Published in final edited form as:
Biol Psychiatry. 2011 July 1; 70(1): 13–18. doi:10.1016/j.biopsych.2011.01.004.

Using Brain Imaging Measures in Studies of Procognitive Pharmacological Agents in Schizophrenia: Psychometric and Quality Assurance Considerations

Deanna M. Barch and
Washington University, Departments of Psychology, Psychiatry and Radiology

Daniel H. Mathalon
University of California San Francisco, Department of Psychiatry

Abstract

The first phase of the CNTRICs initiative focused on the identification of cognitive constructs from human and animal neuroscience that were relevant to understanding cognitive deficits in schizophrenia, as well as promising task paradigms that could be used to assess these constructs behaviorally. The current phase of CNTRICs has the goal of expanding this initial work by including measures of brain function that can augment these behavioral tasks as biomarkers to be used in the drug development processing. Here we review many of the psychometric issues that need to be addressed in regards to the development and inclusion of such methods in the drug development process. In addition, we review quality assurance concerns, issues associated with multi-center trials, concerns associated with potential pharmacological confounds on imaging measures, as well as power and analysis considerations. Although review is couched in the context of the use of biomarkers for treatment studies in schizophrenia, we believe the issues and suggestions included are relevant to the entire range of neuropsychiatric disorders as well as to a wide range of imaging modalities (i.e., fMRI, PET, ERP, EEG, TMS, NIR, etc.), and are relevant to both pharmacological and psychological intervention approaches.

The first phase of the CNTRICs initiative focused on the identification of cognitive constructs from human and animal neuroscience that were relevant to understanding cognitive deficits in schizophrenia, as well as promising task paradigms that could be used to assess these constructs behaviorally. The current phase of CNTRICs has the goal of expanding this initial work by including measures of brain function that can augment these behavioral tasks as biomarkers to be used in the drug development process. A relatively recent review by Breier (1) highlights the fact that biomarker development is one of the most pressing needs for current central nervous system drug development, as it is a critical

Corresponding Author: Deanna M. Barch, Ph.D. Departments of Psychology, Psychiatry and Radiology Washington University Box 1125, One Brookings Drive St. Louis, MO. 63130 Phone: 314-935-8729 Fax: 314-935-8790 dbarch@artsci.wustl.edu.

component of timely and efficient drug evaluation (1). The NIH website defines a biomarker as “A characteristic that is objectively measured and evaluated as an indicator of normal biologic or pathogenic processes or pharmacological responses to a therapeutic intervention.” This definition is broad and makes it clear that such measures could include receptor binding studies, functional magnetic resonance imaging (fMRI), electroencephalography (EEG), event-related potentials (ERP), and magneto-electroencephalography (MEG).

Breier's review highlighted three important roles for biomarkers in drug development. A first role for biomarkers is in dosage development and target validation. Historically, the pharmaceutical industry has focused on the use of PET technology for this type of biomarker, with an emphasis on examining dose-related occupancy of brain receptor sites as evidence that the drug has reached its intended target, an often invaluable piece of information in interpreting the results of negative trials. However, for many targets, there are no appropriate ligands available for such studies. As such, measures of changes in functional brain activation (fMRI, ERP, etc.) in response to task conditions ranging from simple sensory stimulation to cognitive challenge can also serve as indicators that the drug has modulated a target brain region(s) of interest (2). Further such fMRI (or ERP) measures can also serve in a second role for biomarkers, namely as surrogate outcome measures that may allow early evaluation of the eventual efficacy of a drug on longer term outcomes (1). In principle, valid biomarkers should enhance the drug discovery process by providing critical “proof of principle” evidence in early phase 2 trials. Recent work by Lewis and colleagues highlighted the promise of cognitive neuroscience measures in this role, and demonstrated that a novel GABAergic agent enhanced performance and cortical oscillations during performance of tasks designed and validated in the cognitive neuroscience literature (3). Yet a third important role for biomarkers is to help identify which people are most likely to benefit from a drug, or to identify more homogenous samples of participants that enhance power to detect significant effects.

