Ann. N.Y. Acad. Sci. ISSN 0077-8923

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Issue: *The Year in Cognitive Neuroscience*

How reliable are the results from functional magnetic resonance imaging?

Craig M. Bennett and Michael B. Miller

Department of Psychology, University of California at Santa Barbara, Santa Barbara, California

Address for correspondence: Craig M. Bennett, Department of Psychology, University of California, Santa Barbara, Santa Barbara, CA 93106, USA. bennett@psych.ucsb.edu

Functional magnetic resonance imaging (fMRI) is one of the most important methods for *in vivo* investigation of cognitive processes in the human brain. Within the last two decades, an explosion of research has emerged using fMRI, revealing the underpinnings of everything from motor and sensory processes to the foundations of social cognition. While these results have revealed the potential of neuroimaging, important questions regarding the reliability of these results remain unanswered. In this paper, we take a close look at what is currently known about the reliability of fMRI findings. First, we examine the many factors that influence the quality of acquired fMRI data. We also conduct a review of the existing literature to determine if some measure of agreement has emerged regarding the reliability of fMRI. Finally, we provide commentary on ways to improve fMRI reliability and what questions remain unanswered. Reliability is the foundation on which scientific investigation is based. How reliable are the results from fMRI?

Keywords: fMRI; statistics; reliability

Introduction

Reliability is the cornerstone of any scientific enterprise. Issues of research validity and significance are relatively meaningless if the results of our experiments are not trustworthy. It is the case that reliability can vary greatly depending on the tools being used and what is being measured. Therefore, it is imperative that any scientific endeavor be aware of the reliability of its measurements.

Surprisingly, most functional magnetic resonance imaging (fMRI) researchers have only a vague idea of how reliable their results are. Reliability is not a typical topic of conversation between most investigators and only a small fraction of papers investigating fMRI reliability have been published. This became an important issue in 2009 as a paper by Vul and colleagues[1] set the stage for debate. Their paper, originally entitled "Voodoo Correlations in Social Neuroscience," was focused on a statistical problem known as the "nonindependence error." Critical to their argument was the reliability of functional imaging results. Vul and colleagues argued that test–retest variability of fMRI results placed an "upper bound" on the strength of possible correlations between fMRI data and behavioral measures:

$$r_{(\text{ObservedA,ObservedB})} = r_{(\text{A,B})} * \text{sqrt}(\text{reliability}_{\text{A}} * \text{reliability}_{\text{B}})$$

This calculation reflects that the strength of a correlation between two measures is a product of the measured relationship and the reliability of the measurements.[1,2] Vul and colleagues specified that behavioral measures of personality and emotion have a reliability of around 0.8 and that fMRI results have a reliability of around 0.7. Not everyone agreed. Across several written exchanges multiple research groups debated what the "actual reliability" of fMRI was. Jabbi and colleagues[3] stated that the reliability of fMRI could be as high as 0.98. Lieberman and colleagues split the difference and argued that fMRI reliability was likely around 0.90.[4] While much ink was spilled debating the reliability of fMRI results, very little consensus was reached regarding an appropriate approximation of its value.

The difficulty of detecting signal (what we are trying to measure) from among a sea of noise (everything else we do not care about) is a constant struggle for all scientists. It influences what effects can be examined and is directly tied to the reliability of research results. What follows in this paper is

doi: 10.1111/j.1749-6632.2010.05446.x

a multifaceted examination of fMRI reliability. We examine why reliability is a critical metric of fMRI data, discuss what factors influence the quality of the blood oxygen level dependent (BOLD) signal, and investigate the existing reliability literature to determine if some measure of agreement has emerged across studies. Fundamentally, there is one critical question that this paper seeks to address: if you repeat your fMRI experiment, what is the likelihood you will get the same result?

Pragmatics of reliability

Why worry about reliability at all? As long as investigators are following accepted statistical practices and being conservative in the generation of their results, why should the field be bothered with how reproducible the results might be? There are, at least, four primary reasons why test–retest reliability should be a concern for all fMRI researchers.

Scientific truth

Although it is a simple statement that can be taken straight out of an undergraduate research methods course, an important point must be made about reliability in research studies: it is the foundation on which scientific knowledge is based. Without reliable, reproducible results no study can effectively contribute to scientific knowledge. After all, if a researcher obtains a different set of results today than they did yesterday, what has really been discovered? To ensure the long-term success of functional neuroimaging it is critical to investigate the many sources of variability that impact reliability. It is a strong statement, but if results do not generalize from one set of subjects to another or from one scanner to another then the findings are of little value scientifically.

Clinical and diagnostic applications

The longitudinal assessment of changes in regional brain activity is becoming increasingly important for the diagnosis and treatment of clinical disorders. One potential use of fMRI is for the localization of specific cognitive functions before surgery. A good example is the localization of language function prior to tissue resection for epilepsy treatment.[5] This is truly a case where an investigator does not want a slightly different result each time they conduct the scan. If fMRI is to be used for surgical planning or clinical diagnostics then any issues of reliability must be quantified and addressed.

Evidentiary applications

The results from functional imaging are increasingly being submitted as evidence into the United States legal system. For example, results from a commercial company called No Lie MRI (San Diego, CA, USA; http://www.noliemri.com/) were introduced into a juvenile sex abuse case in San Diego during the spring of 2009. The defense was attempting to introduce the fMRI results as scientific justification of their client's claim of innocence. A concerted effort from imaging scientists, including in-person testimony from Marc Raichle, eventually forced the defense to withdraw the request. Although the fMRI results never made it into this case, it is clear that fMRI evidence will be increasingly common in the courtroom. What are the larger implications if the reliability of this evidence is not as trustworthy as we assume?

Scientific collaboration

A final pragmatic dimension of fMRI reliability is the ability to share data between researchers. This is already a difficult challenge, as each scanner has its own unique sources of error that become part of the data.[6] Early evidence has indicated that the results from a standard cognitive task can be quite similar across scanners.[7,8] Still, concordance of results remains an issue that must be addressed for large-scale, collaborative intercenter investigations. The ultimate level of reliability is the reproducibility of results from any equivalent scanner around the world and the ability to integrate this data into larger investigations.

What factors influence fMRI reliability?

The ability of fMRI to detect meaningful signals is limited by a number of factors that add error to each measurement. Some of these factors include thermal noise, system noise in the scanner, physiological noise from the subject, non-task-related cognitive processes, and changes in cognitive strategy over time.[9,10] The concept of reliability is, at its core, a representation of the ability to routinely detect relevant signals from this background of meaningless noise. If a voxel timeseries contains a large amount of signal then the primary sources of variability are actual changes in blood flow related to

neural activity within the brain. Conversely, in a voxel containing a large amount of noise the measurements are dominated by error and would not contain meaningful information. By increasing the amount of signal, or decreasing the amount of noise, a researcher can effectively increase the quality and reliability of acquired data.

The quality of data in magnetic resonance imaging is typically measured using the signal-to-noise ratio (SNR) of the acquired images. The goal is to maximize this ratio. Two kinds of SNRs are important for functional MRI. The first is the image SNR. It is related to the quality of data acquired in a single fMRI volume. Image SNR is typically computed as the mean signal value of all voxels divided by the standard deviation of all voxels in a single image:

$$\mathrm{SNR}_{\mathrm{image}} = \mu_{\mathrm{image}}/\sigma_{\mathrm{image}}$$

Increasing the image SNR will improve the quality of data at a single point in time. However, most important for functional neuroimaging is the amount of signal present in the data across time. This makes the temporal SNR (tSNR) perhaps the most important metric of data for functional MRI. It represents the SNR of the timeseries at each voxel:

$$\mathrm{SNR}_{\mathrm{temporal}} = \mu_{\mathrm{timeseries}}/\sigma_{\mathrm{timeseries}}$$

The tSNR is not the same across all voxels in the brain. Some regions will have higher or lower tSNR depending on location and constitution. For example, there are documented differences in tSNR between gray matter and white matter.[11] The typical tSNR of fMRI can also vary depending on the same factors that influence image SNR.

Another metric of data quality is the contrast-to-noise ratio (CNR). This refers to the ability to maximize differences between signal intensity in different areas in an image (image CNR) or to maximize differences between different points in time (temporal CNR). With regard to functional neuroimaging, the temporal CNR represents the maximum relative difference in signal intensity that is represented within a single voxel. In a voxel with low CNR there would be very little difference between two conditions of interest. Conversely, in a voxel with high CNR there would be relatively large differences between two conditions of interest. The image CNR is not critical to fMRI, but having a high temporal CNR is very important for detecting task effects.

It is generally accepted that fMRI is a rather noisy measurement with a characteristically low tSNR, requiring extensive signal averaging to achieve effective signal detection.[12] The following sections provide greater detail on the influence of specific factors on the SNR/tSNR of functional MRI data. We break these factors down by the influence of differences in image acquisition, the image analysis pipeline, and the contribution of the subjects themselves.

SNR influences of MRI acquisition

The typical high-field MRI scanner is a precision superconducting device constructed to very exact manufacturing tolerances. Still, the images it produces can be somewhat variable depending on a number of hardware and software variables. With regard to hardware, one well-known influence on the SNR of MRI is the strength of the primary B0 magnetic field.[13,14] Doubling this field, such as moving from 1.5 to 3.0 Tesla field strength, can theoretically double the SNR of the data. The B0 field strength is especially important for fMRI, which relies on magnetic susceptibility effects to create the BOLD signal.[15] Hoenig and colleagues showed that, relative to a 1.5 Tesla magnet, a 3.0 Tesla fMRI acquisition had 60–80% more significant voxels.[16] They also demonstrated that the CNR of the results was 1.3 times higher than those obtained at 1.5 Tesla. The strength and slew rate of the gradient magnets can have a similar impact on SNR. Advances in head coil design are also notable, as parallel acquisition head coils have increased radiofrequency reception sensitivity.

It is important to note that there are negative aspects of higher field strength as well. Artifacts due to physiological effects and susceptibility are all increasingly pronounced at higher fields. The increased contribution of physiological noise reduces the expected gains in SNR at high field.[9] The increasing contribution of susceptibility artifacts can virtually wipe out areas of orbital prefrontal cortex and inferior temporal cortex.[17] Also, in terms of tSNR there are diminishing returns with each step up in B0 field strength. At typical fMRI spatial resolution values tSNR approaches an asymptotic limit between 3 and 7 Tesla.[9,18]

Looking beyond the scanner hardware, the parameters of the fMRI acquisition can also have a significant impact on the SNR/CNR of the final images. For example, small changes in the voxel size of a sequence can dramatically alter the final SNR. Moving from 1.5 to 3.0 mm^3 voxels can potentially increase the acquisition SNR by a factor of 8, but at a cost of spatial resolution. Some other acquisition variables that will influence the acquired SNR/CNR are: repetition time (TR), echo time (TE), bandwidth, slice gap, and k-space trajectory. For example, Moser and colleagues found that optimizing the flip angle of their acquisition could approximately double the SNR of their data in a visual stimulation task.[19] Further, the effect of each parameter varies according to the field strength of the magnet.[18] The optimal parameter set for a 3 Tesla system may not be optimal with a 7 Tesla system.

