SAMPLE SIZE AND POWER CALCULATIONS IN FMRI STUDIES

IDENTIFYING AND OVERCOMING OBSTACLES TO SAMPLE SIZE AND POWER CALCULATIONS IN FMRI STUDIES

By

QING GUO, B.Sc., M.Sc., M.Math.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2013) McMaster University

(Health Research Methodology – Biostatistics Specialization) Hamilton, Ontario

TITLE: Identifying and Overcoming Obstacles to Sample Size and Power Calculations in fMRI Studies

AUTHOR: Qing Guo, B.Sc. (Hunan Normal University),
M.Sc. (Lakehead University),
M.Math. (University of Waterloo)

SUPERVISOR: Dr. Eleanor Pullenayegum

NUMBER OF PAGES: xii, 125

ABSTRACT

Functional magnetic resonance imaging (fMRI) is a popular technique to study brain function and neural networks. Functional MRI studies are often characterized by small sample sizes and rarely consider statistical power when setting a sample size. This could lead to data dredging, and hence false positive findings. With the widespread use of fMRI studies in clinical disorders, the vulnerability of participants points to an ethical imperative for reliable results so as to uphold promises typically made to participants that the study results will help understand their conditions. While important, power-based sample size calculations can be challenging. The majority of fMRI studies are observational, i.e., are not designed to randomize participants to test efficacy and safety of any therapeutic intervention. My PhD thesis therefore addresses two objectives: firstly, to identify potential obstacles to implementing sample size calculations, and secondly to provide solutions to these obstacles in observational clinical fMRI studies. This thesis contains three projects.

Implementing a power-based sample size calculation requires specifications of effect sizes and variances. Typically in health research, these input parameters for the calculation are estimated from results of previous studies, however these often seem to be lacking in the fMRI literature. Project 1 addresses the first objective through a systematic review of 100 fMRI studies with clinical participants, examining how often observed input parameters were reported in the results section so as to help design a new well-powered study. Results confirmed that both input estimates and sample size calculations

were rarely reported. The omission of observed inputs in the results section is an impediment to carrying out sample size calculations for future studies.

Uncertainty in input parameters is typically dealt with using sensitivity analysis; however this can result in a wide range of candidate sample sizes, leading to difficulty in setting a sample size. Project 2 suggests a cost-efficiency approach as a short-term strategy to deal with the uncertainty in input data and, through an example, illustrates how it narrowed the range to choose a sample size on the basis of maximizing return on investment.

Routine reporting of the input estimates can thus facilitate sample size calculations for future studies. Moreover, increasing the overall quality of reporting in fMRI studies helps reduce bias in reported input estimates and hence helps ensure a rigorous sample size calculation in the long run. Project 3 is a systematic review of overall reporting quality of observational clinical fMRI studies, highlighting under-reported areas for improvement and suggesting creating a shortened version of the checklist which contains essential details adapted from the guidelines proposed by Poldrack et al. (2008) to accommodate strict word limits for reporting observational clinical fMRI studies.

In conclusion, this PhD thesis facilitates future sample size and power calculations in the fMRI literature by identifying impediments, by providing a short-term solution to overcome the impediments using a cost-efficiency approach in conjunction with conventional methods, and by suggesting a long-term strategy to ensure a rigorous sample size calculation through improving the overall quality of reporting.

In memory of my father, who taught me to believe in diligence, honesty and perseverance, and to follow my dreams. His spirit, enduring wisdom and strength remain.

ACKNOWLEDGEMENTS

First, I thank my family for their love, support and constant inspiration in all my pursuits to fulfill my dreams. Thank you to my mother for selfless devotion, understanding and belief in me, and to my daughter for bringing joy and making it all worthwhile.

I extend the deepest gratitude to my thesis supervisor, Eleanor Pullenayegum, for her support, guidance, challenges, persistent suggestions, and academic attitude and way of training. Moreover, her efforts of coaching have helped me improve the level of academic writing and facilitate my growth as an independent scholar and collaborative researcher. I am also very grateful to other members of my thesis committee: Lehana Thabane, Geoffrey Hall, Margaret McKinnon, and Ron Goeree for their availability, guidance and insight. They have been more than willing to offer their help and wisdom whenever necessary. In addition, I thank all co-authors of my thesis manuscripts for their contributions. My sincere thanks go to Harry Shannon for his genuine advice and encouragements. I thank Stephen Walter, Noori Akhtar-Danesh and Dr. P.J. Devereaux, who offered me the opportunities to serve as a tutor and teaching assistant in the graduate courses that they have coordinated. I cannot sufficiently thank Joanne Buckley for her proof-reading; the remaining errors can be certainly attributed to me. I am also thankful to the Canadian Institute of Health Research training award and Ontario Graduate Scholarship for the partial funding on this work.

I appreciate the friendship and support from fellow students and the administrative staff in the Health Research Methodology program and Department of Clinical Epidemiology and Biostatistics at McMaster University, and St. Joseph's Healthcare Hamilton. I would like to thank many others including but not limited to Jing Wu, Pat Wilson (Taylor), and Lynda Pettit for their consistent friendship along this journey.

Finally, I thank God, for His trials and blessings.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Appendices

PREFACE

This "sandwich" thesis consists of five chapters. Chapter 1 introduces the objectives and themes of the work, and chapter 5 concludes the significance of study findings with implications for future work. Two published projects and one project currently under revision but not yet published in peer-reviewed journals, are contained in chapters 2, 3, and 4 respectively. Qing Guo's contributions to all the articles in the Thesis include: developing the research ideas and research questions; designing the studies; collecting the data; developing the analysis plans; conducting all statistical analyses; writing up all manuscripts; submitting the manuscripts; and responding to reviewers' comments. All the work presented in the chapters was conducted during my PhD studies.

CHAPTER ONE

INTRODUCTION

Functional magnetic resonance imaging (fMRI) studies have been increasingly used in clinical conditions. Clinical fMRI studies fall into two main categories. One category uses imaging to develop a tool to help predict the development of psychiatric disorders (e.g., schizophrenia) or to predict which patients may recover from brain injury or benefit from rehabilitation. The second category, which is the focus of this thesis, is to understand neural activities by use of cognitive tasks underlying core disorders. The disorders can be neurological (e.g., stroke, traumatic head injury, and neurodegenerative disease), psychiatric (e.g., depression, schizophrenia, and Alzheimer's), or developmental (e.g., dyslexia) disorders (D'Esposito, 2006; Carter et al., 2008). When recruiting clinical participants, investigators have made promises to their participants that their participation will help society to understand their conditions; in order to fulfill these promises, we need to provide valid results. Sample size calculation plays an important role in helping ensure reliable results for the proposed study. This is often conducted through a pre-study power analysis to determine a sample size that has a desired power to detect a minimally important difference at a given level of significance (Muller and Benignus, 1992; Lenth, 2001; Chow et al., c2003). While important, sample size calculations are not always conducted in the fMRI field (Carp, 2012).

As the majority of fMRI studies are concerned with relationships between cognitive tasks and brain functions, these studies are observational (i.e., this type of study is not

designed to randomize participants to test efficacy and safety of any therapeutic intervention). This thesis uses systematic review methodology to identify obstacles to calculating sample sizes in observational clinical fMRI studies, and provides solutions and guidance for overcoming these obstacles. Three core issues are: (i) the documentation of observed input estimates to a sample size calculation in the results section of manuscripts, (ii) a short-term strategy to handle uncertainty in input parameter estimates to set a sample size with cost efficiency and statistical power, and (iii) a long-term strategy to help improve the overall quality of reporting to ensure an effective sample size calculation for future studies.

Issue 1: Reporting of observed input parameters in the results section

Generally, fMRI studies are small, and the number of subjects is determined based upon operational constraints (e.g., access to scanning time, limited resources, and medical conditions) rather than statistical power (Murphy and Garavan, 2004; Lazar, 2008). Setting sample size this way could result in a study that has inadequate power to detect the effect of interest.

Statistical methods for estimating sample sizes in fMRI studies have been developed over the past decade (Desmond and Glover, 2002; Murphy and Garavan, 2004; Hayasaka et al., 2007; Mumford and Nichols, 2008). Nevertheless, sample size calculations are rarely reported in the fMRI literature as a whole (Carp, 2012). As a sample size calculation needs estimates of input parameters (e.g., effect size, within- and between-subject variances, and temporal autocorrelation) (Cohen, 1977; Lenth, 2001), omission of

estimated values of these input parameters in the results section is an obstacle to implementing power calculations. The frequency of reporting of the input parameters needed for such calculations remains unknown in current clinical fMRI literature.

Issue 2: Short-term strategy to calculate sample sizes given the substantial uncertainty in input parameters

The innovative nature of the fMRI field coupled with limited reporting leads to much uncertainty in the value of input data necessary for power-based sample size calculations. Sensitivity analyses may help calculate a range of candidate sample sizes by varying the input parameters over their plausible ranges but can be suboptimal when the range is wide.

Furthermore, functional MRI studies are costly. Maximizing the information gained from the study per unit cost while maintaining the study's power is an important consideration. A practical strategy is lacking in the fMRI literature to reduce the uncertainty of conventional power analysis by narrowing the wide range of sample sizes and to set a sample size by maximizing cost efficiency.

Issue 3: Long-term strategy to help improve the overall quality of reporting to facilitate future sample size calculations

Sample size and power calculations are critical to help design a study yielding good quality of results; however these calculations cannot be implemented without estimates of the necessary input parameters.

Routine reporting of observed input parameters in the results section of manuscripts can facilitate an adequately powered new study. Insufficient reporting (e.g., selective reporting) in fMRI studies makes it difficult to evaluate the methodological rigor and the scope for bias in reported results, and hence could introduce bias in reported input estimates. Consequently, better overall reporting of fMRI studies would help improve sample size calculations in the long run. Guidelines have been created by Poldrack et al. (Poldrack et al., 2008) to enhance reporting of fMRI studies. Even though fMRI scientists have increasingly recognized that the guidelines help improve the quality, transparency and consistency of results (Carter et al., 2008; MacDonald et al., 2009; Carp, 2012; Huang et al., 2012), the overall quality of reporting of observational clinical fMRI studies based on these guidelines has not been systematically investigated.

Outline of the thesis

This is a sandwich thesis containing three papers, each mapping onto one of the issues discussed above. The three papers are in chapters 2 to 4.

Chapter 2 delineates a checklist of input parameters such as effect size, within-subject variance, between-subject variance, and temporal autocorrelation matrix that are needed to calculate sample sizes based on three key methods in this field. Through a systematic review, we assessed how often the observed input parameters were reported in the results section and how often sample size calculations were reported in the methods section. We make recommendations to help enhance reporting of the input parameters in the results section.

Chapter 3 introduces a cost-efficiency method and demonstrates through an example how cost-efficiency in conjunction with conventional methods can determine a sample size with adequate power to detect at least one of the proposed effect sizes while being cost efficient. We propose using cost-efficiency as a supplement to conventional sample size methods to deal with substantial uncertainty arising from poor availability of input parameters.

Chapter 4 provides empirical evidence of the quality of reporting in observational clinical fMRI studies through a systematic review. The under- and well- reported areas are identified. Adhering to the guidelines developed by Poldrack et al. (2008) would help improve reporting of observational clinical fMRI studies. Given that the guidelines are lengthy and there is no consensus regarding which items on the list are indeed essential to report, we suggest that a shortened version of the checklist encompassing essential items adapted from the guidelines be created for consideration and discussion.

Chapter 5 synthesizes the key points from the above three chapters, outlines the limitations, and points to practical implications of this thesis work.

References

Carp, J., 2012. The secret lives of experiments: Methods reporting in the fMRI literature. Neuroimage 63, 289-300.

Carter, C.S., Heckers, S., Nichols, T., Pine, D.S., Strother, S., 2008. Optimizing the design and analysis of clinical functional magnetic resonance imaging research studies. Biol. Psychiatry 64, 842-849.

Chow, S., Shao, J., Wang, H., c2003. Sample Size Calculations in Clinical Research. Marcel Dekker, New York.

Cohen, J., 1977. Statistical Power Analysis for the Behavioral Sciences, Rev. ed. Academic Press, New York.

Desmond, J.E., Glover, G.H., 2002. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. J. Neurosci. Methods 118, 115-128.

D'Esposito, M., 2006. Functional MRI: Applications in Clinical Neurology and Psychiatry, 1st ed. Informa Healthcare, United Kingdom.

Hayasaka, S., Peiffer, A.M., Hugenschmidt, C.E., Laurienti, P.J., 2007. Power and sample size calculation for neuroimaging studies by non-central random field theory. Neuroimage 37, 721-730.

Huang, W., Pach, D., Napadow, V., Park, K., Long, X., Neumann, J., Maeda, Y., Nierhaus, T., Liang, F., Witt, C.M., 2012. Characterizing acupuncture stimuli using brain imaging with FMRI--a systematic review and meta-analysis of the literature. PLoS One 7, e32960.

Lazar, N.A., 2008. The Statistical Analysis of Functional MRI Data, 1st ed. Springer-Verlag New York, New York, NY.

Lenth, R.V., 2001. Some Practical Guidelines for Effective Sample Size Determination. The American Statistician 55, 187-193.

MacDonald, A.W.,3rd, Thermenos, H.W., Barch, D.M., Seidman, L.J., 2009. Imaging genetic liability to schizophrenia: systematic review of FMRI studies of patients' nonpsychotic relatives. Schizophr. Bull. 35, 1142-1162.

Muller, K.E., Benignus, V.A., 1992. Increasing scientific power with statistical power. Neurotoxicol. Teratol. 14, 211-219.

Mumford, J.A., Nichols, T.E., 2008. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. Neuroimage 39, 261-268.

Murphy, K., Garavan, H., 2004. An empirical investigation into the number of subjects required for an event-related fMRI study. Neuroimage 22, 879-885.

Poldrack, R.A., Fletcher, P.C., Henson, R.N., Worsley, K.J., Brett, M., Nichols, T.E., 2008. Guidelines for reporting an fMRI study. Neuroimage 40, 409-414.

CHAPTER TWO

A Systematic Review of the Reporting of Sample Size Calculations and Corresponding Data Components in Observational Functional Magnetic Resonance Imaging Studies

Qing Guo[a,b,*], Lehana Thabane[a,b,c], Geoffrey Hall[d], Margaret McKinnon[d,e,f], Ron Goeree[a,c,g], Eleanor Pullenayegum[a,b,c,h]

[a]*Department of Epidemiology and Biostatistics, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada*

[b]*Biostatistics Unit, St Joseph's Healthcare Hamilton, Hamilton, ON, Canada*

[c]*Centre for Evaluation of Medicine, St Joseph's Healthcare Hamilton, Hamilton, ON, Canada*

[d]*Department of Psychiatry and Behavioural Neurosciences, McMaster University, ON, Canada*

[e]*Mood Disorders Program, St. Joseph's Healthcare Hamilton, ON, Canada*

[f]*Kunin-Lunenfeld Applied Research Unit, Baycrest, Toronto, ON, Canada*

[g]*PATH Research Institute, St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada*

[h]*The Hospital for Sick Children, Toronto, ON, Canada*

---

* Corresponding author.

E-mail address: guoq@mcmaster.ca (Qing Guo).

This article is cited as:

Guo Q., Thabane L., Hall G., McKinnon M., Goeree R., Pullenayegum E., A systematic review of the reporting of sample size calculations and corresponding data components in observational functional magnetic resonance imaging studies, NeuroImage (2013), *http://dx.doi.org/10.1016/j.neuroimage.2013.08.012*

ABSTRACT

Background

Anecdotal evidence suggests that functional magnetic resonance imaging (fMRI) studies rarely consider statistical power when setting a sample size. This raises concerns since undersized studies may fail to detect effects of interest and encourage data dredging. Although sample size methodology in this field exists, implementation requires specifications of estimated effect size and variance components.

Methods

We therefore systematically evaluated how often estimates of effect size and variance components were reported in observational fMRI studies involving clinical human participants published in six leading journals between January 2010 and December 2011. A random sample of 100 eligible articles was included in data extraction and analyses. Two independent reviewers assessed the reporting of sample size calculations and the data components required to perform the calculations in the fMRI literature.

Results

One article (1%) reported sample size calculations. The reporting of parameter estimates for effect size (8%), between-subject variance (4%), within-subject variance (1%) and temporal autocorrelation matrix (0%) was uncommon. Three articles (3%) reported Cohen's *d* or *F* effect sizes. The majority (83%) reported peak or average *t*, *z* or *F* statistic. The inter-rater agreement was very good, with a prevalence-adjusted bias-adjusted kappa (PABAK) value greater than 0.88.

Conclusions

Sample size calculations were seldom reported in fMRI studies. Moreover, omission of parameter estimates for effect size, between- and within-subject variances, and temporal autocorrelation matrix could limit investigators' ability to perform power analyses for new studies. We suggest routine reporting of these quantities, and recommend strategies for reducing bias in their reported values.

*Keywords***:** Statistical power; Sample size; Functional magnetic resonance imaging (fMRI); Effect size; Variance components

1. Introduction

Studies using functional magnetic resonance imaging (fMRI) have proliferated in the past decade. Anecdotal evidence suggests that most often fMRI studies involve 12 to 16 subjects per group, and rarely consider statistical power when setting a sample size. Instead, the number of subjects is often determined by practical constraints such as access to scanning time and costs (Murphy and Garavan, 2004). This raises concerns as such studies may have inadequate power to detect effects of interest and thus encourage data dredging (i.e., one simply tests multiple hypotheses on the same dataset until statistical significance is found.) (Smith and Ebrahim, 2002) leading to spurious effects (Yarkoni, 2009) and inflated false positive findings (Simmons et al., 2011; Carp, 2012a). Therefore, it is critical to calculate power-based sample sizes prior to fMRI data collection and to report the calculations in manuscripts to ensure appropriate numbers of subjects. While important, sample size and power calculations are often challenging. In addition to clear scientific objectives, they require specifications of effect sizes (i.e., mean activation), variances, type I and type II error rates (Lenth, 2001; Mumford, 2012).

Approaches for sample size calculations have been developed in this field (Desmond and Glover, 2002; Hayasaka et al., 2007; Mumford and Nichols, 2008). Specifically, implementation requires estimates for i) the size of effect to be detected (e.g., mean activation or percent signal change between two conditions), ii) between-subject variance, iii) within-subject variance, and iv) the temporal auto-correlation variance-covariance matrix. In other fields, input parameters for sample size calculations are often estimated

from previously published studies (Wilkinson and Task Force Stat Inference, 1999; Lenth, 2001; Zaslavsky, 2010). Our experience suggests that estimates of data components necessary to compute sample sizes are difficult to find in the fMRI literature. Moreover, a recent survey has demonstrated lack of reporting of power analysis in the fMRI literature (Carp, 2012b). We hypothesized that effect sizes, between- and within-subject variances, and temporal autocorrelations are similarly rarely reported, meaning that investigators would not be able to conduct sample size calculations even if they wished to do so. Because the majority of fMRI studies are observational (i.e., this type of study is not designed to assess the efficacy or safety of any therapeutic intervention), and fMRI is increasingly applied in clinical disorders (Sheline et al., 2001; Siegle et al., 2002; Glahn et al., 2005; Snitz et al., 2005; Monk et al., 2008; Yoon et al., 2008), the vulnerability of clinical participants points to an ethical imperative for rigorous methodology and better reporting. We conducted a systematic review to assess the reporting of effect sizes and variance components in observational fMRI studies involving clinical human participants published in 2010 and 2011 among six leading journals with high impact factors. Clinical human participants here refer to those who either have a disease or who are at risk of developing a disease. Specifically, we evaluated how often sample size calculations were reported, and quantified the percentage of articles that reported estimates of size of effect and variance components in the results section.

2. Methods

A literature search for fMRI studies was conducted in OVID MEDLINE (1946 to January 2012) by using the keyword search term, "functional magnetic resonance imaging", combined with the acronym "fmri". In the *Journal Citation Report* 2010, we selected four journals with a high impact factor (IF) in the category "Neurosciences", namely, *Neuron* (IF 14.9), *Nature Neuroscience* (IF 14.2), *Brain* (IF 9.2), *Journal of Neuroscience* (IF 7.3), one journal with the highest impact factor in the category, "Neuroimaging" (*Neuroimage*, IF 5.94), and one journal with a good proportion and high quality of fMRI studies (*Proceedings of the National Academy of Sciences of the United States of America*, IF 9.8). The results were limited to a two-year period (from January 2010 through December 2011), English language, and involving human imaging studies in the selected six journals (see Appendix A). Duplicate articles were removed.

2.1 *Eligibility Criteria and Study Selection*

To be included in this review, publications had to meet the predefined inclusion and exclusion criteria. Inclusion criteria were full reports of observational fMRI studies involving clinical human participants, and block or event-related design for the fMRI paradigm. Articles were excluded if they were published only in abstract form, or if they were editorials, letters, comments or reviews. Genetic, resting-state observational fMRI studies, fMRI studies other than observational studies (e.g., randomized clinical trials), and studies of connectivity were also excluded.

We set out to include 100 eligible articles in data extraction and evaluation. After removing the duplicates, we reviewed citations randomly until the target sample size of 100 eligible articles was reached.

2.2 *Data Extraction and Review Process*

Electronic data extraction forms (see Appendix B) were created to abstract data from each citation. We piloted and tested the forms using a random sample of six papers from six journals with two steps: First, two reviewers (QG and EP) independently assessed reporting of three articles among the six papers using the developed data extraction forms, made modifications on the forms, and obtained the same perception, interpretation and definitions of responses to each evaluated item. The between-reviewer agreement was thus potentially increased. Second, the two reviewers independently evaluated the other three articles based on the modified abstraction forms. The observed percentage of agreement on judgments between the two reviewers was 0.70 or higher. Final abstraction forms were devised prior to use. Eligibility of articles, and characteristics of eligible articles including the type of journal where the articles were published, article publication year, study design, study sample size, and funding sources were collected. We also examined whether sample size calculations were reported, and whether the estimated values that are required in the existing approaches for power-based sample size calculations were reported in the results section.

Two authors (QG and EP), blinded to each other's assessment, extracted and reviewed the reporting of each article independently. QG randomly screened unique articles from

the initial search strategy for eligible studies until the target number of 100 was reached. Among the initial articles, 50 were randomly selected for EP to assess eligibility. Of the100 eligible articles that the first reviewer extracted, 50 articles were randomly selected for EP to abstract data for quality assurance. The sample size of 50 was chosen so as to estimate the kappa for the inter-rater agreement (Altman, 1991) within a margin of error of 0.3 with 95% confidence, assuming that the true kappa would be 0.6 or more and that the proportion of agreements by chance was 0.7 or less. Any disagreements were resolved through consensus.

2.3 *Parameters Needed to Report for Future Sample Size Calculations*

Here we focused on three approaches for sample size and power calculations developed in fMRI studies (Desmond and Glover, 2002; Hayasaka et al., 2007; Mumford and Nichols, 2008). These were reviewed briefly below; specifically outlining parameters needed to perform power analyses for a new study (see Table 1 for a summary).

1) Mumford and Nichols (2008)

Mumford and Nichols' method (2008) for group-level fMRI studies incorporates temporal autocorrelation into the within-subject variance estimate. The power calculation is based on a non-central *T* or *F* distribution. The implementation requires estimates of Δ (size of effect or mean percent signal change between two conditions), σ_w^2 (within-subject variance), σ_b^2 (between-subject variance) and V (temporal autocorrelation matrix). These estimates are calculated by averaging over all voxels in a ROI.

2) Hayasaka, Peiffer, Hugenschmidt, and Laurienti (2007)

Hayasaka and his colleagues (2007) presented a method, based on non-central random field theory (RFT), of calculating statistical power to detect signals among spatially correlated voxels. In particular, this method can calculate power at participated areas of the brain in a 3D image to enable visualizing of spatially varying power over the brain. This method adjusts for multiple comparisons and accounts for spatial correlation among voxels. The power calculation is based on the distribution of the maximum of non-central T- or F-random fields. The parameters needed to perform power-based sample size calculations are Cohen's *d* or Cohen's *f* effect sizes.