The current paper will focus on psychometric and quality assurances issues related to using non-invasive measures of brain function such as fMRI, EEG, ERP and MEG as biomarkers during one or more phases of the drug development process, with the goal of reviewing the extent literature and offering suggestions regarding potentially useful approaches for addressing these challenges in future studies. In large part, the issues raised in this review pertain to all types of measures of brain function. For ease of presentation, we will couch our discussion primarily in terms of fMRI. However, we in no way mean to imply that these issues are any more or less important for fMRI than for any of the other methods available for measuring brain function associated with cognitive performance. Further, we believe the issues and suggestions included here are relevant to the entire range of neuropsychiatric disorders, and are relevant to both pharmacological and psychological intervention approaches.

We think that the discussion of various psychometric and quality assurance considerations will make more sense if couched in the context of a specific concrete example. Our example will be the use of fMRI to measure brain function associated with one of the constructs and tasks identified as relevant for translation in the first phase of the CNTRICs initiative: goal

maintenance as measured by the Dot Probe Expectancy (DPX) task, a variant of the AX-CPT task (4).

Reliability

One of the most critical psychometric considerations for a measure to be used as a biomarker is reliability. The term “reliability” refers to the repeatability or consistency of a measure's values across a set of observations from research participants. It is important to emphasize that reliability refers to the ability of a measure to consistently distinguish among, or reproduce the rank ordering of, individuals on a particular trait over repeated assessments, assuming that individuals do not undergo any true change between assessments. As such, reliability is only defined with respect to a set of observations from a sample or population—it does not describe the precision or accuracy of a measure for a single individual (5, 6). Moreover, reliability is not an absolute property of a measure; rather it is a property of the measurements obtained using that measure in a sample drawn from a particular population. Because the true variation among individuals varies across populations, the same measure can yield reliable measurements in one population but not in another.

The supplemental data section contains a formal definition of reliability using classical test theory (CTT). In practice, reliability coefficients are calculated as intraclass correlation coefficients using variance component estimates derived from analysis of variance models (7). In such models, a factor is included to estimate “person” variance based on scores averaged over the various measurement conditions (e.g., test occasions, raters), and this person variance provides an estimate of true score variance. When person variance is large relative to the total variance of the measurements, measurement error is relatively small and reliability is high. Based on these definitions, it should be clear that poor reliability can result from large error variance, small person variance, or both.

Based on classical test theory, several types of reliability have been developed, each based on different approaches to estimating measurement error variance. Reliability coefficients based on internal consistency, such as coefficient alpha (8), estimate measurement error based on variation across the instances of measurement, or items, for measures that employ multiple items. Test-retest reliability coefficients estimate measurement error based on variation across measurement occasions. In the context of our DPX example, internal consistency reliability would refer to the degree to which the different instances of high and low goal maintenance trials provide similar values of Blood Oxygen Level Dependent (BOLD) activity across subjects. Test-retest reliability would refer to the degree to which similar overall estimates of BOLD activity associated with high or low goal maintenance trials are obtained from a group of subjects across different timepoints, assuming that the cognitive function being measured has not undergone any true change over the test-retest interval examined.

Reliability is critical because it sets an upper limit on a measure's validity as defined by its concurrent association with other measures of the same or theoretically related constructs or by its predictive association with related constructs measured in the future. This is because it

is impossible for a measure to correlate more highly with a validity criterion (e.g., some other measure of goal maintenance besides the DPX, or BOLD activity associated with a different task that requires goal maintenance or related executive functions) than it correlates with itself (i.e., reliability), unless the same measurement error has contaminated both the measure of interest and the validity criterion with which it is being correlated (i.e., correlated measurement error). Thus, a measure with low reliability will be limited in its capacity to predict individual differences on other measures of interest. See supplemental materials for an additional discussion of reliability and measurement of change over time.