The ugly truth is that any number of factors in the control room or magnet suite can increase noise in the images. A famous example from one imaging center was when the broken filament from a light bulb in a distant corner of the magnet suite started causing visible sinusoidal striations in the acquired EPI images. This is an extreme example, but it makes the point that the scanner is a precision device that is designed to operate in a narrow set of well-defined circumstances. Any deviation from those circumstances will increase noise, thereby reducing SNR and reliability.

SNR considerations of analysis methods

The methods used to analyze fMRI data will affect the reliability of the final results. In particular, those steps taken to reduce known sources of error are critical to increasing the final SNR/CNR of preprocessed images. For example, spatial realignment of the EPI data can have a dramatic effect on lowering movement-related variance and has become a standard part of fMRI preprocessing.[20,21] Recent algorithms can also help remove remaining signal variability due to magnetic susceptibility induced by movement.[22] Temporal filtering of the EPI timeseries can reduce undesired sources of noise by frequency. The use of a high-pass filter is a common method to remove low-frequency noise, such as signal drift due to the scanner.[23] Spatial smoothing of the data can also improve the SNR/CNR of an image. There is some measure of random noise added to the true signal of each voxel during acquisition. Smoothing across voxels can help to average out error across the area of the smoothing filter.[24] It can also help account for local differences in anatomy across subjects. Smoothing is most often done using a Gaussian kernel of approximately 6–12 mm^3 full width at half maximum.

There has been some degree of standardization regarding preprocessing and statistical approaches in fMRI. For instance, Mumford and Nichols found that approximately 92% of group fMRI results were computed using an ordinary least squares estimation of the general linear model.[25] Comparison studies with carefully standardized processing procedures have shown that the output of standard software packages can be very similar.[26,27] However, in actual practice, the diversity of tools and approaches in fMRI increases the variability between sets of results. The functional imaging analysis contest in 2005 demonstrated that prominent differences existed between fMRI results generated by different groups using the same original data set. On reviewing the results, the organizers concluded that brain regions exhibiting robust signal changes could be quite similar across analysis techniques, but the detection of areas with lower signal was highly variable.[28] It remains the case that decisions made by the researcher regarding how to analyze the data will impact what results are found.

Strother and colleagues have done a great deal of research into the influence of image processing pipelines using a predictive modeling framework.[29–31] They found that small changes in the processing pipeline of fMRI images have a dramatic impact on the final statistics derived from that data. Some steps, such as slice timing correction, were found to have little influence on the results from experiments with a block design. This is logical, given the relative insensitivity of block designs to small temporal shifts. However, the steps of motion correction, high-pass filtering, and spatial smoothing were found to significantly improve the analysis. They reported that the optimization of preprocessing pipelines improved both intrasubject and between-subject reproducibility of results.[31] Identifying an optimal set of processing steps and parameters can dramatically improve the sensitivity of an analysis.

SNR influences of participants

The MRI system and fMRI analysis methods have received a great deal of attention with regard to SNR. However, one area that may have the greatest contribution to fMRI reliability is how stable/unstable the patterns of activity within a single subject can be. After all, a test–retest methodology involving human beings is akin to hitting a moving target. Any discussion of test–retest reliability in fMRI has to take into consideration the fact that the cognitive state of a subject is variable over time.

There are two important ways that a subject can influence reliability within a test–retest experimental design. The first involves within-subject changes that take place over the course of a single session. For instance, differences in attention and arousal can significantly modulate subsequent responses to sensory stimulation.[32–34] Variability can also be caused by evolving changes in cognitive strategy used during tasks such as episodic retrieval.[35,36] If a subject spontaneously shifts to a new decision criterion midway during a session then the resulting data may reflect the results of two different cognitive processes. Finally, learning will take place with continued task experience, shifting the pattern of activity as brain regions are engaged and disengaged during task-relevant processing.[37–39] For studies investigating learning this is a desired effect, but for others this is an undesired source of noise.

The second influence on reliability is related to physiological and cognitive changes that may take place within a subject between the test and retest sessions. Within 24 h an infinite variety of reliability-reducing events can take place. All of the above factors may show changes over the days, weeks, months, or years between scans. These changes may be even more dramatic depending on the amount of time between scanning sessions.

Estimates of fMRI reliability

A diverse array of methods has been created for measuring the reliability of fMRI. What differs between them is the specific facet of reliability they are intended to quantify. Some methods are only concerned with significant voxels. Other methods address similarity in the magnitude of estimated activity across all voxels. The choice of how to calculate reliability often comes down to which aspect of the results are desired to remain stable over time.

Measuring stability of super-threshold extent

Do you want the voxels that are significant during the test scan to still be significant during the retest scan? This would indicate that super-threshold voxels are to remain above the threshold during subsequent sessions. The most prevalent method to quantify this reliability is the cluster overlap method. The cluster overlap method is a measure revealing what set of voxels are considered to be super-threshold during both test and retest sessions.

Two approaches have been used to calculate cluster overlap. The first, and by far most prevalent, is a measure of similarity known as the Dice coefficient. It was first used to calculate fMRI cluster overlap by Rombouts and colleagues and has become a standard measure of result similarity.[40] It is typically calculated by the following equation:

$$R_{\text{overlap}} = 2(V_{\text{overlap}})/(V_1 + V_2)$$

Results from the Dice equation can be interpreted as the number of voxels that will overlap divided by the average number of significant voxels across sessions. Another approach to calculating similarity is the Jaccard index. The Jaccard index has the advantage of being readily interpretable as the percent of voxels that are shared, but is infrequently used in the investigation of reliability. It is typically calculated by the following equation:

$$R_{\text{overlap}} = V_{\text{overlap}}/(V_1 + V_2 - V_{\text{overlap}})$$

Results from the Jaccard equation can be interpreted as the number of overlapping voxels divided by the total number of unique voxels in all sessions. For both the Dice and Jaccard methods, a value of 1.0 would indicate that all super-threshold voxels identified during the test scan were also active in the retest scan, and vice-versa. A value of 0.0 would indicate that no voxels in either scan were shared between the test and retest sessions. See Figure 1, for a graphical representation of overlapping results from two runs in an example data set.

The main limitation of all cluster overlap methods is that they are highly dependent on the statistical threshold used to define what is "active." Duncan and colleagues demonstrated that the reported reliability of the cluster overlap method decreases as the significance threshold is increased.[41] Similar results were reported by Rombouts and colleagues, who

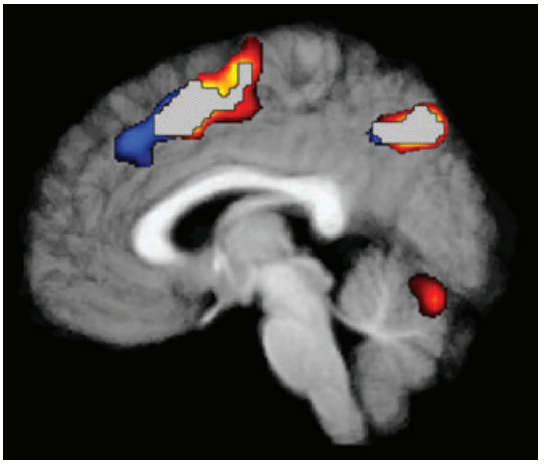

Figure 1. Visualization of cluster overlap using two runs of data from a two-back working memory task. The regions in red represent significant clusters from the first run and regions in blue represent significant clusters from the second run. The crosshatched region represents the overlapping voxels that were significant in both runs. Important to note is that not all significant voxels remained significant across the two runs. One cluster in the cerebellum did not replicate at all. Data is from Bennett *et al.*[49]

found nonlinear changes in cluster overlap reliability across multiple levels of significance.[42]

These overlap statistics seek to represent the proportion of voxels that remain significant across repetitions relative to the proportion that are significant in only a subset of the results. Another, similar approach would be to conduct a formal conjunction analysis between the repetitions. The goal of this approach would be to uniquely identify those voxels that are significant in all sessions. One example of this approach would be the "Minimum Statistic compared to the Conjunction Null" (MS/CN) of Nichols and colleagues[43] Using this approach a researcher could threshold the results, allowing for the investigation of reliability with a statistical criterion.

A method similar to cluster overlap, called voxel counting, was reported in early papers. The use of voxel counting simply evaluated the total number of activated voxels in the test and retest images. This has proven to be a suboptimal approach for the examination of reliability, as it is done without regard to the spatial location of significant voxels.[44] An entirely different set of results could be observed in each image yet they could contain the same number of significant voxels. As a consequence this method is no longer used.

Measuring stability of activity in significant clusters

Do you want the estimated magnitude of activity in each cluster to be stable between the test scan and the retest scan? This is a more stringent criteria than simple extent reliability, as it is necessary to replicate the exact degree of activation and not simply what survives thresholding. The most standard method to quantify this reliability is through an intraclass correlation (ICC) of the time1–time2 cluster values. The ICC is different from the traditional Pearson product–moment correlation as it is specialized for data of one type, or class. Although there are many versions of the ICC, it is typically taken to be a ratio of the variance of interest divided by the total variance.[45,46] The ICC can be computed as follows:

$$\mathrm{ICC} = \sigma^2_{\mathrm{between}} / (\sigma^2_{\mathrm{between}} + \sigma^2_{\mathrm{within}})$$

One of the best reviews of the ICC was completed by Shrout and Fleiss,[46] who detailed six types of ICC calculation and when each is appropriate to use. One advantage of the ICC is that it can be interpreted similarly to the Pearson correlation. A value of 1.0 would indicate near-perfect agreement between the values of the test and retest sessions, as there would be no influence of within-subject variability. A value of 0.0 would indicate that there was no agreement between the values of the test and retest sessions, because within-subject variability would dominate the equation.

Studies examining reliability using ICCs are often computed based on summary values from regions of interest (ROIs). Caceras and colleagues[47] compared four methods commonly used to compute ROI reliability using ICCs. The median (ICC) is the median of the ICC values from within an ROI. $\mathrm{ICC}_{\mathrm{med}}$ is the median ICC of the contrast values. $\mathrm{ICC}_{\mathrm{max}}$ is the calculation of ICC values at the peak-activated voxel within an activated cluster. $\mathrm{ICC}_{\mathrm{v}}$ is defined the intravoxel reliability, a measure of the total variability that can be explained by the intravoxel variance.