3) Desmond and Glover (2002)

Desmond and Glover (2002) suggested a simulation-based method to calculate statistical power for group-level fMRI studies. It allows only for block-design on task presentation and does not account for temporal autocorrelation when estimating the within-subject variance. The power calculation needs to specify the estimates of Δ (size of effect or mean percent signal change between two conditions), σ_w^2 (within-subject variance) and σ_b^2 (between-subject variance). These parameter estimates are averaged over all voxels within some specific ROIs or in a whole-brain.

2.4 *Statistical Analysis*

Categorical variables were reported as percentages. Continuous variables were expressed as mean and standard deviation for symmetric distributions, or median, 1st

quartile (Q1) and 3rd quartile (Q3) for skewed distributions. The percentage of studies that reported each evaluation item and a 95% confidence interval (CI) was calculated by using an exact binomial method. All statistical analyses were conducted using the SAS 9.2 software (Cary, NC).

Inter-rater agreement was assessed through the observed percent agreement, Cohen's Kappa coefficient (κ), and prevalence-adjusted bias-adjusted kappa (PABAK) (Byrt et al., 1993). When the prevalence of a rating is very high or low, the value of kappa may indicate a low level of agreement while the observed percentage of agreement is high, known as the kappa paradox (Feinstein and Cicchetti, 1990). Hence, we reported percent agreement and the PABAK in addition to the kappa to address this paradox and to better interpret the inter-rater agreement. Kappa coefficient results were interpreted based on the scale as proposed by Byrt (Byrt, 1996): 0.00 or less (No agreement), 0.01-0.20 (Poor agreement), 0.21-0.40 (Slight agreement), 0.41-0.60 (Fair agreement), 0.61-0.80 (Good agreement), 0.81-0.92 (Very good agreement), 0.93-1.00 (Excellent agreement).

2.5 *Sample Size*

We performed a sample size calculation to determine the number of articles to be included in this study. A sample size of 100 was chosen so that with 95% confidence we would be able to quantify the true percentage of articles that reported the data components necessary to calculate sample sizes for future studies within 10% (see Appendix C).

3. Results

3.1 *Study Selection*

A total of 1120 unique articles were retrieved from the initial search strategy. Among these, a random sample of 1100 articles was screened for eligibility until we reached the target sample size. Therefore, a target number of 100 eligible articles were included for final review and analysis (see Figure 1 for a flow diagram). Based on a random sample of 50 articles, prevalence-adjusted bias-adjusted inter-rater agreement (PABAK) on evaluating articles for eligibility was excellent, with PABAK=1. The reference list of the 100 eligible articles is included in Appendix D.

3.2 *Study Characteristics*

Among the included 100 eligible articles published in six leading journals in 2010 and 2011, about 60% of studies came from the journal *Neuroimage*. The major study design was cross-sectional (94%). The funding source was reported in 78% of the citations. The funding came mostly from two or more different sources (77%) rather than completely from industry (1%). The numbers of articles published were 53% in 2010 and 47% in 2011. The median total number of subjects was 34 (Q1=26, Q3=48) ranging from 8 to 126, and most studies (79%) had a sample size of no more than 50 (see Table 2).

3.3 *Sample Size Calculations*

Out of the 100 articles, 1% (95% CI: 0.03% to 5.45%) reported sample size and power calculations, 8% (95% CI: 3.52% - 15.16%) reported size of effect, 4% (95% CI: 1.10% -

9.93%) reported between-subject variance, and 1% (95% CI: 0.03% to 5.45%) reported within-subject variance in the results section. None (95% CI: 0 to 3.62%) reported a temporal autocorrelation variance-covariance matrix. Three articles reported Cohen's *d* or *F* effect sizes. The majority (83%, 95% CI: 74.18% to 89.77%) reported peak or average *t*, *z* or *F* statistic (see Table 3).

The inter-rater agreement on evaluating the percentage reporting the peak or average *t*-, *z*-, or *F*- statistic, how often the sample size calculations were reported, the percentage reporting within-subject variance, and the temporal autocorrelation variance-covariance matrix was excellent when adjusted by prevalence and bias, with PABAK=1 individually. Agreement on evaluating the reporting of the size of effect, between-subject variance, and the Cohen's *d* or *f* effect size was very good, excellent, and excellent (PABAK=0.88, 0.96, and 0.96, respectively) (Table 4).

4. Discussion

Articles published in six leading journals rarely reported sample size calculations. Our result agrees with the observation in a recent study (Muller et al., 2007; Carp, 2012b). Poor reporting of observed effect sizes, between- and within- subject variances, and a temporal autocorrelation structure in the results section poses serious obstacles to implementing power analyses for future studies.

None of the included articles reported temporal covariance structure, a required component which can increase the accuracy of sample size estimates in the method proposed by Mumford and Nichols (2008). To overcome the difficulty that the number of

parameters to characterize the autocorrelation can be quite large, Mumford and Nichols propose a covariance model with three parameters (i.e., the first-order autoregressive correlation, ρ; autoregressive variance, σ^2_{AR}; and white noise variance, σ^2_{WN}). The three parameters are estimated by averaging over all voxels in a whole-brain or within a region of interest (ROI) and can be calculated by most fMRI software. One limitation with the method developed by Mumford and Nichols is that they have not discussed consequences of mis-specifying the autocorrelation structure.

In practice, temporal autocorrelation may decay faster or slower than given by the first order autoregressive, AR (1), correlation structure. The AR (1) structure is in fact a special case of a damped exponential structure (Munoz et al., 1992), and so the damped exponential is a more flexible choice. Proper specification of the correlation structure is important for accurate inference; a study by Muller et al. (2007) found that a mis-specified correlation structure can lead to misleading effect sizes. We therefore suggest that authors model the data with different covariance structures, choose the one with the best fit and report relevant parameters of that covariance structure rather than choosing an AR(1) + WN by default.

Eight articles reported effect sizes. In contrast, 83 articles reported *t*-, *z*- or *F*-statistics. Using the reported *t*-, *z*- or *F*- statistics will usually be insufficient to calculate sample sizes for future studies, as future studies often differ in design characteristics (e.g., different stimulus presentation, inter-stimulus interval, trials per run, numbers and length of runs etc.) and statistical modeling within and between subjects. These differences will

lead to different within- and between-subject variances and thus to different *t*-, *z*- or *F*-statistics. Additionally, t-, z- or F-statistics can be particularly susceptible to bias, notably the use of peak *t*-, *z*- or *F*- statistics in statistically significant regions, and ROIs that are defined in a "non-independence" fashion meaning that choices of ROI depend on the data that are also used for power analysis (Kriegeskorte et al., 2009; Vul et al., 2009). Thus reported t-, z- or F-statistics are likely to be overestimates and, if used to power future studies, could lead to sample sizes that are too small.

Given how rarely effect sizes and variance components are reported in the literature, we suggest the following strategies when performing a power analysis: 1) conducting a sensitivity analysis by varying values of effect size and variances over their plausible ranges, and then using the cost efficiency approach to narrow the range (Guo et al., 2012); 2) using pilot data that have the same or similar study design as the future study, as pilot data have been reported to provide sensible estimates on effect size and variances particularly when limited data are available from previous studies to implement power calculations (Thabane et al., 2010). Meanwhile, caution should be taken when using pilot data so as to calculate well-powered sample sizes for future studies (Kraemer et al., 2006); 3) If pilot data are not available, one alternative is to use open data, which are made free and public in web-based databases such as the open fMRI project (http://www.openfmri.org) and the open fMRI database (http://fmridc.org/f/fmridc), open science framework (http://opensciencefframework.org), brain map database (http://www.brainmap.org), and the neuroscience informatics tools and resources clearing

house (http://www.nitrc.org); 4) To avoid a biased selection of ROIs, one alternative would be to define ROIs anatomically (e.g., using anatomical atlas) (Poldrack and Mumford, 2009) and which are functionally specific (Friston et al., 1999; Friston et al., 2000), from published similar studies, or from independent data such as the Brain Map database; 5) using meta-analyses: as many fMRI studies are small-sized and conducted in a single site, meta-analysis is known to help increase precision of results and to generalize findings from individual studies (Cohn and Becker, 2003). However, if a large portion of included studies in meta-analysis are underpowered and biased (e.g., publication bias or selective reporting bias), the meta-analysis will be biased as well. This indicates that there is room for improvement in designing well-powered individual studies and reducing reporting bias.

Furthermore, some editorial policies would make it easier for investigators to carry out valid sample size calculations in the future. Firstly, editors and reviewers may consider requiring authors to report observed effect sizes and variance components in the results section, which would enable future studies to be more appropriately powered. These details can be put in supplementary online documents if publication space is limited. Similar to trial registration which improved reporting quality of RCTs (Reveiz et al., 2010), registering fMRI studies with power calculations and making study protocols publicly available prior to study completion may more likely reduce reporting bias, ensure reporting transparency and thus improve the overall reporting quality. Secondly, requiring manuscripts to report results from both statistically significant and non-significant regions

would reduce the risk of bias from reporting only statistically significant regions (Maxwell, 2004). One practical option is to use brain mapping (Jernigan et al., 2003). The other alternative is to report mean effect sizes and variance from the regions discovered from previous studies and related to the testing hypothesis. Thirdly, publishing well-powered studies with negative results (Whitley and Ball, 2002) would reduce publication bias and thus increase the reliability of published effect sizes.

In terms of reporting standards generally, the STROBE statement (von Elm et al., 2007) has been designed to guide reporting of observational epidemiological research. It has been demonstrated that journals that adopted a standard reporting guideline (i.e., the CONSORT statement) had better quality of reporting than those that did not (Moher et al., 2001; Plint et al., 2006), suggesting that use of standard guidelines in the fMRI field would yield better quality of reporting as well. As none of the six journals included in this review have adopted standard guidelines so far, we encourage journal editors to consider endorsing the STROBE statement for reporting observational fMRI studies, especially in terms of how the study size was arrived at, and statistical methods used to control for confounding and reporting the patient data flow diagram and sampling strategy (von Elm et al., 2007). They can also consider endorsing reporting guidelines proposed by Poldrack et al. (2008), particularly in describing experimental methods. Realizing that the standard guidelines cannot cover all necessary details for all studies (Carp, 2012b) and knowing the risk of sample size manipulation to obtain desired results (Schulz and Grimes, 2005),

we suggest requiring investigators to justify choices of effect sizes and variances used in their sample size calculations.

This study had several limitations. First, the articles we reviewed were selected from six leading journals with high impact factors. We suspect that our findings would be poorer in journals with a lower impact factor; that is, our results may underestimate the true percentage of not reporting effect sizes, between- and within-subject variances in general publications. Second, it is possible that the sample size calculations were conducted but simply not reported. However, evidence suggests that the absence of sample size calculations is mainly due to not having performed the calculations rather than simply omitting the reporting (Moher et al., 1994).

5. Conclusion

Our study finds that reporting of observational fMRI studies involving clinical human participants published in six leading journals seldom includes sample size calculations. Moreover, omission of the estimates for sizes of effect, between- and within-subject variances, and temporal autocorrelation matrix limits investigators' ability to calculate sample sizes for new studies. We suggest that, at a minimum, routine reporting of these four quantities in the results section is necessary, and strategies of reducing bias associated with the reported results should be considered, including using pilot data, sharing data in web-based warehouse, defining regions of interest independently, and registering study and public availability of protocol prior to study completion. Once these

practices become the norm, well-powered and reliable sample size calculations can be implemented for future studies.

Acknowledgments

We thank Joshua Carp and two anonymous reviewers for useful comments on the earlier version of the manuscript. This work was supported by an Ontario Graduate Scholarship to QG and a Discovery Grant from the Natural Sciences and Engineering Research council to EP.

References

Altman, D.G., 1991. Practical Statistics for Medical Research, 1st ed. Chapman and Hall/CRC.

Byrt, T., 1996. How good is that agreement? Epidemiology 7, 561.

Byrt, T., Bishop, J., Carlin, J.B., 1993. Bias, prevalence and kappa. J. Clin. Epidemiol. 46, 423-429.

Carp, J., 2012a. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. Front. Neurosci. 6, 149.

Carp, J., 2012b. The secret lives of experiments: Methods reporting in the fMRI literature. Neuroimage 63, 289-300.

Cohn, L.D., Becker, B.J., 2003. How meta-analysis increases statistical power. Psychol. Methods 8, 243-253.

Desmond, J.E., Glover, G.H., 2002. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. J. Neurosci. Methods 118, 115-128.

Feinstein, A.R., Cicchetti, D.V., 1990. High agreement but low kappa: I. The problems of two paradoxes. J. Clin. Epidemiol. 43, 543-549.

Friston, K., Phillips, J., Chawla, D., Buchel, C., 2000. Nonlinear PCA: characterizing interactions between modes of brain activity. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 355, 135-146.

Friston, K., Phillips, J., Chawla, D., Buchel, C., 1999. Revealing interactions among brain systems with nonlinear PCA. Hum. Brain Mapp. 8, 92-97.

Glahn, D.C., Ragland, J.D., Abramoff, A., Barrett, J., Laird, A.R., Bearden, C.E., Velligan, D.I., 2005. Beyond hypofrontality: a quantitative meta-analysis of functional neuroimaging studies of working memory in schizophrenia. Hum. Brain Mapp. 25, 60-69.

Guo, Q., Hall, G., McKinnon, M., Thabane, L., Goeree, R., Pullenayegum, E., 2012. Setting sample size using cost efficiency in fMRI studies. Open Access Medical Statistics 2, 33-41.

Hayasaka, S., Peiffer, A.M., Hugenschmidt, C.E., Laurienti, P.J., 2007. Power and sample size calculation for neuroimaging studies by non-central random field theory. Neuroimage 37, 721-730.

Jernigan, T.L., Gamst, A.C., Fennema-Notestine, C., Ostergaard, A.L., 2003. More "mapping" in brain mapping: statistical comparison of effects. Hum. Brain Mapp. 19, 90-95.

Kraemer, H.C., Mintz, J., Noda, A., Tinklenberg, J., Yesavage, J.A., 2006. Caution regarding the use of pilot studies to guide power calculations for study proposals. Arch. Gen. Psychiatry 63, 484-489.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12, 535-540.

Lenth, R.V., 2001. Some Practical Guidelines for Effective Sample Size Determination. The American Statistician 55, 187-193.

Maxwell, S.E., 2004. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. Psychol. Methods 9, 147-163.

Moher, D., Dulberg, C.S., Wells, G.A., 1994. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 272, 122-124.

Moher, D., Jones, A., Lepage, L., CONSORT Grp, 2001. Use of the CONSORT statement and quality of reports of randomized trials - A comparative before-and-after evaluation. Jama-Journal of the American Medical Association 285, 1992-1995.

Monk, C.S., Klein, R.G., Telzer, E.H., Schroth, E.A., Mannuzza, S., Moulton, J.L.,3rd, Guardino, M., Masten, C.L., McClure-Tone, E.B., Fromm, S., Blair, R.J., Pine, D.S., Ernst, M., 2008. Amygdala and nucleus accumbens activation to emotional facial expressions in children and adolescents at risk for major depression. Am. J. Psychiatry 165, 90-98.

Muller, K.E., Edwards, L.J., Simpson, S.L., Taylor, D.J., 2007. Statistical tests with accurate size and power for balanced linear mixed models. Stat. Med. 26, 3639-3660.

Mumford, J.A., 2012. A power calculation guide for fMRI studies. Soc. Cogn. Affect. Neurosci. 7, 738-742.

Mumford, J.A., Nichols, T.E., 2008. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. Neuroimage 39, 261-268.

Munoz, A., Carey, V., Schouten, J.P., Segal, M., Rosner, B., 1992. A parametric family of correlation structures for the analysis of longitudinal data. Biometrics 48, 733-742.

Murphy, K., Garavan, H., 2004. An empirical investigation into the number of subjects required for an event-related fMRI study. Neuroimage 22, 879-885.

Plint, A.C., Moher, D., Morrison, A., Schulz, K., Altman, D.G., Hill, C., Gaboury, I., 2006. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. Med. J. Aust. 185, 263-267.

Poldrack, R.A., Mumford, J.A., 2009. Independence in ROI analysis: where is the voodoo? Soc. Cogn. Affect. Neurosci. 4, 208-213.

Reveiz, L., Cortes-Jofre, M., Asenjo Lobos, C., Nicita, G., Ciapponi, A., Garcia-Dieguez, M., Tellez, D., Delgado, M., Sola, I., Ospina, E., Iberoamerican Cochrane Network, 2010. Influence of trial registration on reporting quality of randomized trials: study from highest ranked journals. J. Clin. Epidemiol. 63, 1216-1222.

Schulz, K., Grimes, D., 2005. Epidemiology 1 - Sample size calculations in randomised trials: mandatory and mystical. Lancet 365, 1348-1353.

Sheline, Y.I., Barch, D.M., Donnelly, J.M., Ollinger, J.M., Snyder, A.Z., Mintun, M.A., 2001. Increased amygdala response to masked emotional faces in depressed subjects resolves with antidepressant treatment: an fMRI study. Biol. Psychiatry 50, 651-658.

Siegle, G.J., Steinhauer, S.R., Thase, M.E., Stenger, V.A., Carter, C.S., 2002. Can't shake that feeling: event-related fMRI assessment of sustained amygdala activity in

response to emotional information in depressed individuals. Biol. Psychiatry 51, 693-707.

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22, 1359-1366.

Smith, G.D., Ebrahim, S., 2002. Data dredging, bias, or confounding. BMJ 325, 1437-1438.

Snitz, B.E., MacDonald, A.,3rd, Cohen, J.D., Cho, R.Y., Becker, T., Carter, C.S., 2005. Lateral and medial hypofrontality in first-episode schizophrenia: functional activity in a medication-naive state and effects of short-term atypical antipsychotic treatment. Am. J. Psychiatry 162, 2322-2329.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L.P., Robson, R., Thabane, M., Giangregorio, L., Goldsmith, C.H., 2010. A tutorial on pilot studies: the what, why and how. BMC Med. Res. Methodol. 10, 1-2288-10-1.

von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gotzsche, P.C., Vandenbroucke, J.P., STROBE, I., 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann. Intern. Med. 147, 573-577.

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009. Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. Perspectives on Psychological Science 4, 274-290.

Whitley, E., Ball, J., 2002. Statistics review 4: sample size calculations. Crit. Care 6, 335-341.

Wilkinson, L., Task Force Stat Inference, 1999. Statistical methods in psychology journals - Guidelines and explanations. Am. Psychol. 54, 594-604.

Yarkoni, T., 2009. Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power-Commentary on Vul et al. (2009). Perspectives on Psychological Science 4, 294-298.

Yoon, J.H., Minzenberg, M.J., Ursu, S., Ryan Walter, B.S., Wendelken, C., Ragland, J.D., Carter, C.S., 2008. Association of dorsolateral prefrontal cortex dysfunction with disrupted coordinated brain activity in schizophrenia: relationship with impaired cognition, behavioral disorganization, and global function. Am. J. Psychiatry 165, 1006-1014.

Zaslavsky, B.G., 2010. Empirical Bayes models of Poisson clinical trials and sample size determination. Pharm. Stat. 9, 133-141.

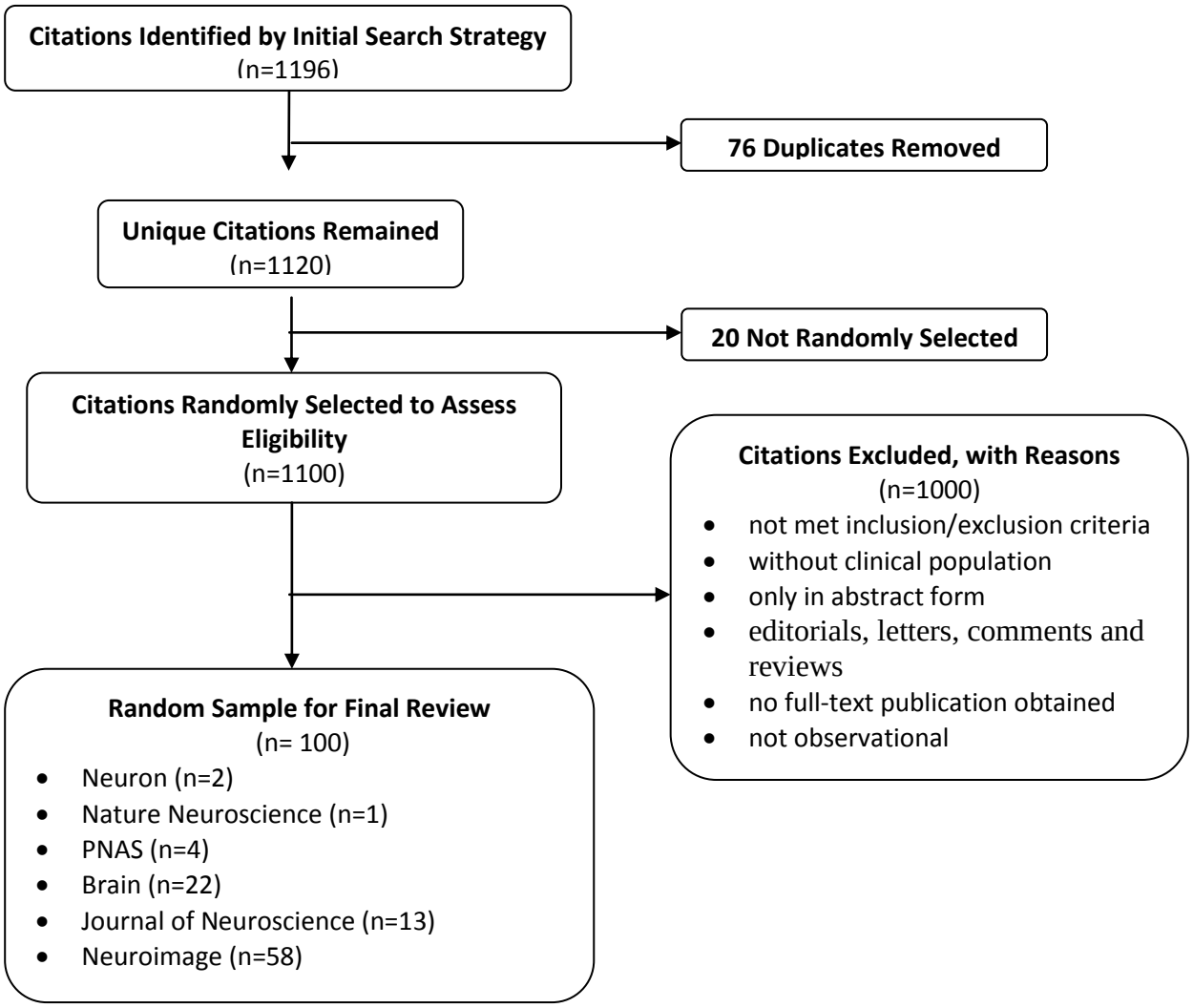

Figure1. Flow diagram of citation selection process.

Table 1. Sample Size Approaches and Parameters Required to Report

Approach	Parameters
Mumford & Nichols (*Neuroimage*. 2008; 39(1):261-268)	Δ, σ_w^2, σ_b^2, V
Desmond & Glover (*J.Neurosci.Methods*. 2002;118(2):115-128)	Δ, σ_w^2, σ_b^2
Hayasaka et al. (*Neuroimage*. 2007; 37(3): 721-730)	Cohen's *d* or *f* effect size

Note: Δ(size of effect or mean percent signal change between two conditions); σ_w^2 (within-subject variance); V (Temporal autocorrelation matrix); σ_b^2(between-subject variance);

Table 2. Characteristics of Included fMRI Studies (Information Extracted from Each Article)

Study Feature	All articles (n=100) Median (Q1, Q3) or %
Publication Journal	
Neuron	2
Nature Neuroscience	1
Proceedings of the National Academy of Sciences of the United States of America	4
Brain	22
Journal of Neuroscience	13
Neuroimage	58
Publication Year	
2010	53
2011	47
Study Design	
Case-control	0
Cohort	6
Cross-sectional	94
Number of Subjects	34 (26, 48)
Up to 10	2
10-50	77
51-100	17
More than 100	4
Funding Sources	
Completely funded by industry	1
Others	77
Not reported	22

Note: Q1=first quartile or 25th percentile, Q3=third quartile or 75th percentile.