Framing Questions About Reliability

It does not make sense to ask a question such as “what is the reliability of fMRI, or MEG, or ERP, in general”. This is because reliability will vary as a function of the specific task (e.g., DPX versus another measure of goal maintenance), the specific contrast within the task (e.g., high goal maintenance trials alone, the comparison of high versus low goal maintenance trials, etc.), the specific voxels or brain regions assessed (e.g., responses in visual cortex versus dorsolateral prefrontal cortex) and the specific dependent measure (e.g., average activity in a specific region, the peak voxel with a region, etc.). Further, reliability will differ as a function of the methods or analysis approaches used. Finally, as already noted, reliability is sample and population specific—one cannot assume that a measure that is reliable in a sample of normal healthy subjects will be equally reliable in a patient population.

A growing number of studies have examined ways of assessing the test-retest reliability of imaging data (9-11). Such studies have used a wide variety of statistics and approaches to this question, with a consequently wide variety of results, ranging from good to excellent reliability (12-15) to moderate to poor reliability (16-20). Many such studies have used various forms of correlation coefficients, including Pearson and intraclass correlation coefficients (ICCs)(7), to assess test-retest reliability (based on Classical Test Theory assumptions) across scan sessions. Although these measures can be useful in many contexts, as pointed out by Zandbelt (21), such measures (Pearson's r, relative ICCs) can give high values when the rank ordering of subjects is stable across sessions, even when large changes in the absolute value of an estimate has occurred. If one wishes to establish reliability for purposes in which the absolute value of the activation is important (e.g., a situation in which the absolute value has some specific interpretation in terms of impairment, etc.), then one should use an absolute value ICC estimate that allows “main effects” that influence the measures for all subjects to count against a measure's reliability estimate (see (7)). However, in many situations such main effects (e.g., those due to practice effects or equipment changes that influence all individuals equally) do influence the measures reliability in the context of the scientific questions being addressed (e.g., is there a greater change in activation in one treatment group versus another). In such cases, a relative ICC estimate is appropriate n the context of classical test theory, reliability is a unitary construct that does not explicitly distinguish among different sources of measurement error. This can result in considerable variability in the reliability coefficients estimated for a measure if there are different sources of error across different studies. One way to address this concern is to adopt the framework of Generalizability Theory (G-Theory), an approach that allows one to

explicitly address multiple sources of measurement error that affect the reliability – or generalizability – of fMRI activations (22, 23).

Generalizability Theory Framework

Generalizability Theory (G-Theory) was developed as an extension and expansion of classical test theory to explicitly recognize and model the multiple sources of measurement error that can influence a measure's reliability, providing the flexibility to assess a number of sources of error but to estimate reliability with respect to only those sources that will be relevant to one's particular research question and study design (22-25). Please see supplemental materials for more discussion of the relationship between G-Theory and ICCs based on classical test theory. In G-theory, these different sources of error variation are referred to as “facets.” Each facet is defined by a set of similar measurement conditions, sampled from some larger “universe of admissible observations”. While we typically employ only a subset of measurement conditions for a given facet to estimate a person's true score (called a “universe score” in G-theory), our interest is in generalizing this estimate to the person's mean score over the entire universe of admissible observations for this measurement facet. In G-theory, the objects of measurement (typically persons) are sampled from a “population”, and the variability among persons is referred to as “universe score variance”. G-theory explicitly makes use of an analysis of variance framework, treating as factors, and estimating variance components, for: 1) the main effect of each facet of measurement; 2) the main effect of persons (or whatever one considers to be the object of measurement); and 3) the various interaction effects. In the context of a typical fMRI reliability or “generalizability” study, such facets could include task run, test session (e.g., time 1 versus time 2), or even site if one is conducting a multi-site trial. In our DPX example, we might have 10 participants (the objects of measurement sampled from a specific population), 8 “runs” of the DPX task with high and low goal maintenance trials interleaved (the task run facet), each person might be tested at each of two different university scanners (the site facet), undergoing a scan session on two occasions separated by a few days at each scanner site (the test occasion facet). An analysis of this type of generalizability study design would focus on a dependent measure of interest, such as activity associated with high goal maintenance trials, either in specific brain regions or on a voxel-by-voxel basis. Such an analysis would allow one to estimate the variance in the dependent measure associated with the effect of persons, the effect of each of the measurement facets, and all of their interactions (except for the highest order interaction, persons × run × occasion × site, which is confounded with residual error in the ANOVA model). Ideally, the variance associated with persons would be high in relation to the proportion of the total variance in the data, and the variance associated with factors such as task run, occasion, site, and the various interactions, would be low. Such results would indicate that the fMRI paradigm applied to the population of interest is reliable across task runs, occasions, and sites.