There are several notable weaknesses to the use of ICC in calculating reliability. First, the generalization of ICC results is limited because calculation is specific to the data set under

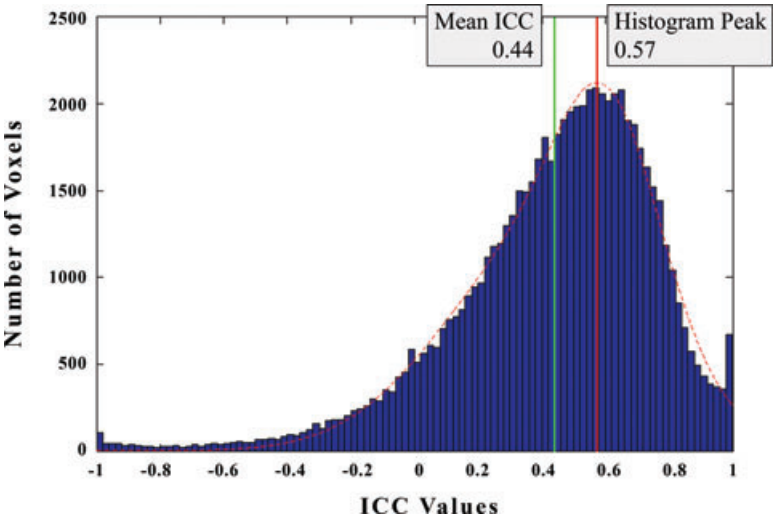

Figure 2. **Histogram showing the frequency of voxelwise ICC values during a two-back working memory task. The histogram was computed from a data set of 16 subjects using 100 bins between ICC values of 1.0 and −1.0. The distribution of values is negatively skewed, with a mean ICC value of ICC = 0.44 and the most frequently occurring value of ICC = 0.57. Data is from Bennett *et al.*[49]**

investigation. An experiment with high intersubject variability could have different ICC values relative to an experiment with low intersubject variability, even if the stability of values over time is the same. As discussed later in this chapter, this can be particularly problematic when comparing the reliability of clinical disorders to that of normal controls. Second, because of the variety of ICC subtypes there can often be confusion regarding which one to use. Using an incorrect subtype can result in quite different reliability estimates.[48]

Measuring voxelwise reliability of the whole brain

Do you want to know the reliability of results on a whole-brain, voxelwise basis? Completing a voxelwise calculation would indicate that the level of activity in all voxels should remain consistent between the test and retest scans. This is the strictest criterion for reliability. It yields a global measure of concordance that indicates how effectively activity across the whole brain is represented in each test–retest pairing. Very few studies have examined reliability using this approach, but it may be one of the most valuable metrics of fMRI reliability. This is one of the few methods that gives weight to the idea that the estimated activity should remain consistent between test and retest, even if the level of activity is close to zero.

Figure 2 is an example histogram plot from our own data that shows the frequency of ICC values for all voxels across the whole brain during a two-back working memory task.[49] The mean and mode of the distribution is plotted. It is quickly apparent that there is a wide range of ICC reliability values across the whole brain, with some voxels having almost no reliability and others approaching near perfect reliability.

Other reliability methods

Numerous other methods have also been used to measure the reliability of estimated activity. Some of these include maximum likelihood (ML), coefficient of variation (CV), and variance decomposition. Although these methods are in the minority by frequency of use, this does not diminish their utility in examining reliability. This is especially true with regard to identifying the sources of test–retest variability that can influence the stability of results.

One particularly promising approach for the quantification of reliability is predictive modeling. Predictive modeling measures the ability of a training set of data to predict the structure of a testing set of data. One of the best established

modeling techniques within functional neuroimaging is the nonparametric prediction, activation, influence, and reproducibility sampling (NPAIRS) approach by Strother and colleagues[29,30] Within the NPAIRS modeling framework separate metrics of prediction and reproducibility are generated.[50] The first, prediction accuracy, evaluates classification in the temporal domain, predicting which condition of the experiment each scan belongs to. The second metric, reproducibility, evaluates the model in the spatial domain, comparing patterns of regional brain activity over time. Although this approach is far more complicated than the relatively simple cluster overlap or ICC metrics, predictive modeling does not suffer from many of the drawbacks that these methods have. NPAIRS, and other predictive modeling approaches, enable a much more thorough examination of fMRI reliability.

Some studies have investigated fMRI reliability using the Pearson product–moment (r) correlation. Intuitively, this is a logical method to use, as it measures the relationship between two variables. However, it is generally held that the Pearson product–moment correlation is not an ideal measure of test–retest reliability. Safrit identified three reasons why the product–moment correlation should not be used to calculate reliability.[51] First, the Pearson product–moment correlation is setup to determine the relationship between two variables, not the stability of a single variable. Second, it is difficult to measure reliability with the Pearson product–moment correlation beyond a single test–retest pair. It becomes increasingly awkward to quantify reliability with two or more retest sessions. One can try to average over multiple pairwise Pearson product–moment correlations between the multiple sessions, but it is far easier to take the ANOVA approach of the ICC and examine it from the standpoint of between- and within-subject variability. Third, the Pearson product–moment correlation cannot detect systematic error. This would be the case when the retest values deviate by a similar degree, such as adding a constant value to all of the original test values. The Pearson product–moment correlation would remain the same, while an appropriate ICC would indicate that the test–retest agreement is not exact. Although the use of ICC measures has its own set of issues, it is generally a more appropriate tool for the investigation of test–retest reliability.

Review of existing reliability estimates

Since the advent of fMRI some results have been common and quite easily replicated. For example, activity in primary visual cortex during visual stimulation has been thoroughly studied. Other fMRI results have been somewhat difficult to replicate. What does the existing literature have to say regarding the reliability of fMRI results?

There have been a number of individual studies investigating the test–retest reliability of fMRI results, but few articles have reviewed the entire body of literature to find trends across studies. To obtain a more effective estimate of fMRI reliability, we conducted a survey of the existing literature on fMRI reliability. To find papers for this investigation, we searched for "test–retest fMRI" using the NCBI PubMed database (www.pubmed.gov). This search yielded a total of 183 papers, 37 of which used fMRI as a method of investigation, used a general linear model to compute their results, and provided test–retest measures of reliability. To broaden the scope of the search, we then went through the reference section of the 37 papers found using PubMed to look for additional works not identified in the initial search. There were 26 additional papers added to the investigation through this secondary search method. The total number of papers retrieved was 63. Each paper was examined with regard to the type of cognitive task, kind of fMRI design, number of subjects, and basis of reliability calculation.

We have separated out the results into three groups: those that used the voxel overlap method, those that used ICC, and papers that used other calculation methods. The results of this investigation can be seen in Tables 1–3. In the examination of cluster overlap values in the literature, we attempted to only include values that were observed at a similar significance threshold across all of the papers. The value we chose as the standard was P(uncorrected) < 0.001. Deviations from this standard approach are noted in the tables.

Conclusions from the reliability review

What follows are some general points that can be taken away from the reliability survey. Some of the conclusions that follow are quantitative results from the review and some are qualitative descriptions of trends that were observed as we conducted the review.

Table 1. Results of examined papers using intraclass correlation as a reliability metric

First author	Year	Task	Design	Type	Basis	Contrast	No. of Subs	Approximate T–R interval	Min ICC	Mean ICC	Max ICC
Caceres[47]	2009[a]	Auditory target detection	Block	Sig. voxels	Contrast values	Task vs. rest	10	3 months	–	0.35	–
Caceres[47]	2009[a]	N-back working memory	Block	Sig. voxels	Contrast values	Task vs. control	10	3 months	–	0.49	–
Freyer[94]	2009[b]	Probabilistic reversal learning	Event	All voxels	Contrast values	Task vs. control	10	16 weeks	–	–	–
Gountouna[95]	2009	Finger tapping	Block	ROI	Contrast values	Task vs. rest	14	Unknown	0.23	0.53	0.72
Bosnell[71]	2008	Hand tapping	Block	ROI	Percent signal change	Task vs. rest	22	<1 day	–	0.82	–
Friedman[8]	2008[a,c]	Finger tapping	Block	ROI	Percent signal change	Task vs. rest	5	1 day	0.47	0.74	0.85
Schunck[96]	2008	Anticipatory anxiety	Block	ROI	Percent signal change	Task vs. rest	14	10 days	−0.06	0.34	0.66
Kong[97]	2007	Finger tapping	Block	ROI	Percent signal change	Task vs. rest	8	1 week	0.00	0.37	0.76
Kong[97]	2007	Acupuncture	Block	ROI	Percent signal change	Task vs. rest	8	1 week	0.00	0.16	0.54
Raemaekers[58]	2007	Prosaccade/ antisaccade	Event	All voxels	*t*-Statistic values	Task vs. rest	12	1 week	−0.08	–	0.79
Aron[52]	2006	Probabilistic classification learning	Event	ROI	Contrast values	Task vs. rest	8	59 weeks	0.76	0.88	0.99
Johnstone[98]	2005	Amygdala-facial affect localizer	Block	Amygdala ROI	Contrast values	Task vs. rest	15	8 weeks	0.02	0.38	0.63
Wei[99]	2004	Auditory two-back	Block	ROI	Activation index	Task vs. rest	8	9 weeks	0.14	0.43	0.71
Specht[100]	2003[b]	Visual attention	Event	Sig. voxels	Percent signal change	Task vs. rest	5	8 weeks	–	–	–
Manoach[66]	2001	Sternberg item recognition	Event	ROI	Percent signal change	Task vs. control	7	14 weeks	0.23	0.52	0.81
Mean value									0.17	0.50	0.75

[a]Median value given.
[b]Data presented as graphs or figures, unable to quantify values.
[c]Data acquired from multiple scanners.

A diverse collection of methods have been used to assess fMRI reliability

The first finding mirrors the earlier discussion on reliability calculation. A very diverse collection of methods has been used to investigate fMRI reliability. This list includes: ICC, cluster overlap, voxel counts, receiver operating characteristic (ROC) curves, ML, conjunction analysis, Cohen's kappa index, CV, Kendall's *W*, laterality index (LI), variance component decomposition, Pearson correlation, predictive modeling, and still others. Although this diversity of methods has created converging evidence of fMRI reliability, it has also limited the ability to compare and contrast the results of existing reliability studies.

