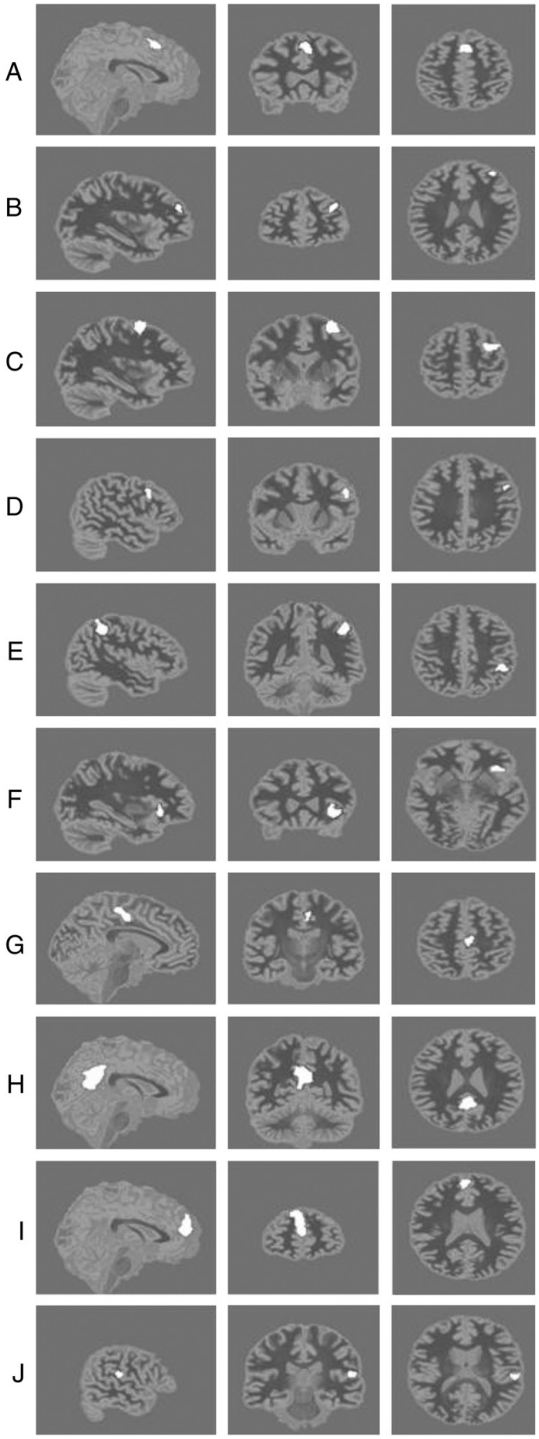

Table 3
Variance components and intraclass correlation coefficients.

Region	Subject	Scanner	Subject*Scanner	Unexplained	ICC_{bet}	ICC_{wit}
Activations						
dAC	8.243	0.000	0.000	13.680	0.376	0.376
right laPFC (BA 10)	19.175	0.000	0.000	22.398	0.461	0.461
left laPFC (BA 10)	4.624	0.000	0.000	22.959	0.168	0.168
right dlPFC (BA 6)	16.122	0.000	0.000	11.860	0.576	0.576
left dlPFC (BA 6)	24.069	0.000	0.269	9.226	0.717	0.725
right dlPFC (BA 6-8-44)	17.251	0.000	0.000	20.299	0.459	0.459
left dlPFC (BA 6-8-44)	15.029	0.000	0.000	11.845	0.559	0.559
right PPL (BA 7-39-40)	11.468	2.119	2.259	13.167	0.395	0.546
r*ight* Insula	8.166	0.000	0.000	8.540	0.489	0.489
left Insula	9.116	0.151	1.839	6.603	0.515	0.627
Deactivations						
*d*PC	1.812	0.255	2.392	7.991	0.146	0.358
RSC	28.782	0.010	0.000	19.166	0.600	0.600
maPFC	14.695	0.000	5.388	15.766	0.410	0.560
right Auditory cortex	1.390	0.000	0.000	14.555	0.087	0.087

dAC, dorsal anterior cingulate; laPFC, lateral anterior prefrontal cortex; dlPFC, dorsolateral prefrontal cortex; PPL, posterior parietal lobe; dPC, dorsal posterior cingulate; RSC, retrosplenial cortex; maPFC, medial anterior prefrontal cortex. Brodmann areas (BA) are indicated. Intraclass Correlation Coefficients (ICC) between (bet) and within (wit) scanners.

Reliability estimates may depend in part on the size of the ROIs and how the ROIs are defined. Friedman reported increased between scanner ICC estimates (particularly for 3 T scanners) with increasing size of ROIs (Friedman et al., 2008). Choosing ROIs based on common activation loci across all sites was found to result in increased reliability estimates (Friedman et al., 2008). However, low reliability values do not only reflect scanner-related effects. For example, we found the right auditory cortex to have a low ICC. This could be due to the timing of the BOLD volume acquisitions, and therefore associated auditory stimuli being desynchronised with the task, by design. Considering another location, whilst left BA10 ICC was low at 0.168, the right was 0.461. This clearly suggests a neural lateralisation (subject) rather than scanner effect, consistent with our general finding above, that the *subject* factor (and not *scanner*) strongly affected reliability estimates. Of note, our between-scanner ICCs for the *n*-back cognitive task are in the same range as the values obtained by Friedman for simpler sensori-motor tasks.

Considering the existing literature on fMRI within-scanner studies, 33 repeat scans of the same single subject doing simple motor visual and cognitive (random number generation) tasks have been reported (McGonigle et al., 1999). The authors concluded that correct inference about subject neural responses required use of a statistical model that accounted for both within- and between-session variance, such as a random effects analysis. Between-session variability is not necessarily high (Smith et al., 2005). A study of five subjects, each being scanned on two occasions during a visual cognitive task, concluded that there was evidence for good reliability of activation in visual processing and frontal areas (Specht et al., 2003). Nine elderly subjects were scanned 3 times during both a finger-tapping and an *n*-back task (Marshall et al., 2004).

Fig. 6. Ten of the 14 ROIs used in the reliability analysis. The remaining ROIs are similar to the corresponding contralateral ROIs shown in the figure. Sagittal, coronal and axial views are displayed for every ROI. (A) Dorsal anterior cingulate; (B) right lateral anterior prefrontal cortex; (C) right dorsolateral prefrontal cortex; (D) right dorsolateral prefrontal cortex; (E) right posterior parietal lobe; (F) right insula; (G) dorsal posterior cingulate; (H) retrosplenial cortex; (I) medial anterior prefrontal cortex; (J) right auditory cortex.

Table 4
Estimated image smoothness at each scanner on each visit.

Dimension	Scanner A		Scanner B		Scanner C	
	Visit 1	Visit 2	Visit 1	Visit 2	Visit 1	Visit 2
x	10.38 ± 0.28	10.54 ± 0.52	11.56 ± 0.96	11.42 ± 1.37	10.68 ± 1.18	11.15 ± 1.51
y	10.53 ± 0.33	10.73 ± 0.60	11.72 ± 0.94	11.61 ± 1.31	10.88 ± 1.34	11.29 ± 1.55
z	10.78 ± 0.32	10.86 ± 0.35	10.99 ± 0.39	11.15 ± 0.38	10.49 ± 0.95	10.68 ± 0.96

Mean and standard deviation full width half maximum values provided (millimetres).

The authors concluded that centres of activation loci were highly reproducible (within 3 mm) and that patterns of activation were qualitatively repeatable, but there was substantial variability in the amplitude and extent of the activated regions. An analysis of nine repeat scans of eight subjects during a finger-tapping task reported a consistent pattern of brain activation between scanning sessions (Yoo et al., 2005). In another study, 15 subjects were scanned on 3 occasions during an emotional processing task (Johnstone et al., 2005) and stability of the left amygdala neural response to fearful faces was reported. Considering between-scanner fMRI studies, four scanners of the same field strength (three 1.5 T GE scanners, one 1.5 T Siemens scanner) were used to investigate five to eight subjects from each site doing a spatial working memory motor task on one occasion at each site (Casey et al., 1998). Highly consistent findings across sites were reported. Differences in subject behaviour across sites were reported. Ten scanners (five 1.5 T, four 3 T and one 4 T) were used to scan five subjects on one occasion during a finger-tapping task (Zou et al., 2005). They concluded that imaging at the higher field strengths of 3 and 4 T yielded better reproducibility than at 1.5 T. The effects of subject, scanning sequence and field strength on reproducibility were significant. Whilst within-scanner reliability was high, between-scanner reliability was low (Friedman et al., 2008). It is notable that the study of Zou et al. included data from scanners with a range of field strengths and reported poor between-scanner reliability, whilst the study of Casey et al. included data from the same field strength scanners and reported highly consistent findings. Similarly, five identical 1.5 T systems were used to scan five subjects doing a finger-tapping task (Costafreda et al., 2007). These authors concluded that their results supported the feasibility of multi-site fMRI studies using identical scanner systems.

Limitations of this report should be noted. First, the data set is large and the literature demonstrates that there are many ways to examine such data. The methods chosen here may not generalize entirely to other methods. It is not possible to comprehensively report every possible method of investigating the data in this article. Third, as above, a different choice of image threshold may affect reliability estimates; however, we used a threshold very commonly chosen so as to be of most use to other workers. Fourth, differences in residual smoothness of images between scanners were found and future smoothness equalization may reduce estimated between-scanner variability (Friedman et al., 2006). Fifth, the REML method resulted in some zero variance estimates; however, the true values, if not precisely zero, are very small.

In summary, most previous studies of reliability and reproducibility have focused on simple sensory or motor paradigms. In contrast, we focused on the *n*-back task because it engages higher cognitive processes of relevance to psychiatric research (Rose et al., 2006a,b). Previous studies typically involved small numbers of subjects (e.g., five or six). Here we obtained a data set with larger numbers of subjects in each group, to better focus on group level analyses, which are of particular interest to psychiatric studies. Differences in scanner field strength, pulse sequence differences (Zou et al., 2005) and image smoothness (Friedman et al., 2006) have been reported to affect reproducibility, so we used the same field strength scanners with as similar pulse sequences as possible, and implemented the paradigms in an identical way at each site. Consistent with our hypotheses, we found reproducibility values for the *n*-back task in the range reported for single scanner studies, and within-scanner reliability was similar to between-scanner reliability, which supports data pooling.

Acknowledgement

The CaliBrain study was funded by a Chief Scientist Office (Scotland) Project Grant (CZB/4/427). Victoria Gradin and Gordon Waiter are funded by SINAPSE (http://www.sinapse.ac.uk). The Edinburgh University component of the imaging was performed at the SFC Brain Imaging Research Centre (www.sbirc.ed.ac.uk). The funding sources had no involvement in study design, data collection, analysis and interpretation, writing of the report, and the decision to submit the paper for publication. The authors thank Dr. Lorna Alcott for advice on statistical issues.