Table 3. Parameters Reported in Results Section (Information Extracted from Each Article)

		All articles (n=100)	
Parameters	**Description**	**% reported**	**(95% CI)**
δ	Size of the effect or percent signal change between two conditions or groups	8	(3.52, 15.16)
σ_b	Between-subject variability	4	(1.10, 9.93)
σ_w	Within-subject variability	1	(0.03, 5.45)
COV_{AR} (σ_{ar}, ρ, σ_{wn})	Temporal autocorrelation variance-covariance matrix (Auto-regressive variability, AR(1) correlation coefficient, white-noise variability)	0	(0, 3.62)
t, z, or F	Peak or average *t*-, *z*-, or *F*- statistic	83	(74.18, 89.77)
d or f	Cohen's *d* or *f* effect size	3	(0.62, 8.52)

Table 4. Inter-rater Agreement on Evaluated Items

Item	**Observed Agreement (%)**	**κ (95% CI)**	**PABAK (95% CI)**
Eligibility criteria	100	-	1 (1, 1)
Reported Sample size Calculations	100	-	1 (1, 1)
δ	94	0.37 (-0.19, 0.93)	0.88 (0.82, 0.94)
σ_b	98	0 (0, 0)	0.96 (0.92, 1.00)
σ_w	100	-	1 (1, 1)
COV_{AR}	100	-	1 (1, 1)
t, z, or F statistics	100	1 (1, 1)	1 (1, 1)
d or f	98	0.66 (0.03, 1.00)	0.96 (0.92, 1.00)

Note: PABAK (prevalence-adjusted bias-adjusted κ).

Appendix A: Search strategy for Medline using OVID database (Ovid MEDLINE in-Process & Other Non-indexed Citations and Ovid MEDLINE 1948 – Present). This search was conducted on February 12, 2012*.

Step Number	Search Strategy
1	functional magnetic resonance imaging.mp.
2	fmri.mp.
3	1 or 2
4	limit 3 to (english language and humans and yr="2010-2011")
5	"neuron".jn.
6	"nature neuroscience".jn.
7	“proceedings of the national academy of science of the united states of america”.jn.
8	”brain”.jn.
9	”journal of neuroscience”.jn.
10	"neuroimage".jn.
11	4 and 5
12	4 and 6
13	4 and 7
14	4 and 8
15	4 and 9
16	4 and 10
17	11 or 12 or 13 or 14 or 15 or 16

** This search was limited to the years of 2010 and 2011, six journals, English language and humans

Appendix B: Data extraction forms.

1) Eligibility Form: The inclusion and exclusion criteria (information extracted from each article by initial search strategy). The coding for the article is entered in the column namely "Criteria-Met Statusu_Article#" †.

Eligibility Criteria	Item No	Item Description	Article#
Inclusion Criteria	Inc1	Involved clinical population	
Inclusion Criteria	Inc2	Study design: block or event-related	
Inclusion Criteria	Inc3	Observational study	
Inclusion Criteria	Inc4	Full reports	
Exclusion Criteria	Exc1	Only in abstract form	
Exclusion Criteria	Exc2	Editorials	
Exclusion Criteria	Exc3	Letters	
Exclusion Criteria	Exc4	Reviews or comments	
Exclusion Criteria	Exc5	Genetic studies	
Exclusion Criteria	Exc6	Resting-state studies	
Exclusion Criteria	Exc7	Studies of connectivity	
Exclusion Criteria	Exc8	Non-observational studies	

† Coding instruction: 1 – "Meet the criteria", 0 – "Don't meet the criteria" or "Unclear".

2) Study Characteristic Form: Information extracted from each eligible article. The coding for the article is entered in the column, namely "Article#"‡.

Characteristics of Included Articles	Article#
Publication Journal	
Publication Year	
Study Design	
Number of Subjects	
Funding Sources	

‡ Coding instruction: Publication Journal: 1 – "Neuron", 2 – "Nature Neuroscience", 3 – "Proceedings of the National Academy of Sciences of the United States of America", 4 – "Brain", 5 – "Journal of Neuroscience", 6 – "Neuroimage"; Publication Year: 0 – "2010", 1 – "2011"; Study Design: 1 – "Case-control", 2 – "Cohort", 3 – "Cross-sectional"; Number of Subjects: 1 – "Up to 10", 2 – "10-50", 3 – "51-100", 4 – "More than 100"; Funding Sources: 1 – "Completely funded by industry", 2 – "Others", 3 – "Not reported".

3) Primary Outcome Form: Parameters required for sample size and power calculation reported in results section (information extracted from each eligible article). The coding for the article is entered in the column, namely "Article#"*.

Parameters	Description	Article#
δ	Size of the effect or mean activation (or percent signal change) between two conditions or groups	
COV_{AR} (σ_{ar}, ρ, σ_{wn})	Temporal autocorrelation variance-covariance matrix (autoregressive variability, AR(1) correlation coefficient, white-noise variability)	
σ_b	Between-subject variability	
σ_w	Within-subject variability	
d or f	Cohen's *d* or *f* effect size	
t, z, or F	Peak or average *t-,z-, or F* statistic	

* Coding instruction: 1 – "Reported", 0 – "Not reported" or "Unclear".

Appendix C. Sample Size Calculation for Estimating a Single Proportion with a Desired Width of Confidence Interval

The primary outcome of this paper is the proportion of reviewed articles that reported parameter estimates necessary for future sample size determination. Since no previous similar studies have provided the estimates of the proportion, we determined sample sizes by varying the proportion estimates and margin of error (MOE) over its plausible ranges (See the results below) through sensitivity analysis. The mathematical formula for sample size calculation with an expected estimate and precision is as follows:

The estimated range of p is $p \pm Z_{1-\frac{\alpha}{2}}\sqrt{\frac{p\times(1-p)}{n}}$, MOE= $Z_{1-\frac{\alpha}{2}}\sqrt{\frac{p\times(1-p)}{n}}$, then

$$n = \frac{Z^2_{1-\frac{\alpha}{2}} \times p(1-p)}{MOE^2}$$

As shown in the table below, we noticed that the sample size of 96 could achieve any estimate of proportion of reporting at a MOE of 10% and also could reach its extreme estimates of proportion with a value less than 5% or greater than 95% and with an MOE of 5% at a 95% confidence level. We therefore chose a sample size of 100 by rounding up from 96.

Table C.1 Sample Size Calculations by Varying Estimated Proportion and Margin of Error

Estimated % of reporting (*p*)	**Margin of Error (MOE)**	
	5%	**10%**
5%	73	18
10%	138	35
15%	196	49
20%	246	61
25%	288	72
30%	323	81
35%	350	87
40%	369	92
45%	380	95
50%	384	96
55%	380	95
60%	369	92
65%	350	87
70%	323	81
75%	288	72
80%	246	61
85%	196	49
90%	138	35
95%	73	18

Appendix D. Reference of 100 Eligible Articles

Agam, Y., Joseph, R.M., Barton, J.J., Manoach, D.S., 2010. Reduced cognitive control of response inhibition by the anterior cingulate cortex in autism spectrum disorders. Neuroimage 52, 336-347.

Agosta, F., Henry, R.G., Migliaccio, R., Neuhaus, J., Miller, B.L., Dronkers, N.F., Brambati, S.M., Filippi, M., Ogar, J.M., Wilson, S.M., Gorno-Tempini, M.L., 2010. Language networks in semantic dementia. Brain 133, 286-299.

Allen, P., Stephan, K.E., Mechelli, A., Day, F., Ward, N., Dalton, J., Williams, S.C., McGuire, P., 2010. Cingulate activity and fronto-temporal connectivity in people with prodromal signs of psychosis. Neuroimage 49, 947-955.

Ances, B., Vaida, F., Ellis, R., Buxton, R., 2011. Test-retest stability of calibrated BOLD-fMRI in HIV- and HIV+ subjects. Neuroimage 54, 2156-2162.

Arichi, T., Moraux, A., Melendez, A., Doria, V., Groppo, M., Merchant, N., Combs, S., Burdet, E., Larkman, D.J., Counsell, S.J., Beckmann, C.F., Edwards, A.D., 2010. Somatosensory cortical activation identified by functional MRI in preterm and term infants. Neuroimage 49, 2063-2071.

Bach, S., Brandeis, D., Hofstetter, C., Martin, E., Richardson, U., Brem, S., 2010. Early emergence of deviant frontal fMRI activity for phonological processes in poor beginning readers. Neuroimage 53, 682-693.

Bai, X., Vestal, M., Berman, R., Negishi, M., Spann, M., Vega, C., Desalvo, M., Novotny, E.J., Constable, R.T., Blumenfeld, H., 2010. Dynamic time course of typical childhood absence seizures: EEG, behavior, and functional magnetic resonance imaging. J. Neurosci. 30, 5884-5893.

Bardin, J.C., Fins, J.J., Katz, D.I., Hersh, J., Heier, L.A., Tabelow, K., Dyke, J.P., Ballon, D.J., Schiff, N.D., Voss, H.U., 2011. Dissociations between behavioural and functional magnetic resonance imaging-based evaluations of cognitive function after brain injury. Brain 134, 769-782.

Beisteiner, R., Robinson, S., Wurnig, M., Hilbert, M., Merksa, K., Rath, J., Hollinger, I., Klinger, N., Marosi, C., Trattnig, S., Geissler, A., 2011. Clinical fMRI: evidence for a 7T benefit over 3T. Neuroimage 57, 1015-1021.

Belleville, S., Clement, F., Mellah, S., Gilbert, B., Fontaine, F., Gauthier, S., 2011. Training-related brain plasticity in subjects at risk of developing Alzheimer's disease. Brain 134, 1623-1634.

Bird, G., Silani, G., Brindley, R., White, S., Frith, U., Singer, T., 2010. Empathic brain responses in insula are modulated by levels of alexithymia but not autism. Brain 133, 1515-1525.

Blau, V., Reithler, J., van Atteveldt, N., Seitz, J., Gerretsen, P., Goebel, R., Blomert, L., 2010. Deviant processing of letters and speech sounds as proximate cause of reading failure: a functional magnetic resonance imaging study of dyslexic children. Brain 133, 868-879.

Bonelli, S.B., Powell, R.H., Yogarajah, M., Samson, R.S., Symms, M.R., Thompson, P.J., Koepp, M.J., Duncan, J.S., 2010. Imaging memory in temporal lobe epilepsy: predicting the effects of temporal lobe resection. Brain 133, 1186-1199.

Brune, M., Ozgurdal, S., Ansorge, N., von Reventlow, H.G., Peters, S., Nicolas, V., Tegenthoff, M., Juckel, G., Lissek, S., 2011. An fMRI study of "theory of mind" in at-risk states of psychosis: comparison with manifest schizophrenia and healthy controls. Neuroimage 55, 329-337.

Campbell-Sills, L., Simmons, A.N., Lovero, K.L., Rochlin, A.A., Paulus, M.P., Stein, M.B., 2011. Functioning of neural systems supporting emotion regulation in anxiety-prone individuals. Neuroimage 54, 689-696.

Cantin, S., Villien, M., Moreaud, O., Tropres, I., Keignart, S., Chipon, E., Le Bas, J.F., Warnking, J., Krainik, A., 2011. Impaired cerebral vasoreactivity to CO2 in Alzheimer's disease using BOLD fMRI. Neuroimage 58, 579-587.

Celone, K.A., Thompson-Brenner, H., Ross, R.S., Pratt, E.M., Stern, C.E., 2011. An fMRI investigation of the fronto-striatal learning system in women who exhibit eating disorder behaviors. Neuroimage 56, 1749-1757.

Chang, Y., Lee, J.J., Seo, J.H., Song, H.J., Kim, J.H., Bae, S.J., Ahn, J.H., Park, S.J., Jeong, K.S., Kwon, Y.J., Kim, S.H., Kim, Y., 2010. Altered working memory process in the manganese-exposed brain. Neuroimage 53, 1279-1285.

Chua, H.F., Ho, S.S., Jasinska, A.J., Polk, T.A., Welsh, R.C., Liberzon, I., Strecher, V.J., 2011. Self-related neural response to tailored smoking-cessation messages predicts quitting. Nat. Neurosci. 14, 426-427.

Coman, I.L., Gnirke, M.H., Middleton, F.A., Antshel, K.M., Fremont, W., Higgins, A.M., Shprintzen, R.J., Kates, W.R., 2010. The effects of gender and catechol O-methyltransferase (COMT) Val108/158Met polymorphism on emotion regulation in velo-cardio-facial syndrome (22q11.2 deletion syndrome): An fMRI study. Neuroimage 53, 1043-1050.

Daunizeau, J., Vaudano, A.E., Lemieux, L., 2010. Bayesian multi-modal model comparison: a case study on the generators of the spike and the wave in generalized spike-wave complexes. Neuroimage 49, 656-667.

de Guibert, C., Maumet, C., Jannin, P., Ferre, J.C., Treguier, C., Barillot, C., Le Rumeur, E., Allaire, C., Biraben, A., 2011. Abnormal functional lateralization and activity of language brain areas in typical specific language impairment (developmental dysphasia). Brain 134, 3044-3058.

Dima, D., Dietrich, D.E., Dillo, W., Emrich, H.M., 2010. Impaired top-down processes in schizophrenia: a DCM study of ERPs. Neuroimage 52, 824-832.

Dinstein, I., Thomas, C., Humphreys, K., Minshew, N., Behrmann, M., Heeger, D.J., 2010. Normal movement selectivity in autism. Neuron 66, 461-469.

Dziobek, I., Preissler, S., Grozdanovic, Z., Heuser, I., Heekeren, H.R., Roepke, S., 2011. Neuronal correlates of altered empathy and social cognition in borderline personality disorder. Neuroimage 57, 539-548.

Ellmore, T.M., Beauchamp, M.S., Breier, J.I., Slater, J.D., Kalamangalam, G.P., O'Neill, T.J., Disano, M.A., Tandon, N., 2010. Temporal lobe white matter asymmetry and language laterality in epilepsy patients. Neuroimage 49, 2033-2044.

Emmorey, K., Xu, J., Gannon, P., Goldin-Meadow, S., Braun, A., 2010. CNS activation and regional connectivity during pantomime observation: no engagement of the mirror neuron system for deaf signers. Neuroimage 49, 994-1005.

Erk, S., Mikschl, A., Stier, S., Ciaramidaro, A., Gapp, V., Weber, B., Walter, H., 2010. Acute and sustained effects of cognitive emotion regulation in major depression. J. Neurosci. 30, 15726-15734.

Freund, P., Weiskopf, N., Ward, N.S., Hutton, C., Gall, A., Ciccarelli, O., Craggs, M., Friston, K., Thompson, A.J., 2011. Disability, atrophy and cortical reorganization following spinal cord injury. Brain 134, 1610-1622.

Germine, L.T., Garrido, L., Bruce, L., Hooker, C., 2011. Social anhedonia is associated with neural abnormalities during face emotion processing. Neuroimage 58, 935-945.

Golarai, G., Hong, S., Haas, B.W., Galaburda, A.M., Mills, D.L., Bellugi, U., Grill-Spector, K., Reiss, A.L., 2010. The fusiform face area is enlarged in Williams syndrome. J. Neurosci. 30, 6700-6712.

Goulden, N., McKie, S., Suckling, J., Williams, S.R., Anderson, I.M., Deakin, J.F., Elliott, R., 2010. A comparison of permutation and parametric testing for between group effective connectivity differences using DCM. Neuroimage 50, 509-515.

Gradin, V.B., Kumar, P., Waiter, G., Ahearn, T., Stickle, C., Milders, M., Reid, I., Hall, J., Steele, J.D., 2011. Expected value and prediction error abnormalities in depression and schizophrenia. Brain 134, 1751-1764.

Grefkes, C., Nowak, D.A., Wang, L.E., Dafotakis, M., Eickhoff, S.B., Fink, G.R., 2010. Modulating cortical connectivity in stroke patients by rTMS assessed with fMRI and dynamic causal modeling. Neuroimage 50, 233-242.

Greimel, E., Schulte-Ruther, M., Kircher, T., Kamp-Becker, I., Remschmidt, H., Fink, G.R., Herpertz-Dahlmann, B., Konrad, K., 2010. Neural mechanisms of empathy in adolescents with autism spectrum disorder and their fathers. Neuroimage 49, 1055-1065.

Halko, M.A., Datta, A., Plow, E.B., Scaturro, J., Bikson, M., Merabet, L.B., 2011. Neuroplastic changes following rehabilitative training correlate with regional electrical field induced with tDCS. Neuroimage 57, 885-891.

Hu, W., Lee, H.L., Zhang, Q., Liu, T., Geng, L.B., Seghier, M.L., Shakeshaft, C., Twomey, T., Green, D.W., Yang, Y.M., Price, C.J., 2010. Developmental dyslexia in Chinese and English populations: dissociating the effect of dyslexia from language differences. Brain 133, 1694-1706.

Ibarretxe-Bilbao, N., Zarei, M., Junque, C., Marti, M.J., Segura, B., Vendrell, P., Valldeoriola, F., Bargallo, N., Tolosa, E., 2011. Dysfunctions of cerebral networks precede recognition memory deficits in early Parkinson's disease. Neuroimage 57, 589-597.

Jiang, Z., Krainik, A., David, O., Salon, C., Tropres, I., Hoffmann, D., Pannetier, N., Barbier, E.L., Bombin, E.R., Warnking, J., Pasteris, C., Chabardes, S., Berger, F., Grand, S., Segebarth, C., Gay, E., Le Bas, J.F., 2010. Impaired fMRI activation in patients with primary brain tumors. Neuroimage 52, 538-548.

Jollant, F., Lawrence, N.S., Olie, E., O'Daly, O., Malafosse, A., Courtet, P., Phillips, M.L., 2010. Decreased activation of lateral orbitofrontal cortex during risky choices under uncertainty is associated with disadvantageous decision-making and suicidal behavior. Neuroimage 51, 1275-1281.

Jones, T.B., Bandettini, P.A., Kenworthy, L., Case, L.K., Milleville, S.C., Martin, A., Birn, R.M., 2010. Sources of group differences in functional connectivity: an investigation applied to autism spectrum disorder. Neuroimage 49, 401-414.

Kaiser, M.D., Hudac, C.M., Shultz, S., Lee, S.M., Cheung, C., Berken, A.M., Deen, B., Pitskel, N.B., Sugrue, D.R., Voos, A.C., Saulnier, C.A., Ventola, P., Wolf, J.M., Klin, A., Vander Wyk, B.C., Pelphrey, K.A., 2010. Neural signatures of autism. Proc. Natl. Acad. Sci. U. S. A. 107, 21223-21228.

Kareken, D.A., Bragulat, V., Dzemidzic, M., Cox, C., Talavage, T., Davidson, D., O'Connor, S.J., 2010. Family history of alcoholism mediates the frontal response to alcoholic drink odors and alcohol in at-risk drinkers. Neuroimage 50, 267-276.

Khursheed, F., Tandon, N., Tertel, K., Pieters, T.A., Disano, M.A., Ellmore, T.M., 2011. Frequency-specific electrocorticographic correlates of working memory delay period fMRI activity. Neuroimage 56, 1773-1782.

Killory, B.D., Bai, X., Negishi, M., Vega, C., Spann, M.N., Vestal, M., Guo, J., Berman, R., Danielson, N., Trejo, J., Shisler, D., Novotny, E.J.,Jr, Constable, R.T., Blumenfeld, H., 2011. Impaired attention and network connectivity in childhood absence epilepsy. Neuroimage 56, 2209-2217.

Klinge, C., Eippert, F., Roder, B., Buchel, C., 2010. Corticocortical connections mediate primary visual cortex responses to auditory stimulation in the blind. J. Neurosci. 30, 12798-12805.

Klinge, C., Roder, B., Buchel, C., 2010. Increased amygdala activation to emotional auditory stimuli in the blind. Brain 133, 1729-1736.

Krawitz, A., Braver, T.S., Barch, D.M., Brown, J.W., 2011. Impaired error-likelihood prediction in medial prefrontal cortex in schizophrenia. Neuroimage 54, 1506-1517.

Kumari, V., Fannon, D., Peters, E.R., Ffytche, D.H., Sumich, A.L., Premkumar, P., Anilkumar, A.P., Andrew, C., Phillips, M.L., Williams, S.C., Kuipers, E., 2011. Neural changes following cognitive behaviour therapy for psychosis: a longitudinal study. Brain 134, 2396-2407.

Li, W., Howard, J.D., Gottfried, J.A., 2010. Disruption of odour quality coding in piriform cortex mediates olfactory deficits in Alzheimer's disease. Brain 133, 2714-2726.

Li, X., Mullen, K.T., Thompson, B., Hess, R.F., 2011. Effective connectivity anomalies in human amblyopia. Neuroimage 54, 505-516.

Lombardo, M.V., Chakrabarti, B., Bullmore, E.T., MRC AIMS, C., Baron-Cohen, S., 2011. Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. Neuroimage 56, 1832-1838.

Lombardo, M.V., Chakrabarti, B., Bullmore, E.T., Sadek, S.A., Pasco, G., Wheelwright, S.J., Suckling, J., MRC AIMS, C., Baron-Cohen, S., 2010. Atypical neural self-representation in autism. Brain 133, 611-624.

Longe, O., Maratos, F.A., Gilbert, P., Evans, G., Volker, F., Rockliff, H., Rippon, G., 2010. Having a word with yourself: neural correlates of self-criticism and self-reassurance. Neuroimage 49, 1849-1856.

Lu, J., Liu, H., Zhang, M., Wang, D., Cao, Y., Ma, Q., Rong, D., Wang, X., Buckner, R.L., Li, K., 2011. Focal pontine lesions provide evidence that intrinsic functional connectivity reflects polysynaptic anatomical pathways. J. Neurosci. 31, 15065-15071.

Lueken, U., Kruschwitz, J.D., Muehlhan, M., Siegert, J., Hoyer, J., Wittchen, H.U., 2011. How specific is specific phobia? Different neural response patterns in two subtypes of specific phobia. Neuroimage 56, 363-372.

Luijten, M., Veltman, D.J., van den Brink, W., Hester, R., Field, M., Smits, M., Franken, I.H., 2011. Neurobiological substrate of smoking-related attentional bias. Neuroimage 54, 2374-2381.

Lundell, H., Christensen, M.S., Barthelemy, D., Willerslev-Olsen, M., Biering-Sorensen, F., Nielsen, J.B., 2011. Cerebral activation is correlated to regional atrophy of the spinal cord and functional motor disability in spinal cord injured individuals. Neuroimage 54, 1254-1261.

Ma, Y., Han, S., 2011. Neural representation of self-concept in sighted and congenitally blind adults. Brain 134, 235-246.

MacDonald, P.A., MacDonald, A.A., Seergobin, K.N., Tamjeedi, R., Ganjavi, H., Provost, J.S., Monchi, O., 2011. The effect of dopamine therapy on ventral and dorsal striatum-mediated cognition in Parkinson's disease: support from functional MRI. Brain 134, 1447-1463.

Maurer, U., Schulz, E., Brem, S., der Mark, S., Bucher, K., Martin, E., Brandeis, D., 2011. The development of print tuning in children with dyslexia: evidence from longitudinal ERP data supported by fMRI. Neuroimage 57, 714-722.

Meulenbroek, O., Rijpkema, M., Kessels, R.P., Rikkert, M.G., Fernandez, G., 2010. Autobiographical memory retrieval in patients with Alzheimer's disease. Neuroimage 53, 331-340.

Mizuno, A., Liu, Y., Williams, D.L., Keller, T.A., Minshew, N.J., Just, M.A., 2011. The neural basis of deictic shifting in linguistic perspective-taking in high-functioning autism. Brain 134, 2422-2435.

Mohr, H.M., Roder, C., Zimmermann, J., Hummel, D., Negele, A., Grabhorn, R., 2011. Body image distortions in bulimia nervosa: investigating body size overestimation and body size satisfaction by fMRI. Neuroimage 56, 1822-1831.

Mourao-Miranda, J., Hardoon, D.R., Hahn, T., Marquand, A.F., Williams, S.C., Shawe-Taylor, J., Brammer, M., 2011. Patient classification as an outlier detection problem: an application of the One-Class Support Vector Machine. Neuroimage 58, 793-804.

Nestor, L., McCabe, E., Jones, J., Clancy, L., Garavan, H., 2011. Differences in "bottom-up" and "top-down" neural activity in current and former cigarette smokers: Evidence for neural substrates which may promote nicotine abstinence through increased cognitive control. Neuroimage 56, 2258-2275.