ICC and cluster overlap methods

Although there have been a number of methods used to investigate reliability, the two that stand out by frequency of use are cluster overlap and ICC. One advantage of these methods is that they are easy to calculate. The equations are simple to understand, easy to implement, and fast to process. A second advantage of these methods is their easy interpretation by other scientists. Even members of the general public can understand the concept behind the

Table 2. Results of examined papers using cluster overlap as a reliability metric

First author	Year	Task	Design	Calculation	Basis	Contrast	Threshold	No. of Subs	Approximate T–R interval	Dice overlap		
										Min overlap	Avg. overlap	Max overlap
Duncan[41]	2009[a]	One-back object/word localizer	Block	Dice	ROI	Task vs. rest	P(uncorr) < 0.001	45	<1 h	0.380	0.435	0.490
Gountouna[95]	2009	Finger tapping	Block	Dice	Sig. voxels	Task vs. rest	P(corr) < 0.05	14	Not given	0.410	0.455	0.500
Meindl[101]	2009	Default mode	Free	Dice	ICA	Component	P(uncorr) < 0.05	18	1 week	0.080	0.390	0.760
Feredoes[102]	2007[b,e]	Button press	Event	Custom	Sig. voxels	Task vs. rest	P(corr) < 0.05	6	1 week	–	0.245	–
Feredoes[102]	2007[b,e]	Delayed recognition	Event	Custom	Sig. voxels	Task vs. control	P(corr) < 0.05	6	1 week	0.000	0.210	0.413
Raemaekers[58]	2007	Prosaccade/ Antisaccade	Event	Dice	Sig. voxels	Task vs. rest	P(corr) < 0.005	12	1 week	0.760	0.785	0.810
Rau[59]	2007	Naming/noun generation	Block	Dice	Sig. voxels	Task vs. rest	P(corr) < 0.05	13	9 days	0.000	0.350	0.820
Harrington[103]	2006a[c]	Multiple language	Event	Dice	ROI	Task vs. rest	P(corr) < 0.05	10	4 weeks	–	–	–
Harrington[104]	2006b[c]	Multiple encoding	Block	Dice	ROI	Task vs. rest	P(corr) < 0.05	9	10 weeks	–	–	–
Havel[105]	2006	Motor movement	Block	Dice	Sig. voxels	Task vs. rest	P(uncorr) < 0.001	15	6 days	0.000	0.230	0.710
Wagner[106]	2005	Verbal encoding	Block	Dice	Sig. voxels	Task vs. control	Individualized	20	33 weeks	–	0.362	–
Wagner[106]	2005	Verbal recognition	Block	Dice	Sig. voxels	Task vs. control	Individualized	20	33 weeks	–	0.420	–
Yoo[107]	2005[c]	Finger tapping	Block	Dice	ROI	Task vs. rest	P(uncorr) < 0.005	8	8 weeks	–	–	–
Specht[100]	2003	Visual attention	Event	Dice	Voxelwise	Task vs. rest	P(uncorr) < 0.01	5	2 weeks	0.420	0.583	0.692
Swallow[108]	2003[b]	Visual FEF and MT localizers	Block	Jaccard	ROI	Task vs. rest	z(uncorr) > 4.5	11	Not given	0.416	0.463	0.507
Maldjian[57]	2002[b]	Word generation	Block	Jaccard	Sig. voxels	Task vs. rest	P(uncorr) < 0.005	8	1 week	0.748	0.856	0.993
Maldjian[57]	2002[b]	Forward–backward listening	Block	Jaccard	Sig. voxels	Task vs. rest	P(uncorr) < 0.005	8	1 week	0.410	0.662	0.817
Rutten[109]	2002[b,d]	Combined language tasks	Block	Custom	Sig. voxels	Task vs. rest	z(uncorr) > 4.5	9	5 months	–	0.420	–
Miki[110]	2001	Visual checkerboard	Block	Dice	Sig. voxels	Task vs. rest	z(uncorr) > 4.5	4	<1 h	0.560	0.610	0.660
Machielsen[111]	2000	Visual encoding	Block	Dice	Sig. voxels	Task vs. control	P(corr) < 0.05	10	14 days	–	0.507	–
Machielsen[111]	2000	Visual encoding	Block	Dice	ROI	Task vs. control	P(corr) < 0.05	10	14 days	0.211	0.374	0.514
Miki[112]	2000	Visual light stimulation	Block	Dice	Sig. voxels	Task vs. rest	z(uncorr) > 4.5	7	5 days	0.020	0.480	0.770
Tegeler[113]	1999	Finger tapping	Block	Dice	Sig. voxels	Task vs. rest	Top 2% of voxels	6	<1 h	–	0.410	–
Rombouts[42]	1998	Visual light stimulation	Block	Dice	Sig. voxels	Task vs. rest	P(corr) < 0.05	10	2 weeks	0.460	0.640	0.760
Rombouts[40]	1997	Visual light stimulation	Block	Dice	Sig. voxels	Task vs. rest	r(uncorr) > 0.50	14	2 weeks	0.150	0.310	0.500

Continued.

Table 2. *Continued*

First author	Year	Task	Design	Calculation	Basis	Contrast	Threshold	No. of Subs	Approximate T–R interval	Dice overlap Min overlap	Dice overlap Avg. overlap	Dice overlap Max overlap
Ramsey[114]	1996[b]	Finger tapping	Block	Jaccard	Sig. voxels	Task vs. rest	$P(\text{corr}) < 0.05$	7	11 weeks	–	0.333	–
Yetkin[115]	1996[b]	Finger tapping	Block	Jaccard	Sig. voxels	Task vs. rest	$r(\text{uncorr}) > 0.60$	4	<1 h	–	0.742	–
Yetkin[115]	1996[b]	Somatosensory touch	Block	Jaccard	Sig. voxels	Task vs. rest	$r(\text{uncorr}) > 0.60$	4	<1 h	–	0.621	–
Mean value										0.314	0.476	0.670

[a]Overlap values estimated from figure.
[b]Results recalculated to represent Dice statistic.
[c]Data presented as graphs or figures, unable to quantify values.
[d]Overlap only calculated for a single region.
[e]Calculated using total voxels in first session only, not average.

overlapping of clusters and most everyone is familiar with correlation values. Although these techniques certainly have limitations and caveats, they seem to be the emerging standard for the analysis of fMRI reliability.

Previous studies of reliability

What sample size is necessary to conduct effective reliability research? Most of the studies that were reviewed used less than 10 subjects to calculate their reliability measures, with 11 subjects being the overall average across the investigation. Should reliability studies have more subjects? Because a large amount of the error variance is coming from subject-specific factors it may be wise to use larger sample sizes when assessing study reliability, as a single anomalous subject could sway study reliability in either direction. Another notable factor is that a large percentage of studies using fMRI are completed with a restricted range of subjects. Most samples will typically be recruited from a pool of university undergraduates. These samples may have a different reliability than a sample pulled at random from the larger population. Because of sample restriction the results of most test–retest investigations may not reflect the true reliability of other populations such as children, the elderly, and individuals with clinical disorders.

Reliability varies by test–retest interval

Generally, increased amounts of time between the initial test scan and the subsequent retest scan will lower reliability. Still, even back-to-back scans are not perfectly reliable. The average Jaccard overlap of studies where the test and retest scans took place within the same hour was 33%. Many studies with intervals lasting 3 months or more had a lower overlap percentage. This is a somewhat loose guideline though. Notably, the results reported by Aron and colleagues had one of the longest test–retest intervals but also possessed the highest average ICC score.[52]

Reliability varies by cognitive task and experimental design

Motor and sensory tasks seem to have greater reliability than tasks involving higher cognition. Caceras and colleagues found that the reliability of an N-back task was higher than that of an auditory target detection task.[47] Differences in the design of an fMRI experiment also seem to affect the reliability of results. Specifically, block designs appear to have a slight advantage over event-related designs in terms of reliability. This may be a function of the greater statistical power inherent in a block design and its increased SNR.

Significance is related to reliability, but it is not a strong correlation

Several studies have illustrated that super-threshold voxels are not necessarily more reliable than sub-threshold voxels. Caceras and colleagues examined the joint probability distribution of significance and reliability.[47] They found that there were some highly activated ROIs with low reliability and some sub-threshold regions that had high reliability. These ICC results fit in well with the data from cluster overlap studies. The average cluster overlap was 29%.

Table 3. Results of examined papers using other forms of reliability calculation

First author	Year	Task	Design	Method	Type	No. of Subs	Approximate T–R interval
Liou[116]	2009	Multiple tasks	Event	Cohen's kappa index	Voxelwise	12	<1 h
Magon[117]	2009	Breath holding	Block	Coefficient of variation	ROI	11	3 weeks
Maitra[118]	2009	Finger tapping	Block	Maximum likelihood	Voxelwise	1	5 days
Miller[55]	2009	Multiple memory tasks	Block/event	Pearson correlation	Voxelwise	14	12 weeks
Shehzad[119]	2009	Resting state	Free	Connectivity ICC	ROI	26	1 h/5 months
Shehzad[119]	2009	Resting state	Free	Kendall's *W*	ROI	26	1 h/5 months
Zhang[31]	2009	Static force	Block	NPAIRS	Components	16	<1 h
Zandbelt[120]	2008	Go/NoGo	Block	Signal change SD	ROI	10	1 week
Chen[121]	2007	Language imaging	Block	Reliability maps/ROC	Voxelwise	12	Variable
Leontiev[122]	2007	Retinotopic mapping	Block	Coefficient of variation	ROI	10	1 day
Yoo[123]	2007	Motor imagery	Block	Signal change SD	ROI	10	<1 h
Jansen[124]	2006	Multiple tasks	Block	Laterality index	ROI	10	2 h
Mayer[125]	2006	Covert word generation	Event	Active voxel count	ROI	8	<1 h
Peelen[126]	2005	Visual categorization	Block	Sign test	ROI	6	3 weeks
Smith[127]	2005	Visual, motor, cognitive	Block	Variance components	ROI	1	1 day
Wagner[106]	2005	Verbal memory	Block	Pearson correlation	All voxels	20	33 weeks
Liu[128]	2004	Handgrip task	Block	General linear model	Voxelwise	8	1 month
Stark[129]	2004	Emotional pictures	Block	Kappa index	Sig. voxels	24	1 week
Strother[30]	2004	Static force	Block	NPAIRS	Components	16	<1 h
Phan[130]	2003	Aversive images	Block	General Linear model	ROI	8	<1 h
Kiehl[131]	2003	Auditory oddball	Event	Conjunction analysis	Voxelwise	10	6 weeks
Neumann[132]	2003	Stroop Task	Event	BOLD dynamics	ROI	4	1 week
Maitra[133]	2002	Finger tapping	Block	Maximum likelihood	All voxels	1	~1 week
Miller[36]	2002	Episodic retrieval	Block	Pearson correlation	All voxels	6	6 months
Loubinoux[134]	2001	Sensorimotor	Block	Coefficient of variation	Sig. voxels	21	Variable
Salli[135]	2001	Wrist flexing	Block	Reliability maps	Voxelwise	1	<1 h
White[136]	2001	Finger tapping	Block	ROI discrimination	ROI	6	3 weeks
McGonigle[137]	2000	Visual, motor, cognitive	Block	General linear model	Voxelwise	1	1 day
Waldvogel[138]	2000	Tapping/ checkerboard	Block	Signal change stability	All voxels	6	1 week
Cohen[44]	1999[a]	Visual and motor	Block	Voxel counting	Sig. voxels	6	<1 h
Tegeler[113]	1999	Finger tapping	Block	Pearson correlation	All voxels	6	<1 h
Moser[19]	1996	Visual stimulation	Block	Signal change SD	Single slice	18	<1 h

[a]Cohen *et al.* conducted the experiment to argue against voxel counting.