References

Brown, H., Prescott, R., 2006. Applied Mixed Models in Medicine. Wiley, West Sussex.
Casey, B.J., Cohen, J.D., O'Craven, K., Davidson, R.J., Irwin, W., Nelson, C.A., Noll, D.C., Hu, X., Lowe, M.J., Rosen, B.R., Trwitt, C.L., Turski, P.A., 1998. Reproducibility of fMRI results across four institutions using a spatial working memory task. NeuroImage 8, 249–261.
Costafreda, S.G., Brammer, M.J., Vencio, R.Z., Mourao, M.L., Portela, L.A., de Castro, C.C., Giampietro, V.P., Amaro Jr., E., 2007. Multisite fMRI reproducibility of a motor task using identical MR systems. Journal of Magnetic Resonance Imaging 26, 1122–1126.
Friedman, L., Glover, G.H., Krenz, D., Magnotta, V., 2006. Reducing inter-scanner variability of activation in a multicenter fMRI study: role of smoothness equalization. Neuroimage 32 (4), 1656–1668.
Friedman, L., Stern, H., Brown, G.G., Mathalon, D.H., Turner, J., Glover, G.H., Gollub, R.L., Lauriello, J., Lim, K.O., Cannon, T. Greve, Douglas, N., Bockholt, H.J., Belger, A., Mueller, B., Doty, M.J., He, J., Wells, W., Smyth, P., Pieper, S., Kim, S., Kubicki, M., Vangel, M., Potkin, S.G., 2008. Test-retest and between-site reliability in a multicenter fMRI study. Human Brain Mapping 29 (8), 958–972.
Friston, K., 2004. Introduction to Statistical Parametric Mapping. In: Frackowiak, R.S.J. (Ed.), Human Brain Function. Academic Press, London.
Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., Penny, W.D., 2007. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press, London.
Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15 (4), 870–878.
Goldberg, T.E., Egan, M.F., Gscheidle, T., Coppola, R., Weickert, T., Kolachana, B.S., Goldman, D., Weinberger, D.R., 2003. Executive subprocesses in working memory: relationship to catechol-O-methyltransferase Val158Met genotype and schizophrenia. Archives of General Psychiatry 60 (9), 889–896.
Jansma, J.M., Ramsey, N.F., Coppola, R., Kahn, R.S., 2000. Specific versus nonspecific brain activity in a parametric N-back task. Neuroimage 12 (6), 688–697.
Jensen, O., Tesche, C.D., 2002. Frontal theta activity in humans increases with memory load in a working memory task. European Journal of Neuroscience 15 (8), 1395–1399.
Job, D.E., Whalley, H.C., McIntosh, A.M., Owens, D.G., Johnstone, E.C., Lawrie, S.M., 2006. Grey matter changes can improve the prediction of schizophrenia in subjects at high risk. BMC Medicine 4, 29.
Johnstone, T., Somerville, L.H., Alexander, A.L., Oakes, T.R., Davidson, R.J., Kalin, N.H., Whalen, P.J., 2005. Stability of amygdala BOLD response to fearful faces over multiple scan sessions. Neuroimage 25 (4), 1112–1123.
Marshall, I., Simonotto, E., Deary, I.J., Maclullich, A., Ebmeier, K.P., Rose, E.J., Wardlaw, J.M., Goddard, N., Chappell, F.M., 2004. Repeatability of motor and working-memory tasks in healthy older volunteers: assessment at functional MR imaging. Radiology 233 (3), 868–877.
McGonigle, D.J., Howseman, A.M., Athwal, B.S., Friston, K.J., Frackowiak, R.S.J., Holmes, A.P., 1999. Variability in fMRI: an examination of intersession differences. Neuroimage 11, 708–734.
McMillan, K.M., Laird, A.R., Witt, S.T., Meyerand, M.E., 2007. Self-paced working memory: validation of verbal variations of the n-back paradigm. Brain Research 1139, 133–142.

Nystrom, L.E., Braver, T.S., Sabb, F.W., Delgado, M.R., Noll, D.C., Cohen, J.D., 2000. Working memory for letters, shapes, and locations: fMRI evidence against stimulus-based regional organization in human prefrontal cortex. Neuroimage 11 (5 Pt 1), 424–446.

Owen, A.M., McMillan, K.M., Laird, A.R., Bullmore, E., 2005. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. Human Brain Mapping 25 (1), 46–59.

Rombouts, S.A., Barkhof, F., Hoogenraad, F.G., Sprenger, M., Scheltens, P., 1998. Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. Magnetic Resonance Imaging 16 (2), 105–113.

Rose, E.J., Simonotto, E., Ebmeier, K.P., 2006a. Limbic over-activity in depression during preserved performance on the n-back task. Neuroimage 29 (1), 203–215.

Rose, E.J., Simonotto, E., Spencer, E.P., Ebmeier, K.P., 2006b. The effects of escitalopram on working memory and brain activity in healthy adults during performance of the n-back task. Psychopharmacology 185 (3), 339–347.

Seymour, B., O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., Frackowiak, R.S., 2004. Temporal difference models describe higher-order learning in humans. Nature 429 (6992), 664–667.

Smith, S.M., Beckmann, C.F., Ramnani, N., Woolrich, M.W., Bannister, P.R., Jenkinson, M., Matthews, P.M., McGonigle, D.J., 2005. Variability in fMRI: a re-examination of inter-session differences. Human Brain Mapping 24 (3), 248–257.

Specht, K., Willmes, K., Shah, N.J., Jancke, L., 2003. Assessment of reliability in functional imaging studies. Journal of Magnetic Resonance Imaging 17 (4), 463–471.

Suckling, J., Ohlssen, D., Andrew, C., Johnson, G., Williams, S.C., Graves, M., Chen, C.H., Spiegelhalter, D., Bullmore, E., 2008. Components of variance in a multicentre functional MRI study and implications for calculation of statistical power. Human Brain Mapping 29 (10), 1111–1122.

Watter, S., Geffen, G.M., Geffen, L.B., 2001. The n-back as a dual-task: P300 morphology under divided attention. Psychophysiology 38 (6), 998–1003.

Whalley, H.C., Simonotto, E., Moorhead, W., McIntosh, A., Marshall, I., Ebmeier, K.P., Owens, D.G., Goddard, N.H., Johnstone, E.C., Lawrie, S.M., 2006. Functional imaging as a predictor of schizophrenia. Biological Psychiatry 60 (5), 454–462.

Yekutieli, D., Benjamini, Y., 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. Journal of Statistical Planning and Inference 82, 171–196.

Yoo, S.S., Wei, X., Dickey, C.C., Guttmann, C.R., Panych, L.P., 2005. Long-term reproducibility analysis of fMRI using hand motor task. International Journal of Neuroscience 115 (1), 55–77.

Zou, K.H., Greve, D.N., Wang, M., Pieper, S.D., Warfield, S.K., White, N.S., Manandhar, S., Brown, G.G., Vangel, M.G., Kikinis, R., Williams, M.W., 2005. Reproducibility of functional MR imaging: preliminary results of prospective multi-institutional study performed by Biomedical Informatics Research Network. Radiology 237 (3), 781–789.

Appendix E

Voxel Based Morphometry publication

BMC Medical Imaging

BioMed Central

Technical advance

Open Access

Prospective multi-centre Voxel Based Morphometry study employing scanner specific segmentations: Procedure development using CaliBrain structural MRI data

T William J Moorhead*[1,6], Viktoria-Eleni Gountouna[1,6], Dominic E Job[1,6], Andrew M McIntosh[1,6], Liana Romaniuk[1,6], G Katherine S Lymer[4,6], Heather C Whalley[1,6], Gordon D Waiter[2,6], David Brennan[3,6], Trevor S Ahearn[2,6], Jonathan Cavanagh[5,6], Barrie Condon[3,6], J Douglas Steele[2,6], Joanna M Wardlaw[4,6] and Stephen M Lawrie[1,6]

Address: [1]The Division of Psychiatry, Centre for Clinical Brain Sciences (CCBS), School of Molecular and Clinical Medicine, University of Edinburgh, Edinburgh, UK, [2]Aberdeen Biomedical Imaging Centre, Division of Applied Medicine University of Aberdeen, Aberdeen, UK, [3]The Department of Clinical Physics and Bioengineering, NHS Greater Glasgow South University Hospitals Division, Glasgow, UK, [4]SFC Brain Imaging Research Centre, SINAPSE Collaboration http://www.sinapse.ac.uk, Division of Clinical Neurosciences, University of Edinburgh, Western General Hospital, Edinburgh, UK, [5]Sackler Institute of Psychological Research, Faculty of Medicine, University of Glasgow, Glasgow, UK and [6]Centre for Neuroscience, Division of Medical Sciences, University of Dundee, Dundee, UK

Email: T William J Moorhead* - bill.moorhead@ed.ac.uk; Viktoria-Eleni Gountouna - e.gountouna@ed.ac.uk; Dominic E Job - djob@staffmail.ed.ac.uk; Andrew M McIntosh - andrew.mcintosh@ed.ac.uk; Liana Romaniuk - l.romaniuk@sms.ed.ac.uk; G Katherine S Lymer - katherine.lymer@ed.ac.uk; Heather C Whalley - hwhalley@staffmail.ed.ac.uk; Gordon D Waiter - g.waiter@abdn.ac.uk; David Brennan - d.brennan@clinmed.gla.ac.uk; Trevor S Ahearn - t.ahearn@abdn.ac.uk; Jonathan Cavanagh - jc199d@clinmed.gla.ac.uk; Barrie Condon - barrie.condon@udcf.gla.ac.uk; J Douglas Steele - d.steele@dundee.ac.uk; Joanna M Wardlaw - jmw@skull.dcn.ed.ac.uk; Stephen M Lawrie - s.lawrie@ed.ac.uk

* Corresponding author

Published: 15 May 2009

BMC Medical Imaging 2009, **9**:8 doi:10.1186/1471-2342-9-8

Received: 17 November 2008
Accepted: 15 May 2009

This article is available from: http://www.biomedcentral.com/1471-2342/9/8

Abstract

Background: Structural Magnetic Resonance Imaging (sMRI) of the brain is employed in the assessment of a wide range of neuropsychiatric disorders. In order to improve statistical power in such studies it is desirable to pool scanning resources from multiple centres. The CaliBrain project was designed to provide for an assessment of scanner differences at three centres in Scotland, and to assess the practicality of pooling scans from multiple-centres.

Methods: We scanned healthy subjects twice on each of the 3 scanners in the CaliBrain project with T_1-weighted sequences. The tissue classifier supplied within the Statistical Parametric Mapping (SPM5) application was used to map the grey and white tissue for each scan. We were thus able to assess within scanner variability and between scanner differences. We have sought to correct for between scanner differences by adjusting the probability mappings of tissue occupancy (tissue priors) used in SPM5 for tissue classification. The adjustment procedure resulted in separate sets of tissue priors being developed for each scanner and we refer to these as scanner specific priors.

Results: Voxel Based Morphometry (VBM) analyses and metric tests indicated that the use of scanner specific priors reduced tissue classification differences between scanners. However, the

metric results also demonstrated that the between scanner differences were not reduced to the level of within scanner variability, the ideal for scanner harmonisation.

Conclusion: Our results indicate the development of scanner specific priors for SPM can assist in pooling of scan resources from different research centres. This can facilitate improvements in the statistical power of quantitative brain imaging studies.

Background

Structural Magnetic Resonance Imaging (sMRI) of the brain is employed in the assessment of a wide range of neuropsychiatric disorders. Voxel Based Morphometry (VBM) has been established as a leading method for analysing large sMRI studies. VBM is a fully automated process that is used to localise differences in brain parenchyma [1,2]. The VBM implementation segments T_1-weighted MRI scans into voxel-wise mappings of grey and white tissue and Cerebrospinal Fluid (CSF). It provides for statistical comparisons of these mappings within clinical studies. VBM requires good quality co-registration at the voxel level and it is sensitive to differences between MRI scanners. Reports of VBM analyses that pool scans from different sites for analysis have employed validity assessments [3,4]. In a validity assessment, a VBM contrast of control subjects between the contributing sites is used to map the regions of significant difference between scanners. A masking image that charts these regions is formed. These masked regions are excluded from VBM reporting as results in these regions could be driven by artifactual scanner differences [5,6].

In VBM the use of validity masking is undesirable because it limits the analyses to less than whole brain coverage. As VBM draws its inferences from voxel-wise comparisons it is necessary to apply fine grain corrections of the sMRI tissue classification in order to avoid validity masking. We investigated making such corrections by scanning the same fourteen healthy subjects, twice, at three scanning sites in Scotland with T_1-weighted sequences. These acquisitions were implemented as part of the CaliBrain study, and we have complete sets of scans for thirteen subjects. The three scanners in the CaliBrain project were matched on the basis of vendor, field strength and head coil type. In this investigation we used an established sMRI analysis tool Statistical Parametric Mapping (SPM5) [1] to segment the T1 scans into grey and white tissue maps and CSF maps.

Baseline analyses of the Calibrain T_1 segmentations revealed significant differences between the scanners and these differences were of an order that would require validity masking. We investigated the practicality of compensating for scanner differences through the adjustments in the SPM5 segmentation procedure. The SPM segmentation protocol employs spatial priors that map the probable distributions of grey and white tissue and CSF. These priors account for the low frequency variability in tissue presentation across the brain. We adjusted these prior mappings to compensate for the scanner differences. In this process we developed separate sets of priors for each scanner. As above, we refer to these as scanner specific priors. In VBM analyses of the segmentations based upon scanner specific priors, we found that the baseline differences which had indicated a requirement for validity masking were removed.