In G-Theory, the results of an initial generalizability study are intended to be used to inform the design of the subsequent primary research study - referred to as a “decision study” -for which one needs to make choices about issues such as the number of task runs, the number of measurement occasions, etc. A researcher can use the data from a generalizability study

to determine which choices would give sufficient reliability for the question at hand. There are several important advantages in conducting such a generalizability study, despite the added time, effort, and expense. An obvious advantage is that it could be a waste of money to conduct a trial with measures of unknown reliability, as failure to find significant effects in a clinical trial may be due to an ineffective treatment or it may be due to unreliable outcome measures that cannot detect treatment effects over and above the random fluctuations in the measures over time. Indeed, allowing the task design to be informed by data from a generalizability study could potentially lead to design of a shorter task (e.g., fewer task runs) than one might otherwise have used, thereby saving time and money and minimizing subject burden. For example, in our DPX example described above, although 8 task runs were employed in a generalizability study, the generalizability analyses could indicate that 4 runs will yield a sufficiently reliable measure for the purposes of the planned substantive (i.e., decision) study. An excellent example of such an initial generalizability study is the one by Yendiki and colleagues, which was designed to inform the design of a much larger multi-site study by clarify the choice of tasks and the number of runs per task (26).

Conducting reliability (generalizability) studies

There are a number of important considerations in designing generalizability studies. Here we focus on the issues of group versus individual levels of activation, and the unit of measurement. See the supplemental materials for a discussion of the choice of participants and test-retest time frame.

Group Versus Individual Level of Activation

Some studies intended to demonstrate “reliability” of a measure have simply shown that group means do not show significant change from test to retest. This does not explicitly address the question of measurement reliability. For example, with respect to fMRI, Caceres and others (9, 21) point out that stable group level activations can be observed over time even when there is a great deal of change in how individual subjects contribute to that group activation. In planning treatment studies or other longitudinal studies, our interest is typically focused on measuring individual differences and their changes over time; group level activations are not particularly informative about the dependability or reliability of our measurements for the purposes of tracking individual differences over time.

The Unit of Measurement

A number of studies have examined some sort of summary statistic of brain activation for each individual, such as the mean or median beta weight in a functionally defined or *a priori* ROI (e.g., 15) for fMRI studies, or a component's peak amplitude or latency assessed at one or more leads or sensors in an EEG or MEG study. With respect to fMRI, a second possibility is to compute reliability for individual voxels, generating voxel-wise maps of reliability coefficients for the entire brain. Caceres and colleagues recently completed a study that compared four different approaches to assessing ROI level reliability in an fMRI study. The first was a novel approach that used the median of the ICC distribution across voxels within an ROI (medICC). The second was the beta weight (contrast) value for an

individual subject at the voxel with the largest group level statistic (17) (ICCmax). The third was the median contrast value within an ROI for a subject (15) (ICCmed). The fourth was an intra-ROI measurement that provided information on the consistency of the spatial distribution of activation within an ROI (16)(ICCv). The results of these comparisons suggested that the ICCv was strongly influenced by smoothing in the data and that ICCmax had the lowest ICC values and was strongly influenced by smoothing and cluster size (see (9) for more detail). MedICC and ICCmed had relatively similar ICC values, but ICCmed was more influenced by smoothing and cluster size than medICC, while medICC was more influenced by head movement than ICCmed. See supplemental materials for a discussion of additional approaches to assessing reliability that treats voxels rather than subjects as the objects of measurement.