This means that, across studies, the average number of significant voxels that will replicate is roughly one third. This evidence speaks against the assumption that significant voxels will be far more reliable in an investigation of test–retest reliability.

An optimal threshold of reliability has not been established

There is no consensus value regarding what constitutes an acceptable level of reliability in fMRI. Is an ICC value of 0.50 enough? Should studies be required to achieve an ICC of 0.70? All of the studies in the review simply reported what the reliability values were. Few studies proposed any kind of criteria to be considered a "reliable" result. Cicchetti and Sparrow did propose some qualitative descriptions of data based on the ICC-derived reliability of results.[53] They proposed that results with an ICC above 0.75 be considered "excellent," results between 0.59 and 0.75 be considered "good," results between 0.40 and 0.58 be considered "fair," and results lower than 0.40 be considered "poor." More specifically to neuroimaging, Eaton and colleagues used a threshold of ICC >0.4 as the mask value for their study, whereas Aron and colleagues [52] used an ICC cutoff of ICC >0.5 as the mask value.[54]

Interindividual variability is consistently greater than intraindividual variability

Many studies reported both within- and between-subject reliability values in their results. In every case, the within-subject reliability far exceeded the between-subjects reliability. Miller and colleagues explicitly examined variability across subjects and concluded that there are large-scale, stable differences between individuals on almost any cognitive task.[35,36] More recently, Miller and colleagues directly contrasted within- and between-subject variability.[55] They concluded that between-subject variability was far higher than any within-subject variability. They further demonstrated that the results from one subject completing two different cognitive tasks are typically more similar than the data from two subjects doing the same task. These results are mirrored by those of Costafreda and colleagues, who found that well over half (57%) of the variability in their fMRI data was due to between-subject variation.[56] It seems to be the case that within-subject measurements over time may vary, but they vary far less than differences in the overall pattern of activity between individuals.

There is little agreement regarding the true reliability of fMRI results

Although we mention this as a final conclusion from the literature review, it is perhaps the most important point. Some studies have estimated the reliability of fMRI data to be quite high, or even close to perfect for some tasks and brain regions.[52,57,58] Other studies have been less enthusiastic, showing fMRI reliability to be relatively low.[41,59] Across the survey of fMRI test–retest reliability we found that the average ICC value was 0.50 and the average cluster overlap value was 29% of voxels (Dice overlap $= 0.45$, Jaccard overlap $= 0.29$). This represents an average across many different cognitive tasks, fMRI experimental designs, test–retest time periods, and other variables. While these numbers may not be representative of any one experiment, they do provide an effective overview of fMRI reliability.

Other issues and comparisons

Test–retest reliability in clinical disorders

There have been few examinations of test–retest reliability in clinical disorders relative to the number of studies with normal controls. A contributing factor to this problem may be that the scientific understanding of brain disorders using neuroimaging is still in its infancy. It may be premature to examine clinical reliability if there is only a vague understanding of anatomical and functional abnormalities in the brain. Still, some investigators have taken significant steps forward in the clinical realm. These few investigations suggest that reliability in clinical disorders is typically lower than the reliability of data from normal controls. Some highlights of these results are listed later, categorized by disorder.

Epilepsy

Functional imaging has enormous potential to aid in the clinical diagnosis of epileptiform disorders. Focusing on fMRI, research by Di Bonaventura and colleagues found that the spatial extent of activity associated with fixation off sensitivity was stable over time in epileptic patients.[60] Of greater research interest for epilepsy has been the reliability of combined EEG/fMRI imaging. Symms and colleagues reported that they could reliably localize interictal epileptiform discharges using

EEG-triggered fMRI.[61] Waites and colleagues also reported the reliable detection of discharges with combined EEG/fMRI at levels significantly above chance.[62] Functional imaging also has the potential to assist in the localization of cognitive function prior to resection for epilepsy treatment. One possibility would be to use noninvasive fMRI measures to replace cerebral sodium amobarbital anesthetization (Wada test). Fernandez and colleagues reported good reliability of lateralization indices (whole-brain test–retest, $r = 0.82$) and cluster overlap measures (Dice overlap = 0.43, Jaccard overlap = 0.27).[5]

Stroke

Many aspects of stroke recovery can impact the results of functional imaging data. The lesion location, size, and time elapsed since the stroke event each have the potential to alter function within the brain. These factors can also lead to increased between-subject variability relative to groups of normal controls. This is especially true when areas proximal to the lesion location contribute to specific aspects of information processing, such as speech production. Kimberley and colleagues found that stroke patients had generally higher ICC values relative to normal controls.[63] This mirrors the findings of Eaton and colleagues, who showed that the average reliability of aphasia patients was approximately equal to that of normal controls as measured by ICC.[54] These results may be indicative of equivalent fMRI reliability in stroke victims, or it may be an artifact of the ICC calculation. Kimberly and colleagues state that increased between-subject variability of stroke patients can lead to inflated ICC estimates.[63] They argue that fMRI reliability in stroke patients likely falls within the moderate range of values ($0.4 < \text{ICC} < 0.6$).

Schizophrenia

Schizophrenia is a multidimensional mental disorder characterized by a wide array of cognitive and perceptual dysfunctions.[64,65] Although there have been a number of studies on the reliability of anatomical measures in schizophrenia there have been few that have focused on function. Manoach and colleagues demonstrated that the fMRI results from schizophrenic patients on a working memory task were less reliable overall than that of normal controls.[66] The reliability of significant ROIs in the schizophrenic group ranged from ICC values of −0.20 to 0.57. However, the opposite effect was found by Whalley and colleagues in a group of subjects at high genetic risk for schizophrenia (no psychotic symptoms).[67] The ICC values for these subjects were equally reliable relative to normal controls on a sentence completion task. More research is certainly needed to find consensus on reliability in schizophrenia.

Aging

The anatomical and functional changes that take place during aging can increase the variability of fMRI results at all levels.[68] Clement and colleagues reported that cluster overlap percentages and the cluster-wise ICC values were not significantly different between normal elderly controls and patients with mild cognitive impairment (MCI).[69] On an episodic retrieval task, healthy controls had ICC values averaging 0.69 whereas patients diagnosed with MCI had values averaging 0.70. However, they also reported that all values for the older samples were lower than those reported for younger adults on similar tasks. Marshall and colleagues found that although the qualitative reproducibility of results was high, the reliability of activation magnitude during aging was quite low.[70]

It is clear that the use of ICCs in clinical research must be approached carefully. As mentioned by Bosnell and colleagues and Kimberly and colleagues, extreme levels of between-subject variability will artificially inflate the resulting ICC reliability estimate.[63,71] Increased between-subject variability is a characteristic found in many clinical populations. Therefore, it may be the case that comparing two populations with different levels of between-subject variability may be impossible when using an ICC measure.

Reliability across scanners/multicenter studies

One area of increasing research interest is the ability to combine the data from multiple scanners into larger, integrative data sets.[72] There are two areas of reliability that are important for such studies. The first is subject-level reliability, or how stable the activity of one person will be scan-to-scan. The second is group-level reliability, or how stable the group fMRI results will be from one set of subjects

to another or from one scanner to another. Given the importance of multicenter collaboration it is critical to evaluate how results will differ when the data comes from a heterogeneous group of MRI scanners as opposed to a single machine. Generally, the concordance of fMRI results from center to center is quite good, but not perfect.

Casey and colleagues was one of the first groups to examine the reliability of results across scanners.[7] Between three imaging centers they found a "strong similarity" in the location and distribution of significant voxel clusters. More recently, Friedman and colleagues found that intercenter reliability was somewhat worse than test–retest reliability across several centers with an identical hardware configuration.[8] The median ICC of their intercenter results was ICC = 0.22. Costafreda and colleagues also examined the reproducibility of results from identical fMRI setups.[56] Using a variance components analysis, they determined that the MR system accounted for roughly 8% of the variation in the BOLD signal. This compares favorably relative to the level of between-subject variability (57%).

The reliability of results from one scanner to another seems to be approximately equal to or slightly less than the values of test–retest reliability with the same MRI hardware. Special calibration and quality control steps can be taken to ensure maximum concordance across scanners. For instance, before conducting anatomical MRI scans in the Alzheimer's disease neuroimaging initiative (ADNI, http://www.loni.ucla.edu/ADNI/) a special MR phantom is typically scanned. This allows for correction of magnet-specific field inhomogeneity and maximizes the ability to compare data from separate scanners. Similar calibration measures are being discussed for functional MRI.[73–75] It may be the case that as calibration becomes standardized it will lead to increased intercenter reliability.

Other statistical issues in fMRI

It is important to note that a number of important fMRI statistical issues have gone unmentioned in this paper. First, there is the problem of conducting thousands of statistical comparisons without an appropriate threshold adjustment. Correction for multiple comparisons is a necessary step in fMRI analysis that is often skipped or ignored.[76] Another statistical issue in fMRI is temporal autocorrelation in the acquired timeseries. This refers to the fact that any single timepoint of data is not necessarily independent of the acquisitions that came before and after.[77,78] Autocorrelation correction is widely available, but is not implemented by most investigators. Finally, throughout the last year the "nonindependence error" has been discussed at length. Briefly, this refers to selecting a set of voxels to create an ROI and then using the same measure to evaluate some statistical aspect of that region. Ideally, an independent data set should be used after the ROI has been initially defined. It is important to address these issues because they are still debated within the field and often ignored in fMRI analysis. Their correction can have a dramatic impact on how reproducible the results will be from study to study.

Conclusions

How can a researcher improve fMRI reliability?

The generation of highly reliable results requires that sources of error be minimized across a wide array of factors. An issue within any single factor can significantly reduce reliability. Problems with the scanner, a poorly designed task, or an improper analysis method could each be extremely detrimental. Conversely, elimination of all such issues is necessary for high reliability. A well-maintained scanner, well-designed tasks, and effective analysis techniques are all prerequisites for reliable results.