In addition to VBM analyses we applied metric tests to quantify within scanner variability and between scanner differences. These metrics were applied at baseline and on the adjusted segmentations. The metric results demonstrated that the use scanner specific priors can reduce the tissue classification differences between scanners. However, these reductions were not sufficient to bring the between scanner differences down to the level of within scanner variability.

Methods

Study Design

The CaliBrain project was designed to allow for the assessment of differences between scanners and for these differences to be considered in the context of within scanner variability. For this, healthy subjects travelled twice to each of the three scanning centres within a six months period. At each visit the subjects received a T_1-weighted structural MRI scan. We used SPM5 [1] to segment the structural scans into baseline grey and white tissue maps and CSF maps. The SPM priors used in these baseline tissue classifications were taken from a study of psychosis which employed a scan sequence that was equivalent to that used for the CaliBrain acquisitions [7]. The priors in the psychosis study were drawn from scans of young adults with a family history of schizophrenia and control subjects with no family history of psychosis. All 93 subjects in this study were well at time of scanning.

The practice of adjusting the SPM tissue priors specifically for a cohort acquired at one scanning centre is well established [8-12] and the importance of matching the spatial priors to the investigated population was demonstrated in a study of healthy young adults [13]. We have extended the practice of adjusting the SPM priors adjustment so that they provide compensation for scanner differences. To

achieve scanner level compensation we implemented an iterative adjustment protocol that employed proportional feedback to develop scanner specific priors for each scanner. We illustrate the operation of this protocol by randomly selecting six subjects from the CaliBrain project. We used the 1st round scans of these subjects to develop the scanner specific priors. These priors were then used to segment all the CaliBrain T_1 scans and this gave the segmentations for our adjusted analyses. VBM contrast analyses and metrics were used to assess the scanner differences at baseline and for adjusted segmentations. The seven subjects that were excluded from the scanner specific process formed a test group upon which we could assess the viability of our protocol.

Data Acquisition

The CaliBrain project acquired MRI brain scans from three imaging research centres: The Department of Radiology, University of Aberdeen; The Division of Psychiatry and The SFC Brain Imaging Research Centre within The Centre for Clinical Brain Sciences (CCBS) at The University of Edinburgh; and The Department of Clinical Physics, NHS Greater Glasgow South University Hospitals Division. The three scanners used were manufactured by General Electric (GE Healthcare, Milwaukee, Wisconsin) and had primary field strengths of 1.5T. The scanners are nominated within this report as scanners 'A, B and C'.

Fourteen healthy participants (10 male, mean age 36.3, age range 22–51 years) took part in the study. All participants were native English speakers, right-handed (self reported), met the standard MRI safety criteria and had no history of diagnosed neurological disorder, major psychiatric disorder or treatment with psychotropic medication, including treatment for substance misuse. The participants were not paid, but they were reimbursed for expenses. All participants provided written informed consent and the study was approved by the local research ethics committees. The scan records were incomplete for one subject and thus the harmonisation methods in the CaliBrain project were conducted on the basis of 13 healthy participants.

Three General Electric 1.5T scanners were used in this study, with some inevitable differences in hardware and software versions. In site A scanning was conducted with a General Electrics (GE) 1.5T Signa NVi/CVi scanner (software version 9.1; gradients with max. amplitude 40 mT/m and max. slew rate 150 T/m/s; standard quadrature head coil). In site B scanning was conducted with a General Electrics (GE) 1.5T Signa LX scanner (software version 9.1M4; Echo-speed gradients with max. amplitude 22 mT/m and max. slew rate 120 T/m/s; standard quadrature head coil). In site C scanning was conducted with a General Electrics (GE) 1.5T Signa scanner (software version 11M3/11M4SP1; gradients with max. amplitude 40 mT/m and max. slew rate 150 T/m/s; standard quadrature head coil).

All subjects participated in six scanning sessions, two at each of the three sites. The time lapse between scans at each site was nominally two weeks. The scanning parameters were kept constant across the three scanners, allowing for minor deviations arising from differences in scanner hardware and software. A high resolution T_1-weighted scan was acquired using a 3D inversion recovery-prepared fast gradient echo volume sequence with the following parameters: orientation coronal; repetition time (TR) of 5.9 ms (sites A and C) or 8.2 ms (site B); echo time (TE) of 1.9 ms (site A) or 3.3 ms (site B) or 1.4 ms (site C); slice thickness = 1.7 mm without a gap; inversion time (TI) 600 ms; matrix = 256 × 256, voxel within slice dimension = 0.86 mm square; field of view (FOV) = 220 mm^2; flip angle = 15°; 128 slices.

VBM Preprocessing and Segmentation

Prior to SPM5 segmentation, co-registration and reslicing procedures were applied to ensure that all the scans were aligned to the anterior-posterior commissure axis (AC-PC) in the standard MNI template space. As part of this process the scans were re-sampled to a resolution of 1 × 1 × 1 mm. The SPM5 segmentation at baseline was implemented using a study specific priors set that had been previously developed for a study of psychosis [7]. The SPM5 adjusted segmentations were obtained using the scanner specific priors derived in our priors adjustment procedure. The SPM5 segmentations were run using the default settings, the 'Number of Gaussians per class' was set to [2 2 2] and the 'Bias regularization' was set to 'medium'. In keeping with the established practice in psychosis research the segmented results were output as unmodulated and normalized to the MNI template. The normalization employed the SPM5 default normalization with the 'Nonlinear Frequency Cutoff = 25'. Also, in keeping with established VBM practice for psychosis in tissue density analyses the SPM5 segmentations were smoothed using an isotropic 12 mm Full Width Half Maximum (FWHM) kernel.

Procedure for creating Scanner Specific Priors

Our iterative procedure that compensates for scanner differences employs proportional feedback to develop sets of scanner specific priors for use in the adjusted SPM5 segmentations. The process flow diagram in Figure 1 gives an overview of this procedure. We designate one scanner as the target scanner and a second as the object scanner. The scans from the target scanner are segmented using SPM5 and for this segmentation the priors were taken from our psychosis study [7]. The priors applied to the target scanner remain unchanged throughout the run of the iterative

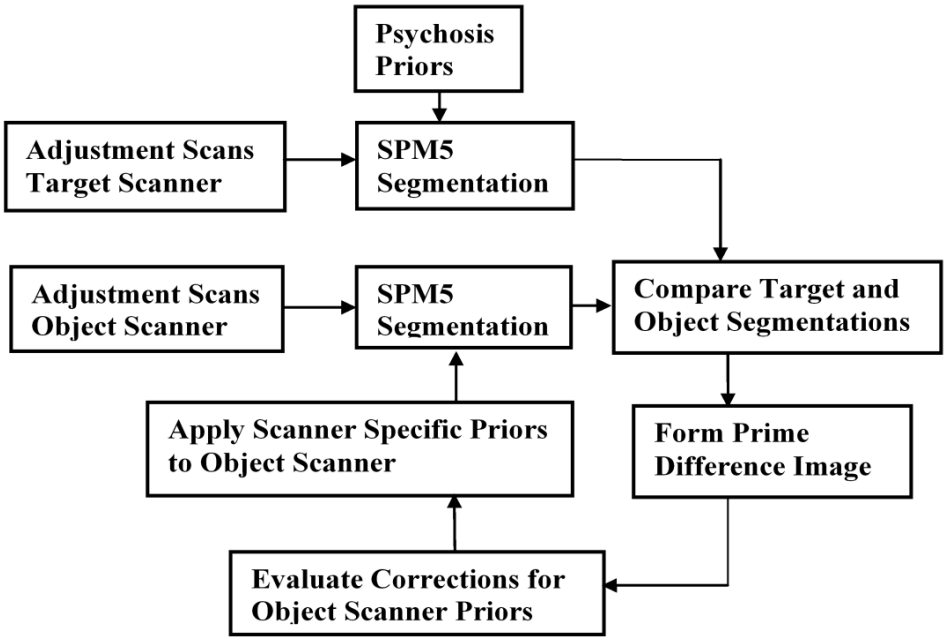

Figure 1
Process Flow diagram. Process Flow for procedure that develops scanner specific priors to correct for segmentation differences between the object and target scanners. Adjustment of the object scanner priors is used to minimise the difference between the scanners. The final adjusted object scanner priors are output as the scanner specific priors.

procedure. The object scanner segmentation is also initialised with the priors taken from our psychosis study [7]. Then through our iterative procedure these object scanner priors are incrementally adjusted. These adjustments are set to compensate for the segmentation differences between the target and object scanners.

The process illustrated in Figure 1 adjusts the object priors through comparisons based upon grey and white segmentations. This dual dependence of the adjusted priors on grey and white tissue is accommodated by allowing the iteration process to alternate the prime comparison between the grey and white tissue types. Thus for every other iteration the prime tissue comparison is applied to the grey segment and this is interspersed with the prime tissue comparison being made on the white segment.

When grey is the prime comparison segment we adjust the grey prior to correct for the voxel level differences found between the target and object scanners and also at the voxel level we apply a balancing adjustment to the white or CSF prior to ensure that the sum of the priors at the voxel level is maintained at its nominal sum of unity. Similarly when the prime comparison is made upon the white segment we adjust the white prior to correct for the differences between the target and object scanner and we apply a balancing adjustment to the grey or CSF prior. In VBM assessments of psychosis CSF presentation is not an established measure of interest, thus we do not assign CSF the prime status within the priors adjustment process.

On a subject by subject basis the prime comparison segmentations obtained from the target and object scanners are subtracted. These subtractions were implemented at the voxel-level and the differences were averaged across the subjects included in the priors adjustment process. The averaged voxel-level differences were used to form a difference image that was then smoothed to suppress sampling noise and reduce subject bias. Next a proportion of the smoothed difference image is used to adjust the grey, white and CSF priors applied to the object scanner. These adjusted priors are then made available for the next itera-

tion of the procedure. This process is repeated until the segmentations given by the object scanner converge with those given by the target scanner. We assess this convergence through the use of metrics described below.

The evaluation of the prime difference image *Pgdiff* for the grey segment *G* is given in equation (1). When the prime comparison segment is white *W*, the difference image *Pwdiff* is given by equation (2). In these calculations of the prime difference images, the processed scans are designated by subscripts *(subject, visit, scanner)*, with *N* = 6, the number of compared subjects. These comparisons were limited to the 1st round scans. The averaging across the adjustment subjects suppresses individual differences.

$$Pgdiff = \frac{1}{N}\sum_{n=1}^{N}\left(G_{(n,\text{visit},Obje.ct)} - G_{(n,\text{visit},Target)}\right) \quad (1)$$

$$Pwdiff = \frac{1}{N}\sum_{n=1}^{N}\left(W_{(n,\text{visit},Obje.ct)} - W_{(n,\text{visit},Target)}\right) \quad (2)$$

$$adjGprior = curGprior - sPgdiff * beta \quad (3)$$

$$adjWprior = \begin{Bmatrix} curWprior + (sPgdiff * beta) & if(curWprior > curCprior) \\ 0 & if(curCprior >= curWprior) \end{Bmatrix} \quad (4)$$

$$adjCprior = \begin{Bmatrix} curCprior + (sPgdiff * beta) & if(curCprio >= curWprior) \\ 0 & if(curWprio > curCprior) \end{Bmatrix} \quad (5)$$

When the primary segment is grey the adjusted prior *adjGprior* is given by equation (3). In this voxel-level process the current grey prior *curGprior* has a proportion *beta* of the smoothed prime difference *sPgdiff* image subtracted. When the prime comparison segment is grey the evaluations of the adjusted white *adjWprior* and CSF priors *adjCprio*r are given by equations (4) and (5). In these the changes applied to the grey prior are balanced by equivalent additions to the white or CSF priors. At the voxel level we test the relative occupancy of the white and CSF priors and assign the balancing adjustment to which ever prior exhibits greater occupancy. When the prime comparison segment is white the adjusted priors evaluations are equivalent to those given in equations (3), (4) and (5) with the exceptions that the prime difference image is given by *Pwdiff* and the grey and white priors are interchanged. This averaged difference images *Pgdiff* and *Pwdiff* are smoothed using an isotropic kernel, with a FWHM of 10 mm. Smoothing at this level suppresses sampling noise and limits the subject bias that results from the relatively small number of subjects that we have used to create the scanner specific priors

The value of *beta* determines the proportion of the difference image that is used to adjust the priors for the following iteration of the protocol. The setting of *beta* has an important bearing on this protocol. A high setting for *beta* could lead to instability whilst using value that is too low could result in sluggish convergence. As part of the development of this method, we experimented with the *beta* setting and found that setting *beta* to 0.33 or greater could lead to instability in the priors adjustment process. We found that a *beta* setting of 0.15 allowed for stable convergence of the segmentations from different scanners. We found that further reductions of the *beta* value did not improve the degree of convergence that was obtained from the adjustment process. The reductions in *beta* did increase the number iterations required to attain convergence. We ran a between scanner distance metric to assess the degree of convergence between the target and object scanners. We terminated the adjustment procedure when the incremental change in the between scanner distance was less than 0.1% and held at this level in subsequent iterations.