Newman, A.J., Supalla, T., Hauser, P.C., Newport, E.L., Bavelier, D., 2010. Prosodic and narrative processing in American Sign Language: an fMRI study. Neuroimage 52, 669-676.

Oertel, V., Knochel, C., Rotarska-Jagiela, A., Schonmeyer, R., Lindner, M., van de Ven, V., Haenschel, C., Uhlhaas, P., Maurer, K., Linden, D.E., 2010. Reduced laterality as a trait marker of schizophrenia--evidence from structural and functional neuroimaging. J. Neurosci. 30, 2289-2299.

Papagni, S.A., Mechelli, A., Prata, D.P., Kambeitz, J., Fu, C.H., Picchioni, M., Walshe, M., Toulopoulou, T., Bramon, E., Murray, R.M., Collier, D.A., Bellomo, A., McGuire, P., 2011. Differential effects of DAAO on regional activation and functional connectivity in schizophrenia, bipolar disorder and controls. Neuroimage 56, 2283-2291.

Pompei, F., Jogia, J., Tatarelli, R., Girardi, P., Rubia, K., Kumari, V., Frangou, S., 2011. Familial and disease specific abnormalities in the neural correlates of the Stroop Task in Bipolar Disorder. Neuroimage 56, 1677-1684.

Radua, J., Phillips, M.L., Russell, T., Lawrence, N., Marshall, N., Kalidindi, S., El-Hage, W., McDonald, C., Giampietro, V., Brammer, M.J., David, A.S., Surguladze, S.A.,

2010. Neural response to specific components of fearful faces in healthy and schizophrenic adults. Neuroimage 49, 939-946.

Roberts, G.M., Garavan, H., 2010. Evidence of increased activation underlying cognitive control in ecstasy and cannabis users. Neuroimage 52, 429-435.

Roussotte, F.F., Bramen, J.E., Nunez, S.C., Quandt, L.C., Smith, L., O'Connor, M.J., Bookheimer, S.Y., Sowell, E.R., 2011. Abnormal brain activation during working memory in children with prenatal exposure to drugs of abuse: the effects of methamphetamine, alcohol, and polydrug exposure. Neuroimage 54, 3067-3075.

Rytsar, R., Fornari, E., Frackowiak, R.S., Ghika, J.A., Knyazeva, M.G., 2011. Inhibition in early Alzheimer's disease: an fMRI-based study of effective connectivity. Neuroimage 57, 1131-1139.

S, C., Villien, M., Moreaud, O., Tropres, I., Keignart, S., Chipon, E., Le Bas, J.F., Warnking, J., Krainik, A., 2011. Impaired cerebral vasoreactivity to CO2 in Alzheimer's disease using BOLD fMRI. Neuroimage 58, 579-587.

Saur, D., Ronneberger, O., Kummerer, D., Mader, I., Weiller, C., Kloppel, S., 2010. Early functional magnetic resonance imaging activations predict language outcome after stroke. Brain 133, 1252-1264.

Schacht, J.P., Anton, R.F., Randall, P.K., Li, X., Henderson, S., Myrick, H., 2011. Stability of fMRI striatal response to alcohol cues: a hierarchical linear modeling approach. Neuroimage 56, 61-68.

Schonberg, T., O'Doherty, J.P., Joel, D., Inzelberg, R., Segev, Y., Daw, N.D., 2010. Selective impairment of prediction error signaling in human dorsolateral but not ventral striatum in Parkinson's disease patients: evidence from a model-based fMRI study. Neuroimage 49, 772-781.

Schulte, T., Muller-Oehring, E.M., Rohlfing, T., Pfefferbaum, A., Sullivan, E.V., 2010. White matter fiber degradation attenuates hemispheric asymmetry when integrating visuomotor information. J. Neurosci. 30, 12168-12178.

Schweckendiek, J., Klucken, T., Merz, C.J., Tabbert, K., Walter, B., Ambach, W., Vaitl, D., Stark, R., 2011. Weaving the (neuronal) web: fear learning in spider phobia. Neuroimage 54, 681-688.

Sepede, G., Ferretti, A., Perrucci, M.G., Gambi, F., Di Donato, F., Nuccetelli, F., Del Gratta, C., Tartaro, A., Salerno, R.M., Ferro, F.M., Romani, G.L., 2010. Altered

brain response without behavioral attention deficits in healthy siblings of schizophrenic patients: an event-related fMRI study. Neuroimage 49, 1080-1090.

Sharp, D.J., Beckmann, C.F., Greenwood, R., Kinnunen, K.M., Bonnelle, V., De Boissezon, X., Powell, J.H., Counsell, S.J., Patel, M.C., Leech, R., 2011. Default mode network functional and structural connectivity after traumatic brain injury. Brain 134, 2233-2247.

Shilyansky, C., Karlsgodt, K.H., Cummings, D.M., Sidiropoulou, K., Hardt, M., James, A.S., Ehninger, D., Bearden, C.E., Poirazi, P., Jentsch, J.D., Cannon, T.D., Levine, M.S., Silva, A.J., 2010. Neurofibromin regulates corticostriatal inhibitory networks during working memory performance. Proc. Natl. Acad. Sci. U. S. A. 107, 13141-13146.

Stice, E., Yokum, S., Blum, K., Bohon, C., 2010. Weight gain is associated with reduced striatal response to palatable food. J. Neurosci. 30, 13105-13109.

Stice, E., Yokum, S., Burger, K.S., Epstein, L.H., Small, D.M., 2011. Youth at risk for obesity show greater activation of striatal and somatosensory regions to food. J. Neurosci. 31, 4360-4366.

Tadic, S.D., Griffiths, D., Murrin, A., Schaefer, W., Aizenstein, H.J., Resnick, N.M., 2010. Brain activity during bladder filling is related to white matter structural changes in older women with urinary incontinence. Neuroimage 51, 1294-1302.

Toyomura, A., Fujii, T., Kuriki, S., 2011. Effect of external auditory pacing on the neural activity of stuttering speakers. Neuroimage 57, 1507-1516.

Tu, P., Buckner, R.L., Zollei, L., Dyckman, K.A., Goff, D.C., Manoach, D.S., 2010. Reduced functional connectivity in a right-hemisphere network for volitional ocular motor control in schizophrenia. Brain 133, 625-637.

Umarova, R.M., Saur, D., Kaller, C.P., Vry, M.S., Glauche, V., Mader, I., Hennig, J., Weiller, C., 2011. Acute visual neglect and extinction: distinct functional state of the visuospatial attention system. Brain 134, 3310-3325.

van der Mark, S., Klaver, P., Bucher, K., Maurer, U., Schulz, E., Brem, S., Martin, E., Brandeis, D., 2011. The left occipitotemporal system in reading: disruption of focal fMRI connectivity to left inferior frontal and inferior parietal language areas in children with dyslexia. Neuroimage 54, 2426-2436.

van Oers, C.A., Vink, M., van Zandvoort, M.J., van der Worp, H.B., de Haan, E.H., Kappelle, L.J., Ramsey, N.F., Dijkhuizen, R.M., 2010. Contribution of the left and

right inferior frontal gyrus in recovery from aphasia. A functional MRI study in stroke patients with preserved hemodynamic responsiveness. Neuroimage 49, 885-893.

Voon, V., Gao, J., Brezing, C., Symmonds, M., Ekanayake, V., Fernandez, H., Dolan, R.J., Hallett, M., 2011. Dopamine agonists and risk: impulse control disorders in Parkinson's disease. Brain 134, 1438-1446.

Wagner, D.D., Dal Cin, S., Sargent, J.D., Kelley, W.M., Heatherton, T.F., 2011. Spontaneous action representation in smokers when watching movie characters smoke. J. Neurosci. 31, 894-898.

Wang, L., Metzak, P.D., Honer, W.G., Woodward, T.S., 2010. Impaired efficiency of functional networks underlying episodic memory-for-context in schizophrenia. J. Neurosci. 30, 13171-13179.

Wang, W.C., Lazzara, M.M., Ranganath, C., Knight, R.T., Yonelinas, A.P., 2010. The medial temporal lobe supports conceptual implicit memory. Neuron 68, 835-842.

Wildenberg, J.C., Tyler, M.E., Danilov, Y.P., Kaczmarek, K.A., Meyerand, M.E., 2011. High-resolution fMRI detects neuromodulation of individual brainstem nuclei by electrical tongue stimulation in balance-impaired individuals. Neuroimage 56, 2129-2137.

Wu, T., Chan, P., Hallett, M., 2010. Effective connectivity of neural networks in automatic movements in Parkinson's disease. Neuroimage 49, 2581-2587.

Wu, T., Wang, L., Hallett, M., Li, K., Chan, P., 2010. Neural correlates of bimanual anti-phase and in-phase movements in Parkinson's disease. Brain 133, 2394-2409.

Yassa, M.A., Stark, S.M., Bakker, A., Albert, M.S., Gallagher, M., Stark, C.E., 2010. High-resolution structural and functional MRI of hippocampal CA3 and dentate gyrus in patients with amnestic Mild Cognitive Impairment. Neuroimage 51, 1242-1252.

Zempleni, M.Z., Michels, L., Mehnert, U., Schurch, B., Kollias, S., 2010. Cortical substrate of bladder control in SCI and the effect of peripheral pudendal stimulation. Neuroimage 49, 2983-2994.

CHAPTER THREE

Setting Sample Size Using Cost Efficiency in fMRI Studies

Qing Guo[1,3], Geoffrey Hall[2], Margaret McKinnon[2,4,7], Lehana Thabane[1,3,5], Ron Goeree[1,5,6], Eleanor Pullenayegum[1,3,5]

[1]Department of Epidemiology and Biostatistics, [2]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, ON, Canada; [3]Biostatistics Unit, [4]Mood Disorders Program, [5]Centre for Evaluation of Medicine, [6]Programs for Assessment of Technology in Health (PATH) Research Institute, St Joseph's Healthcare Hamilton, Hamilton, ON, Canada; St. Joseph's Healthcare, Hamilton, ON, Canada; [7]Kunin-Lunenfeld Applied Research Unit, Baycrest, Toronto, ON, Canada

Correspondence:

Qing Guo
Department of Epidemiology and Biostatistics
Faculty of Health Sciences
McMaster University, Hamilton, ON, Canada
Email: guoq@mcmaster.ca

This article is cited as:

Reproduced from Guo Q, Hall G, McKinnon M, Thabane L, Goeree R, Pullenayegum E. Setting sample size using cost efficiency in fMRI studies. Open Access Medical Statistics 2012:2 33-41.

ABSTRACT

Background

Sample size calculations are rarely performed for functional magnetic resonance imaging studies involving clinical populations. This may be due to uncertainty as to the size of expected effect and the variance of the blood oxygenation level dependent response. Moreover, existing sample size methods ignore the costs associated with performing the proposed study. The current paper describes how cost efficiency, a recently proposed method, can be used in conjunction with existing methods to address these issues.

Methods

Cost efficiency is the ratio of a study's value to its cost, and sample size is chosen to maximize cost efficiency (i.e., to maximize return on investment). It is suggested that sample size calculations begin by calculating the sample sizes required to achieve a given power, through varying the input parameters to the calculation over their plausible ranges. Cost efficiency can then help narrow the resulting range of sample sizes and help choose one sample size. The approach is illustrated through a recent functional magnetic resonance imaging study of autobiographical memory retrieval in patients with major depressive disorder.

An example

Setting power to 80% and type 1 error rate to 5%, the method of Mumford and Nichols was used to calculate sample size. There were no reported effect sizes for similar studies in the literature; consequently, this parameter was varied over its plausible range (Cohen's

d varying from 0.2 to 0.8). This yielded sample sizes ranging from 50 to 800. Within these, cost efficiency gave a sample size of 88.

Conclusions

Poor reporting of the input parameters to power-based methods of sample size determination results in a wide range of candidate sample sizes. The cost efficiency approach supplies a way of narrowing this range and choosing a sample size from that.

Keywords: cost efficiency, sample size, power, fMRI studies

Introduction

Functional magnetic resonance imaging (fMRI) has been widely used to examine patterns of neural activation at rest and while performing motor and cognitive tasks. Despite the widespread use of fMRI technology, sample size calculations for fMRI studies have proved challenging. fMRI studies commonly focus on the effect of a stimulus or effects of different stimuli on the blood-oxygenation-level-dependent (BOLD) response of neural regions of interest, often contrasting stimulus-specific (eg, high versus low working memory load) and group effects (eg, patients versus controls). Data are collected sequentially during a scanning session, and at each time point the BOLD response is measured for each voxel in the brain. Given that the change in BOLD in response to a stimulus will vary across brain regions and time, analyzing fMRI data is more complicated than the simple comparisons of means that are required for many clinical studies. Here, the correlation in responses over time and space and the multiplicity inherent in fMRI data make sample size determination difficult.[1]

Despite these challenges, there has been progress in sample size estimation in fMRI studies.[2–5] However, these calculations are subject to the same limitations as conventional sample size calculations. The input parameters (eg, minimal clinically important difference, standard deviation) are often unknown. The common approach to uncertainty in input parameters is a sensitivity analysis in which the parameter values are varied over a plausible range, thus producing a range of candidate sample sizes. The difficulty with this approach in fMRI studies is that the innovative nature of the field, coupled with limited reporting, often leads to a high degree of uncertainty as to the values of the input

parameters, resulting in a very large range of sample sizes. Thus, sensitivity analysis can be of limited use when planning studies in which a budget needs to be determined a priori.

A further limitation of existing methods is that cost, which investigators cannot ignore in practice, has no role in conventional sample size approaches. Two methods consider cost in sample size estimation, yet neither is currently used in fMRI studies. One is value of information and the other is cost efficiency. Value of information chooses the sample size to maximize the expected value of information gained through the trial minus the expected cost incurred. Value of information methods are most widely used with randomized controlled trials that are conducted to inform a decision on whether to adopt a new intervention (eg, a treatment or diagnostic tool) into routine practice. Here, study value is often measured in terms of the quality-adjusted life years gained by society as a result of the information gathered in the study. For example, the study might show that there are serious side effects associated with the new intervention, leading to a decision to retain the standard treatment, thus saving quality-adjusted life years by avoiding the introduction of a potentially harmful intervention. In fMRI studies, quantifying the expected study value (ie, quality-adjusted life years saved) is difficult since fMRI is often used at an early stage of discovery when it is unclear how the information will be used in improving patient care. Cost efficiency, recently proposed by Bacchetti et al,[6] aims to maximize the value-per-unit cost and can be implemented without quantifying the projected study value. While the concept of cost efficiency is not widely used in health

research, maximizing expected return on investment is a criterion most people consider when investing their own money.

This paper discusses the potential use of cost efficiency for setting sample size in fMRI studies. The mathematical basis for the cost efficiency method is explained, and its use is illustrated through an fMRI study. The potential role of cost efficiency in fMRI studies is discussed and some conclusions are offered.

Methods

The cost efficiency approach suggested by Bacchetti et al focuses on the ratio of the value of information to the cost, thereby aiming to maximize return on investment.[6] Let v_n be the expected scientific, clinical, or practical value of the study if the sample size is n, and let c_n be the corresponding cost of the study. Cost efficiency chooses n to maximize the ratio of v_n to c_n (ie, to maximize v_n/c_n).

Cost in this context is measured from the perspective of the investigator and thus includes all financial expenditures: the fixed cost to set up and administer the study, perform data analysis, and disseminate the findings, as well as the cost per patient to cover scanning and travel costs. The study value v_n is measured in terms of the information gained from the study results. As discussed above, study value is most easily quantified in studies that will be used to inform a decision as to whether to adopt a new intervention. In contrast, fMRI studies tend to be used at an earlier stage of discovery, and the information gained could ultimately benefit patients in many possible ways. Although this information is certainly valuable to society, it is difficult to quantify its value. The

advantage of the cost efficiency approach is that study value does not need to be quantified.

The key concept in cost efficiency is that the study value v_n can often be replaced by a simple stand-in function of sample size. It has been shown that if a function $f(n)$ can be found such that $v_n/f(n)$ does not increase as n increases, then choosing the sample size n to minimize $c_n/f(n)$ provides more cost efficiency (ie, a higher ratio of v_n/c_n) than any larger choice of n.

Two widely applicable choices of $f(n)$ have been proposed. The first option is $f(n) = n$ with a resulting sample size $n_{\min}$, which minimizes total study cost divided by sample size (average cost per subject). Using $f(n) = n$ requires that v_n/n be nonincreasing in n, and this condition is denoted as $C_{\min}$. $C_{\min}$ has been shown to hold under a wide range of definitions of study value, such as value proportional to study power,[6,7] inversely proportional to confidence interval width, proportional to reduction in Bayesian credible interval width from its prior width,[8–10] proportional to the reduction in squared error loss versus using the prior mean, and proportional to gain in Shannon information.[11] Thus, Bacchetti et al conclude that it is reasonable to use $n_{\min}$ without verifying condition $C_{\min}$ for each specific study.[6] The second choice is to take $f(n) = \sqrt{n}$ so that n_{root} minimizes total study cost divided by the square root of the sample size. When $f(n) = \sqrt{n}$, it is required that $v_n/\sqrt{n}$ be nonincreasing in n, and this condition is denoted as C_{root}. The condition C_{root} is more stringent than Cmin and generally holds for all $n > 4$ when there is low prior information (in the sense that the Bayesian priors have equal prior means for the

two groups, and at least one prior having a standard deviation at least as large as its mean). Bacchetti et al also suggest using n_{root}, with no need to verify C_{root} for each specific study provided that there is low prior information.[6]

Bacchetti et al argue that the sample sizes $n_{\min}$ and n_{root} are more cost-efficient than any larger sample size calculated by another sample size determination method, and therefore cannot be considered inadequate, regardless of the power they achieve.[6] In contrast, the current paper suggests using cost efficiency alongside conventional power calculations. Studies with low power may not detect the effect of interest. Since studies without statistically significant results are less likely to be published than studies with significant findings, investigators may feel driven to produce significant results. When power on the primary comparison is low, this encourages data dredging with a danger of spuriously significant findings.[12–15] Thus, it is suggested that in fMRI studies, cost efficiency considerations be used in conjunction with power-based methods.

Specifically, it is proposed that investigators begin with sample size calculations based on achieving a given statistical power. Due to the uncertainty of input parameters, a range of candidate sample sizes are calculated through a sensitivity analysis. Cost efficiency can then be used to choose a sample size from this range. This approach is now illustrated through an example.

An example

The example is a 2-year neuroimaging study of autobiographical memory retrieval in patients with major depressive disorder. This study examined patterns of neural activation

during autobiographical memory retrieval of negative events compared to positive and neutral events in patients with recurrent major depressive disorder and matched healthy controls. It was hypothesized that the difference in activation of the hippocampus for negative events compared to positive and neutral events would be greater for patients than controls. In a pre-scan interview, participants identified six events from the past 2 years: two highly positive events (eg, a birthday party), two highly negative events (eg, receiving bad news), and two comparatively neutral events (eg, swimming). During the scanning session, over a 20-second period, participants were presented with event period titles describing one of the events identified in the interview and asked to recall the event, or an incomplete sentence to be completed. This was followed by 10 seconds during which patients rated their degree of autobiographical re-experiencing and a 5-second fixation. The stimuli was presented in a six-block format in a random order, with two runs per event type and each run containing five stimuli corresponding to one of two memories generated in the pre-scan interview. The total scan time, including set up and localization, was about 1 hour.

How sample size can be determined will now be illustrated. To begin with, the sample size method of Mumford and Nichols was used,[5] ie, using sensitivity analysis to deal with input parameters whose values are uncertain. Cost efficiency was then applied to help narrow the range of sample sizes.

Mumford and Nichols' sample size calculation requires estimates of within- and between- subject variances, and the size of the effect to be detected.[5] No information about these estimates was reported in prior studies in this population, and so for the

purposes of illustration the values presented by Mumford and Nichols were used: a first-order autoregressive correlation of 0.73, a first-order autoregressive total variance of 0.98, white noise variance of 1.313, and between-subject variance of 0.421.

There was substantial uncertainty about the likely size of the difference between patients and controls in the relative activation of the hippocampus for positive versus neutral events. Values of Cohen's *d* effect size ranging from 0.2 (small) to 0.8 (large) were plausible, and this parameter was varied in sensitivity analyses. Using values of 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8 yielded sample sizes of 786, 350, 198, 126, 88, 66, and 51, respectively, to achieve 80% power at a significance level of 0.05. The type 1 error rate was not adjusted for multiple comparisons because the primary region of interest was restricted a priori to the hippocampus. The relationships between power and sample size for different effect sizes are displayed in Figure 1.

Since these calculations yielded sample sizes ranging from 50 to 800, additional criteria were needed to select an appropriate sample size. Cost efficiency can help with this. In this study, the cost components included fixed costs such as part-time research assistant's salary, which covered the responsibilities of scheduling, subject recruitment, data entry, data analyses, conference travel fees, and test administration (in total, $39,671). The cost per patient included participant reimbursement fees to accommodate study participants' parking, traveling and testing time ($50 per person), and scanner time ($400 per person). Therefore the total cost was $39,671 plus $450 per patient. Since this was the first neuroimaging study of autobiographical memory retrieval in patients with major depressive disorder, n_{root} was a reasonable choice. It is easy to show (Appendix 1)

that n_{root} is equal to fixed costs divided by variable costs (ie, $39,671 is divided by $450), which in the current study came to 88 patients (Figure 2).

Here it is demonstrated step-by-step how n_{root} helps to choose a cost-efficient sample size. As studies with low power will have small sample sizes, yielding results with a wide confidence interval (ie, low precision), using the reduction from the width of confidence interval is a measure of study value. Therefore, the study value was defined as inversely proportional to confidence interval width in this demonstration. The parameter of interest was the difference between patients and controls in the mean activation of the hippocampus for negative events compared to positive and neutral events (ie, the emotional valence by group interaction). It can be observed in Figure 3 that study value divided by the square root of the sample size was monotone, decreasing as sample size increased. Figure 4 illustrates that cost divided by the square root of the sample size was minimized at the sample size of 88, and as can be seen from Figure 5, cost efficiency at the sample size of 88 was bigger than any larger choice among the range of sample sizes.

Study value was chosen to be inversely proportional to confidence interval width to illustrate how the cost efficiency method works in practice. Bacchetti et al have shown that similar results hold when using other definitions of study value.[6]

Discussion

Conventional methods of sample size calculation in fMRI studies involving clinical populations are limited by substantial uncertainty in the values of input parameters.

Varying these parameters over their plausible ranges can result in a very large range of candidate sample sizes (in the current example, the range was 50 to 800). It is argued that cost efficiency can help choose a sample size from this range. In the current example, a sample size of 88 was more cost-efficient than any larger choice. Thus, cost efficiency provides an upper bound to the sample size that the investigator should consider.

The current approach to using cost efficiency differs from Bacchetti et al's approach. Arguing that the standard choice of 80% for statistical power is arbitrary, Bacchetti et al proposes cost efficiency as a stand-alone method for choosing sample size.[6] The current authors agree that cost efficiency is a reasonable criterion: although an unfamiliar concept in health research, most individuals aim to maximize their cost efficiency (return on investment) when investing their own money. This difference of opinion with Bacchetti et al comes from the belief that the need for significant results to achieve publication is likely stronger in the fMRI literature than in other areas of medicine. This, coupled with the enormous scope for multiple testing in fMRI studies, makes the danger of false positive findings in underpowered fMRI studies particularly acute. It can thus be argued that studies with very low power may in fact have negative value. Thus, it is suggested that cost efficiency considerations be used in conjunction with power-based methods.

The suggested approach of using cost efficiency as a supplement to traditional power-based methods can overcome the potential conflict between a sample size calculated from the traditional method and the cost efficiency sample size. While larger sample sizes will always have greater power, they cost more as well. Consequently, there is a trade-off between the study power and cost. Cost efficiency provides a way of compromising

between statistical power and cost. Since the authors propose choosing the initial range of sample sizes by using traditional power calculation, and applying the cost efficiency criterion in order to choose a sample size among this range, the current approach will have at least 80% power to detect at least one of the proposed effect sizes.