There are a number of practical ways that fMRI researchers can improve the reliability of their results. For example, Friedman and Glover reported that simply increasing the number of fMRI runs improved the reliability of their results from ICC = 0.26 to ICC = 0.58.[73] That is quite a large jump for an additional 10 or 15 min of scanning. Some general areas where reliability can be improved are given later.

Increase the SNR and CNR of the acquisition

One area of attention is to improve the SNR and CNR ratios of the data collection. An easy way to do this would be to simply acquire more data. It is a zero-sum game, as increasing the number of TRs that are acquired will help improve the SNR but will also increase the task length. Subject fatigue, scanner time limitations, and the diminishing

returns with each duration increase will all play a role in limiting the amount of time that can be dedicated to any one task. Still, a researcher considering a single 6-min EPI scan for their task might add additional data collection to improve the SNR of the results. With regard to the magnet, every imaging center should verify acquisition quality before scanning. Many sites conduct quality assurance scans at the beginning of each day to ensure stable operation. This has proven to be an effective method of detecting issues with the MR system before they cause trouble for investigators. It is a hassle to cancel a scanning session when there are subtle artifacts present, but this is a better option than acquiring noisy data that does not make a meaningful contribution to the investigation. As a final thought, research groups can always start fundraising to purchase a new magnet with improved specifications. If data acquisition is being done on a 1.5 Tesla magnet with a quadrature head coil enormous gains in SNR can be made by moving to 3.0 Tesla or higher and using a parallel-acquisition head coil.[79,80]

Minimize individual differences in cognitive state, both across subjects and over time

Because magnet time is expensive and precious the critical component of effective task instruction can often be overlooked. Researchers would rather be acquiring data as opposed to spending additional time giving detailed instructions to a subject. However, this is a very easy way to improve the quality of the final data set. If it takes 10 trials for the participant to really "get" the task then those trials have been wasted, adding unnecessary noise to the final results. Task training in a separate laboratory session in conjunction with time in a mock MRI scanner can go a long way toward homogenizing the scanner experience for subjects. It may not always be possible to fully implement these steps, but they should not be avoided simply to reduce the time spent per subject.

For multisession studies, steps can be taken to help stabilize intrasubject changes over time. Scanning test and retest session at the same time of day can help due to circadian changes in hormone level and cognitive performance.[81–83] A further step to consider is minimizing the time between sessions to help stabilize the results. Much more can change over the course of a month than over the course of a week.

Maximize the experiment's statistical power

Power represents the ability of an experiment to reject the null hypothesis when the null hypothesis is indeed false.[84] For fMRI this ability is often discussed in terms of the number of subjects that will be scanned and the design of the task that will be administered, including how many volumes of data will be acquired from each subject. More subjects and more volumes almost always contribute to increasing power, but there are occasions when one may improve power more than the other. For example, Mumford and Nichols demonstrated that, when scanner time was limited, different combinations of subjects and trials could be used to achieve high levels of power.[85] For their hypothetical task it would take only five 15 sec blocks to achieve 80% power if there were 23 subjects, but it would take 25 blocks if there were only 18 subjects. These kinds of power estimations are quite useful in determining the best use of available scanner time. Tools such as fmripower (http://fmripower.org) can use data from existing experiments to yield new information on how many subjects and scans a new experiment will require to reach a desired power level.[85–87]

The structure of the stimulus presentation has a strong influence on an experiment's statistical power. The dynamic interplay between stimulus presentation and interstimulus jitter are important, as is knowing what contrasts will be completed once the data has been acquired. Each of these parameters can influence the power and efficiency of the experiment, impacting the reliability of the results. Block designs tend to have greater power relative to event-related designs. One can also increase power by increasing block length, but care should be exercised not to make blocks so long that they approach the low frequencies associated with scanner drift. There are several good software tools available that will help researchers create an optimal design for fMRI experiments. OptSeq is a program that helps to maximize the efficiency of an event-related fMRI design.[88] OptimizeDesign is a set of Matlab scripts that utilize a genetic search algorithm to maximize specific aspects of the design.[89] Using this tool, researchers can separately weight statistical power, HRF estimation efficiency, stimulus counterbalancing, and maintenance of stimulus frequency. These programs, and others like them, are valuable tools for ensuring that the

ability to detect meaningful signals is effectively maximized.

It is important to state that the reliability of a study in no way implies that an experiment has accurately assessed a specific cognitive process. The validity of a study can be quite orthogonal to its reliability—it is possible to have very reliable results from a task that mean little with regard to the cognitive process under investigation. No increase in SNR or optimization of event timing can hope to improve an experiment that is testing for the wrong thing. This makes task selection of paramount importance in the planning of an experiment. It also places a burden on the researcher in terms of effective interpretation of fMRI results once the analysis is done.

Where does neuroimaging go next?

In many ways cognitive neuroscience is still at the beginning of fMRI as a research tool. Looking back on the last two decades it is clear that functional MRI has made enormous gains in both statistical methodology and popularity. However, there is still much work to do. With specific regard to reliability, there are some specific next steps that must be taken for the continued improvement of this method.

Better characterization of the factors that influence reliability

Additional research is necessary to effectively understand what factors influence the reliability of fMRI results. The field has a good grasp of the acquisition and analysis factors that influence SNR. Still, there is relatively little knowledge regarding how stable individuals are over time and what influences that stability. Large-scale studies specifically investigating reliability and reproducibility should therefore be conducted across several cognitive domains. The end goal of this research would be to better characterize the reliability of fMRI across multiple dimensions of influence within a homogeneous set of data. Such a study would also create greater awareness of fMRI reliability in the field as a whole. The direct comparison of reliability analysis methods, including predictive modeling, should also be completed.

Meta/mega analysis

The increased pooling of data from across multiple studies can give a more generalized view of important cognitive processes. One method, meta-analysis, refers to pooling the statistical results of numerous studies to identify those results that are concordant and discordant with others. For example, one could obtain the MNI coordinates of significant clusters from several studies having to do with response inhibition and plot them in the same stereotaxic space to determine their concordance. One popular method of performing such an analysis is the creation of an activation likelihood estimate (ALE).[90,91] This method allows for the statistical thresholding of meta-analysis results, making it a powerful tool to examine the findings of many studies at once. Another method, mega-analysis, refers to reprocessing the raw data from numerous studies in a new statistical analysis with much greater power. Using this approach any systematic error introduced by any one study will contribute far less to the final statistical result.[92] Mega-analyses are far more difficult to implement because the raw imaging data from multiple studies must be obtained and reprocessed. Still, the increase in detection power and the greater generalizability of the results are strong reasons to engage in such an approach.

One roadblock to collaborative multicenter studies is the lack of data provenance in functional neuroimaging. Provenance refers to complete detail regarding the origin of a data set and the history of operations that have been preformed on the data. Having a complete history of the data enables analysis by other researchers and provides information that is critical for replication studies.[93] Moving forward there will be an additional focus on provenance to enable increased understanding of individual studies and facilitate integration into larger analyses.

New emphasis on replication

The nonindependence debate of 2009 was less about effect sizes and more about reproducibility. The implicit argument made about studies that were "non-independent" was that if researchers ran a nonindependent study over again the resulting correlation would be far lower with a new, independent data set. There should be a greater emphasis on the replicability of studies in the future. This can be frustrating because it is expensive and time consuming to acquire and process a replication study. However, moving forward this may become

increasingly novel to validate important results and conclusions.

General conclusions

One thing is abundantly clear: fMRI is an effective research tool that has opened broad new horizons of investigation to scientists around the world. However, the results from fMRI research may be somewhat less reliable than many researchers implicitly believe. Although it may be frustrating to know that fMRI results are not perfectly replicable, it is beneficial to take a longer-term view regarding the scientific impact of these studies. In neuroimaging, as in other scientific fields, errors will be made and some results will not replicate. Still, over time some measure of truth will accrue. This paper is not intended to be an accusation against fMRI as a method. Quite the contrary, it is meant to increase the understanding of how much each fMRI result can contribute to scientific knowledge. If only 30% of the significant voxels in a cluster will replicate then that value represents an important piece of contextual information to be aware of. Likewise, if the magnitude of a voxel is only reliable at a level of ICC = 0.50 then that value represents important information when examining scatter plots comparing estimates of activity against a behavioral measure.

There are a variety of methods that can be used to evaluate reliability, and each can provide information on unique aspects of the results. Our findings speak strongly to the question of why there is no agreed-upon average value for fMRI reliability. There are so many factors spread out across so many levels of influence that it is almost impossible to summarize the reliability of fMRI with a single value. Although our average ICC value of 0.50 and our average overlap value of 30% are effective summaries of fMRI as a whole, these values may be higher or lower on a study-to-study basis. The best characterization of fMRI reliability would be to give a window within which fMRI results are typically reliable. Breaking up the range of 0.0–1.0 into thirds, it is appropriate to say that most fMRI results are reliable in the ICC = 0.33–0.66 range.

To conclude, functional neuroimaging with fMRI is no longer in its infancy. Instead, it has reached a point of adolescence, where knowledge and methods have made enormous progress but there is still much development left to be done. Our growing pains from this point forward are going to be a more complete understanding of its strengths, weaknesses, and limitations. A working knowledge of fMRI reliability is key to this understanding. The reliability of fMRI may not be high relative to other scientific measures, but it is presently the best tool available for *in vivo* investigation of brain function.

Acknowledgments

The authors thank George Wolford for his input and mentorship. The authors thank the many researchers who have been investigating reliability in fMRI over the years. This paper owes much to their hard work and diligence. This work on fMRI reliability is supported by the Institute for Collaborative Biotechnologies through contract no. W911NF-09-D-0001 from the U.S. Army Research Office.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Vul, E. *et al*. 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psycholog. Sci*. **4:** 274–290.
2. Nunnally, J. 1970. *Introduction to Psychological Measurement*. McGraw Hill. New York.
3. Jabbi, M. *et al*. 2009. *Response to "Voodoo Correlations in Social Neuroscience" by Vul et al*.
4. Lieberman, M.D., E.T. Berkman & T.D. Wager. 2009. Correlations in social neuroscience aren't voodoo: Commentary on Vul et al. (2009). *Perspect. Psycholog. Sci*. **4:** 299–307.
5. Fernandez, G. *et al*. 2003. Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. *Neurology* **60:** 969–975.
6. Jovicich, J. *et al*. 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. *Neuroimage* **30:** 436–443.
7. Casey, B. J. *et al*. 1998. Reproducibility of fMRI results across four institutions using a spatial working memory task. *Neuroimage* **8:** 249–261.
8. Friedman, L. *et al*. 2008. Test-retest and between-site reliability in a multicenter fMRI study. *Hum. Brain Mapp*. **29:** 958–972.
9. Kruger, G. & G.H. Glover. 2001. Physiological noise in oxygenation-sensitive magnetic resonance imaging. *Magn. Reson. Med*. **46:** 631–637.