Testing Scanner Specific Priors Procedure

We tested our priors adjustment procedure by randomly selecting six subjects from the CaliBrain project. We applied the scanner specific priors adjustment to the 1st round scans of these subjects. We designated the scanners in the CaliBrain project as scanners A, B and C. Scanner A was set as the target scanner and scanner B as the object scanner and we developed a set of scanner specific priors for scanner B. We also developed scanner specific priors for scanner C with scanner A set as the target scanner. Throughout these adjustment procedures the priors set used for scanner A was fixed as the priors drawn from our study of psychosis [7]. The scanner specific priors developed for scanners B and C were initialised with the priors from our psychosis study. The choice of scanner A for as the target scanner was based upon the baseline metric results that indicated that scanner A has a low within scanner variability and that it exhibited the lowest overall between scanner differences.

VBM Statistical Analysis

Using the SPM5 application we implemented VBM statistical analyses of the grey and white segmentations at baseline and for our adjusted segmentations. In these we treated the visits and scanners as separate grouping components and thus formed a factorial analysis matrix that was composed of six groups. We designated the Independence variable as 'NO' to account for the fact that we have repeated measures on the same subjects. We reported the overall F-test for main effect of scanner in the CaliBrain

study. Also, we used this design matrix to report t-test contrast results for within scanner variability and between scanner differences. The t-tests for between scanner differences were made by combining the two visits at each scanner. All t-tests and the F-test were carried out with an uncorrected threshold of 0.001, and we reported Family wise error (FWE) correction for multiple comparisons. All groups were composed of the same subjects and as the scans were all acquired within a six month period there was no requirement to covary for age or gender.

Voxel-wise distance metrics

We employed a percentage distance metric to quantify the within scanner variability and between scanner differences. In SPM, tissue occupancy is assigned at the voxel level for grey, white and CSF as full occupancy or as partial volumes. At full occupancy the voxel is assigned as either grey or white or CSF with occupancy of 1.0. At the interface between tissue types partial occupancy is assigned on a continuous scale from 0.0 to 1.0 and the sum of the assigned occupancy for each voxel does not exceed 1.0.

In order to evaluate the distance between two tissue classifications we computed as a percentage the absolute distance. The general form of the absolute percentage distance computation is illustrated in equation (6) where we compare two voxels V1 and V2. This reports the percentage absolute difference with respect to the average value of the compared voxels. We chose this metric because it accentuates the differences in the compared segmentations.

$$AbsolutePercentageDistance = \frac{200*|V1-V2|}{(V1+V2)} \qquad (6)$$

In keeping with established VBM analyses the metrics were applied to the smoothed segmentations and limited to valid-voxels where the compared segments had occupancy of greater than 0.05. The summary value reported by the metric is an average of the absolute percentage difference found at the valid voxels in the normalised and smoothed segmentations. The metrics are applied on a subject basis and for each subject we evaluate the within scanner variability for scanners A, B and C and we evaluate the between scanner differences for the scanner pairs AB, AC and BC. A paired sample t-test is used to compare the baseline and adjusted metric results and to report the mean difference and its significance.

Results

Metric Results

We applied the percentage distance metric to the grey matter segmentations to obtain measures of within scanner variability and between scanner differences. The metric was applied at baseline and after adjustment using the scanner specific priors. Table 1 gives the grey matter metric results averaged across the six subjects used to generate the scanner specific priors. Table 2 gives the grey matter metric results averaged across the seven subjects who were excluded from the process that developed the scanner specific priors. In Tables 1 and 2 we note the mean difference between baseline and adjusted analyses and we give the p_value for the paired sample t-test as measure of significance in the adjustment process.

VBM Analyses

We ran VBM baseline and adjusted analyses on the normalised and smoothed segmentations obtained for the seven subjects who were excluded from the prior's adjustment procedure. In these analyses we treat the visits and scanners as separate grouping components and thus form a design matrix composed of six groups. We applied an F-test to consider the main effect of scanner and t-tests to investigate within and between scanner differences. All groups are composed of the same seven subjects and as the scans were all conducted within a six month period there is no requirement to co-vary for age or gender.

Baseline VBM Results

The F-test main effect Maximum Intensity Projection (MIP) for the baseline grey matter analysis is illustrated in Figure 2. The F-test was carried out with an uncorrected

Table 1: Grey matter metric results for the subjects used in priors generation

Scanner Comparison	Baseline*	Adjusted*	Mean Difference (paired sample significance)
AA	2.1 (0.5)	2.1 (0.5)	0.00 ($p < 1.00$)
BB	3.0 (0.5)	3.0 (0.4)	0.02 ($p < 0.79$)
CC	2.1 (0.6)	2.1 (0.5)	0.02 ($p < 0.36$)
AB	7.2 (0.8)	3.7 (0.4)	3.5 ($p < 0.001$)
BC	8.0 (0.8)	4.0 (0.6)	4.0 ($p < 0.001$)
AC	3.1 (0.4)	2.3 (0.3)	0.8 ($p < 0.001$)

Grey matter metric results averaged over the six subjects used to generate the scanner specific priors. For the within and between scanner comparison both baseline and adjusted absolute percentage distances are recorded.
*Absolute Percentage distance % (std dev)

Table 2: Grey matter metric results for the subjects excluded from priors generation

Scanner Comparison	Baseline*	Adjusted*	Mean Difference (paired sample significance)
AA	2.8 (0.6)	2.8 (0.6)	0.00 ($p < 1.00$)
BB	3.4 (0.6)	3.4 (0.6)	0.00 ($p < 1.00$)
CC	2.6 (0.7)	2.5 (0.7)	0.05 ($p < 0.078$)
AB	7.3 (0.6)	5.1 (0.7)	2.2 ($p < 0.001$)
BC	8.2 (0.8)	5.5 (0.8)	2.7 ($p < 0.001$)
AC	4.0 (0.6)	3.6 (0.5)	0.36 ($p < 0.003$)

Grey matter metric results averaged over the seven subjects excluded from the scanner specific priors adjustment procedure. For the within and between scanner comparisons both baseline and adjusted absolute percentage distances are recorded.
*Absolute Percentage distance % (std dev)

threshold of ($p < 0.001$). This reveals significant scanner effects in the frontal lobes, temporal poles, the thalamus, brain-stem, parietal lobes and occipital lobes. The results of the baseline grey matter F-test are given in Table 3. In this we report the significant maximal voxels giving the MNI coordinate and the anatomical location. We report the Family Wise Error (FWE) corrected p-value of maximal voxel and we also report the extent of the cluster associated with the maximal voxel.

We investigated the source of these baseline grey matter differences by applying t-test contrasts. t-test comparisons of 1st and 2nd round scans of each scanner revealed that there were no significant within scanner differences. t-test comparisons between the scanners revealed that there were no significant differences between scanners A and C. t-test comparisons between scanners B and C were found to replicate the differences reported in the F-test for main effect of scanner. t-test comparisons between scanners B and A also demonstrated replication of the differences reported in the F-test for main effect of scanner.

The F-test results for the baseline white matter VBM analysis are given in Figure 3 and Table 4. This illustrates the spatial distribution of the between scanner differences with significant differences in the right middle frontal gyrus and the thalamus. We investigated sources of these baseline white matter differences by applying t-test contrasts. Comparing 1st and 2nd round scans demonstrated that there were no within scanner differences. In the between scanner tests we found no significant differences for the A-C contrasts and we found significant differences in the A-B and B-C contrasts. The B-C white matter differences were more extensive that those found in the A-B contrasts.

Adjusted VBM Results

We applied VBM analyses to the adjusted segmentations obtained from the seven subjects who were excluded from the scanner specific priors development process. The VBM analyses applied were a direct replication of the baseline tests. In these F-tests for the main effect of scanner we found that there were no significant differences for either the grey or white matter analyses. We repeated the adjusted VBM analyses with all 13 CaliBrain subjects for whom we had complete records and these analyses confirmed that no significant differences remained between the pooled scanners.

Table 3: Grey Matter Baseline maximal voxel results

F-test Cluster Anatomical location	Maximal voxel MNI coordinate	FWE p_corrected
Right Temporal Pole	34, 15, -33	0.001
Left Temporal Pole	-35, 16, -36	0.001
Left Inferior Parietal lobule	-55, -48, 47	0.001
Left Inferior frontal gyrus	-42, 45, -11	0.001
Thalamus	13, -10, -1	0.006
Right Inferior Parietal lobule	56, -48, 40	0.009
Left Middle frontal gyrus	-43, 20, 46	0.014
Left Middle frontal gyrus	-44, 44, 24	0.023
Cingulate gyrus	0, 46, 32	0.040
Brain stem	-1, -30, -28	0.045
Right Superior frontal gyrus	15, 40, 49	0.052

VBM Grey matter baseline tests for the effect of scanner. Reporting the extent of the F-test cluster for an uncorrected threshold of $p < 0.001$. Giving the anatomical location of the maximal voxel, the MNI coordinate of this maximal voxel and the p_corrected significance.

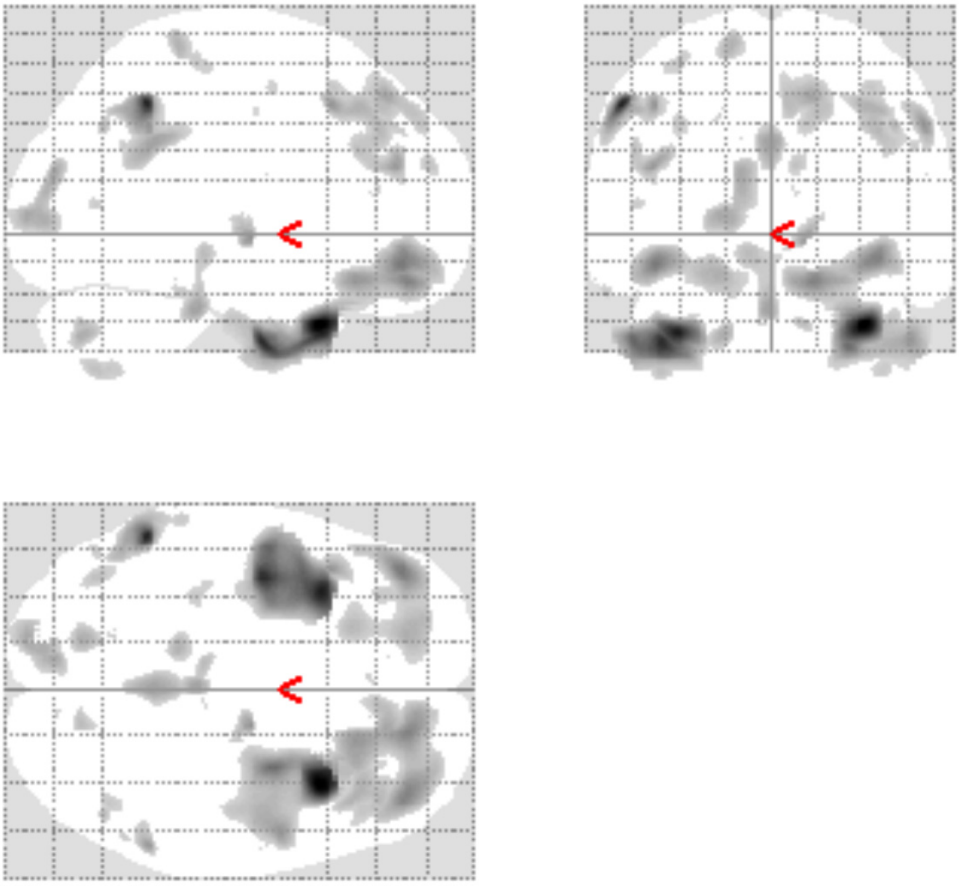

Figure 2
Grey Matter Baseline Results. Grey Matter Baseline Maximum Intensity Projection for the CaliBrain Project. Illustrates the regions where the scanners differ when the uncorrected threshold is $p < 0.001$.