The above discussion focuses on evaluating the reliability of activation in individual ROIs or single voxels on the one hand.. However, the field has been increasingly interested in multivariate analyses of fMRI data. Caceres examined reliability for a “network” measure of activation (defined as the network activated in a group analysis), and found that the reliability of this metric was higher than for many (though not all) of the individual brain regions (9). As such, it is possible that analysis approaches that focus on identifying brain networks associated with task performance may have higher reliability than assessments of individual ROIs, something that deserves greater examination in future studies. For example, measures of functional connectivity (either task or resting state) may have higher reliability than activation of individual regions, although this hypothesis awaits empirical evaluation.

Quality assurance considerations

Some aspects of quality assurance for imaging biomarkers are relatively obvious, such as ensuring that the correct acquisition parameters and procedures are used for all protocols. However, there are a number of other considerations that are equally important, but for which clear inclusion/exclusion criteria are less obvious. Here we discuss behavioral performance and multi-site issues. See the Supplemental Materials for a discussion of signal to noise, movement, equipment stability, and how such issues should be conceptualized from the perspective of Generalizability Theory.

Behavioral Performance

Variation in behavioral performance levels can influence the level and pattern of brain activity even in the absence of any pharmacological or psychological intervention (27-29). As such, it is highly important to take task difficulty issues into consideration when choosing paradigms and control groups, and to have good practice and training procedures that ensure that participants understand the task and the response demands prior to scanning. Further, one may need to set a criterion for the level of performance participants must achieve in order for their data to be included in any analyses. One quantitative approach to this issue is to formally compute the level of chance performance for one's paradigm and to determine what level of performance an individual needs to achieve in order to perform significantly *above* chance. The potential disadvantage to this approach that it is possible that one's intervention for cognition may actually help participants better understand how to

perform a cognitive task. Thus, excluding those individuals performing poorly at the start of the trial could eliminate precisely those individuals who are most likely to benefit from the intervention.

A second concern associated with behavioral performance is that changes in performance across the course of a trial may confound interpretations of changes in brain activity (27-29). This is of particular concern if practice effects are confounded with events within the clinical trial, such that the first testing session is always off drug and the second is on drug (or post intervention). Of course, including a no-intervention control group and using parallel task versions when available helps address this concern. However, it would help to reduce any confound between practice effects and the intervention of interest in order to maximize sensitivity to true change and to reduce the influence of this confound. One way to do this is to use something akin to a multiple baseline approach. The largest practice effects in many cognitive paradigms tend to occur between the first and second time that a person completes a task, as the first session allows them to understand the task and establish a strategy. In such a multiple-baseline approach, participants would come in for a full behavioral testing session prior to the first imaging session in order to increase the likelihood that participants are on a more stable part of the learning curve at the start of the first imaging session. In addition, one can use behavioral performance as a covariate in analyses of the imaging data in order to elucidate the degree to which changes over time in functional brain activity relate to changes in behavioral performance.

Multi-site Considerations

Ideally, one would use identical equipment for task presentation and data acquisition across sites, although the degree to which differences in such factors influence the imaging data may depend on the nature of the task paradigm. In addition, training procedures and task presentation procedures need to be highly standardized across sites, with little room for individual experimenter or testor variability. A number of groups have also published papers on power and sample size in multisite studies (30), ways in which data can be analyzed to reduce site differences, (31, 32) and ways that site differences in signal characteristics (e.g., SNR, smoothness) can be taken into account in statistical analyses (26, 33, 34). One important question for multi-site studies is whether calibration across sites is only necessary for an initial study, or whether such calibration would need to be conducted anew for each new study, should the same set of sites conduct multiple studies. If money and time were no consideration, it would likely be wisest to conduct all calibration procedures anew at the start of each study, as personnel and equipment can change over time in a way that could influence site performance.