10. Huettel, S.A., A.W. Song & G. McCarthy. 2008. *Functional Magnetic Resonance Imaging*. Sinauer Associates. Sunderland, MA.
11. Bodurka, J. *et al*. 2005. Determination of the brain tissue-specific temporal signal to noise limit of 3 T BOLD-weighted time course data. *Presented at Proceedings of the International Society for Magnetic Resonance in Medicine, Miami*.
12. Murphy, K., J. Bodurka & P.A. Bandettini. 2007. How long to scan? The relationship between fMRI temporal signal to noise ratio and necessary scan duration. *Neuroimage* **34:** 565–574.
13. Ogawa, S. *et al*. 1993. Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging. A comparison of signal characteristics with a biophysical model. *Biophys. J*. **64:** 803–812.
14. Bandettini, P.A. *et al*. 1994. Spin-echo and gradient-echo EPI of human brain activation using BOLD contrast: a comparative study at 1.5 T. *NMR Biomed*. **7:** 12–20.
15. Turner, R. *et al*. 1993. Functional mapping of the human visual cortex at 4 and 1.5 tesla using deoxygenation contrast EPI. *Magn. Reson. Med*. **29:** 277–279.
16. Hoenig, K., C.K. Kuhl & L. Scheef. 2005. Functional 3.0-T MR assessment of higher cognitive function: are there advantages over 1.5-T imaging? *Radiology* **234:** 860–868.
17. Jezzard, P. & S. Clare. 1999. Sources of distortion in functional MRI data. *Hum. Brain Mapp*. **8:** 80–85.
18. Triantafyllou, C. *et al*. 2005. Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters. *Neuroimage* **26:** 243–250.
19. Moser, E., C. Teichtmeister & M. Diemling. 1996. Reproducibility and postprocessing of gradient-echo functional MRI to improve localization of brain activity in the human visual cortex. *Magn. Reson. Imaging* **14:** 567–579.
20. Zhilkin, P. & M.E. Alexander. 2004. Affine registration: a comparison of several programs. *Magn. Reson. Imaging* **22:** 55–66.
21. Oakes, T.R. *et al*. 2005. Comparison of fMRI motion correction software tools. *Neuroimage* **28:** 529–543.
22. Andersson, J.L. *et al*. 2001. Modeling geometric deformations in EPI time series. *Neuroimage* **13:** 903–919.
23. Kiebel, S. & A. Holmes. 2007. The general linear model. In *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. K. Friston et al., Eds.: Academic Press. London.
24. Mikl, M. *et al*. 2008. Effects of spatial smoothing on fMRI group inferences. *Magn. Reson. Imaging* **26:** 490–503.
25. Mumford, J.A. & T. Nichols. 2009. Simple group fMRI modeling and inference. *Neuroimage* **47:** 1469–1475.
26. Gold, S. *et al*. 1998. Functional MRI statistical software packages: a comparative analysis. *Hum. Brain Mapp*. **6:** 73–84.
27. Morgan, V.L. *et al*. 2007. Comparison of fMRI statistical software packages and strategies for analysis of images containing random and stimulus-correlated motion. *Comput. Med. Imaging Graph*. **31:** 436–446.
28. Poline, J.B. *et al*. 2006. Motivation and synthesis of the FIAC experiment: reproducibility of fMRI results across expert analyses. *Hum. Brain Mapp*. **27:** 351–359.
29. Strother, S.C. *et al*. 2002. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *Neuroimage* **15:** 747–771.
30. Strother, S. *et al*. 2004. Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. *Neuroimage* **23**(Suppl 1): S196–S207.
31. Zhang, J. *et al*. 2009. Evaluation and optimization of fMRI single-subject processing pipelines with NPAIRS and second-level CVA. *Magn. Reson. Imaging*. **27:** 264–278.
32. Peyron, R. *et al*. 1999. Haemodynamic brain responses to acute pain in humans: sensory and attentional networks. *Brain.* **122**(Pt 9): 1765–1780.
33. Sterr, A. *et al*. 2007. Activation of SI is modulated by attention: a random effects fMRI study using mechanical stimuli. *Neuroreport* **18:** 607–611.
34. Munneke, J., D.J. Heslenfeld & J. Theeuwes. 2008. Directing attention to a location in space results in retinotopic activation in primary visual cortex. *Brain Res*. **1222**:184–191.
35. Miller, M. B. *et al*. 2001. Brain activations associated with shifts in response criterion on a recognition test. *Can. J. Exp. Psychol*. **55:** 162–173.
36. Miller, M.B. *et al*. 2002. Extensive individual differences in brain activations associated with episodic retrieval are reliable over time. *J. Cogn. Neurosci*. **14:** 1200–1214.
37. Grafton, S., E. Hazeltine & R. Ivry. 1995. Functional mapping of sequence learning in normal humans. *J. Cogn. Neurosci*. **7:** 497–510.
38. Poldrack, R.A. *et al*. 1999. Striatal activation during acquisition of a cognitive skill. *Neuropsychology* **13:** 564–574.
39. Rostami, M. *et al* 2009. Neural bases of goal-directed implicit learning. *Neuroimage* **48**:303–310.

40. Rombouts, S.A. *et al*. 1997. Test-retest analysis with functional MR of the activated area in the human visual cortex. *AJNR Am. J. Neuroradiol*. **18:** 1317–1322.
41. Duncan, K.J. *et al*. 2009. Consistency and variability in functional localisers. *Neuroimage* **46:** 1018–1026.
42. Rombouts, S.A. *et al*. 1998. Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. *Magn. Reson. Imaging*. **16:** 105–113.
43. Nichols, T. *et al*. 2005. Valid conjunction inference with the minimum statistic. *Neuroimage* **25:** 653–660.
44. Cohen, M.S. & R.M. DuBois. 1999. Stability, repeatability, and the expression of signal magnitude in functional magnetic resonance imaging. *J. Magn. Reson. Imaging*. **10:** 33–40.
45. Bartko, J. 1966. The intraclass correlation coefficient as a measure of reliability. *Psycholog. Rep.* **19:** 3–11.
46. Shrout, P. & J. Fleiss. 1979. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psycholog. Bull*. **86:** 420–428.
47. Caceres, A. *et al*. 2009. Measuring fMRI reliability with the intra-class correlation coefficient. *Neuroimage* **45:** 758–768.
48. Muller, R. & P. Buttner. 1994. A critical discussion of intraclass correlation coefficients. *Stat. Med*. **13:** 2465–2476.
49. Bennett, C.M., S.A. Guerin & M.B. Miller. 2009. The impact of experimental design on the detection of individual variability in fMRI. *Presented at Cognitive Neuroscience Society, San Francisco, CA*.
50. Zhang, J. *et al*. 2008. A Java-based fMRI processing pipeline evaluation system for assessment of univariate general linear model and multivariate canonical variate analysis-based pipelines. *Neuroinformatics* **6**:123–134.
51. Safrit, M. 1976. Reliability Theory. American Alliance for Health, Physical Education, and Recreation. Washington, DC.
52. Aron, A.R., M.A. Gluck & R.A. Poldrack. 2006. Long-term test-retest reliability of functional MRI in a classification learning task. *Neuroimage* **29:** 1000–1006.
53. Cicchetti, D. & S. Sparrow. 1981. Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *Am. J. Ment. Defic*. **86:** 127–137.
54. Eaton, K.P. *et al*. 2008. Reliability of fMRI for studies of language in post-stroke aphasia subjects. *Neuroimage* **41:** 311–322.
55. Miller, M.B. *et al*. 2009. Unique and persistent individual patterns of brain activity across different memory retrieval tasks. *Neuroimage* **48:** 625–635.
56. Costafreda, S.G. *et al*. 2007. Multisite fMRI reproducibility of a motor task using identical MR systems. *J. Magn. Reson. Imaging*. **26:** 1122–1126.
57. Maldjian, J.A. *et al*. 2002. Multiple reproducibility indices for evaluation of cognitive functional MR imaging paradigms. *AJNR Am. J. Neuroradiol*. **23:** 1030–1037.
58. Raemaekers, M. *et al*. 2007. Test-retest reliability of fMRI activation during prosaccades and antisaccades. *Neuroimage* **36:** 532–542.
59. Rau, S. *et al*. 2007. Reproducibility of activations in Broca area with two language tasks: a functional MR imaging study. *AJNR Am. J. Neuroradiol*. **28:** 1346–1353.
60. Di Bonaventura, C. *et al.* 2005. Long-term reproducibility of fMRI activation in epilepsy patients with Fixation Off Sensitivity. *Epilepsia* **46:** 1149–1151.
61. Symms, M.R. *et al*. 1999. Reproducible localization of interictal epileptiform discharges using EEG-triggered fMRI. *Phys. Med. Biol*. **44:** N161–N168.
62. Waites, A.B. *et al*. 2005. How reliable are fMRI-EEG studies of epilepsy? A nonparametric approach to analysis validation and optimization. *Neuroimage* **24:** 192–199.
63. Kimberley, T.J., G. Khandekar & M. Borich. 2008. fMRI reliability in subjects with stroke. *Exp. Brain. Res*. **186:** 183–190.
64. Freedman, R. 2003. Schizophrenia. *N. Engl. J. Med.* **349:** 1738–1749.
65. Morrison, P.D. & R.M. Murray. 2005. Schizophrenia. *Curr. Biol.* **15:** R980–R984.
66. Manoach, D.S. *et al*. 2001. Test-retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. *Am. J. Psychiatry*. **158:** 955–958.
67. Whalley, H.C. *et al*. 2009. fMRI changes over time and reproducibility in unmedicated subjects at high genetic risk of schizophrenia. *Psychol. Med*. **39:** 1189–1199.
68. MacDonald, S.W., L. Nyberg & L. Backman. 2006. Intra-individual variability in behavior: links to brain structure, neurotransmission and neuronal activity. *Trends Neurosci*. **29:** 474–480.
69. Clement, F. & S. Belleville. 2009. Test-retest reliability of fMRI verbal episodic memory paradigms in healthy older adults and in persons with mild cognitive impairment. *Hum. Brain Mapp* **30:** 4033–47.
70. Marshall, I. *et al*. 2004. Repeatability of motor and working-memory tasks in healthy older volunteers: assessment at functional MR imaging. *Radiology* **233:** 868–877.