Discussion

Combining structural MRI scans from different scanners presents the possibility of increasing the statistical power in VBM analyses of neuropsychiatric disorders. Our aim was to refine the application of SPM segmentation processes and reduce the effects of scanner differences which currently limit multi-centre MRI pooling [3,4]. We have examined the application of the SPM5 priors based segmentation to scans sourced from three scanners. The scanners were matched by vendor, primary field strength, and head coil type, and equivalent sequences were used at each scanner. Although these scanners are well matched we found in our baseline analyses significant between scanner differences in the tissue segmentations. We have demonstrated that if we employ scanner specific priors in our application of SPM that these between scanner differences are reduced.

In VBM analyses the harmonisation constraints for the use of multiple scanners are onerous as VBM requires the pro-

Table 4: White Matter Baseline results

F-test Cluster Anatomical location	Maximal voxel MNI coordinate	FWE p_corrected
Right Middle frontal gyrus	18, 45, -19	0.021
Thalamus	15, -10, 0	0.053

VBM White matter baseline tests for the effect of scanner. Reporting the extent of the F-test cluster for an uncorrected threshold of $p < 0.001$. Giving the anatomical location of the maximal voxel, the MNI coordinate of this maximal voxel and the p_corrected significance.

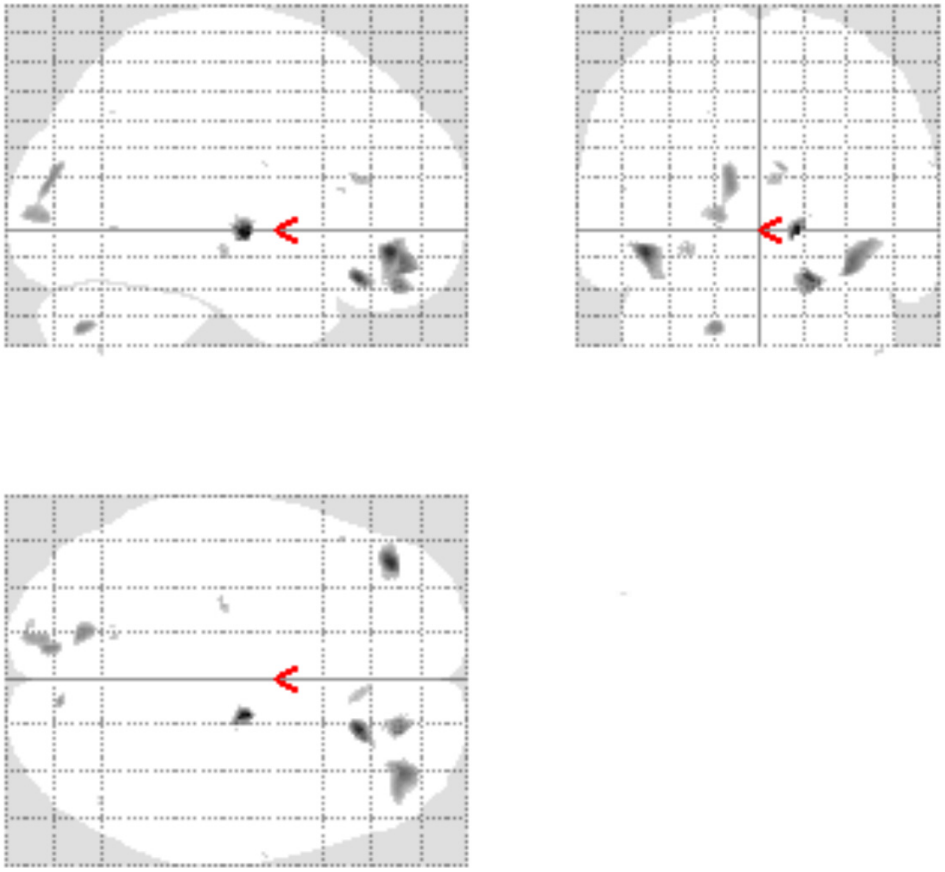

Figure 3
White Matter baseline Results. White matter baseline Maximum Intensity Projection for the CaliBrain project, Illustrates the regions where the scanners differ when the uncorrected threshold is $p < 0.001$.

vision of corrections for scanner differences at the voxel level. Previous work has shown that it is possible to pool scans from multiple centres in parcellated volumetric studies. In a volumetric analysis of images from multiple scanners [14], their semi-automated method applied global corrections for the tissue classification which were computed separately for each scanner. The methods reported summary volumes for grey and white matter in the cerebrum, and cerebellum and lateral ventricle volumes [15]. This set-level segmentation method employed global estimates of the intensity values that marked the transitions between tissue types and CSF. These globally applied transitions were adjusted for each scanning site.

A methodology that seeks to minimise the differences between scanners through an integration of scan sequence parameters into the segmentation functions was proposed by [16] and gives global adjustment in the intensity to tissue mapping. These global corrections are appropriate in studies where the inferences drawn are limited to lobar tissue occupancy. A volumetric method that addresses the localised intensity to tissue mappings has been proposed by [17] and recognises that localised adjustments for the intensity to tissue mapping within the brain are necessary for scan pooling to be valid for parcellation studies. The method that we have proposed is in keeping with this existing work as we have implemented corrections at a scale that is close to the analysis scale for VBM.

Research for the Alzheimer's Disease Neuroimaging Initiative (ADNI) project demonstrated that to pool scans from multiple sites, it is important to minimise differences between pooled scanners [18-20]. The ADNI project is a longitudinal analysis of ageing and in this within site MRI reproducibility was tested on a range of scanners and sequences. Based upon this research an MR-RAGE sequence was recommended for multiple site scanning and a scheme of corrections that includes field mapping and geometry correction is applied. ADNI researchers investigated the use of B1 field mapping to correct for within scanner variation in the RF inhomogenity for phased array head coils [20]. The results indicated that this technique has limitations. However, B1 field mapping can be applied as an addition to the priors adjustments protocol that we have developed. It is possible that the inclusion of field mapping would further reduce the between scanner differences in the CaliBrain project. However, the scan time acquisitions necessary for correction of the B1 field are not available in the CaliBrain project.

Recent reports of VBM analyses that sourced scans from multiple scanners employed validation masks to limit the reporting of results to regions where the scanner segmentations were equivalent [3,4]. Meda [4] demonstrated in a VBM study of psychosis at four centres that is possible to limit the effects of scanner differences by validity masking and ensuring that in the pooled analysis that the subjects and controls are drawn equally from all contributing centres [4]. A VBM analysis of scans taken from six scanners [3] reported that through the use of equivalent scan sequences and good quality control, the extent of validation masking required could be limited to a single region in the thalamus.

In the CaliBrain project we consider within scanner variability and between scanner differences and our aim was to reduce the between scanner differences to the level of within scanner variability. In keeping with the ADNI recommendations we have sought to minimise the scanner differences in terms of vendor, field strength, head coil and sequences. However, scanner B in the CaliBrain project does differ from scanners A and C in terms of maximum gradient amplitude and maximum slew rate. Our baseline results indicate that scanners A and C are well matched and scans from these two sites could be pooled without further adjustment or compensation. However, our baseline results also demonstrate that scanner B exhibits significant differences with respect to both scanners A and C.

In order to reduce the differences between the scanners in the CaliBrain project we have developed a procedure that employs proportional feedback to adjust the priors for each of the scanners. We have scan records for 13 healthy subjects who were scanned twice at three scanners within a six month period. We demonstrate our protocol for creating scanner specific priors using the 1st round scans of six subjects. We test the adequacy of these scanner specific priors through metric and VBM analyses. The tests for adequacy are applied to the seven subjects who were excluded from the priors adjustment protocol. These tests are limited by the number of subject scans available and we are unable to evaluate the full effects of subject variation expected in a multi-centre clinical study.

Clinical studies that could benefit from the scanner specific priors method are expected to have subject numbers considerably greater than those available for the CaliBrain project. In a multi-centre clinical study, with the exception of the travelling subjects used to develop the scanner specific priors, the subjects would be recruited and scanned independently at the contributing centres. In such a clinical study a test for adequacy of scanner harmonisation could be implemented through comparisons of the healthy control scans recruited from the contributing centres [3,4].

The metric that we report assesses the absolute distance between segmentations. The metrics are applied at the

voxel level and are averaged to report an overall distance inclusive of noise and systematic differences. We report in Table 1 on the scans that were used to implement the scanner specific priors procedure. This indicates that the within scanner variability ranges from 3.0% in scanner B to 2.1% in scanners A and C. The adjustment process gives rise to a reduction in the within scanner variability. However, the paired-t tests reveal that these within scanner adjustments do not represent a significant change. In Table 1 the baseline between scanner differences are at a maximum for the BC comparison. Here the adjustment procedure gave rise to significant reductions in all three scanner comparison metrics.

In Table 2 we consider the effects of the scanner specific priors on the scans of seven subjects who were excluded from the priors adjustment process. At baseline the within scanner variability and between scanner distances were equivalent to the baseline results reported in Table 1. Consequently, the use of the scanner specific priors resulted in significant reductions in all three scanner comparisons. However, for the comparisons that include scanner B, the reductions are not sufficient to bring the between scanner difference down to the level of within scanner variability.

The VBM analyses that we applied demonstrated that at baseline there are no significant differences between scanners A and C, However, we found that comparisons of scanners A and C with scanner B gave rise to differences that would require validity mapping such as that employed in VBM analyses by [3,4]. After developing scanner specific priors for scanners B and C and re-segmenting the scans we found that the requirement for validity mapping was removed, because we recorded no significant differences in the grey and white matter F-tests for scanner effect.

Conclusion

Our results indicate the development of scanner specific priors for the SPM application can assist in the pooling of scan resources from different research centres. This development can facilitate scan pooling and allow for improvements in the statistical power of multi-centre brain imaging studies. Our results indicate that six subjects were adequate for the purpose of matching the scanners in the CaliBrain project. In the typical clinical study the range of tissue presentations would be expected to be greater than that seen in our study of healthy controls. Thus it is likely that in a clinical study that more than six travelling subjects would be required. The number of travelling subjects required would depend upon the diversity of tissue presentation in the study and upon the nature of the differences in the scanners pooled. The method that we have suggested may be limited to multi-site studies in which there are no major hardware and acquisition protocol differences across sites. The CaliBrain project uses scanners from the same vendor all with the same field strengths and head coils with matched sequences. This provides an optimal environment for multiple site scan pooling. Different field strengths and image acquisition protocols could have very different tissue contrasts that would lead to marked differences in segmentation results. In such cases the differences in tissue classification may well be beyond the scope of our compensatory method.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TWJM developed the scanner correction procedure, played a leading role in the design of CaliBrain study and is the principal author of the manuscript; VEG is the main working researcher on the CaliBrain project, they developed the links with the three contributing scanning centres, established the CaliBrain scanning protocols and contributed to the manuscript; DEJ played a major role in the design of the CaliBrain project, assisted in the development of the scanning protocols at the three centres, and contributed to the final text; AMM assisted in the design of the clinical objects of the CaliBrain project and assisted in the preparation of the manuscript. LR contributed to the statistical analyses and contributed to the final manuscript; GKSL assisted in the development of the multi-centre links, the development of the scanning protocols and contributed to the final manuscript; HCW contributed to the development of the CaliBrain protocols and assisted in the drafting of the manuscript; GDW contributed to the development of the CaliBrain protocols supported the scanning at one of the centres and contributed to the manuscript; DB contributed to the development of the CaliBrain protocols, supported the scanning at one of the centres, and contributed to the manuscript; TSA contributed to the development of the CaliBrain protocols and contributed to the manuscript; JC assisted in the design of the clinical objectives of the CaliBrain project and contributed to the manuscript; BC contributed to the development of the CaliBrain protocols and contributed to the manuscript; JDS assisted in the design of the clinical and methodological objects of the CaliBrain project and contributed to the manuscript; JMW assisted in the design of the clinical and methodological objects of the CaliBrain project and contributed to the manuscript; SML is the principal architect of the CaliBrain Project. All authors contributed to the drafting of the manuscript and all authors approved the final version for publication.