There are some practical limitations to cost efficiency. First, only financial cost is considered. Societal costs such as inconvenience and risks to study participants are not included. Moreover, in the face of limited resources, funding one study means that another study is not funded; the costs used in cost efficiency calculations do not include this opportunity cost. Second, financial costs must be estimated accurately. This is a particular concern when writing grant applications for agencies that routinely cut budgets, since it is common for investigators to pad their budget against the cut. This will distort cost efficiency calculations. Third, it is possible that the study's projected value is less than the cost incurred. When this happens, the study does not add additional value[16] and should not be undertaken. The cost efficiency approach does not consider this. Fourth, whilst n and $\sqrt{n}$ are chosen as two widely applicable choices of $f(n)$, the authors do not claim that they are optimal, but rather that they are useful because their properties have been evaluated extensively through theory and simulations. It is therefore possible that some other forms of $f(n)$ may exist and do a better job in terms of identifying the sample size with the optimal cost efficiency. Future research may be helpful. Finally, cost efficiency may produce a sample size that is beyond the investigator's budget. This is a problem shared by other methods of sample size determination, in particular methods based on achieving a given power. In the current approach, budgetary constraints could be

incorporated by narrowing the range of sample sizes produced by sensitivity analysis down to those that are feasible.

In the current example, sample size calculations were based on hypothetical values for the variance components of the BOLD response. Thus, the results are for the purposes of illustration only. An alternative approach to this might be a Bayesian approach in which a prior is placed on these unknown parameters. Then the probability of rejecting the null hypothesis is calculated as the classical power averaged over the prior distribution.[17] In the current example, it was decided not to adopt the Bayesian approach; the results of such a power calculation are sensitive to the choice of prior, and the lack of reported values of the variance components in the literature makes it extremely difficult to place a realistic prior on them. However, if these variance components were to be more widely reported in the future, Bayesian sample size determination methods would become more viable.

This lack of reported variance components in the literature is a serious concern. If studies do not include these estimates in their results sections, there will remain substantial uncertainty as to their value, leaving investigators unable to power their studies accurately. Moreover, the uncertainty in variance parameters and effect sizes is not because the data to quantify them does not exist, but rather because it is simply not reported. There is a pressing need for a reporting guideline for fMRI studies outlining which values investigators should report to facilitate adequately powered studies in the future.

Conclusion

There is often substantial uncertainty in the effect sizes and variance components that form the input parameters to conventional sample size calculations. For any given study, this can lead to a wide range of sample size estimates. Authors should be encouraged to report effect sizes and estimates of within- and between-subject variances in their manuscripts to facilitate sample size calculations for future studies. Until this practice is widespread, the authors have argued that cost efficiency can supplement conventional sample size methods by narrowing the range of sample sizes under consideration on the basis of maximizing the expected return on investment.

Acknowledgments

Qing Guo was partly supported by funding from the Canadian Institute of Health Research (CIHR) training award and Ontario Graduate Scholarship (OGS).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Lazar N. *The Statistical Analysis of Functional MRI Data*. New York, NY: Springer; 2008.
2. Desmond JE, Glover GH. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *J Neurosci Methods*. 2002; 18(2):115–128.
3. Murphy K, Garavan H. An empirical investigation into the number of subjects required for an event-related fMRI study. *Neuroimage*. 2004;22(2):879–885.

4. Hayasaka S, Peiffer AM, Hugenschmidt CE, Laurienti PJ. Power and sample size calculation for neuroimaging studies by non-central random field theory. *Neuroimage*. 2007; 37(3):721–730.
5. Mumford JA, Nichols TE. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage*. 2008; 39(1):261–268.
6. Bacchetti P, McCulloch CE, Segal MR. Simple, defensible sample sizes based on cost efficiency. *Biometrics*. 2008; 64(2):577–594.
7. Bacchetti P, Wolf LE, Segal MR, McCulloch CE. Ethics and sample size. Am J Epidemiol. 2005; 161(2):105–110.
8. Joseph L, Belisle P. Bayesian sample size determination for normal means and differences between normal means. Statistician. 1997; 46(2):209–226.
9. Lindley DV. The choice of sample size. Statistician. 1997; 46(2): 129–138.
10. Pham-Gia T. On Bayesian analysis, Bayesian decision theory and the sample size problem. Statistician. 1997; 46(2):139–144.
11. Bernardo JM. Statistical inference as a decision problem: the choice of sample size. Statistician. 1997; 46(2):151–153.
12. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H Jr. Publication bias and clinical trials. Control Clin Trials. 1987; 8(4):343–353.
13. Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA. 1990;263(10):1385–1389.
14. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet. 1991;337(8746):867–872.
15. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124.
16. Briggs A, Sculpher M, Claxton K. Decision Modelling for Health Economic Evaluation. Oxford, UK: Oxford University Press; 2006.
17. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester, UK: John Wiley and Sons Ltd; 2004.

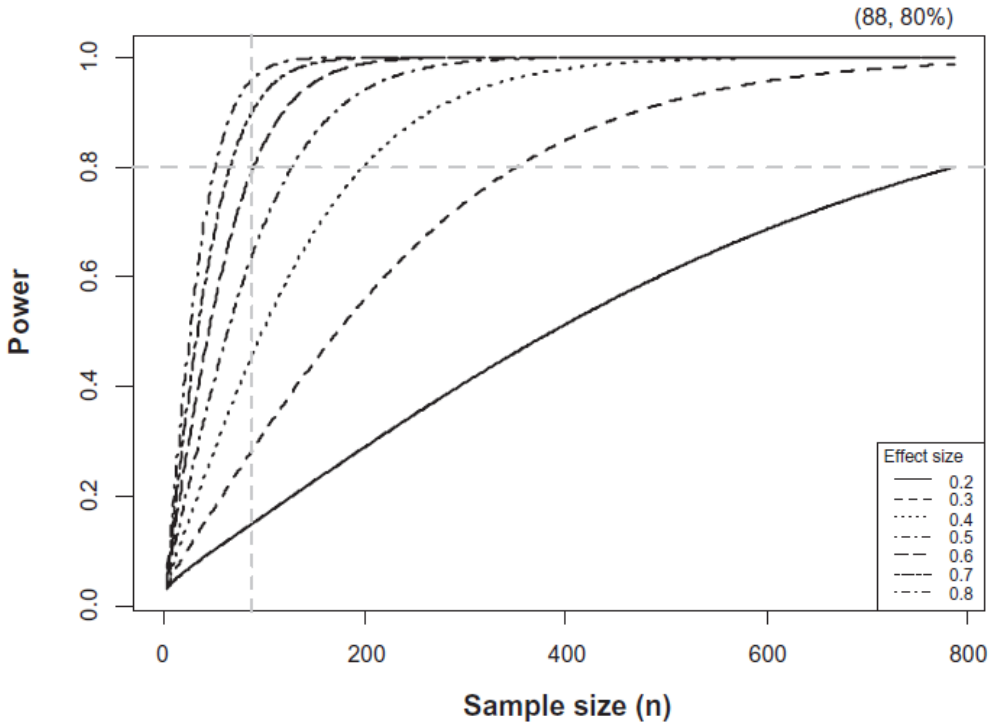

Figure 1 Power estimates for a block study for different sample sizes.
Notes: Each curve is for different group effect sizes (Cohen's *d*). The horizontal grey dotted line indicates 80% power and the vertical grey dotted line represents most cost-efficient sample size ($n_{root} = 88$).

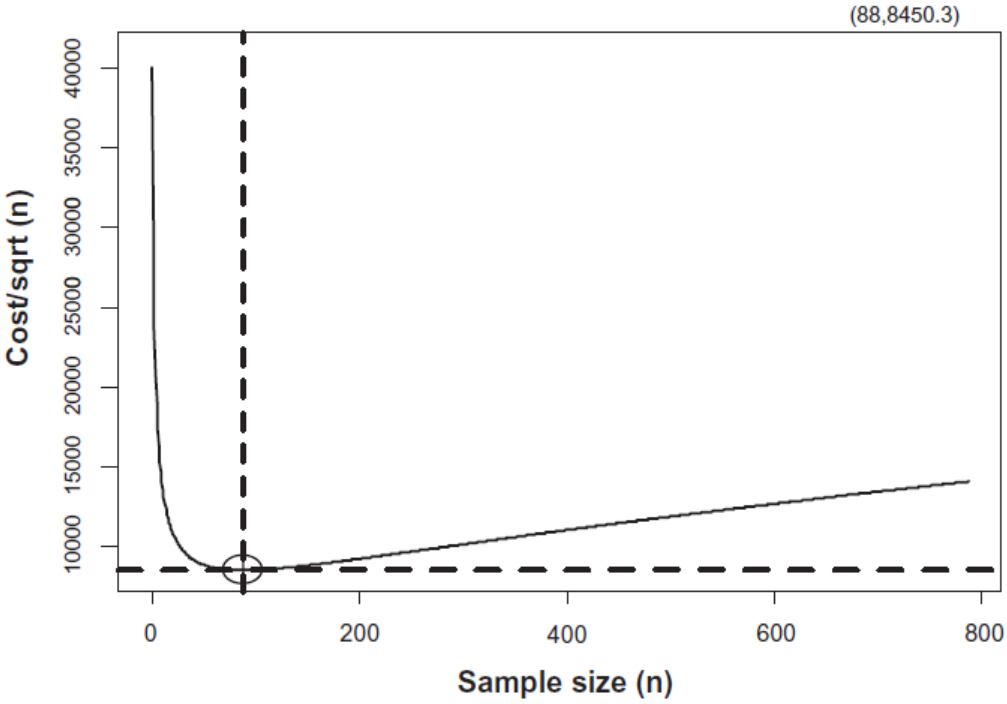

Figure 2 The relationship between the cost divided by the square root of the sample size and sample size for all assumed group effect sizes.

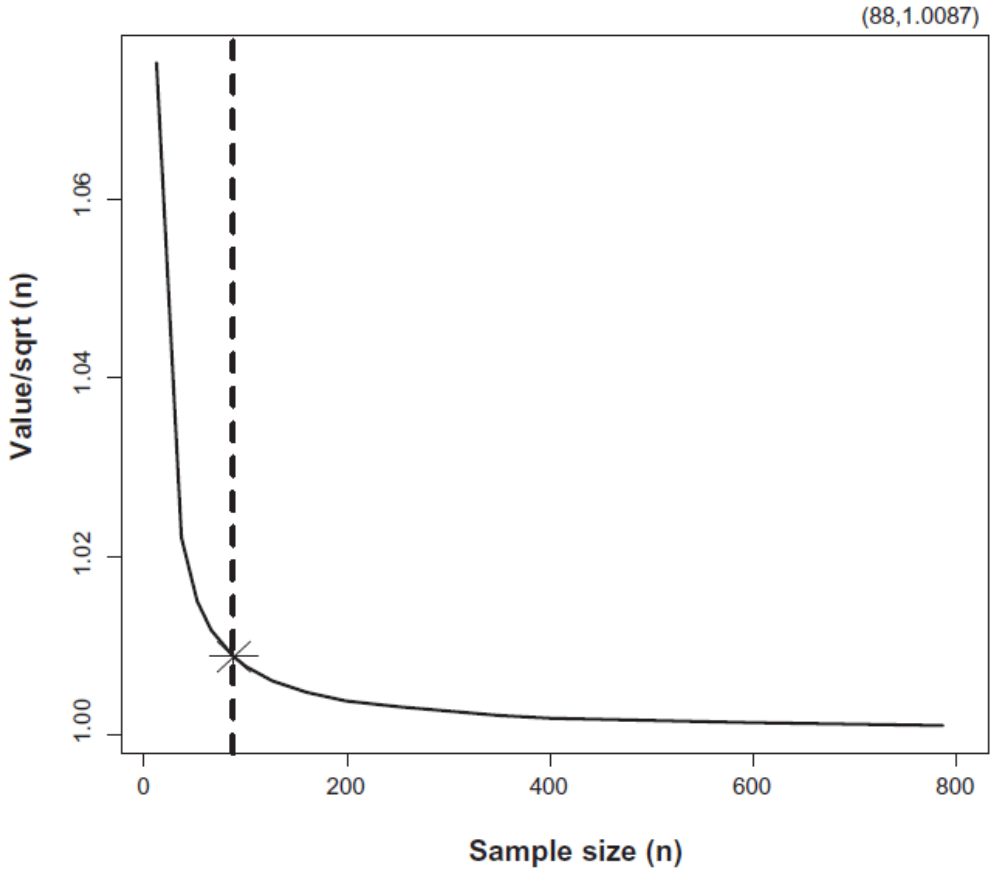

Figure 3 The relationship of the study value (which is inversely proportional to confidence interval width) divided by the square root of the sample size versus sample size.

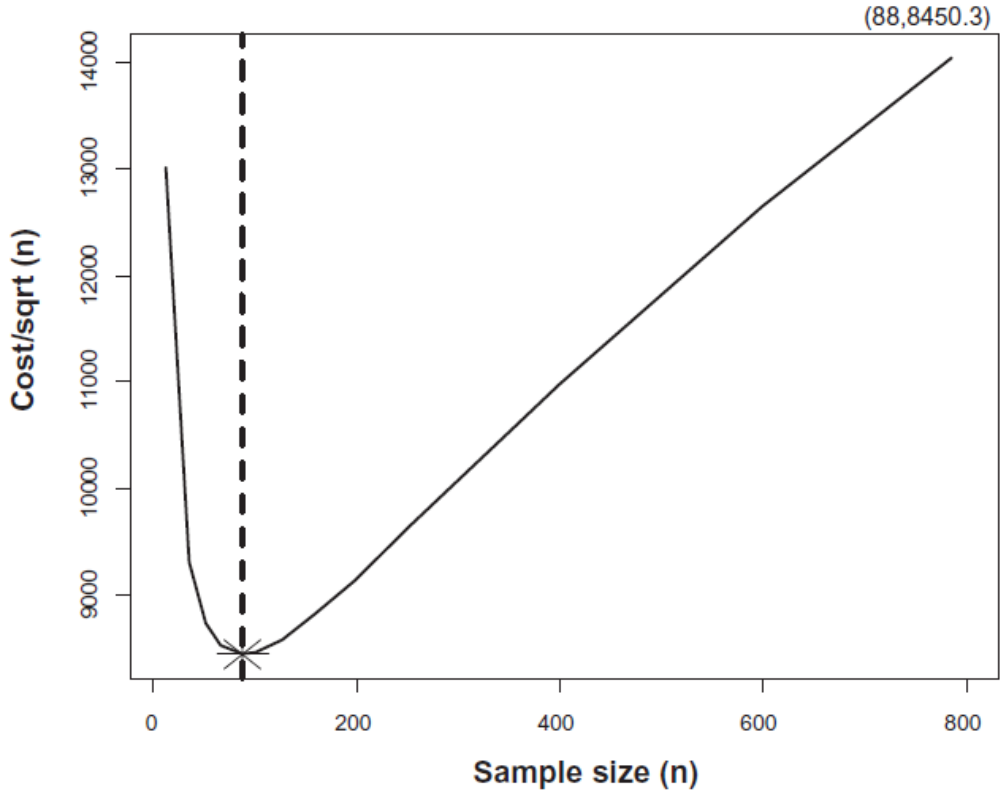

Figure 4 The relationship of cost divided by the square root of the sample size versus sample size.

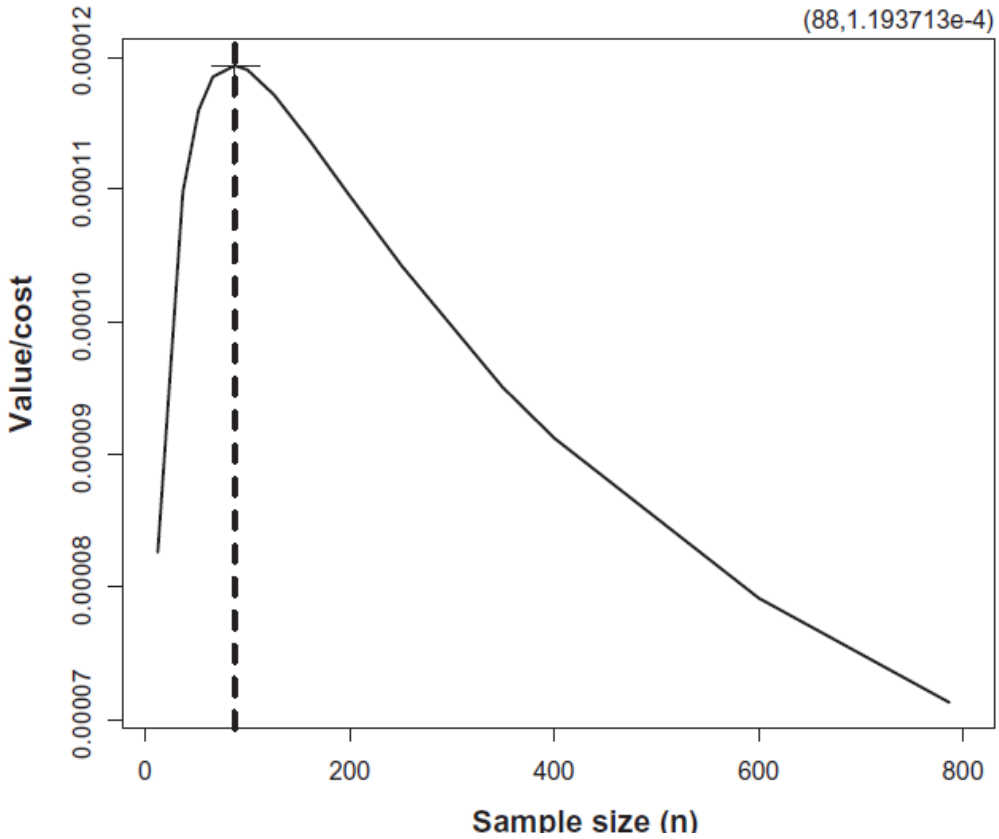

Figure 5 The relationship between cost efficiency and sample size.

Appendix I

If the study is composed of linear structure of a fixed cost plus a per-patient cost, that is, $C_n = C_f + C_s \times n$, where C_f, the fixed cost independent of n, is greater than zero and C_s, the cost per subject, is greater than zero, then n_{root} is the ratio of C_f to C_s, that is,

$n_{root} = C_f / C_s$.

Proof: If $\frac{\partial(\frac{C_n}{\sqrt{n}})}{\partial n} = \frac{C_s \times \sqrt{n} - (C_s \times n + C_f) \times \frac{1}{2} \times \frac{1}{\sqrt{n}}}{n} = 0,$ then $n = C_f / C_s$;

and $\frac{\partial^2(\frac{C_n}{\sqrt{n}})}{\partial n^2} = \frac{3 \times C_f - C_s \times n}{4 \times n^{\frac{5}{2}}} > 0$ at $n = C_f / C_s$. Therefore, $C_n / \sqrt{n}$ reaches the local minimum at $n = C_f / C_s$. Thus, $n_{root} = C_f / C_s$.

CHAPTER FOUR

The Reporting of Observational Clinical Functional Magnetic Resonance Imaging Studies: A Systematic Review

Qing Guo[1,2*], Melissa Parlar[9], Wanda Truong[4], Geoffrey Hall[3,6], Lehana Thabane[1,2,5], Margaret McKinnon[6,7,9], Ron Goeree[1,5,8], Eleanor Pullenayegum[1,2,5]

1 Department of Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada, **2** Biostatistics Unit, St Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada, **3** Department of Psychology, Neuroscience and Behaviour, McMaster University, Ontario, Canada, **4** Department of Psychology, University of Calgary, Calgary, Alberta, Canada, **5** Centre for Evaluation of Medicine, St Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada, **6** Mood Disorders Program, St. Joseph's Healthcare Hamilton, Ontario, Canada, **7** Kunin-Lunenfeld Applied Research Unit, Baycrest, Toronto, Ontario, Canada, **8** PATH Research Institute, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada, **9** Department of Psychiatry and Behavioural Neurosciences, McMaster University, Ontario, Canada

Corresponding author:

Qing Guo
Department of Epidemiology and Biostatistics
Faculty of Health Sciences
McMaster University, Hamilton, ON, Canada
Email: guoq@mcmaster.ca

This article is currently under revision for publication in *PLos ONE*.

ABSTRACT

Background

Complete reporting assists readers in confirming the methodological rigor and validity of findings and allows replication. The reporting quality of observational functional magnetic resonance imaging (fMRI) studies involving clinical participants is unclear. We sought to determine the quality of reporting in observational fMRI studies involving clinical participants.

Methods

We searched OVID MEDLINE for fMRI studies in six leading journals between January 2010 and December 2011.Three independent reviewers abstracted data from articles using an 83-item checklist adapted from the guidelines proposed by Poldrack et al. (Neuroimage 2008; 40: 409-14). We calculated the percentage of articles reporting each item of the checklist and the percentage of reported items per article.

Results

A random sample of 100 eligible articles was included in the study. Thirty-one items were reported by fewer than 50% of the articles and 13 items were reported by fewer than 20% of the articles. The median percentage of reported items per article was 51% (ranging from 30% to 78%). Although most articles reported statistical methods for within-subject modeling (92%) and for between-subject group modeling (97%), none of the articles reported observed effect sizes for any negative finding (0%). Few articles reported justifications for fixed-effect inferences used for group modeling (3%) and temporal

autocorrelations used to account for within-subject variances and correlations (18%). Other under-reported areas included whether and how the task design was optimized for efficiency (22%), distribution of inter-trial intervals (23%), and determination of scaling factor for percentage signal change (34%).

Conclusions

This study indicates that substantial improvement in the reporting of observational clinical fMRI studies is required. Poldrack's guidelines provide a means of improving overall reporting quality. Nonetheless, these guidelines are lengthy and at odds with strict word limits for publication; creation of a shortened-version of Poldrack's checklist may be useful in this regard.

*Keywords***:** Observational clinical fMRI; Reporting quality; Guidelines; Systematic review

Introduction

In the past decade, publication of functional MRI (fMRI) studies has increased a great deal. Given that fMRI is increasingly applied to the study of clinical disorders (e.g., [1-6]), and considering the vulnerability of clinical participants, there is an ethical imperative for scientists to apply rigorous methodology and provide adequate reporting. Rigorous methodology is required in order to uphold the promises typically made to participants during the consent process, namely that the study will help investigators to understand their conditions. Complete reporting with sufficient details permits readers to ensure the methodological rigor of a study [7], consider the validity of the methodology and findings [8-12], and extend and replicate the findings [7-11,13-15]. However, empirical evidence indicates that many publications lack key details, such as sample size calculations, whether temporal autocorrelations were modeled, descriptions of slice-timing and motion correction, slice order and coverage of functional brain images in their methods section [16].

There are some standard guidelines developed to aid authors on the reporting of scientific research, such as the Consolidated Standards for Reporting Trials (CONSORT) [8] and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) initiative [7]. Recently, Poldrack and his colleagues have proposed guidelines specifically for reporting fMRI studies [12]. Although many authors have suggested endorsing the guidelines proposed by Poldrack et al. in reporting fMRI studies to improve the quality, transparency and consistency of results [16-19], few systematic reviews have been conducted to appraise the quality of reporting based on these guidelines. Although a

study by Carp (2012) recently examined adherence to Poldrack et al.'s guidelines in randomly selected fMRI studies published since 2007, it included few studies involving clinical populations. Thus, the reporting quality in clinical fMRI studies remains unclear.

Moreover, as the majority of fMRI studies are observational (i.e., the type of study is not designed to randomize participants to test efficacy and safety of any therapeutic intervention), these studies are less scrutinized than randomized clinical trials with experimental interventions; for example, randomized trials have to be registered with clinicaltrials.gov. Therefore, we aimed to systematically evaluate the quality of reporting in observational fMRI studies involving clinical human participants (i.e., individuals who either have a disease or are at risk of developing a disease) using a checklist adapted from the guidelines proposed by Poldrack et al. In this study, we set out to address the following two questions: (1) what percentage of articles reported each item of the fMRI-specific guideline, and (2) what percentage of items was reported per article?