Potential Drug-Related Confounds

Pharmacologically induced confounds may be more of an issue with some methods (e.g., fMRI) than with other methods (e.g., EEG/MEG). Iannetti and Wise have provided a cogent and comprehensive review of these issues (35). These researchers outlined three processes that can be influenced by pharmacological manipulations that could mediate observed changes in BOLD activity: 1) a direct influence on neural activity (what we typically hope to measure); 2) an influence on the processes that signal blood vessels that control cerebral

blood flow (CBF): 3) an influence on the processes that modify vascular reactivity. To assess and estimate which of these influences is driving any observed change in BOLD, Iannetti and Wise suggest including the following in any pharmacological fMRI study: 1) a control task not expected to be influenced by the pharmacological intervention; 2) measures of changes in cerebral blood flow and vascular reactivity; and 3) assessments of arousal, cardiac pulsation and respiration. Please see the supplemental materials for a more detailed discussion.

Power and Analysis Considerations

Sample Size and Power

There are major advantages in terms of power to within-subject designs that compare the same individual on and off the intervention. This approach is highly feasible in proof of concept type trials that use a single dose design in which the order of placebo versus drug can be counterbalanced across participants. However, this approach is less feasible in studies with longer term administration of a pharmacological agent or intervention in which it may be not be possible to counterbalance the “on” and “off” imaging days. In these situations, it may be necessary to have a control group that does not receive the intervention. In such designs, power concerns need to take into account the between subject nature of the design. Many of the published discussions of power in fMRI studies focus only on within-subject designs, and reliance on those sample size suggestions (36, 37) will greatly overestimate the power than one has for a between-subject design (38). Fortunately, recent work has outlined the power considerations and sample sizes necessary for such between group designs (30). However, it should be noted that effect sizes are sometimes difficult to determine in complex designs where one is looking at changes across conditions and time. As such, one may also need to design a study to be able to detect the *minimum* effect size that is likely to be clinically relevant and power the study to detect this effect size, even if one hopes the obtained effect size will be larger. This concern is particularly pertinent in studies conducted as part of the drug development process, as failing to find a significant and potentially clinical important effect due to low power could derail the development of a promising mechanism.

Analysis Approach and Power

In studies using measures of functional brain activity, the choice of analysis approach can have a major impact on power. With almost all imaging methods, exploratory analyses can involve examining many different voxels or electrodes. The need to control Type 1 error in such analyses can often greatly reduce power and increase the risk of Type II errors. As such, researchers should consider taking a tiered approach to data analysis that starts with the most highly powered approach. In fMRI studies, a highly powered approach is to use *a priori* identified regions of interest (ROIs) and to analyze only the mean values for all voxels included in any individual ROI. The advantage of this approach is that it necessitates relatively few comparisons, requires relatively little correction for multiple comparisons, and is very theoretically driven. However, the disadvantage is that you can choose the wrong size or location of the ROIs, and miss detecting real changes in functional activation in response to the intervention.

The next most powerful approach would be to use such *a priori* ROIs as masks, but to examine each voxel within the mask using some appropriate correction for the number of comparison within or between ROIs. This approach is still more highly powered than a whole brain exploratory analysis, as one is conducting fewer comparisons than in a whole-brain analysis. Further, this approach may be less sensitive to misspecification than the "whole" ROI approach since it allows for the detection of significant changes in subregions of the masking ROIs. The least powerful approach is to do some type of whole brain exploratory analysis, which typically requires a rather stringent correction for multiple comparisons (39), but avoids any pitfalls associated with an inaccurate prediction of the location of brain activity changes in response to the intervention.

Summary

The inclusion of measures of brain activity during cognitive performance has the promise of enhancing the drug discovery process and potentially allowing us to develop effective and targeted treatments for impaired cognition in schizophrenia. However, there are a number of psychometric and methodological challenges that need to be addressed in order to make the use of these measures feasible, easy to implement, and resistant to confounds in interpretation. While these challenges will necessitate time and effort on the part of the field, they are not insurmountable and are well worth the effort in terms of the potential payoff. We hope that this review helps to highlight some of the key methodological and conceptual challenges that remain, as well as providing useful suggestions for subsequent studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

Dr. Barch has received grants from the NIMH, NIA, NARSAD, Allon, Novartis, and the McDonnel Center for Systems Neuroscience. Dr. Mathalon has received research grants from the NIMH, NARSAD, AstraZeneca, and GlaxoSmithKline and has been a paid consultant for Pfizer.