Acknowledgements

VEG was supported by an MRC studentship. The CaliBrain study was funded by a Chief Scientist Office (Scotland) Project Grant (CZB/4/427), Chief Investigator SML. DEJ was supported by the MRC (NeuroGrid and

NeuroPsyGrid), AMI is supported by The Health Foundation, GKSL and SML were supported by the Sir Mortimer and Theresa Sackler Foundation. The Edinburgh imaging was carried out at the SFC Brain Imaging Research Centre http://www.sbirc.ed.ac.uk, Division of Clinical Neurosciences, University of Edinburgh, a core area of the Wellcome Trust Clinical Research Facility and part of the SINAPSE (Scottish Imaging Network – A Platform for Scientific Excellence) collaboration http://www.sinapse.ac.uk funded by the Scottish Funding Council and the Chief Scientist Office

References

1. Ashburner J, Friston KJ: **Unified segmentation.** *NeuroImage* 2005, **26(3):**839-51.
2. Ashburner J, Friston KJ: **Voxel-based morphometry – The methods.** *NeuroImage* 2000, **40(4):**1429-35.
3. Stonnington CM, Tan G, Klöppel S, Chu C, Draganski B, Jack CR Jr, Chen K, Ashburner J, Frackowiak RS: **Interpreting scan data acquired from multiple scanners: a study with Alzheimer's disease.** *Neuroimage* 2008, **39(3):**1180-5.
4. Meda SA, Giuliani NR, Calhoun VD, Jagannathan K, Schretlen DJ, Pulver A, Cascella N, Keshavan M, Kates W, Buchanan R, Sharma T, Pearlson GD: **A large scale (N = 400) investigation of gray matter differences in schizophrenia using optimized voxel-based morphometry.** *Schizophr Res* 2008, **101(1–3):**95-105.
5. Tofts PS: **Standardisation and optimisation of magnetic resonance techniques for multicentre studies.** *J Neurol Neurosurg Psychiatry* 1998, **64(Suppl 1):**S37-43. Review.
6. Jovicich J, Czanner S, Greve D, Haley E, Kouwe A van der, Gollub R, Kennedy D, Schmitt F, Brown G, Macfall J, Fischl B, Dale A: **Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data.** *NeuroImage* 2006, **30(2):**436-43.
7. McIntosh AM, Baig BJ, Hall J, Job D, Whalley HC, Lymer GK, Moorhead TW, Owens DG, Miller P, Porteous D, Lawrie SM, Johnstone EC: **Relationship of catechol-O-methyltransferase variants to brain structure and function in a population at high risk of psychosis.** *Biol Psychiatry* 2007, **61(10):**1127-34.
8. Wilke M, Schmithorst VJ, Holland SK: **Normative pediatric brain data for spatial normalization and segmentation differs from standard adult data.** *Magn Reson Med* 2003, **50(4):**749-57.
9. Moorhead TW, Job DE, Whalley HC, Sanderson TL, Johnstone EC, Lawrie SM: **Voxel-based morphometry of comorbid schizophrenia and learning disability: analyses in normalized and native spaces using parametric and nonparametric statistical methods.** *Neuroimage* 2004, **22(1):**188-202.
10. Job DE, Whalley HC, McIntosh AM, Owens DG, Johnstone EC, Lawrie SM: **Grey matter changes can improve the prediction of schizophrenia in subjects at high risk.** *BMC Med* 2006, **4:**29.
11. Spencer MD, Moorhead TW, Gibson RJ, McIntosh AM, Sussmann JE, Owens DG, Lawrie SM, Johnstone EC: **Low birthweight and preterm birth in young people with special educational needs: a magnetic resonance imaging analysis.** *BMC Med* 2008, **6:**1.
12. Wilke M, Holland SK, Altaye M, Gaser C: **Template-O-Matic: a toolbox for creating customized pediatric templates.** *Neuroimage* 2008, **41(3):**903-13.
13. Shen S, Szameitat AJ, Sterr A: **VBM lesion detection depends on the normalization template: a study using simulated atrophy.** *Magn Reson Imaging* 2007, **25(10):**1385-96.
14. VanHaren NEM, Cahn W, Hulshoff Pol HE, Schnack HG, Caspers E, Lemstra A, Sitskoorn MM, Wiersma D, Bosch RJ van den, Dingemans PM, Schene AH, Kahn RS: **Brain volumes as predictor of outcome in recent-onset schizophrenia: a multi-center MRI study.** *Schizophr Res* 2003, **64(1):**41-52.
15. Schnack HG, van Haren NEM, Hulshoff Pol HE, Picchioni M, Weisbrod M, Sauer H, Cannon T, Huttunen M, Murray R, Kahn RS: **Reliability of brain volumes from multicenter MRI acquisition: a calibration study.** *Hum Brain Mapp* 2004, **22(4):**312-20.
16. Fischl B, Salat DH, Kouwe AJ van der, Makris N, Ségonne F, Quinn BT, Dale AM: **Sequence-independent segmentation of magnetic resonance images.** *Neuroimage* 2004, **23(Suppl 1):**S69-84.
17. Han X, Fischl B: **Atlas renormalization for improved brain MR image segmentation across scanner platforms.** *IEEE Trans Med Imaging* 2007, **26(4):**479-86.
18. Hua X, Leow AD, Parikshak N, Lee S, Chiang MC, Toga AW, Jack CR Jr, Weiner MW, Thompson PM, The Alzheimer's Disease Neuroimaging Initiative: **Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: An MRI study of 676 AD, MCI, and normal subjects.** *Neuroimage* 2008, **43(3):**458-69.
19. Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, L Whitwell J, Ward C, Dale AM, Felmlee JP, Gunter JL, Hill DL, Killiany R, Schuff N, Fox-Bosetti S, Lin C, Studholme C, DeCarli CS, Krueger G, Ward HA, Metzger GJ, Scott KT, Mallozzi R, Blezek D, Levy J, Debbins JP, Fleisher AS, Albert M, Green R, Bartzokis G, Glover G, Mugler J, Weiner MW: **The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods.** *J Magn Reson Imaging* 2008, **27(4):**685-91.
20. Leow AD, Klunder AD, Jack CR Jr, Toga AW, Dale AM, Bernstein MA, Britson PJ, Gunter JL, Ward CP, Whitwell JL, Borowski BJ, Fleisher AS, Fox NC, Harvey D, Kornak J, Schuff N, Studholme C, Alexander GE, Weiner MW, Thompson PM, ADNI Preparatory Phase Study: **Longitudinal stability of MRI for mapping brain change using tensor-based morphometry.** *Neuroimage* 2006, **31(2):**627-40.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2342/9/8/prepub

Bibliography

Aron, A. R., Gluck, M. A., and Poldrack, R. A. (2006). Long-term test-retest reliability of functional mri in a classification learning task. *Neuroimage*, 29(3):1000–1006.

Ashburner, J. and Friston, K. J. (2000). Voxel-based morphometry–the methods. *Neuroimage*, 11(6 Pt 1):805–21.

Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *Neuroimage*, 26(3):839–51.

Bosnell, R., Wegner, C., Kincses, Z. T., Korteweg, T., Agosta, F., Ciccarelli, O., Stefano, N. D., Gass, A., Hirsch, J., Johansen-Berg, H., Kappos, L., Barkhof, F., Mancini, L., Manfredonia, F., Marino, S., Miller, D. H., Montalban, X., Palace, J., Rocca, M., Enzinger, C., Ropele, S., Rovira, A., Smith, S., Thompson, A., Thornton, J., Yousry, T., Whitcher, B., Filippi, M., and Matthews, P. M. (2008). Reproducibility of fmri in the clinical setting: implications for trial designs. *Neuroimage*, 42(2):603–10.

Brown, G. G. and Eyler, L. T. (2006). Methodological and conceptual issues in functional magnetic resonance imaging: Applications to schizophrenia research. *Annual Review of Clinical Psychology*, 2(1):51–81.

Brown, G. G., Mathalon, D. H., Stern, H., Ford, J., Mueller, B., Greve, D. N., McCarthy, G., Voyvodic, J., Glover, G., Diaz, M., Yetter, E., Ozyurt, I. B., Jorgensen, K. W., Wible, C. G., Turner, J. A., Thompson, W. K., Potkin, S. G., and Network, F. B. I. R. (2011). Multisite reliability of cognitive bold data. *Neuroimage*, 54(3):2163–75.

Caceres, A., Hall, D. L., Zelaya, F. O., Williams, S. C. R., and Mehta, M. A. (2009). Measuring fmri reliability with the intra-class correlation coefficient. *Neuroimage*, 45(3):758–68.

Carrier, J. and Monk, T. H. (2000). Circadian rhythms of performance: new trends. *Chronobiology International*, 17(6):719–32.

Casey, B. J., Cohen, J. D., O'Craven, K., Davidson, R. J., Irwin, W., Nelson, C. A., Noll, D. C., Hu, X., Lowe, M. J., and Rosen, B. R. (1998). Reproducibility of fmri results across four institutions using a spatial working memory task. *Neuroimage*, 8(3):249–261.

Chee, M. W. L., Lee, H. L., Soon, C. S., Westphal, C., and Venkatraman, V. (2003). Reproducibility of the word frequency effect: comparison of signal change and voxel counting. *Neuroimage*, 18(2):468–482.

Chiarelli, P. A., Bulte, D. P., Piechnik, S., and Jezzard, P. (2007). Sources of systematic bias in hypercapnia-calibrated functional mri estimation of oxygen metabolism. *Neuroimage*, 34(1):35–43.

Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements. *J Clin Exp Neuropsychol*, 23(5):695–700.

Cohen, E. R., Rostrup, E., Sidaros, K., Lund, T. E., Paulson, O. B., Ugurbil, K., and Kim, S.-G. (2004). Hypercapnic normalization of bold fmri: comparison across field strengths and pulse sequences. *Neuroimage*, 23(2):613–624.

Cohen, M. S. and DuBois, R. M. (1999). Stability, repeatability, and the expression of signal magnitude in functional magnetic resonance imaging. *Journal of Magnetic Resonance Imaging*, 10(1):33–40.

Colombo, P., Baldassarri, A., Corona, M. D., Mascaro, L., and Strocchi, S. (2004). Multicenter trial for the set-up of a mri quality assurance programme. *Magn Reson Imaging*, 22(1):93–101.

Costafreda, S. G., Brammer, M. J., Vêncio, R. Z. N., Mourão, M. L., Portela, L. A. P., de Castro, C. C., Giampietro, V. P., and Amaro, E. (2007). Multisite fmri reproducibility of a motor task using identical mr systems. *J Magn Reson Imaging*, 26(4):1122–6.

Diedrichsen, J. and Shadmehr, R. (2005). Detecting and adjusting for artifacts in fmri time series data. *Neuroimage*, 27(3):624–634.

Duncan, K. J., Pattamadilok, C., Knierim, I., and Devlin, J. T. (2009). Consistency and variability in functional localisers. *Neuroimage*, 46(4):1018–26.

Ekman, P. and Friesen, W. (1976). *Pictures of facial affect*. Consulting Psychologists, Palo Alto, CA.

Fera, F., Yongbi, M. N., van Gelderen, P., Frank, J. A., Mattay, V. S., and Duyn, J. H. (2004). Epi-bold fmri of human motor cortex at 1.5 t and 3.0 t: Sensitivity dependence on echo time and acquisition bandwidth. *Journal of Magnetic Resonance Imaging*, 19(1):19–26.

Feredoes, E. and Postle, B. R. (2007). Localization of load sensitivity of working memory storage: quantitatively and qualitatively discrepant results yielded by single-subject and group-averaged approaches to fmri group analysis. *Neuroimage*, 35(2):881–903.