Methods:

Search Strategy and Eligible Journals

We searched OVID MEDLINE (1946 to January 2012) by using key word search terms (e.g., functional magnetic resonance imaging) combined with the acronym (e.g., fMRI) for articles published in 2010 and 2011, in the English language, and involving human participants. Compared with journals in general, top journals are cited more frequently (e.g., higher impact factors) and more scrutinized towards publication (e.g., lower manuscript acceptance rates). Furthermore, studies have indicated that high impact factor

and low manuscript acceptance rates of journals are associated with higher methodological rigor of articles published in the journals [20-24]. In this study, we further set our selection among six leading journals: In the *Journal Citation Report* 2010, we selected four journals with a high impact factor (IF) in the category "Neurosciences", namely, *Neuron* (IF 14.9), *Nature Neuroscience* (IF 14.2), *Brain* (IF 9.2), *Journal of Neuroscience* (IF 7.3), one journal with the highest impact factor in the category "Neuroimaging" (*NeuroImage,* IF 5.94), and one journal with a good proportion and high quality of publications in fMRI studies (*Proceedings of the National Academy of Sciences of the United States of America,* IF 9.8). More details on the search strategy can be found on Appendix A. Duplicate articles were removed.

Eligibility Criteria for Studies and Study Selection

We included articles that were peer-reviewed, full reports of observational fMRI studies involving human clinical participants, and block or event-related design for the fMRI paradigm. We excluded articles that were published only in abstract form or any that were only editorials, letters, comments or reviews. Genetic, resting-state observational fMRI studies, fMRI studies other than observational studies (e.g., randomized clinical trials), and studies of connectivity were also excluded.

We decided to include a target sample size of 100 articles in the review and analysis. The citations, which were identified with the search strategy, were reviewed randomly until 100 articles were selected by applying the eligibility criteria.

Data Extraction

We used an electronic data extraction form containing 83 items to assess the reporting of study articles (see Appendix B), which we piloted using a random selection of four studies. We deleted three items from Poldrack et al.'s original checklist as we found assessing them required too much subjectivity. The data were extracted from each article and any online supplements. Items were answered with "Reported", "Not Reported", or "Not Applicable".

Three authors (QG, MP, and WT), blinded to each other's assessments, abstracted the reporting of each article independently. The first reviewer (QG) evaluated all 100 articles, of which 50 articles were randomly selected for the second reviewer (MP), and the other 50 articles were given to the third reviewer (WT) for abstraction. After completion of independent assessments, any disagreements between reviewers were resolved through consensus.

Statistical Analysis

We calculated the percentage of studies that reported each evaluation item and a 95% confidence interval (CI) using an exact binomial method. We then estimated the median, minimum and maximum percentages of reported items for each article.

Inter-rater agreement was assessed using the prevalence-adjusted bias-adjusted kappa (PABAκ) coefficient [25]. When the prevalence of a rating is very high or low, the value of kappa may indicate a low level of agreement while the observed percentage of agreement is high, known as the kappa paradox [26]. Hence, we used prevalence-adjusted

bias-adjusted kappa to address this paradox and to better interpret the inter-rater agreement. Kappa coefficient results were interpreted based on the scale as proposed by Byrt [27]: 0.00 or less (No agreement), 0.01-0.20 (Poor agreement), 0.21-0.40 (Slight agreement), 0.41-0.60 (Fair agreement), 0.61-0.80 (Good agreement), 0.81-0.92 (Very good agreement), 0.93-1.00 (Excellent agreement).

We performed a sample size calculation to determine the number of articles to be included in the extraction and analysis. A sample size of 100 was chosen so that with 95% confidence, we would be able to quantify the true percentage of articles that reported each item to within 10% (see Appendix C). All statistical analyses were conducted using the SAS 9.2 software (Cary, NC).

Results

Study Selection

The search strategy identified 1120 unique articles. Among these, a target number of 100 articles was reached and included in the final review after screening 1100 articles randomly for eligibility (see Figure 1 for a flow diagram). The list of the 100 eligible articles is included in Appendix D.

Study Characteristics

Among the included 100 eligible articles published in six leading journals in 2010 and 2011, about 60% came from the journal *NeuroImage*. The majority of study designs were cross-sectional (94%). The funding source was reported in 78% of the citations, and came primarily from two or more different sources (77%) rather than from industry alone (1%).

Fifty three percent of included articles were published in 2010 and the remaining forty seven percent in 2011. The median total number of subjects was 34 (first quartile (Q1) = 26, third quartile (Q3) = 48) ranging from 8 to 126, and most studies (79%) had a sample size of no more than 50 (see Table 1).

Items Commonly Reported

Of the 83 items, 22 items were reported by 85% or more of the 100 included articles. Specifically, all of the studies reported sample sizes. Most studies further described the manufacturer, field strength and model name of the scanner and the pulse sequence type (98%), statistical methods used for group modeling (97%), subjects' characteristics such as age and gender (94%), statistical methods used for within-subject modeling (92%), eligibility criteria on selecting subjects (91%), and whether statistical inferences were corrected for multiple comparisons (90%). Similarly, 86% of the articles reported how regions of interest (ROIs) were defined. Of 86 articles that reported analyses not conducted on the whole brain, 80 (93%) explained how regions were determined (see Tables 2a-2i).

Items Not Commonly Reported

A total of 31 items were not often reported; 13 items were reported by fewer than 20% of the included articles. Critically, and in sharp contrast to Poldrack's guidelines, none of the studies reported observed effect sizes if they failed to reject the null hypothesis. Only one article (3%, 1/31) provided justifications for using fixed-effect inferences for group modeling. Other items that were insufficiently reported included slice-timing and motion

corrections (12%), temporal autocorrelation modeling used to account for within-subject variances and correlations (18%), whether and how the task design was optimized for efficiency if it was an event-related design (22%, 8/35), distributions of inter-stimulus intervals (ISI), if ISI was variable (23%, 9/39), statistical methods for repeated measurements (24%), and smoothness and resolution element (RESEL) count if family-wise error (FWE) was found by random field theory (RFT) (25%, 1/4). Moreover, only six articles (28%, 6/21) described whether variances were assumed equal among groups if there were more than two groups. Of the 35 articles that reported percent signal changes, 12 (34%) explained how scaling factors were determined. Similarly, 45% of the articles stated how signal was extracted within ROIs.

Reported Items per Article

The median (minimum, maximum) percentage of reported items per article was 51% (30%, 78%).

The inter-rater agreement was very good (PABAκ > 0.8) for 31 items, good (0.6< PABAκ ≤ 0.8) for 31 items, fair (0.4<PABAκ ≤ 0.6) for 20 items, and slight (PABAκ=0.34) for one item.

Discussion

This study identified some reporting practices in observational clinical fMRI studies that met expectations and other areas where reporting was less than adequate. In particular, only one quarter of the items from the recommended reporting guidelines by Poldrak et

al. (2008) were reported adequately. Indeed, only one half of recommended items were routinely reported in each article. Moreover, one third of the items were reported by less than half of the articles. Less adequately reported items were distributed across the categories: experimental design, inter-subject registration and smoothing, data preprocessing, statistical modeling, and statistical inference on ROI analysis. These results indicate that substantial room for improvement exists in the reporting of observational clinical fMRI studies.

Specifically, improvement in reporting important details is recommended in areas such as observed effect sizes in the results section when study results are negative, justifications for fixed-effect inferences used for group modeling, and temporal autocorrelation matrix used to account for within-subject variance and correlations. As effect sizes observed from statistically significant regions overestimate true effect sizes [28,29], including values from non-significant regions (e.g., that are identified from previous studies to be related to our testing hypothesis) would help provide a more realistic range of effect size estimates and reduce the risk of bias arising from reporting on active regions only. Given the existence of temporal autocorrelation in fMRI time series, incorporating autocorrelation structure increases the accuracy of variance estimates. Reporting temporal autocorrelation estimates enables proper power analyses based on the method proposed by Mumford and Nichols [30]. Whereas findings from fixed-effect inferences particularly reflect the cohort of subjects studied, random-effect inferences generalize findings to the population at large from which the study sample was drawn [31]. It is now the rule of thumb to use random-effect inferences for between-subject

group modeling and fixed-effect inferences for single-subject modeling. Providing justifications for using fixed-effects for group modeling would enhance understanding and interpretation.

This study differed substantially from the one existing review of fMRI reporting [16] in the number of items, definitions of items, study population and study design. For example, although Carp's study used a single reviewer, we conducted a systematic review by using a duplicate abstraction, measuring inter-rater agreement and resolving disagreements through a consensus. Moreover, our study focused on observational studies with clinical participants; in contrast, Carp evaluated fMRI studies in general which may not capture many observational studies involving clinical participants. There are also some notable differences in results between the two studies. For example, in the current study around one third reported the distribution of inter-trial intervals, compared to one-twelfth in Carp's study. About one half reported the number of subjects rejected from analyses with reasons for rejection in our study, which is one quarter greater than that of Carp's study. Similarly, less than one third of the articles in our study reported the following four methodological items but still showed better reporting than those in Carp's study: how potentially confounding variables were matched across groups for group comparisons, whether autocorrelations were modeled, whether equal variance was assumed across groups for multiple group designs, and the number of RESELs and image smoothness for studies using FWE correction. Although different, both studies did detect some common important items that are frequently absent from published reports, indicating that incomplete reporting challenges the evaluation, understanding and

interpretation of study findings, and limits the use of results for synthesis, e.g., for meta analyses.

Complete reporting becomes particularly important for studies involving clinical populations, where ensuring methodological rigor is necessary to uphold investigators' promises to their participants that their participation will help society to better understand the nature of their condition. Our findings point towards the need for substantial improvement in this regard. In several other fields of health research, it has been demonstrated that journals adopting standard reporting guidelines (e.g., CONSORT statement) have better quality of reporting than those that do not [32-34], thus the use of guidelines in the fMRI literature may help improve the quality of reporting as well.

The guidelines for reporting fMRI studies proposed by Poldrack and his colleagues (2008) are lengthy and there is no consensus as to whether each item on the list is in fact essential to report. Given that authors have to work within strict word limits, the length of Poldrack's checklist is problematic. In addition to the first three items that we deleted from the original checklist, resulting in an 83-item list (see Appendix B), we suggest further dropping four items that involve substantial subjectivity or have low inter-rater agreement, leading to a list of 79 items. Reasons for exclusion, where relevant, for items on the initial 83-item list are given in the last column of Tables 2a-2i. Of 79 items, sixty-one items are generally applied in most situations and the other 18 are conditional items that need only to be reported under specific conditions. As reporting guidelines are ever-evolving documents and require continual updates, there will always be room for improvement in existing guidelines [35]. For example, the current guidelines developed

by Poldrack et al. do not cover the following items relevant to the quality of reporting observational clinical studies: sample size calculations in the methods section, observed input estimates necessary to compute future sample sizes in the results section, controlling for confounding variables, sampling strategy (e.g., random sampling or convenience sampling), characteristics of clinical participants, and participation data flow diagrams to better understand potential bias due to non-participation [36]. Future work may be needed to reflect these components in the reporting guidelines. Conversely, the practical concerns of the lengthy guidelines and strict space limitations entail making a shortened checklist containing essential and relevant items.

Our recommendation for creating a shortened version of the checklist is not to supplant the existing guidelines but rather a suggestion to consider during the next update of the guidelines. We hope that our suggestions will lead to more discussion and future consensus regarding what is in fact essential to report in the observational clinical fMRI literature.

The present study has several limitations. First, findings in this study reflect the quality of reporting of observational clinical fMRI studies in six top journals published between 2010 and 2011, results that may not apply to journals in general. Most likely, these results may overestimate true rates of reporting. Second, several items on the checklist used for evaluation in this systematic review involve subjectivity. Using duplicate review and consensus for any disagreements helped to reduce differences in interpretations between reviewers.

Conclusion

This study has highlighted under-reported areas in observational fMRI studies involving clinical participants and points towards a need for improvement. Adherence to the guidelines for fMRI studies proposed by Poldrack and his colleagues could help improve quality of reporting. Considering that the guidelines are lengthy and there are strict word limits for reports, we suggest that there is a need for a consensus meeting to create a shortened version of the checklist.

Author Contributions

Conceived and designed the experiments: QG LT EP GH. Performed the experiments: QG MP WT. Analyzed the data: QG. Wrote and revised the paper: QG. Interpreted the data: QG RG MM GH LT EP. Reviewed the manuscript: EP GH LT MM WT MP.

References

1. Sheline YI, Barch DM, Donnelly JM, Ollinger JM, Snyder AZ, et al. (2001) Increased amygdala response to masked emotional faces in depressed subjects resolves with antidepressant treatment: An fMRI study. Biol Psychiatry 50: 651-658.
2. Siegle GJ, Steinhauer SR, Thase ME, Stenger VA, Carter CS. (2002) Can't shake that feeling: Event-related fMRI assessment of sustained amygdala activity in response to emotional information in depressed individuals. Biol Psychiatry 51: 693-707.
3. Glahn DC, Ragland JD, Abramoff A, Barrett J, Laird AR, et al. (2005) Beyond hypofrontality: A quantitative meta-analysis of functional neuroimaging studies of working memory in schizophrenia. Hum Brain Mapp 25: 60-69.
4. Snitz BE, MacDonald A,3rd, Cohen JD, Cho RY, Becker T, et al. (2005) Lateral and medial hypofrontality in first-episode schizophrenia: Functional activity in a medication-naive state and effects of short-term atypical antipsychotic treatment. Am J Psychiatry 162: 2322-2329.

5. Monk CS, Klein RG, Telzer EH, Schroth EA, Mannuzza S, et al. (2008) Amygdala and nucleus accumbens activation to emotional facial expressions in children and adolescents at risk for major depression. Am J Psychiatry 165: 90-98.

6. Yoon JH, Minzenberg MJ, Ursu S, Ryan Walter BS, Wendelken C, et al. (2008) Association of dorsolateral prefrontal cortex dysfunction with disrupted coordinated brain activity in schizophrenia: Relationship with impaired cognition, behavioral disorganization, and global function. Am J Psychiatry 165: 1006-1014.

7. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, et al. (2007) The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. Ann Intern Med 147: 573-577.

8. Begg C, Cho M, Eastwood S, Horton R, Moher D, et al. (1996) Improving the quality of reporting of randomized controlled trials. the CONSORT statement. JAMA 276: 637-639.

9. Chan AW, Krleza-Jeric K, Schmid I, Altman DG. (2004) Outcome reporting bias in randomized trials funded by the canadian institutes of health research. CMAJ 171: 735-740. 10.1503/cmaj.1041086.

10. Chan AW, Altman DG. (2005) Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. BMJ 330: 753. 10.1136/bmj.38356.424606.8F.

11. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, et al. (2008) Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One 3: e3081. 10.1371/journal.pone.0003081.

12. Poldrack RA, Fletcher PC, Henson RN, Worsley KJ, Brett M, et al. (2008) Guidelines for reporting an fMRI study. Neuroimage 40: 409-414.

13. Young NS, Ioannidis JP, Al-Ubaydli O. (2008) Why current publication practices may distort science. PLoS Med 5: e201. 10.1371/journal.pmed.0050201.

14. Langan S, Schmitt J, Coenraads PJ, Svensson A, von Elm E, et al. (2010) The reporting of observational research studies in dermatology journals: A literature-based study. Arch Dermatol 146: 534-541. 10.1001/archdermatol.2010.87.

15. Papathanasiou AA, Zintzaras E. (2010) Assessing the quality of reporting of observational studies in cancer. Ann Epidemiol 20: 67-73.

16. Carp J. (2012) The secret lives of experiments: Methods reporting in the fMRI literature. Neuroimage 63: 289-300.

17. Carter CS, Heckers S, Nichols T, Pine DS, Strother S. (2008) Optimizing the design and analysis of clinical functional magnetic resonance imaging research studies. Biol Psychiatry 64: 842-849.

18. MacDonald AW,3rd, Thermenos HW, Barch DM, Seidman LJ. (2009) Imaging genetic liability to schizophrenia: Systematic review of FMRI studies of patients' nonpsychotic relatives. Schizophr Bull 35: 1142-1162. 10.1093/schbul/sbn053; 10.1093/schbul/sbn053.

19. Huang W, Pach D, Napadow V, Park K, Long X, et al. (2012) Characterizing acupuncture stimuli using brain imaging with FMRI--a systematic review and meta-analysis of the literature. PLoS One 7: e32960. 10.1371/journal.pone.0032960; 10.1371/journal.pone.0032960.

20. Lee KP, Schotland M, Bacchetti P, Bero LA. (2002) Association of journal quality indicators with methodological quality of clinical research articles. JAMA 287: 2805-2808.

21. Birken CS, Parkin PC. (1999) In which journals will pediatricians find the best evidence for clinical practice? Pediatrics 103: 941-947.

22. Opthof T. (1997) Sense and nonsense about the impact factor. Cardiovasc Res 33: 1-7.

23. Schoonbaert D, Roelants G. (1996) Citation analysis for measuring the value of scientific publications: Quality assessment tool or comedy of errors? Trop Med Int Health 1: 739-752.

24. Bruer JT. (1982) Methodological rigor and citation frequency in patient compliance literature. Am J Public Health 72: 1119-1123.

25. Byrt T, Bishop J, Carlin JB. (1993) Bias, prevalence and kappa. J Clin Epidemiol 46: 423-429.

26. Feinstein AR, Cicchetti DV. (1990) High agreement but low kappa: I. the problems of two paradoxes. J Clin Epidemiol 43: 543-549.

27. Byrt T. (1996) How good is that agreement? Epidemiology 7: 561.

28. Maxwell SE. (2004) The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. Psychol Methods 9: 147-163. 10.1037/1082-989X.9.2.147.

29. Mumford JA. (2012) A power calculation guide for fMRI studies. Soc Cogn Affect Neurosci 7: 738-742. 10.1093/scan/nss059; 10.1093/scan/nss059.

30. Mumford JA, Nichols TE. (2008) Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. Neuroimage 39: 261-268.

31. Frackowiak RSJ, Ashburner JT, Penny WD, Zeki S. (2004) Random effects analysis (chapter 12). In: Ashburner J, Friston K, Penny W, editors. Human Brain Function. London, UK: Academic Press.

32. Moher D, Jones A, Lepage L, CONSORT Grp. (2001) Use of the CONSORT statement and quality of reports of randomized trials - A comparative before-and-after evaluation. Jama-Journal of the American Medical Association 285: 1992-1995. 10.1001/jama.285.15.1992.

33. Plint AC, Moher D, Morrison A, Schulz K, Altman DG, et al. (2006) Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. Med J Aust 185: 263-267.

34. Alvarez F, Meyer N, Gourraud PA, Paul C. (2009) CONSORT adoption and quality of reporting of randomized controlled trials: A systematic analysis in two dermatology journals. Br J Dermatol 161: 1159-1165. 10.1111/j.1365-2133.2009.09382.x; 10.1111/j.1365-2133.2009.09382.x.

35. Moher D, Schulz KF, Simera I, Altman DG. (2010) Guidance for developers of health research reporting guidelines. PLoS Med 7: e1000217. 10.1371/journal.pmed.1000217; 10.1371/journal.pmed.1000217.

36. Young EA, Breslau N. (2004) Cortisol and catecholamines in posttraumatic stress disorder: An epidemiologic community study. Arch Gen Psychiatry 61: 394-401. 10.1001/archpsyc.61.4.394.

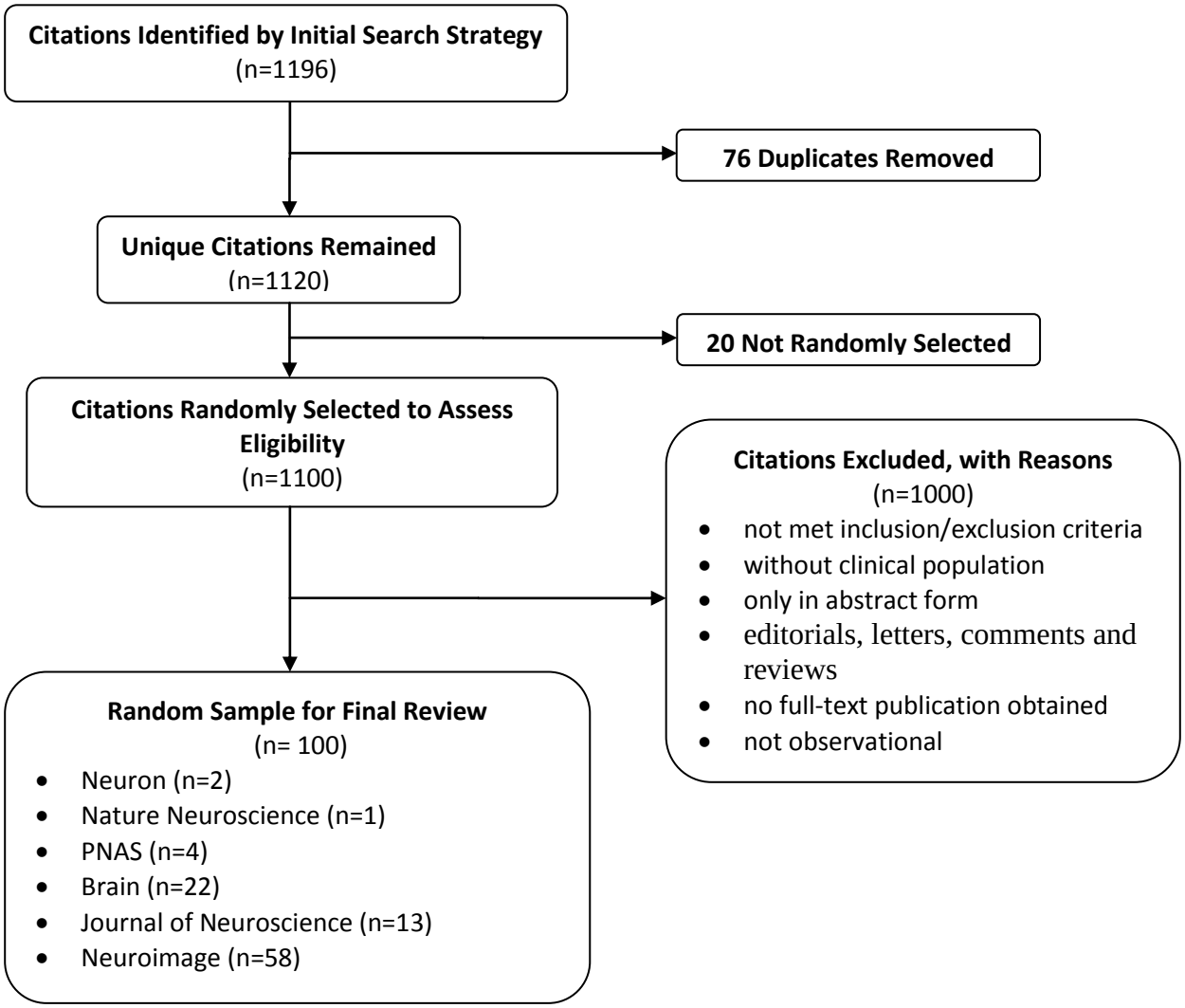

Figure1. Flow diagram of citation selection process.

Table 1. Characteristics of Included fMRI Studies (Information Extracted from Each Article)

Study Feature	All articles (n=100) Median (Q1, Q3) or %
Publication Journal	
Neuron	2
Nature Neuroscience	1
Proceedings of the National Academy of Sciences of the United States of America	4
Brain	22
Journal of Neuroscience	13
Neuroimage	58
Publication Year	
2010	53
2011	47
Study Design	
Case-control	0
Cohort	6
Cross-sectional	94
Number of Subjects	34 (26, 48)
Up to 10	2
10-50	77
51-100	17
More than 100	4
Funding Sources	
Completely funded by industry	1
Others	77
Not reported	22

Note: Q1=first quartile or 25th percentile, Q3=third quartile or 75th percentile.