71. Bosnell, R. *et al*. 2008. Reproducibility of fMRI in the clinical setting: implications for trial designs. *Neuroimage* **42:** 603–610.
72. Van Horn, J.D. & A.W. Toga. 2009. Multisite neuroimaging trials. *Curr. Opin. Neurol.* **22:** 370–378.
73. Friedman, L. & G.H. Glover. 2006. Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. *Neuroimage* **33:** 471–481.
74. Chiarelli, P.A. *et al*. 2007. A calibration method for quantitative BOLD fMRI based on hyperoxia. *Neuroimage* **37:** 808–820.
75. Thomason, M.E., L.C. Foland & G.H. Glover. 2007. Calibration of BOLD fMRI using breath holding reduces group variance during a cognitive task. *Hum. Brain Mapp.* **28:** 59–68.
76. Bennett, C.M., G.L. Wolford & M.B. Miller. 2009. The principled control of false positives in neuroimaging. *Soc. Cogn. Affective Neurosci.* **4:** 417–422.
77. Woolrich, M.W. *et al*. 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage* **14:** 1370–1386.
78. Smith, A.T., K.D. Singh & J.H. Balsters. 2007. A comment on the severity of the effects of non-white noise in fMRI time-series. *Neuroimage* **36:** 282–288.
79. Zou, K.H. *et al*. 2005. Reproducibility of functional MR imaging: preliminary results of prospective multi-institutional study performed by Biomedical Informatics Research Network. *Radiology* **237:** 781–789.
80. Simmons, W.K. *et al*. 2009. The selectivity and functional connectivity of the anterior temporal lobes. *Cereb. Cortex*. doi: 10.1093/cercor/bhp149.
81. Carrier, J. & T.H. Monk. 2000. Circadian rhythms of performance: new trends. *Chronobiol. Int*. **17:** 719–732.
82. Huang, J. *et al*. 2006. Diurnal changes of ERP response to sound stimuli of varying frequency in morning-type and evening-type subjects. *J. Physiol. Anthropol*. **25:** 49–54.
83. Salthouse, T.A., J.R. Nesselroade & D.E. Berish. 2006. Short-term variability in cognitive performance and the calibration of longitudinal change. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **61:** P144–P151.
84. Cohen, J. 1977. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press. New York, NY.
85. Mumford, J.A. & T.E. Nichols. 2008. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage* **39:** 261–268.
86. Mumford, J.A., R.A. Poldrack & T. Nichols. 2007. FMRIpower: a power calculation tool for 2-stage fMRI models. *Presented at Human Brain Mapping, Chicago, IL*.
87. Van Horn, J. D. *et al.* 1998. Mapping voxel-based statistical power on parametric images. *Neuroimage* **7:** 97–107.
88. Dale, A. 1999. Optimal experimental design for event-related fMRI. *Hum. Brain Mapp.* **8:** 109–114.
89. Wager, T.D. & T. Nichols. 2003. Optimization of experimental design in fMRI: a general framework using a genetic algorithm. *Neuroimage* **18:** 293–309.
90. Turkeltaub, P.E. *et al*. 2002. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage* **16:** 765–780.
91. Eickhoff, S.B. *et al*. 2009. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Hum. Brain Mapp.* **30:** 2907–2926.
92. Costafreda, S.G. In press. Pooling fMRI data: meta-analysis, mega-analysis and multi-center studies. *Front. Neuroinformatics*. **3:** 1–8.
93. Mackenzie-Graham, A.J. *et al*. 2008. Provenance in neuroimaging. *Neuroimage* **42:** 178–195.
94. Freyer, T. *et al*. 2009. Test-retest reliability of event-related functional MRI in a probabilistic reversal learning task. *Psychiatry Res*. **174:** 40–46.
95. Gountouna, V.E. *et al*. 2009. Functional magnetic resonance imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task. *Neuroimage*. **49:** 552–560.
96. Schunck, T. *et al*. 2008. Test-retest reliability of a functional MRI anticipatory anxiety paradigm in healthy volunteers. *J. Magn. Reson. Imaging* **27:** 459–468.
97. Kong, J. *et al*. 2007. Test-retest study of fMRI signal change evoked by electroacupuncture stimulation. *Neuroimage* **34:** 1171–1181.
98. Johnstone, T. *et al*. 2005. Stability of amygdala BOLD response to fearful faces over multiple scan sessions. *Neuroimage* **25:** 1112–1123.
99. Wei, X. *et al*. 2004. Functional MRI of auditory verbal working memory: long-term reproducibility analysis. *Neuroimage* **21:** 1000–1008.
100. Specht, K. *et al*. 2003. Assessment of reliability in functional imaging studies. *J. Magn. Reson. Imaging* **17:** 463–471.
101. Meindl, T. *et al*. 2009. Test-retest reproducibility of the default-mode network in healthy individuals. *Hum. Brain Mapp*. **31:** 237–46.
102. Feredoes, E. & B.R. Postle. 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:** 881–903.
103. Harrington, G.S., M.H. Buonocore & S.T. Farias. 2006. Intrasubject reproducibility of functional MR imaging activation in language tasks. *AJNR Am. J. Neuroradiol.* **27:** 938–944.
104. Harrington, G.S. *et al.* 2006. The intersubject and intrasubject reproducibility of FMRI activation during three encoding tasks: implications for clinical applications. *Neuroradiology* **48:** 495–505.
105. Havel, P. *et al.* 2006. Reproducibility of activation in four motor paradigms. An fMRI study. *J. Neurol.* **253:** 471–476.
106. Wagner, K. *et al.* 2005. The reliability of fMRI activations in the medial temporal lobes in a verbal episodic memory task. *Neuroimage* **28:** 122–131.
107. Yoo, S.S. *et al.* 2005. Long-term reproducibility analysis of fMRI using hand motor task. *Int. J. Neurosci.* **115:** 55–77.
108. Swallow, K.M. *et al.* 2003. Reliability of functional localization using fMRI. *Neuroimage* **20:** 1561–1577.
109. Rutten, G.J. *et al.* 2002. Reproducibility of fMRI-determined language lateralization in individual subjects. *Brain Lang.* **80:** 421–437.
110. Miki, A. *et al.* 2001. Reproducibility of visual activation during checkerboard stimulation in functional magnetic resonance imaging at 4 Tesla. *Japan J. Ophthalmol.* **45:** 151–155.
111. Machielsen, W.C. *et al.* 2000. FMRI of visual encoding: reproducibility of activation. *Hum. Brain Mapp.* **9:** 156–164.
112. Miki, A. *et al.* 2000. Reproducibility of visual activation in functional MR imaging and effects of postprocessing. *AJNR Am. J. Neuroradiol.* **21:** 910–915.
113. Tegeler, C. *et al.* 1999. Reproducibility of BOLD-based functional MRI obtained at 4 T. *Hum. Brain Mapp.* **7:** 267–283.
114. Ramsey, N. *et al.* 1996. Reproducibility of human 3D fMRI brain maps acquired during a motor task. *Hum. Brain Mapp.* **4:** 113–121.
115. Yetkin, F.Z. *et al.* 1996. Test-retest precision of functional MR in sensory and motor task activation. *AJNR Am. J. Neuroradiol.* **17:** 95–98.
116. Liou, M. *et al.* 2009. Beyond p-values: averaged and reproducible evidence in fMRI experiments. *Psychophysiology* **46:** 367–378.
117. Magon, S. *et al.* 2009. Reproducibility of BOLD signal change induced by breath holding. *Neuroimage* **45:** 702–712.
118. Maitra, R. 2009. Assessing certainty of activation or inactivation in test-retest fMRI studies. *Neuroimage* **47:** 88–97.
119. Shehzad, Z. *et al.* 2009. The resting brain: unconstrained yet reliable. *Cereb Cortex* **19:** 2209–2229.
120. Zandbelt, B.B. *et al.* 2008. Within-subject variation in BOLD-fMRI signal changes across repeated measurements: quantification and implications for sample size. *Neuroimage* **42:** 196–206.
121. Chen, E.E. & S.L. Small. 2007. Test-retest reliability in fMRI of language: Group and task effects. *Brain Lang.* **102:** 176–185.
122. Leontiev, O. & R.B. Buxton. 2007. Reproducibility of BOLD, perfusion, and CMRO2 measurements with calibrated-BOLD fMRI. *Neuroimage* **35:** 175–184.
123. Yoo, S.S. *et al.* 2007. Reproducibility of trial-based functional MRI on motor imagery. *Int. J. Neurosci.* **117:** 215–227.
124. Jansen, A. *et al.* 2006. The assessment of hemispheric lateralization in functional MRI—robustness and reproducibility. *Neuroimage* **33:** 204–217.
125. Mayer, A.R. *et al.* 2006. Reproducibility of activation in Broca's area during covert generation of single words at high field: a single trial FMRI study at 4 T. *Neuroimage* **32:** 129–137.
126. Peelen, M.V. & P.E. Downing. 2005. Within-subject reproducibility of category-specific visual activation with functional MRI. *Hum. Brain Mapp.* **25:** 402–408.
127. Smith, S.M. *et al.* 2005. Variability in fMRI: a re-examination of inter-session differences. *Hum. Brain Mapp.* **24:** 248–257.
128. Liu, J.Z. *et al.* 2004. Reproducibility of fMRI at 1.5 T in a strictly controlled motor task. *Magn. Reson. Med.* **52:** 751–760.
129. Stark, R. *et al.* 2004. Hemodynamic effects of negative emotional pictures—a test-retest analysis. *Neuropsychobiology* **50:** 108–118.
130. Phan, K.L. *et al.* 2003. Habituation of rostral anterior cingulate cortex to repeated emotionally salient pictures. *Neuropsychopharmacology* **28:** 1344–1350.
131. Kiehl, K.A. & P.F. Liddle. 2003. Reproducibility of the hemodynamic response to auditory oddball stimuli: a six-week test-retest study. *Hum. Brain Mapp.* **18:** 42–52.
132. Neumann, J. *et al.* 2003. Within-subject variability of BOLD response dynamics. *Neuroimage* **19:** 784–796.
133. Maitra, R., S.R. Roys & R.P. Gullapalli. 2002. Test-retest reliability estimation of functional MRI data. *Magn. Reson. Med.* **48:** 62–70.

134. Loubinoux, I. *et al*. 2001. Within-session and between-session reproducibility of cerebral sensorimotor activation: a test–retest effect evidenced with functional magnetic resonance imaging. *J. Cereb. Blood Flow Metab*. **21:** 592–607.
135. Salli, E. *et al*. 2001. Reproducibility of fMRI: effect of the use of contextual information. *Neuroimage*. **13:** 459–471.
136. White, T. *et al*. 2001. Anatomic and functional variability: the effects of filter size in group fMRI data analysis. *Neuroimage* **13:** 577–588.
137. McGonigle, D.J. *et al*. 2000. Variability in fMRI: an examination of intersession differences. *Neuroimage* **11:** 708–734.
138. Waldvogel, D. *et al*. 2000. The variability of serial fMRI data: correlation between a visual and a motor task. *Neuroreport* **11:** 3843–3847.