References

1. Breier A. Developing drugs for cognitive impairment in schizophrenia. Schizophr Bull. 2005; 31:816–822. [PubMed: 16150959]
2. Cho RY, Ford JM, Krystal JH, Laruelle M, Cuthbert BN, Carter CS. Functional neuroimaging and electrophysiology biomarkers for clinical trials for cognition in schizophrenia. Schizophrenia Bulletin. 2005; 31:865–869. [PubMed: 16166611]
3. Lewis DA, Cho RY, Carter CS, Eklund K, Forster S, Kelly MA, et al. Subunit-selective modulation of GABA type A receptor neurotransmission and cognition in schizophrenia. Am J Psychiatry. 2008; 165:1585–1593. [PubMed: 18923067]
4. Barch DM, Berman MG, Engle R, Jones JH, Jonides J, Macdonald A 3rd, et al. CNTRICS final task selection: working memory. Schizophr Bull. 2009; 35:136–152. [PubMed: 18990711]
5. Rogosa DR, Brandt D, Zimowski M. A growth curve approach to the measurement of change. Psychological Bulletin. 1982; 92:726–748.

6. Rogosa DR, Willett JB. Understanding correlates of change by modeling individual differences in growth. Psychometriika. 1983; 50:203–228.
7. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin. 1979; 86:420–428. [PubMed: 18839484]
8. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometriika. 1951; 16:297–334.
9. Caceres A, Hall DL, Zelaya FO, Williams SC, Mehta MA. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage. 2009; 45:758–768. [PubMed: 19166942]
10. Raemaekers M, Vink M, Zandbelt B, van Wezel RJ, Kahn RS, Ramsey NF. Test-retest reliability of fMRI activation during prosaccades and antisaccades. Neuroimage. 2007; 36:532–542. [PubMed: 17499525]
11. Zandbelt BB, Gladwin TE, Raemaekers M, van Buuren M, Neggers SF, Kahn RS, et al. Within-subject variation in BOLD-fMRI signal changes across repeated measurements: quantification and implications for sample size. Neuroimage. 2008; 42:196–206. [PubMed: 18538585]
12. Specht K, Willmes K, Shah NJ, Jancke L. Assessment of reliability in functional imaging studies. J Magn Reson Imaging. 2003; 17:463–471. [PubMed: 12655586]
13. Fernandez G, Specht K, Weis S, Tendolkar I, Reuber M, Fell J, et al. Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. Neurology. 2003; 60:969–975. [PubMed: 12654961]
14. Aron AR, Gluck MA, Poldrack RA. Long-term test-retest reliability of functional MRI in a classification learning task. Neuroimage. 2006; 29:1000–1006. [PubMed: 16139527]
15. Friedman L, Stern H, Brown GG, Mathalon DH, Turner J, Glover GH, et al. Test-retest and between-site reliability in a multicenter fMRI study. Hum Brain Mapp. 2008; 29:958–972. [PubMed: 17636563]
16. Raemaekers M, Vink M, zandbeldt BB, van Wezel RJ, Kahn R, Ramsey NF. Test-retest reliablity of fMRI activation during prosaccades and antisaccades. Neuroimage. 2007; 36:532–542. [PubMed: 17499525]
17. Manoach DS, Halpern EF, Kramer TS, Chang Y, Goff DC, Rauch SL, et al. Test-retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. Am J Psychiatry. 2001; 158:955–958. [PubMed: 11384907]
18. Wei X, Yoo SS, Dickey CC, Zou KH, Guttmann CR, Panych LP. Functional MRI of auditory verbal working memory: long-term reproducibility analysis. Neuroimage. 2004; 21:1000–1008. [PubMed: 15006667]
19. Wagner K, Frings L, Quiske A, Unterrainer J, Schwarzwald R, Spreer J, et al. The reliability of fMRI activations in the medial temporal lobes in a verbal episodic memory task. Neuroimage. 2005; 28:122–131. [PubMed: 16051501]
20. McGonigle DJ, Howseman AM, Athwal BS, Friston KJ, Frackowiak RS, Holmes AP. Variability in fMRI: an examination of intersession differences. Neuroimage. 2000; 11:708–734. [PubMed: 10860798]
21. zandbeldt BB, Gladwin TE, Raemaekers M, van Buuren M, Neggers SF, Kahn R, et al. Within-subject variation in BOLD-fMRI signal changes across repeated measurements: Quantification and implications for sample size. Neuroimage. 2008; 42:196–206. [PubMed: 18538585]
22. Cronbach LJ, Nageswari R, Gleser GC. Theory of generalizability: A liberation of reliablity theory. The British Journal of Statistical Psychology. 1963; 16:137–163.
23. Cronbach, LJ.; Gleser, GC.; Nanda, H.; Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. John WIley; New York: 1972.
24. Brennan, RL. Generalizability Theory. Springer-Verlag; New York: 2000.
25. Shavelson, RJ.; Webb, NM. Generalizability Theory: A Primer. Sage Publications; Newbury Park, CA: 1991.
26. Yendiki A, Greve DN, Wallace S, Vangel M, Bockholt J, Mueller BA, et al. Multi-site characterization of an fMRI working memory paradigm: reliability of activation indices. Neuroimage. 2010; 53:119–131. [PubMed: 20451631]