Table 2a. Percentage of articles reported each item, inter-rater agreement on the item and whether the item should be included in future shortened checklist relating to "Experimental Design"

Item No	Description	% Reported (95% CI)	PABAκ (95% CI)	Item* Selection
1a	Described number of blocks, trials, experimental units per session or per subject	92 (84, 96)	0.90 (0.77, 0.97)	Included
1b	Stated length of each trial and interval between trials described	81 (71, 88)	0.76 (0.60, 0.87)	Included
1c#	If ISIs are variable, reported the mean and range of ISIs and how they were distributed (n=39)	23 (11, 39)	0.76 (0.60, 0.87)	Included
1d#	If block designs, specified the length of blocks (n=73)	79 (67, 87)	0.72 (0.55, 0.84)	Included
1e#	If event-related designs, stated whether the design was optimized for efficiency, and if so, stated how (n=35)	22 (10, 40)	0.70 (0.53, 0.83)	Included
1f#	If mixed design, stated correlation between block and event regressors (n=2)	50 (1, 98)	0.94 (0.83, 0.99)	Included
2a	Stated task instructions on what subjects were asked to do	92 (84, 96)	0.92 (0.80, 0.98)	Included
2b	Described what the Stimuli were and how many there were	69 (58, 77)	0.72 (0.55, 0.84)	Included
2c	Stated whether specific stimuli repeated across trials	49 (38, 59)	0.46 (0.26, 0.63)	Included
3	If the experiment had multiple conditions, stated what the specific planned comparisons were, or whether an omnibus ANOVA test was used	89 (81, 94)	0.90 (0.77, 0.97)	Included

Abbreviations: ISIs, inter-stimulus intervals; ANOVA, analysis of variance
The conditional item which is needed to report when the condition is met.
* To identify whether the item should be included in future shortened checklist. If excluded, the reasons for exclusion are given.

Table 2b. Percentage of articles reported each item, inter-rater agreement on the item and whether the item should be included in future shortened checklist relating to "Study Subjects"

Item No	Description	% Reported (95% CI)	PABAκ (95% CI)	Item* Selection
4a	Stated number of subjects	100 (96, 100)	1.00 (0.93, 1.00)	Included
4b	Stated age (mean and range)	92 (84, 96)	0.90 (0.77, 0.97)	Included
4c	Stated handedness	64 (53, 73)	0.98 (0.89, 0.99)	Included
4d	Stated number of males or females	95 (88, 98)	0.90 (0.77, 0.97)	Included
4e	Stated inclusion and exclusion criteria	91 (83, 95)	0.86 (0.72, 0.94)	Included
4f	If any subjects were scanned but then rejected from analysis after data collection, stated numbers and reasons for rejection	52 (41, 62)	0.82 (0.67, 0.92)	Included
4g#	For group comparisons, stated what variables (if any) were equated across groups (n=90)	70 (59, 79)	0.56 (0.37, 0.71)	Included
5	Stated which IRB approved the protocol	94 (87, 97)	0.94 (0.83, 0.99)	Included
6	Stated how behavioral performance was measured (e.g., response time, accuracy)	56 (45, 65)	0.34 (0.14, 0.52)	Excluded due to much subjectivity and low inter-rater agreement. For example, some standard tools (e.g., E-Prime, Fiber-Optic-Button box) measure response timing and accuracy. If these tools are cited, is it safe to assume that the behavioral performance is measured? If not, what minimum details are required to report so as to score it as 'reported'? Is this item required to report in every study? If not, under what condition?

Abbreviations: IRB, institutional review board
The conditional item which is needed to report when the condition is met.
* To identify whether the item should be included in future shortened checklist. If excluded, the reasons for exclusion are given.

Table 2c. Percentage of articles reported each item, inter-rater agreement on the item and whether the item should be included in future shortened checklist relating to "Image Properties"

Item No	Description	% Reported (95% CI)	PABAκ (95% CI)	Item* Selection
7a	Provided manufacturer, field strength (in Tesla) and model name of MRI system	98 (92, 99)	0.96 (0.86, 0.99)	Included
7b	Gave number of experimental sessions and volumes acquired per session	50 (39, 60)	0.78 (0.62, 0.88)	Included
7c	Stated pulse sequence type (e.g., gradient/spin echo, EPI/spiral)	98 (92, 99)	1.00 (0.93, 1.00)	Included
7d	Stated field of view, matrix size, slice thickness, inter-slice skip	36 (26, 46)	0.76 (0.60, 0.87)	Included
7e	Provided acquisition orientation (axial, sagittal, coronal, oblique)	71 (61, 79)	0.90 (0.77, 0.97)	Included
7f	Stated whether it is on the whole brain. If not, state area of acquisition	65 (54, 74)	0.90 (0.77, 0.97)	Included
7g	Stated order of acquisition of slices (sequential or interleaved)	21 (13, 30)	0.82 (0.67, 0.92)	Included
7h	Stated TE, TR and flip angle	86 (77, 92)	0.92 (0.80, 0.98)	Included

Abbreviations: EPI, Echo Planar Imaging; TE, echo time; TR, repetition time
* To identify whether the item should be included in future shortened checklist. If excluded, the reasons for exclusion are given.

Table 2d. Percentage of articles reported each item, inter-rater agreement on the item and whether the item should be included in future shortened checklist relating to "Data Preprocessing"

Item No	Description	% Reported (95% CI)	PABAκ (95% CI)	Item* Selection
8a	Stated the version number or date of last application for each piece of software used	78 (68, 85)	0.76 (0.60, 0.87)	Included
8b	Specified differences in any subjects who required different processing operations or settings in the analysis (n=78)	3 (1, 10)	0.60 (0.42, 0.75)	Excluded due to much subjectivity. For example, if the study states that all subjects received same operations or settings, making this item not applicable. If not having the similar statements, it is difficult to decide under what condition this item is expected to report.
9a	Specified order of preprocessing operations	26 (17, 35)	0.70 (0.53, 0.83)	Excluded. Each standard software package has a standard order of preprocessing operations. The order is known once the software is specified. It seems less essential to report this item than other important details.
9b	Stated reference slice and interpolation type for slice timing correction	9 (4, 16)	0.94 (0.83, 0.99)	Included
9c	Stated reference scan, image similarity metric, type of interpolation used, degrees-of-freedom , and ideally optimization method for motion correction	15 (8, 23)	0.74 (0.58, 0.86)	Included

* To identify whether the item should be included in future shortened checklist. If excluded, the reasons for exclusion are given.

Table 2e. Percentage of articles reported each item, inter-rater agreement on the item and whether the item should be included in future shortened checklist relating to "Inter-subject Registration and Smoothing"

Item No	Description	% Reported (95% CI)	PABAκ (95% CI)	Item* Selection
10a	Illustrated the voxels presented in all subjects using "mask image"	16 (9, 24)	0.68 (0.51, 0.81)	Included
10b	Described transformation model (linear/affine, nonlinear), type of any non-linear transformations (polynomial, discrete cosine basis), number of parameters (e.g., 12 parameter affine), regularization image-similarity metric, and interpolation method	18 (11, 26)	0.70 (0.53, 0.83)	Included
10c	Stated object anatomical image information used for transformation to Atlas	42 (32, 52)	0.46 (0.26, 0.63)	Included
10d	Stated if anatomical MRI is co-planar with functional acquisition	36 (26, 46)	0.80 (0.65, 0.90)	Included
10e	Stated if functional acquisition is co-registered to anatomical	47 (36, 57)	0.82 (0.67, 0.92)	Included
10f[#]	If functional acquisition is co-registered to anatomical, stated how (n=47)	27 (15, 42)	0.50 (0.31, 0.66)	Included
10g	Provided Atlas/target information	87 (78, 92)	0.66 (0.48, 0.79)	Included
10h	Stated brain image template space, name, modality and resolution (e.g., "FSL's MNI Avg152, T1 2x2x2 mm", "SPM2's MNI gray matter template 2x2x2 mm")	16 (9, 24)	0.64 (0.46, 0.78)	Included
10i	Stated typically MNI, Talairach, or MNI converted to Talairach	85 (76, 91)	0.84 (0.69, 0.93)	Included
10j[#]	If MNI is converted to Talairach, stated the method used (e.g., Brett's mni2tal) (n=13)	61 (31, 86)	0.86 (0.72, 0.94)	Included

10k	State clearly how anatomical locations (e.g., gyral anatomy, Brodmann areas) were determined (e.g., paper atlas, Talairach Daemon, manual inspection of individual's anatomy, etc.)	61 (50, 70)	0.68 (0.50, 0.81)	Included
11	Described size and type of smoothing kernel (e.g., for a group study, "12 mm FHWM Gaussian smoothing applied to ameliorate differences in inter-subject localization"; for single subject fMRI "6 mm FWHM Gaussian smoothing used to reduce noise")	84 (75, 90)	0.96 (0.85, 0.99)	Included

Abbreviations: MRI, magnetic resonance imaging; MNI, Montreal Neurological Institute space
The conditional item which is needed to report when the condition is met.
* To identify whether the item should be included in future shortened checklist. If excluded, the reasons for exclusion are given.

Table 2f. Percentage of articles reported each item, inter-rater agreement on the item and whether the item should be included in future shortened checklist relating to "Statistical Modeling"

Item No	Description	% Reported (95% CI)	PABAκ (95% CI)	Item* Selection
12	For novel methods not described in a separate paper, provided description and validation of method in the text or an appendix (n=2)	50 (1, 98)	0.88 (0.74, 0.96)	Excluded. Giving that methods are continually developing, it involves much subjectivity as to whether or not the reported methods are novel.
13a	Stated statistical model and estimation method for both intra-subject and group modeling	92 (84, 96)	0.80 (0.65, 0.90)	Included
13b	Stated block- or epoch-based or event-related model	97 (91, 99)	0.92 (0.80, 0.98)	Included
13c	Specified hemodynamic response function	58 (47, 67)	0.76 (0.60, 0.87)	Included

13d	Clearly stated additional regressors used (e.g., temporal derivatives, motion, behavioral covariates)	53 (42, 63)	0.58 (0.39, 0.73)	Included
13e	Stated any orthogonalization of regressors	7 (2, 13)	0.86 (0.72, 0.94)	Included
13f	Stated drift modeling or high-pass filtering (e.g., "DCT with cut off of X seconds"; "Gaussian-weighted running line smoother, cut-off 100 seconds", or "cubic polynomial")	55 (44, 64)	0.74 (0.57, 0.86)	Included
13g	Described autocorrelation model (e.g., AR(1), AR(1)+WN, or arbitrary autocorrelation function)	18 (11, 26)	0.80 (0.64, 0.90)	Included
13h	Defined contrast for task or stimulus conditions	90 (82, 95)	0.90 (0.77, 0.97)	Included
14a	Stated statistical model, estimation method and inference type for group modeling (e.g., mixed, random or fixed effects)	97 (91, 99)	0.90 (0.77, 0.97)	Included
14b[#]	If fixed effects inference used for group modeling, provided the justification (n=31)	3 (1, 16)	0.46 (0.26, 0.63)	Included
14c	If the group has more than 2-levels, described the levels and assumptions of the model (e.g., are variances assumed equal between groups)	28 (11, 52)	0.60 (0.41, 0.75)	Included
14d	Stated methods used for repeated measures to account for within subject correlation in group modeling	24 (16, 33)	0.66 (0.48, 0.79)	Included

Abbreviations: DCT, discrete cosine transform; AR(1), first-order Autoregressive Model; WN, white noise.
The conditional item which is needed to report when the condition is met.
* To identify whether the item should be included in future shortened checklist. If excluded, the reasons for exclusion are given.

Table 2g. Percentage of articles reported each item, inter-rater agreement on the item and whether the item should be included in future shortened checklist relating to "Statistical Inference on Statistic Image (thresholding)"

Item No	Description	% Reported (95% CI)	PABAκ (95% CI)	Item* Selection
15a	Stated type of search region for analysis, and the volume in voxels or CC	54 (43, 64)	0.60 (0.41, 0.75)	Included
15b#	If not whole brain, stated how region was determined (n=86)	93 (85, 97)	0.58 (0.39, 0.73)	Included
15c#	Stated and listed each if threshold used for inference and threshold used for visualization in figures is different (n=49)	44 (30, 59)	0.56 (0.37, 0.71)	Included
15d	Stated if inferences are corrected for multiple comparisons	90 (82, 95)	0.80 (0.64, 0.90)	Included
15e#	If correction is limited to a small volume, stated the method for selecting the region (n=73)	72 (60, 82)	0.54 (0.35, 0.70)	Included
15f#	Labeled "uncorrected" if no formal multiple comparisons method is used (n=76)	84 (74, 91)	0.80 (0.64, 0.90)	Included
15g	Stated if it is voxel-wise significance	49 (38, 59)	0.54 (0.35, 0.70)	Included
15h	Stated if inferences are corrected for FWE or FDR	50 (39, 60)	0.78 (0.62, 0.89)	Included
15i#	Listed the smoothness in mm FWHM and the RESEL count if FWE found by random field theory (n=45)	25 (1, 80)	0.70 (0.52, 0.83)	Included
15j#	Provided details of parameters for simulation if FWE found by simulation (e.g., AFNI AphaSim) (n=7)	57 (18, 90)	0.62 (0.43, 0.76)	Included
15k#	If not a standard method, specified the method for finding significance (n=12)	100 (73, 100)	0.72 (0.55, 0.84)	Included
15l	Stated cluster-defining threshold (e.g., *P=0.001*)	51 (40, 61)	0.44 (0.24, 0.61)	Included

15m	Stated the corrected cluster significance level (e.g., "Statistic images were assessed for cluster-wise significance using a cluster-defining threshold of *P*=0.001; the 0.05 FWE-corrected critical cluster size was 103")	55 (44, 64)	0.42 (0.22, 0.59)	Included
15n#	Provided smoothness and RESEL count if significance determined with random field theory (n=8)	12 (1, 52)	0.96 (0.85, 0.99)	Included
15o	Stated correction for multiple planned comparisons based upon each voxel	14 (7, 22)	0.44 (0.24, 0.61)	Included
15p#	Stated observed effect size for any failure to reject the null hypothesis (e.g., lack of activation in a particular region) (n=1)	0 (0, 3)	0.98 (0.89, 0.99)	Included

Abbreviations: CC, cubic centimeter; FWE, family-wise error; FDR, false discovery rate; FWHM, full-width at half-maximum; RESEL, resolution element
The conditional item which is needed to report when the condition is met.
* To identify whether the item should be included in future shortened checklist. If excluded, the reasons for exclusion are given.

Table 2h. Percentage of articles reported each item, inter-rater agreement on the item and whether the item should be included in future shortened checklist relating to "Statistical Inference on ROI Analysis"

Item No	Description	% Reported (95% CI)	PABAκ (95% CI)	Item* Selection
16a	Described how ROIs were defined (e.g., functional or anatomical localizer)	86 (77, 92)	0.54 (0.35, 0.70)	Included
16b	Described how signal was extracted within ROI (e.g., average parameter estimates, FIR deconvolution)	45 (35, 55)	0.46 (0.26, 0.63)	Included
16c#	If percent signal change reported, described how scaling factor was determined (n=35)	34 (19, 52)	0.52 (0.32, 0.68)	Included

16d	Stated if percent signal change is relative to voxel-mean, or whole-brain mean	16 (9, 24)	0.66 (0.48, 0.79)	Included

Abbreviations: ROI, region of interest; FIR, finite impulse response
The conditional item which is needed to report when the condition is met.
* To identify whether the item should be included in future shortened checklist. If excluded, the reasons for exclusion are given.

Table 2i. Percentage of articles reported each item, inter-rater agreement on the item and whether the item should be included in future shortened checklist relating to "Figures and Tables"

Item No	Description	% Reported (95% CI)	PABAκ (95% CI)	Item* Selection
17a	Stated the statistical map that the figure or table is based upon (e.g., *Z*, *t*, *p*)	95 (88, 98)	0.84 (0.69, 0.93)	Included
17b	Provided the thresholds used to create the image or figure (e.g., intensity and cluster extent)	71 (61, 79)	0.60 (0.41, 0.75)	Included
18	Underlying anatomical image stated (e.g., average anatomy, template image)	26 (17, 35)	0.66 (0.48, 0.79)	Included
19a	Locations in stereotactic space provided	73 (63, 81)	0.80 (0.64, 0.90)	Included
19b	Provided statistics for each cluster including maximum and cluster extent	51 (40, 61)	0.86 (0.72, 0.94)	Included
19c	Provided source of anatomical labels (e.g., atlas, automated labeling method)	67 (56, 76)	0.62 (0.43, 0.76)	Included

* To identify whether the item should be included in future shortened checklist. If excluded, the reasons for exclusion are given.

Appendix A

Database: Ovid MEDLINE(R) in-Process & Other Non-indexed Citations and Ovid MEDLINE (R) 1948 – Present. This search was conducted on February 12, 2012*

Step Number	**Search Strategy**
1	functional magnetic resonance imaging.mp.
2	fmri.mp.
3	1 or 2
4	limit 3 to (english language and humans and yr="2010-2011")
5	"neuron".jn.
6	"nature neuroscience".jn.
7	“proceedings of the national academy of science of the united states of america”.jn.
8	”brain”.jn.
9	”journal of neuroscience”.jn.
10	"neuroimage".jn.
11	4 and 5
12	4 and 6
13	4 and 7
14	4 and 8
15	4 and 9
16	4 and 10
17	11 or 12 or 13 or 14 or 15 or 16

** This search was limited to the years 2010-2011, six journals, English language and humans

Appendix B: Data extraction form containing 83 items adapted from Poldrack et al.'s checklist. Three items dropped from the checklist are indicated. Information extracted from each eligible article. The coding for the article is entered in the column, namely "Article#"†.

Category	Item No	Item Description	Instructions	Article#
EXPERIMENTAL DESIGN -design specification	1a	Describe number of blocks, trials, experimental units per session or per subject		
	1b	State length of each trial and interval between trials		
	1c	If ISIs are variable, report the mean and range of ISIs and how they are distributed	If ISIs are constant, it should be recorded 'Not Applicable'.	
	1d	*Block-Designs*: specify the length of blocks	If not a block design, it should be recorded 'Not Applicable'.	
	1e	*Event-related Designs*: state whether the design is optimized for efficiency, and if so, state how	If not an event-related design, it should be recorded 'Not Applicable'.	
	1f	*Mixed designs*: state correlation between block and event regressors	If not a mixed design, it should be recorded 'Not Applicable'.	
	1g	Instructions: state what subjects are asked to do		
EXPERIMENTAL DESIGN - task specification	2a	Stimuli: state whether specific stimuli repeated across trials		
	2b	Stimuli: describe what the Stimuli are and how		

		many there are		
	2c	Stimuli: state whether specific stimuli repeated across trials		
EXPERIMENTAL DESIGN - planned comparison	3	If the experiment has multiple conditions, state what the specific planned comparisons are, or whether an omnibus ANOVA test is used		
HUMAN SUBJECTS - details on subject sample	4a	State number of subjects		
	4b	State age (mean and range)		
	4c	State handedness		
	4d	State number of males or females		
	4e	State inclusion and exclusion criteria,		
	4f	If any subjects were scanned but then rejected from analysis after data collection, state how many and reasons for rejection	Report in either methods or results section	
	4g	For group comparisons, state what variables (if any) were equated across groups.		
HUMAN SUBJECTS - ethics approval	5	State which Institutional Review Board (IRB) approved the protocol		
HUMAN SUBJECTS - behavioral performance	6	State how behavioral performance was measured (e.g., response time, accuracy)		
DATA ACQUISITION - image properties	7a	Describe manufacturer, field strength (in Tesla), model name		

	7b	State the number of experimental sessions and volumes acquired per session		
	7c	State pulse sequence type (gradient/spin echo, EPI/spiral)		
	7d	State field of view, matrix size, slice thickness, inter-slice skip		
	7e	State acquisition orientation (axial, sagittal, coronal, oblique; if axials co-planar with AC-PC, the volume coverage in terms of Z in mm)		
	7f	State clearly whether it is on the whole brain. If not, state area of acquisition		
	7g	State order of acquisition of slices (sequential or interleaved)		
	7h	State TE, TR, flip angle		
DATA ACQUISITION - data preprocessing	8a	For each piece of software used, give the version number. If no version number is available, date of last application of updates)		
	8b	If any subjects required different processing operations or settings in the analysis, those differences should be specified explicitly		
DATA ACQUISITION - preprocessing general	9a	Specify order of preprocessing operations		
		Describe any data quality control measures	*Deleted	
		Unwarping of B0 distortions	*Deleted	
	9b	Slice timing correction: reference slice and type of interpolation used (e.g., “Slice timing correction to the first slice as performed, using		

		SPM5's Fourier phase shift interpolation")		
	9c	Motion correction: reference scan, image similarity metric, type of interpolation used, degrees-of-freedom (If not rigid body) and, ideally, optimization method		
DATA ACQUISITION - intersubject registration	10a	Illustrate the voxels present in all subjects using mask image		
	10b	Describe transformation model (linear/affine, nonlinear), type of any non-linear transformations (polynomial, discrete cosine basis), number of parameters (e.g., 12 parameter affine), regularization image-similarity metric, and interpolation method		
	10c	State object image information (image used to determine transformation to atlas)		
	10d	State if anatomical MRI is co-planar with functional acquisition		
	10e	State whether functional acquisition is co-registered to anatomical		
	10f	If functional acquisition is co-registered to anatomical, state how (such as segmented gray image or functional image)		
	10g	State Atlas/target information		
	10h	State brain image template space, name, modality and resolution (e.g, "FSL's MNI Avg 152, T1 2x2x2 mm"; "SPM2's MNI gray matter template 2x2x2 mm")		

	10i	State typically MNI, Talairach, or MNI converted to Talairach.		
	10j	If MNI is converted to Talairach, state the method used (e.g., Brett's mni2tal)		
	10k	State clearly how anatomical locations (e.g., gyral anatomy, Brodmann areas) were determined (e.g., paper atlas, Talairach Daemon, manual inspection of individuals' anatomy, etc.)		
DATA ACQUISITION - smoothing	11	Describe size and type of smoothing kernel (e.g., for a group study, "12 mm FHWM Gaussian smoothing applied to ameliorate differences in inter-subject localization"; for single subject fMRI "6 mm FWHM Gaussian smoothing used to reduce noise")		
STATISTICAL MODELING - general issues	12	For novel methods that are not described in detail in a separate paper, provide explicit description and validation of method either in the text or as an appendix		
STATISTICAL MODELING - intrasubject fMRI modeling info	13a	Describe statistical model and estimation method: multiple regression is most common statistical model; estimation methods are typically ordinary least squares (OLS), OLS with adjustment for autocorrelation		
	13b	State block/epoch-based or event-related model		
	13c	Specify hemodynamic response function (HRF): assumed HRF model, HRF basis, or estimated HRF		
	13d	Clearly state additional regressors used (e.g.,		

		temporal derivatives, motion, behavioral covariates)		
	13e	State any orthogonalization of regressors		
	13f	State the drift modeling or high-pass filtering (e.g., "DCT with cut off of X seconds"; "Gaussian-weighted running line smoother, cut-off 100 seconds", or "cubic polynomial")		
	13g	Describe the autocorrelation model type, and whether global or local		
	13h	State contrast construction: exactly what terms are subtracted from? Define these in terms of task or stimulus conditions		
STATISTICAL MODELING - group modeling info	14a	State statistical model, estimation method, and inference type		
	14b	If fixed effects inference used, provide the justification		
	14c	If more than 2-levels, describe the levels and assumptions of the model (eg, are variances assumed equal between groups)		
	14d	State if there are repeated measures. If multiple measurements per subject, list method to account for within subject correlation, exact assumptions made about correlation and variance.		
STATISTICAL INFERENCE - inference on	15a	State type of search region for analysis, and the volume in voxels or CC		

statistic image (thresholding)				
	15b	If not whole brain, state how region was determined; method for constructing region should be independent of present statistic image		
	15c	If threshold used for inference and threshold used for visualization in figures is different, clearly state so and list each	If thresholds for inference and visualization were the same, we should record 'Not Applicable'.	
	15d	Explicitly state if inferences are corrected for multiple comparisons		
	15e	If correction is limited to a small volume, the method for selecting the region should be stated explicitly		
	15f	If no formal multiple comparisons method is used, the inference must be explicitly labeled "uncorrected"		
	15g	Describe if it is voxel-wise significance		
	15h	State if inferences are corrected for Family-wise error (FWE) or false discovery rate (FDR)		
	15i	If FWE found by random field theory list the smoothness in mm FWHM and the RESEL count		
	15j	If FWE found by simulation (eg, AFNI AlphaSim), provide details of parameters for simulation		
	15k	If not a standard method, specify the method for finding significance		

	15l	State cluster-defining threshold (eg, *P=0.001*)		
	15m	State the corrected cluster significance level (e.g., "Statistic images were assessed for cluster-wise significance using a cluster-defining threshold of *P*=0.001; the 0.05 FWE-corrected critical cluster size was 103")		
	15n	If significance determined with random field theory, then smoothness and RESEL count must be supplied		
	15o	State correction for multiple planned comparisons based upon each voxel		
	15p	State observed effect size for any failure to reject the null hypothesis (e.g., lack of activation in a particular region)		
STATISTICAL INFERENCE - ROI analysis	16a	Describe how ROIs were defined (eg, functional versus anatomical localizer)		
	16b	Describe how signal was extracted within ROI (e.g., average parameter estimates, FIR deconvolution)		
	16c	If percent signal change reported, describe how scaling factor was determined (eg, height of block regressor or height of isolated event regressor)		
	16d	State if percent signal change is relative to voxel-mean, or whole-brain mean		
FIGURES AND TABLES - general	17a	State the statistical map that the figure or table is based upon (e.g., *Z*, *t*, *p*)		
	17b	Provide the thresholds used to created the		

		image or figure (intensity and cluster extent, where appropriate)		
FIGURES AND TABLES - figures	18	State the underlying anatomical image (e.g., average anatomy, template image)		
		State any additional operations (e.g., masking out parts of the image)	*Deleted	
FIGURES AND TABLES - tables	19a	Provide locations in stereotactic space (with the space described specifically)		
	19b	Provide statistics for each cluster (including maximum and cluster extent)		
	19c	Provide source of anatomical labels (eg, atlas, automated labeling method)		

† Coding for each evaluated item: 0 – "Not Reported", 1 – "Reported", 2 – "Not Applicable"
* Deleted from the Poldrack et al.'s checklist due to substantial subjectivity
Abbreviations: ISIs, inter-stimulus intervals; ANOVA, analysis of variance; EPI, Echo Planar Imaging; TE, echo time; TR, repetition time; MRI, magnetic resonance imaging; MNI, Montreal Neurological Institute space; DCT, discrete cosine transform; CC, cubic centimeter; FWE, family-wise error; FDR, false discovery rate; FWHM, full-width at half-maximum; RESEL, resolution element; ROI, region of interest; FIR, finite impulse response

Appendix C: Sample Size Calculation for Estimating a Single Proportion with a Level of Confidence

The primary outcomes of this paper are the proportion of reviewed articles that reported each item of the STROBE checklist and the proportion that reported estimates of parameters needed for future sample size determination. Given no prior similar studies have provided the estimates of the proportion, we determine sample sizes by varying the proportion estimates and margin of error over its plausible ranges (See the results below) through sensitivity analysis. The mathematics formula for sample size calculation with an expected estimate and precision is as follows:

The estimated range of p is $p \pm Z_{1-\frac{\alpha}{2}}\sqrt{\frac{p\times(1-p)}{n}}$, MOE= $Z_{1-\frac{\alpha}{2}}\sqrt{\frac{p\times(1-p)}{n}}$, then

$$n = \frac{Z^2_{1-\frac{\alpha}{2}} \times p(1-p)}{MOE^2}$$

As shown in the table below, we notice that the sample size of 96 can achieve any estimate of proportion of reporting at a MOE of 10% and also can reach its extreme estimates of proportion with a value less than 5% or greater than 95% and with an MOE of 5% at a 95% confidence level. We therefore chose a sample size of 100 by rounding up from 96.