Firbank, M. J., Harrison, R. M., Williams, E. D., and Coulthard, A. (2000). Quality assurance for mri: practical experience. *British Journal of Radiology*, 73(868):376–383.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B. R., and Dale, A. M. (2004). Automatically parcellating the human cerebral cortex. *Cereb Cortex*, 14(1):11–22.

Freyer, T., Valerius, G., Kuelz, A.-K., Speck, O., Glauche, V., Hull, M., and Voderholzer, U. (2009). Test-retest reliability of event-related functional mri in a probabilistic reversal learning task. *Psychiatry Res*, 174(1):40–6.

Friedman, L., Glover, G. H., and Consortium, T. F. (2006a). Reducing interscanner variability of activation in a multicenter fmri study: controlling for signal-to-fluctuation-noise-ratio (sfnr) differences. *Neuroimage*, 33(2):471–481.

Friedman, L., Glover, G. H., Krenz, D., Magnotta, V. A., and BIRN, F. (2006b). Reducing inter-scanner variability of activation in a multicenter fmri study: role of smoothness equalization. *Neuroimage*, 32(4):1656–1668.

Friedman, L., Stern, H., Brown, G. G., Mathalon, D. H., Turner, J., Glover, G. H., Gollub, R. L., Lauriello, J., Lim, K. O., Cannon, T., Greve, D. N., Bockholt, H. J., Belger, A., Mueller, B., Doty, M. J., He, J., Wells, W. M., Smyth, P., Pieper, S. D., Kim, S., Kubicki, M., Vangel, M. G., and Potkin, S. G. (2008). Test-retest and between-site reliability in a multicenter fmri study. *Hum Brain Mapping*, 29(8):958–972.

Friston, K. (2012). Ten ironic rules for non-statistical reviewers. *NeuroImage*, 61(4):1300–1310.

Fusar-Poli, P., Placentino, A., Carletti, F., Landi, P., Allen, P., Surguladze, S., Benedetti, F., Abbamonte, M., Gasparotti, R., Barale, F., Perez, J., McGuire, P., and Politi, P. (2009). Functional atlas of emotional faces processing: a voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. *Journal of Psychiatry & Neuroscience*, 34(6):418–432.

Gountouna, V.-E., Job, D. E., McIntosh, A. M., Moorhead, T. W. J., Lymer, G. K. L., Whalley, H. C., HALL, J., Waiter, G. D., Brennan, D., McGonigle, D. J., Ahearn, T. S., Cavanagh, J., Condon, B., Hadley, D. M., Marshall, I., Murray, A. D., Steele, J. D., Wardlaw, J. M., and Lawrie, S. M. (2010). Functional magnetic resonance imaging (fmri) reproducibility and variance components across visits and scanning sites with a finger tapping task. *Neuroimage*, 49(1):552–560.

Gradin, V., Gountouna, V.-E., Waiter, G., Ahearn, T. S., Brennan, D., Condon, B., Marshall, I., McGonigle, D. J., Murray, A. D., Whalley, H., Cavanagh, J., Hadley, D., Lymer, K., McIntosh, A., Moorhead, T. W., Job, D., Wardlaw, J., Lawrie, S. M., and Steele, J. D. (2010). Between- and within-scanner variability in the calibrain study n-back cognitive task. *Psychiatry Research: Neuroimaging*, 184(2):86–95.

Hall, J., Whalley, H. C., McKirdy, J. W., Romaniuk, L., McGonigle, D., McIntosh, A. M., Baig, B. J., Gountouna, V.-E., Job, D. E., Donaldson, D. I., SPRENGELMEYER, R., Young, A. W., JOHNSTONE, E. C., and Lawrie, S. M. (2008).

Overactivation of fear systems to neutral faces in schizophrenia. *Biol Psychiatry*, 64(1):70–3.

Han, X. and Fischl, B. (2007). Atlas renormalization for improved brain mr image segmentation across scanner platforms. *IEEE Trans Med Imaging*, 26(4):479–86.

Harrington, G. S., Buonocore, M. H., and Farias, S. T. (2006a). Intrasubject reproducibility of functional mr imaging activation in language tasks. *AJNR American Journal of Neuroradiology*, 27(4):938–944.

Harrington, G. S., Farias, S. T., Buonocore, M. H., and Yonelinas, A. P. (2006b). The intersubject and intrasubject reproducibility of fmri activation during three encoding tasks: implications for clinical applications. *Neuroradiology*, 48(7):495–505.

Hartley, H., Rao, J. L., and Lamotte, L. (1978). A simple 'synthesis'-based method of variance component estimation. *Biometrics*, 34(2):233–242.

Havel, P., Braun, B., Rau, S., Tonn, J.-C., Fesl, G., Brückmann, H., and Ilmberger, J. (2006). Reproducibility of activation in four motor paradigms : An fmri study. *Journal of Neurology*, 253(4):471–476.

Hua, X., Leow, A. D., Parikshak, N., Lee, S., Chiang, M.-C., Toga, A. W., Jack, C. R., Weiner, M. W., Thompson, P. M., and Initiative, A. D. N. (2008). Tensor-based morphometry as a neuroimaging biomarker for alzheimer's disease: an mri study of 676 ad, mci, and normal subjects. *Neuroimage*, 43(3):458–69.

Hüttel, S. A., Song, A. W., and McCarthy, G. (2004). *Functional magnetic resonance imaging*. Sinauer Associates.

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., Whitwell, J. L., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L. G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C. S., Krueger, G., Ward, H. A., Metzger, G. J., Scott, K. T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J. P., Fleisher, A. S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., and Weiner, M. W. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. *J Magn Reson Imaging*, 27(4):685–91.

Job, D. E., Whalley, H. C., McIntosh, A. M., Owens, D. G. C., JOHNSTONE, E. C., and Lawrie, S. M. (2006). Grey matter changes can improve the prediction of schizophrenia in subjects at high risk. *BMC Medicine*, 4:29.

Johnstone, T., Somerville, L. H., Alexander, A. L., Oakes, T. R., Davidson, R. J., Kalin, N. H., and Whalen, P. J. (2005). Stability of amygdala bold response to fearful faces over multiple scan sessions. *Neuroimage*, 25(4):1112–1123.

Jovicich, J., Czanner, S., Greve, D. N., Haley, E., van der Kouwe, A. J. W., Gollub, R. L., Kennedy, D., Schmitt, F., Brown, G. G., Macfall, J., Fischl, B., and Dale, A. M. (2006). Reliability in multi-site structural mri studies: effects of gradient non-linearity correction on phantom and human data. *Neuroimage*, 30(2):436–443.

Koller, C. J., Eatough, J. P., Mountford, P. J., and Frain, G. (2006). A survey of mri quality assurance programmes. *Br J Radiol*, 79(943):592–6.

Kong, J.-T., Gollub, R. L., Webb, J. M., Vangel, M. G., and Kwong, K. (2007). Test-retest study of fmri signal change evoked by electroacupuncture stimulation. *Neuroimage*, 34(3):1171–1181.

Krasnow, B., Tamm, L., Greicius, M. D., Yang, T. T., Glover, G. H., Reiss, A. L., and Menon, V. (2003). Comparison of fmri activation at 3 and 1.5 t during perceptual, cognitive, and affective processing. *Neuroimage*, 18(4):813–826.

Krüger, G. and Glover, G. H. (2001). Physiological noise in oxygenation-sensitive magnetic resonance imaging. *Magn Reson Med*, 46(4):631–7.

Le, T. H. and Hu, X. (1997). Methods for assessing accuracy and reliability in functional mri. *NMR in Biomedicine*, 10(4-5):160–164.

Leontiev, O. and Buxton, R. B. (2007). Reproducibility of bold, perfusion, and cmro2 measurements with calibrated-bold fmri. *Neuroimage*, 35(1):175–184.

Leow, A. D., Klunder, A. D., Jack, C. R., Toga, A. W., Dale, A. M., Bernstein, M. A., Britson, P. J., Gunter, J. L., Ward, C. P., Whitwell, J. L., Borowski, B. J., Fleisher, A. S., Fox, N. C., Harvey, D., Kornak, J., Schuff, N., Studholme, C., Alexander, G. E., Weiner, M. W., Thompson, P. M., and Study, A. P. P. (2006). Longitudinal stability of mri for mapping brain change using tensor-based morphometry. *Neuroimage*, 31(2):627–40.

Lerski, R. A., De Wilde, J., Boyce, D., and Ridgway, J. (1998). *Quality control in magnetic resonance imaging*. Institute of Physics and Engineering in Medicine, York.

Liou, M., Su, H.-R., Savostyanov, A. N., Lee, J.-D., Aston, J. A. D., Chuang, C.-H., and Cheng, P. E. (2009). Beyond p-values: averaged and reproducible evidence in fmri experiments. *Psychophysiology*, 46(2):367–78.

Loubinoux, I., Carel, C., Alary, F., Boulanouar, K., Viallard, G., Manelfe, C., Rascol, O., Celsis, P., and Chollet, F. (2001). Within-session and between-session reproducibility of cerebral sensorimotor activation: A test-retest effect evidenced with functional magnetic resonance imaging. *Journal of Cerebral Blood Flow and Metabolism*, 21(5):592–607.

Machielsen, W. C., Rombouts, S. A. R. B., Barkhof, F., Scheltens, P., and Witter, M. P. (2000). fmri of visual encoding: Reproducibility of activation. *Hum Brain Mapping*, 9(3):156–164.

Magnotta, V. A., Friedman, L., and BIRN, F. (2006). Measurement of signal-to-noise and contrast-to-noise in the fbirn multicenter imaging study. *Journal of Digital Imaging*, 19(2):140–147.

Magon, S., Basso, G., Farace, P., Ricciardi, G. K., Beltramello, A., and Sbarbati, A. (2009). Reproducibility of bold signal change induced by breath holding. *Neuroimage*, 45(3):702–12.

Maldjian, J. A., Laurienti, P. J., Driskill, L., and Burdette, J. H. (2002). Multiple reproducibility indices for evaluation of cognitive functional mr imaging paradigms. *AJNR American Journal of Neuroradiology*, 23(6):1030–1037.

Manoach, D., Halpern, E. F., Kramer, T. S., Chang, Y., Goff, D. C., Rauch, S. L., Kennedy, D., and Gollub, R. L. (2001). Test-retest reliability of a functional mri working memory paradigm in normal and schizophrenic subjects. *American Journal of Psychiatry*, 158(6):955–958.

Marshall, I., Simonotto, E., Deary, I. J., Maclullich, A., Ebmeier, K. P., Rose, E. J., Wardlaw, J. M., Goddard, N. H., and Chappell, F. M. (2004). Repeatability of motor and working-memory tasks in healthy older volunteers: Assessment at functional mr imaging. *Radiology*, 233(3):868–877.

Mattay, V. S., Callicott, J. H., Bertolino, A., Santha, A. K. S., Horn, J. D. V., Tallent, K. A., Frank, J. A., and Weinberger, D. R. (1998). Hemispheric control of motor function: a whole brain echo planar fmri study. *Psychiatry Research: Neuroimaging*, 83(1):7–22.

Matthews, P. M. (2001). *Functional MRI: An Introduction to Methods*, chapter An introduction to functional magnetic resonance imaging of the brain, pages 3–34. Oxford University Press, USA.

McGonigle, D. J., Howseman, A. M., Athwal, B. S., Friston, K. J., Frackowiak, R. S. J., and Holmes, A. P. (2000). Variability in fmri: An examination of intersession differences. *Neuroimage*, 11(6):708–734.

McIntosh, A. M., Baig, B. J., HALL, J., Job, D., Whalley, H. C., Lymer, G. K. S., Moorhead, T. W. J., Owens, D. G. C., Miller, P., Porteous, D., Lawrie, S. M., and JOHNSTONE, E. C. (2007). Relationship of catechol-o-methyltransferase variants to brain structure and function in a population at high risk of psychosis. *Biological Psychiatry*, 61(10):1127–34.

Meda, S. A., Giuliani, N. R., Calhoun, V. D., Jagannathan, K., Schretlen, D. J., Pulver, A., Cascella, N., Keshavan, M., Kates, W., Buchanan, R., Sharma, T., and Pearlson, G. D. (2008). A large scale (n=400) investigation of gray matter differences in schizophrenia using optimized voxel-based morphometry. *Schizophr Res*, 101(1-3):95–105.