27. Van Snellenberg JX, Torres IJ, Thornton AE. Functional neuroimaging of working memory in schizophrenia: task performance as a moderating variable. Neuropsychology. 2006; 20:497–510. [PubMed: 16938013]
28. Callicott JH, Bertolino A, Mattay VS, Langheim FJ, Duyn J, Coppola R, et al. Physiological dysfunction of the dorsolateral prefrontal cortex in schizophrenia revisited. Cerebral Cortex. 2000; 10:1078–1092. [PubMed: 11053229]
29. Gur RC, Gur RE. Hypofrontality in schizophrenia: RIP. Lancet. 1995; 345:1383–1384. [PubMed: 7760605]
30. Suckling J, Barnes A, Job D, Brenan D, Lymer K, Dazzan P, et al. Power calculations for multicenter imaging studies controlled by the false discovery rate. Hum Brain Mapp. 2010; 31:1183–1195. [PubMed: 20063303]
31. Bennett CM, Miller MB. How reliable are the results from functional magnetic resonance imaging? Ann N Y Acad Sci. 2010; 1191:133–155. [PubMed: 20392279]
32. Bosnell R, Wegner C, Kincses ZT, Korteweg T, Agosta F, Ciccarelli O, et al. Reproducibility of fMRI in the clinical setting: implications for trial designs. Neuroimage. 2008; 42:603–610. [PubMed: 18579411]
33. Friedman L, Glover GH. Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. Neuroimage. 2006; 33:471–481. [PubMed: 16952468]
34. Friedman L, Stern H, Brown GG, Mathalon DH, Turner J, Glover G, et al. Test-retest and between-site reliability in a multicenter fMRI study. Human Brain Mapping. 2008; 29:958–972. [PubMed: 17636563]
35. Iannetti GD, Wise RG. BOLD functional MRI in disease and pharmacological studies: room for improvement? Magnetic Resonance Imaging. 2007; 25:978–988. [PubMed: 17499469]
36. Desmond JE, Glover GH. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. J Neurosci Methods. 2002; 118:115–128. [PubMed: 12204303]
37. Mumford JA, Nichols TE. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. Neuroimage. 2008; 39:261–268. [PubMed: 17919925]
38. Suckling J, Ohlssen D, Andrew C, Johnson G, Williams SC, Graves M, et al. Components of variance in a multicentre functional MRI study and implications for calculation of statistical power. Hum Brain Mapp. 2008; 29:1111–1122. [PubMed: 17680602]
39. Lieberman MD, Cunningham WA. Type I and Type II error concerns in fMRI research: re-balancing the scale. Soc Cogn Affect Neurosci. 2009; 4:423–428. [PubMed: 20035017]