Miki, A., Raz, J., van Erp, T. G. M., Liu, C.-S. J., Haselgrove, J. C., and Liu, G. T. (2000). Reproducibility of visual activation in functional mr imaging and effects of postprocessing. *AJNR American Journal of Neuroradiology*, 21(5):910–915.

Miller, M., van Horn, J., Wolford, G. L., Handy, T. C., Valsangkar-Smyth, M., Inati, S., Grafton, S., and Gazzaniga, M. S. (2002). Extensive individual differences in

brain activations associated with episodic retrieval are reliable over time. *Journal of Cognitive Neuroscience*.

Miller, M. B., Donovan, C.-L., Horn, J. D. V., German, E., Sokol-Hessner, P., and Wolford, G. L. (2009). Unique and persistent individual patterns of brain activity across different memory retrieval tasks. *Neuroimage*, 48(3):625–35.

Moorhead, T. W. J., Gountouna, V.-E., Job, D. E., McIntosh, A. M., Romaniuk, L., Lymer, G. K. S., Whalley, H. C., Waiter, G. D., Brennan, D., Ahearn, T. S., Cavanagh, J., Condon, B., Steele, J. D., Wardlaw, J. M., and Lawrie, S. M. (2009). Prospective multi-centre voxel based morphometry study employing scanner specific segmentations: Procedure development using calibrain structural mri data. *BMC Med Imaging*, 9(1):8.

Moorhead, T. W. J., Job, D. E., Whalley, H. C., Sanderson, T. L., Johnstone, E. C., and Lawrie, S. M. (2004). Voxel-based morphometry of comorbid schizophrenia and learning disability: analyses in normalized and native spaces using parametric and nonparametric statistical methods. *NeuroImage*, 22(1):188–202.

Noseworthy, M. D., Alfonsi, J., and Bells, S. (2003). Attenuation of brain bold response following lipid ingestion. *Hum Brain Mapping*, 20(2):116–121.

Owen, A. M., Coleman, M. R., Menon, D. K., Johnsrude, I. S., Rodd, J. M., Davis, M. H., Taylor, K., and Pickard, J. D. (2005). Residual auditory function in persistent vegetative state: a combined pet and fmri study. *Neuropsychol Rehabil*, 15(3-4):290–306.

Plichta, M., Schwarz, A., Grimm, O., Morgen, K., Mier, D., Haddad, L., Gerdes, A., Sauer, C., Tost, H., Esslinger, C., Colman, P., Wilson, F., Kirsch, P., and Meyer-Lindenberg, A. (2012). Test–retest reliability of evoked bold signals from a cognitive–emotive fmri test battery. *Neuroimage*, 60(3):1746–1758.

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R. J. A., Kahn, R. S., and Ramsey, N. F. (2007). Test-retest reliability of fmri activation during prosaccades and antisaccades. *Neuroimage*, 36(3):532–542.

Ramsey, N., Tallent, K., van Gelderen, P., Frank, J., Moonen, C., and Weinberger, D. (1996). Reproducibility of human 3d fmri brain maps acquired during a motor task. *Hum Brain Mapping*, 4(2):113–121.

Rau, S., Fesl, G., Bruhns, P., Havel, P., Braun, B., Tonn, J.-C., and Ilmberger, J. (2007). Reproducibility of activations in broca area with two language tasks: a functional mr imaging study. *AJNR American Journal of Neuroradiology*, 28(7):1346–53.

Raz, A., Lieber, B., Soliman, F., Buhle, J., Posner, J., Peterson, B. S., and Posner, M. I. (2005). Ecological nuances in functional magnetic resonance imaging (fmri): psychological stressors, posture, and hydrostatics. *Neuroimage*, 25(1):1–7.

Rombouts, S. A. R. B., Barkhof, F., Hoogenraad, F. G. C., Sprenger, M., and Scheltens, P. (1998). Within-subject reproducibility of visual activation patterns with functional

magnetic resonance imaging using multislice echo planar imaging. *Magn Reson Imaging*, 16(2):105–113.

Rose, E. J., Simonotto, E., and Ebmeier, K. P. (2006). Limbic over-activity in depression during preserved performance on the n-back task. *Neuroimage*, 29(1):203–15.

Rutten, G. J. M., Ramsey, N. F., van Rijen, P. C., and van Veelen, C. W. M. (2002). Reproducibility of fmri-determined language lateralization in individual subjects. *Brain and Language*, 80(3):421–437.

Schnack, H. G., van Haren, N. E. M., Pol, H. E. H., Picchioni, M., Weisbrod, M., Sauer, H., Cannon, T., Huttunen, M., Murray, R., and Kahn, R. S. (2004). Reliability of brain volumes from multicenter mri acquisition: a calibration study. *Hum Brain Mapping*, 22(4):312–320.

Scholz, V. H., Flaherty, A. W., Kraft, E., Keltner, J. R., Kwong, K. K., Chen, Y. I., Rosen, B. R., and Jenkins, B. G. (2000). Laterality, somatotopy and reproducibility of the basal ganglia and motor cortex during motor tasks. *Brain Research*, 879(1-2):204–215.

Schunck, T., Erb, G., Mathis, A., Jacob, N., Gilles, C., Namer, I. J., Meier, D., and Luthringer, R. (2008). Test–retest reliability of a functional mri anticipatory anxiety paradigm in healthy volunteers. *Journal of Magnetic Resonance Imaging*, 27(3):459–468.

Seghier, M. L., ois Lazeyras, F., Pegna, A. J., Annoni, J.-M., Zimine, I., Mayer, E., Michel, C. M., and Khateb, A. (2004). Variability of fmri activation during a phonological and semantic language task in healthy subjects. *Hum Brain Mapping*, 23(3):140–155.

Smith, S. M., Beckmann, C. F., Ramnani, N., Woolrich, M. W., Bannister, P. R., Jenkinson, M., Matthews, P. M., and McGonigle, D. J. (2005). Variability in fmri: A re-examination of inter-session differences. *Hum Brain Mapping*, 24(3):248–257.

Specht, K., Willmes, K., Shah, N. J., and Jäncke, L. (2003). Assessment of reliability in functional imaging studies. *Journal of Magnetic Resonance Imaging*, 17(4):463–471.

Spencer, M. D., Moorhead, T. W. J., Gibson, R. J., McIntosh, A. M., Sussmann, J. E. D., Owens, D. G. C., Lawrie, S. M., and JOHNSTONE, E. C. (2008). Low birthweight and preterm birth in young people with special educational needs: a magnetic resonance imaging analysis. *BMC Medicine*, 6:1.

Stark, R., Schienle, A., Walter, B., Kirsch, P., Blecker, C., u Ott, Schäfer, A., Sammer, G., Zimmermann, M., and Vaitl, D. (2004). Hemodynamic effects of negative emotional pictures - a test-retest analysis. *Neuropsychobiology*, 50(1):108–118.

Stefanovic, B., Warnking, J. M., Rylander, K. M., and Pike, G. B. (2006). The effect of global cerebral vasodilation on focal activation hemodynamics. *Neuroimage*, 30(3):726–734.

Stonnington, C. M., Tan, G., Klöppel, S., Chu, C., Draganski, B., Jack, C. R., Chen, K., Ashburner, J., and Frackowiak, R. S. J. (2008). Interpreting scan data acquired from multiple scanners: a study with alzheimer's disease. *Neuroimage*, 39(3):1180–5.

Suckling, J., Barnes, A., Job, D., Brennan, D., Lymer, K., Dazzan, P., Marques, T. R., Mackay, C., McKie, S., Williams, S. R., Williams, S. C. R., Deakin, B., and Lawrie, S. (2011). The neuro/psygrid calibration experiment: Identifying sources of variance and bias in multicenter mri studies. *Human brain mapping*.

Suckling, J., Ohlssen, D., Andrew, C., Johnson, G., Williams, S. C., Graves, M., Chen, C.-H., Spiegelhalter, D., and Bullmore, E. (2008). Components of variance in a multicentre functional mri study and implications for calculation of statistical power. *Hum Brain Mapping*, 29(10):1111–1122.

Sutton, B. P., Goh, J., Hebrank, A., Welsh, R. C., Chee, M. W., and Park, D. C. (2008). Investigation and validation of intersite fmri studies using the same imaging hardware. *Journal of Magnetic Resonance Imaging*, 28(1):21–28.

Swallow, K. M., Braver, T. S., Snyder, A. Z., Speer, N. K., and Zacks, J. M. (2003). Reliability of functional localization using fmri. *Neuroimage*, 20(3):1561–1577.

Tegeler, C., Strother, S. C., Anderson, J. R., and Kim, S.-G. (1999). Reproducibility of bold-based functional mri obtained at 4 t. *Hum Brain Mapping*, 7(4):267–283.

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., and Poline, J.-B. (2007). Analysis of a large fmri cohort: Statistical and methodological issues for group analyses. *Neuroimage*, 35(1):105–120.

Thomason, M. E., Foland, L. C., and Glover, G. H. (2007). Calibration of bold fmri using breath holding reduces group variance during a cognitive task. *Hum Brain Mapping*, 28(1):59–68.

Tijms, B. M., Seriès, P., Willshaw, D. J., and Lawrie, S. M. (2011). Similarity-based extraction of individual networks from gray matter mri scans. *Cerebral cortex (New York, NY : 1991)*.

Tofts, P. (1998). Standardisation and optimisation of magnetic resonance techniques for multicentre studies. *Journal of Neurology Neurosurgery and Psychiatry*, 64(Suppl1):S37–43.

Tofts, P. (2003). *Quantitative MRI of the Brain*. John Wiley and Sons.

van Haren, N. E. M., Cahn, W., Pol, H. E. H., Schnack, H. G., Caspers, E., Lemstra, A., Sitskoorn, M. M., Wiersma, D., van den Bosch, R. J., and Dingemans, P. M. (2003). Brain volumes as predictor of outcome in recent-onset schizophrenia: a multi-center mri study. *Schizophrenia Research*, 64(1):41–52.

Vlieger, E.-J., Lavini, C., Majoie, C. B., and den Heeten, G. J. (2003). Reproducibility of functional mr imaging results using two different mr systems. *AJNR American Journal of Neuroradiology*, 24(4):652–657.

Voyvodic, J. T. (2006). Activation mapping as a percentage of local excitation: fmri stability within scans, between scans and across field strengths. *Magn Reson Imaging*, 24(9):1249–1261.

Wagner, K., Frings, L., Quiske, A., Unterrainer, J., Schwarzwald, R., Spreer, J., Halsband, U., and Schulze-Bonhage, A. (2005). The reliability of fmri activations in the medial temporal lobes in a verbal episodic memory task. *Neuroimage*, 28(1):122–131.

Wei, X., Yoo, S.-S., Dickey, C. C., Zou, K. H., Guttmann, C. R. G., and Panych, L. P. (2004). Functional mri of auditory verbal working memory: long-term reproducibility analysis. *Neuroimage*, 21(3):1000–1008.

Wilke, M., Holland, S. K., Altaye, M., and Gaser, C. (2008). Template-o-matic: a toolbox for creating customized pediatric templates. *NeuroImage*, 41(3):903–13.

Wilke, M., Schmithorst, V. J., and Holland, S. K. (2003). Normative pediatric brain data for spatial normalization and segmentation differs from standard adult data. *Magnetic Resonance in Medicine*, 50(4):749–57.

Yetkin, F., McAuliffe, T., Cox, R., and Haughton, V. M. (1996). Test-retest precision of functional mr in sensory and motor task activation. *AJNR American Journal of Neuroradiology*, 17(1):95–98.

Yoo, S.-S., Wei, X., Dickey, C. C., Guttmann, C. R. G., and Panych, L. P. (2005). Long-term reproducibility analysis of fmri using hand motor task. *International Journal of Neuroscience*, 115(1):55–77.

Zou, K. H., Greve, D. N., Wang, M., Pieper, S. D., Warfield, S. K., White, N. S., Manandhar, S., Brown, G. G., Vangel, M. G., Kikinis, R., Wells, W. M., and BIRN, F. (2005). Reproducibility of functional mr imaging: preliminary results of prospective multi-institutional study performed by biomedical informatics research network. *Radiology*, 237(3):781–789.