Sample Size Calculations by Varying Estimated Proportion and Margin of Error

Estimated % of reporting (p)	Margin of Error (MOE) 5%	10%
5%	73	18
10%	138	35
15%	196	49
20%	246	61
25%	288	72
30%	323	81
35%	350	87
40%	369	92
45%	380	95
50%	384	96
55%	380	95
60%	369	92
65%	350	87
70%	323	81
75%	288	72
80%	246	61
85%	196	49
90%	138	35
95%	73	18

Appendix D: Reference of 100 Eligible Articles

1. Agam Y, Joseph RM, Barton JJ, Manoach DS. (2010) Reduced cognitive control of response inhibition by the anterior cingulate cortex in autism spectrum disorders. Neuroimage 52: 336-347.

2. Allen P, Stephan KE, Mechelli A, Day F, Ward N, et al. (2010) Cingulate activity and fronto-temporal connectivity in people with prodromal signs of psychosis. Neuroimage 49: 947-955.

3. Amanzio M, Torta DM, Sacco K, Cauda F, D'Agata F, et al. (2011) Unawareness of deficits in alzheimer's disease: Role of the cingulate cortex. Brain 134: 1061-1076. 10.1093/brain/awr020; 10.1093/brain/awr020.

4. Ances B, Vaida F, Ellis R, Buxton R. (2011) Test-retest stability of calibrated BOLD-fMRI in HIV- and HIV+ subjects. Neuroimage 54: 2156-2162.

5. Arichi T, Moraux A, Melendez A, Doria V, Groppo M, et al. (2010) Somatosensory cortical activation identified by functional MRI in preterm and term infants. Neuroimage 49: 2063-2071.

6. Bach S, Brandeis D, Hofstetter C, Martin E, Richardson U, et al. (2010) Early emergence of deviant frontal fMRI activity for phonological processes in poor beginning readers. Neuroimage 53: 682-693.

7. Bai X, Vestal M, Berman R, Negishi M, Spann M, et al. (2010) Dynamic time course of typical childhood absence seizures: EEG, behavior, and functional magnetic resonance imaging. J Neurosci 30: 5884-5893.

8. Bardin JC, Fins JJ, Katz DI, Hersh J, Heier LA, et al. (2011) Dissociations between behavioural and functional magnetic resonance imaging-based evaluations of cognitive function after brain injury. Brain 134: 769-782.

9. Becerril KE, Repovs G, Barch DM. (2011) Error processing network dynamics in schizophrenia. Neuroimage 54: 1495-1505. 10.1016/j.neuroimage.2010.09.046; 10.1016/j.neuroimage.2010.09.046.

10. Bedard AC, Schulz KP, Cook EH,Jr, Fan J, Clerkin SM, et al. (2010) Dopamine transporter gene variation modulates activation of striatum in youth with ADHD. Neuroimage 53: 935-942. 10.1016/j.neuroimage.2009.12.041; 10.1016/j.neuroimage.2009.12.041.

11. Belleville S, Clement F, Mellah S, Gilbert B, Fontaine F, et al. (2011) Training-related brain plasticity in subjects at risk of developing alzheimer's disease. Brain 134: 1623-1634.

12. Bird G, Silani G, Brindley R, White S, Frith U, et al. (2010) Empathic brain responses in insula are modulated by levels of alexithymia but not autism. Brain 133: 1515-1525.

13. Blau V, Reithler J, van Atteveldt N, Seitz J, Gerretsen P, et al. (2010) Deviant processing of letters and speech sounds as proximate cause of reading failure: A functional magnetic resonance imaging study of dyslexic children. Brain 133: 868-879.

14. Bonelli SB, Powell RH, Yogarajah M, Samson RS, Symms MR, et al. (2010) Imaging memory in temporal lobe epilepsy: Predicting the effects of temporal lobe resection. Brain 133: 1186-1199.

15. Brune M, Ozgurdal S, Ansorge N, von Reventlow HG, Peters S, et al. (2011) An fMRI study of "theory of mind" in at-risk states of psychosis: Comparison with manifest schizophrenia and healthy controls. Neuroimage 55: 329-337.

16. Calautti C, Jones PS, Guincestre JY, Naccarato M, Sharma N, et al. (2010) The neural substrates of impaired finger tapping regularity after stroke. Neuroimage 50: 1-6. 10.1016/j.neuroimage.2009.12.012; 10.1016/j.neuroimage.2009.12.012.

17. Campbell-Sills L, Simmons AN, Lovero KL, Rochlin AA, Paulus MP, et al. (2011) Functioning of neural systems supporting emotion regulation in anxiety-prone individuals. Neuroimage 54: 689-696.

18. Cantin S, Villien M, Moreaud O, Tropres I, Keignart S, et al. (2011) Impaired cerebral vasoreactivity to CO2 in alzheimer's disease using BOLD fMRI. Neuroimage 58: 579-587.

19. Celone KA, Thompson-Brenner H, Ross RS, Pratt EM, Stern CE. (2011) An fMRI investigation of the fronto-striatal learning system in women who exhibit eating disorder behaviors. Neuroimage 56: 1749-1757.

20. Chang Y, Lee JJ, Seo JH, Song HJ, Kim JH, et al. (2010) Altered working memory process in the manganese-exposed brain. Neuroimage 53: 1279-1285.

21. Chua HF, Ho SS, Jasinska AJ, Polk TA, Welsh RC, et al. (2011) Self-related neural response to tailored smoking-cessation messages predicts quitting. Nat Neurosci 14: 426-427.

22. Coman IL, Gnirke MH, Middleton FA, Antshel KM, Fremont W, et al. (2010) The effects of gender and catechol O-methyltransferase (COMT) Val108/158Met polymorphism on emotion regulation in velo-cardio-facial syndrome (22q11.2 deletion syndrome): An fMRI study. Neuroimage 53: 1043-1050.

23. de Guibert C, Maumet C, Jannin P, Ferre JC, Treguier C, et al. (2011) Abnormal functional lateralization and activity of language brain areas in typical specific language impairment (developmental dysphasia). Brain 134: 3044-3058. 10.1093/brain/awr141; 10.1093/brain/awr141.

24. Dinstein I, Thomas C, Humphreys K, Minshew N, Behrmann M, et al. (2010) Normal movement selectivity in autism. Neuron 66: 461-469.

25. Diwadkar VA, Pruitt P, Goradia D, Murphy E, Bakshi N, et al. (2011) Fronto-parietal hypo-activation during working memory independent of structural abnormalities: Conjoint fMRI and sMRI analyses in adolescent offspring of schizophrenia patients. Neuroimage 58: 234-241. 10.1016/j.neuroimage.2011.06.033; 10.1016/j.neuroimage.2011.06.033.

26. Dziobek I, Preissler S, Grozdanovic Z, Heuser I, Heekeren HR, et al. (2011) Neuronal correlates of altered empathy and social cognition in borderline personality disorder. Neuroimage 57: 539-548.

27. Emmorey K, Xu J, Gannon P, Goldin-Meadow S, Braun A. (2010) CNS activation and regional connectivity during pantomime observation: No engagement of the mirror neuron system for deaf signers. Neuroimage 49: 994-1005.

28. Erk S, Mikschl A, Stier S, Ciaramidaro A, Gapp V, et al. (2010) Acute and sustained effects of cognitive emotion regulation in major depression. J Neurosci 30: 15726-15734.

29. Freund P, Weiskopf N, Ward NS, Hutton C, Gall A, et al. (2011) Disability, atrophy and cortical reorganization following spinal cord injury. Brain 134: 1610-1622.

30. Germine LT, Garrido L, Bruce L, Hooker C. (2011) Social anhedonia is associated with neural abnormalities during face emotion processing. Neuroimage 58: 935-945.

31. Golarai G, Hong S, Haas BW, Galaburda AM, Mills DL, et al. (2010) The fusiform face area is enlarged in williams syndrome. J Neurosci 30: 6700-6712.

32. Gradin VB, Kumar P, Waiter G, Ahearn T, Stickle C, et al. (2011) Expected value and prediction error abnormalities in depression and schizophrenia. Brain 134: 1751-1764.

33. Grande M, Meffert E, Huber W, Amunts K, Heim S. (2011) Word frequency effects in the left IFG in dyslexic and normally reading children during picture naming and reading. Neuroimage 57: 1212-1220. 10.1016/j.neuroimage.2011.05.033; 10.1016/j.neuroimage.2011.05.033.

34. Greene DJ, Colich N, Iacoboni M, Zaidel E, Bookheimer SY, et al. (2011) Atypical neural networks for social orienting in autism spectrum disorders. Neuroimage 56: 354-362. 10.1016/j.neuroimage.2011.02.031; 10.1016/j.neuroimage.2011.02.031.

35. Greimel E, Schulte-Ruther M, Kircher T, Kamp-Becker I, Remschmidt H, et al. (2010) Neural mechanisms of empathy in adolescents with autism spectrum disorder and their fathers. Neuroimage 49: 1055-1065.

36. Heim S, Grande M, Meffert E, Eickhoff SB, Schreiber H, et al. (2010) Cognitive levels of performance account for hemispheric lateralisation effects in dyslexic and normally reading children. Neuroimage 53: 1346-1358. 10.1016/j.neuroimage.2010.07.009; 10.1016/j.neuroimage.2010.07.009.

37. Hu W, Lee HL, Zhang Q, Liu T, Geng LB, et al. (2010) Developmental dyslexia in chinese and english populations: Dissociating the effect of dyslexia from language differences. Brain 133: 1694-1706.

38. Ibarretxe-Bilbao N, Zarei M, Junque C, Marti MJ, Segura B, et al. (2011) Dysfunctions of cerebral networks precede recognition memory deficits in early parkinson's disease. Neuroimage 57: 589-597.

39. Jiang Z, Krainik A, David O, Salon C, Tropres I, et al. (2010) Impaired fMRI activation in patients with primary brain tumors. Neuroimage 52: 538-548.

40. Jollant F, Lawrence NS, Olie E, O'Daly O, Malafosse A, et al. (2010) Decreased activation of lateral orbitofrontal cortex during risky choices under uncertainty is associated with disadvantageous decision-making and suicidal behavior. Neuroimage 51: 1275-1281.

41. Kaiser MD, Hudac CM, Shultz S, Lee SM, Cheung C, et al. (2010) Neural signatures of autism. Proc Natl Acad Sci U S A 107: 21223-21228.

42. Kareken DA, Bragulat V, Dzemidzic M, Cox C, Talavage T, et al. (2010) Family history of alcoholism mediates the frontal response to alcoholic drink odors and alcohol in at-risk drinkers. Neuroimage 50: 267-276.

43. King GR, Ernst T, Deng W, Stenger A, Gonzales RM, et al. (2011) Altered brain activation during visuomotor integration in chronic active cannabis users: Relationship to cortisol levels. J Neurosci 31: 17923-17931. 10.1523/JNEUROSCI.4148-11.2011; 10.1523/JNEUROSCI.4148-11.2011.

44. Kleinhans NM, Richards T, Johnson LC, Weaver KE, Greenson J, et al. (2011) fMRI evidence of neural abnormalities in the subcortical face processing system in ASD. Neuroimage 54: 697-704. 10.1016/j.neuroimage.2010.07.037; 10.1016/j.neuroimage.2010.07.037.

45. Klinge C, Eippert F, Roder B, Buchel C. (2010) Corticocortical connections mediate primary visual cortex responses to auditory stimulation in the blind. J Neurosci 30: 12798-12805.

46. Klinge C, Roder B, Buchel C. (2010) Increased amygdala activation to emotional auditory stimuli in the blind. Brain 133: 1729-1736.

47. Krawitz A, Braver TS, Barch DM, Brown JW. (2011) Impaired error-likelihood prediction in medial prefrontal cortex in schizophrenia. Neuroimage 54: 1506-1517.

48. Kucian K, Grond U, Rotzer S, Henzi B, Schonmann C, et al. (2011) Mental number line training in children with developmental dyscalculia. Neuroimage 57: 782-795. 10.1016/j.neuroimage.2011.01.070; 10.1016/j.neuroimage.2011.01.070.

49. Kumari V, Fannon D, Peters ER, Ffytche DH, Sumich AL, et al. (2011) Neural changes following cognitive behaviour therapy for psychosis: A longitudinal study. Brain 134: 2396-2407.

50. Kupers R, Chebat DR, Madsen KH, Paulson OB, Ptito M. (2010) Neural correlates of virtual route recognition in congenital blindness. Proc Natl Acad Sci U S A 107: 12716-12721. 10.1073/pnas.1006199107; 10.1073/pnas.1006199107.

51. Lau JY, Goldman D, Buzas B, Hodgkinson C, Leibenluft E, et al. (2010) BDNF gene polymorphism (Val66Met) predicts amygdala and anterior hippocampus responses to emotional faces in anxious and depressed adolescents. Neuroimage 53: 952-961. 10.1016/j.neuroimage.2009.11.026; 10.1016/j.neuroimage.2009.11.026.

52. Li W, Howard JD, Gottfried JA. (2010) Disruption of odour quality coding in piriform cortex mediates olfactory deficits in alzheimer's disease. Brain 133: 2714-2726.

53. Lombardo MV, Chakrabarti B, Bullmore ET, MRC AIMS C, Baron-Cohen S. (2011) Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. Neuroimage 56: 1832-1838.

54. Lombardo MV, Chakrabarti B, Bullmore ET, Sadek SA, Pasco G, et al. (2010) Atypical neural self-representation in autism. Brain 133: 611-624.

55. Longe O, Maratos FA, Gilbert P, Evans G, Volker F, et al. (2010) Having a word with yourself: Neural correlates of self-criticism and self-reassurance. Neuroimage 49: 1849-1856.

56. Lueken U, Kruschwitz JD, Muehlhan M, Siegert J, Hoyer J, et al. (2011) How specific is specific phobia? different neural response patterns in two subtypes of specific phobia. Neuroimage 56: 363-372.

57. Luijten M, Veltman DJ, van den Brink W, Hester R, Field M, et al. (2011) Neurobiological substrate of smoking-related attentional bias. Neuroimage 54: 2374-2381. 10.1016/j.neuroimage.2010.09.064; 10.1016/j.neuroimage.2010.09.064.

58. Ma Y, Han S. (2011) Neural representation of self-concept in sighted and congenitally blind adults. Brain 134: 235-246.

59. Maurer U, Schulz E, Brem S, der Mark S, Bucher K, et al. (2011) The development of print tuning in children with dyslexia: Evidence from longitudinal ERP data supported by fMRI. Neuroimage 57: 714-722.

60. Meulenbroek O, Rijpkema M, Kessels RP, Rikkert MG, Fernandez G. (2010) Autobiographical memory retrieval in patients with alzheimer's disease. Neuroimage 53: 331-340.

61. Miyake Y, Okamoto Y, Onoda K, Shirao N, Okamoto Y, et al. (2010) Neural processing of negative word stimuli concerning body image in patients with eating disorders: An fMRI study. Neuroimage 50: 1333-1339.

62. Mizuno A, Liu Y, Williams DL, Keller TA, Minshew NJ, et al. (2011) The neural basis of deictic shifting in linguistic perspective-taking in high-functioning autism. Brain 134: 2422-2435.

63. Mohr HM, Roder C, Zimmermann J, Hummel D, Negele A, et al. (2011) Body image distortions in bulimia nervosa: Investigating body size overestimation and body size satisfaction by fMRI. Neuroimage 56: 1822-1831.

64. Nestor L, McCabe E, Jones J, Clancy L, Garavan H. (2011) Differences in "bottom-up" and "top-down" neural activity in current and former cigarette smokers: Evidence for neural substrates which may promote nicotine abstinence through increased cognitive control. Neuroimage 56: 2258-2275.

65. Newman AJ, Supalla T, Hauser PC, Newport EL, Bavelier D. (2010) Prosodic and narrative processing in american sign language: An fMRI study. Neuroimage 52: 669-676.

66. Oertel V, Knochel C, Rotarska-Jagiela A, Schonmeyer R, Lindner M, et al. (2010) Reduced laterality as a trait marker of schizophrenia--evidence from structural and functional neuroimaging. J Neurosci 30: 2289-2299.

67. Papagni SA, Mechelli A, Prata DP, Kambeitz J, Fu CH, et al. (2011) Differential effects of DAAO on regional activation and functional connectivity in schizophrenia, bipolar disorder and controls. Neuroimage 56: 2283-2291.

68. Plotkin A, Sela L, Weissbrod A, Kahana R, Haviv L, et al. (2010) Sniffing enables communication and environmental control for the severely disabled. Proc Natl Acad Sci U S A 107: 14413-14418. 10.1073/pnas.1006746107; 10.1073/pnas.1006746107.

69. Pompei F, Jogia J, Tatarelli R, Girardi P, Rubia K, et al. (2011) Familial and disease specific abnormalities in the neural correlates of the stroop task in bipolar disorder. Neuroimage 56: 1677-1684.

70. Raja Beharelle A, Dick AS, Josse G, Solodkin A, Huttenlocher PR, et al. (2010) Left hemisphere regions are critical for language in the face of early left focal brain injury. Brain 133: 1707-1716. 10.1093/brain/awq104; 10.1093/brain/awq104.

71. Roberts GM, Garavan H. (2010) Evidence of increased activation underlying cognitive control in ecstasy and cannabis users. Neuroimage 52: 429-435.

72. Roussotte FF, Bramen JE, Nunez SC, Quandt LC, Smith L, et al. (2011) Abnormal brain activation during working memory in children with prenatal exposure to drugs of abuse: The effects of methamphetamine, alcohol, and polydrug exposure. Neuroimage 54: 3067-3075.

73. Schacht JP, Anton RF, Randall PK, Li X, Henderson S, et al. (2011) Stability of fMRI striatal response to alcohol cues: A hierarchical linear modeling approach. Neuroimage 56: 61-68.

74. Schonberg T, O'Doherty JP, Joel D, Inzelberg R, Segev Y, et al. (2010) Selective impairment of prediction error signaling in human dorsolateral but not ventral striatum in parkinson's disease patients: Evidence from a model-based fMRI study. Neuroimage 49: 772-781.

75. Schweckendiek J, Klucken T, Merz CJ, Tabbert K, Walter B, et al. (2011) Weaving the (neuronal) web: Fear learning in spider phobia. Neuroimage 54: 681-688.

76. Sepede G, Ferretti A, Perrucci MG, Gambi F, Di Donato F, et al. (2010) Altered brain response without behavioral attention deficits in healthy siblings of schizophrenic patients: An event-related fMRI study. Neuroimage 49: 1080-1090.

77. Shilyansky C, Karlsgodt KH, Cummings DM, Sidiropoulou K, Hardt M, et al. (2010) Neurofibromin regulates corticostriatal inhibitory networks during working memory performance. Proc Natl Acad Sci U S A 107: 13141-13146.

78. Siniatchkin M, Groening K, Moehring J, Moeller F, Boor R, et al. (2010) Neuronal networks in children with continuous spikes and waves during slow sleep. Brain 133: 2798-2813. 10.1093/brain/awq183; 10.1093/brain/awq183.

79. Stankewitz A, Aderjan D, Eippert F, May A. (2011) Trigeminal nociceptive transmission in migraineurs predicts migraine attacks. J Neurosci 31: 1937-1943. 10.1523/JNEUROSCI.4496-10.2011; 10.1523/JNEUROSCI.4496-10.2011.

80. Stice E, Yokum S, Blum K, Bohon C. (2010) Weight gain is associated with reduced striatal response to palatable food. J Neurosci 30: 13105-13109.

81. Stice E, Yokum S, Burger KS, Epstein LH, Small DM. (2011) Youth at risk for obesity show greater activation of striatal and somatosensory regions to food. J Neurosci 31: 4360-4366.

82. Subramanian L, Hindle JV, Johnston S, Roberts MV, Husain M, et al. (2011) Real-time functional magnetic resonance imaging neurofeedback for treatment of parkinson's disease. J Neurosci 31: 16309-16317. 10.1523/JNEUROSCI.3498-11.2011; 10.1523/JNEUROSCI.3498-11.2011.

83. Surguladze SA, Marshall N, Schulze K, Hall MH, Walshe M, et al. (2010) Exaggerated neural response to emotional faces in patients with bipolar disorder and their first-degree relatives. Neuroimage 53: 58-64.

84. Tadic SD, Griffiths D, Murrin A, Schaefer W, Aizenstein HJ, et al. (2010) Brain activity during bladder filling is related to white matter structural changes in older women with urinary incontinence. Neuroimage 51: 1294-1302.

85. Toyomura A, Fujii T, Kuriki S. (2011) Effect of external auditory pacing on the neural activity of stuttering speakers. Neuroimage 57: 1507-1516.

86. Umarova RM, Saur D, Kaller CP, Vry MS, Glauche V, et al. (2011) Acute visual neglect and extinction: Distinct functional state of the visuospatial attention system. Brain 134: 3310-3325.

87. van der Mark S, Klaver P, Bucher K, Maurer U, Schulz E, et al. (2011) The left occipitotemporal system in reading: Disruption of focal fMRI connectivity to left inferior frontal and inferior parietal language areas in children with dyslexia. Neuroimage 54: 2426-2436.

88. van Oers CA, Vink M, van Zandvoort MJ, van der Worp HB, de Haan EH, et al. (2010) Contribution of the left and right inferior frontal gyrus in recovery from aphasia. A functional MRI study in stroke patients with preserved hemodynamic responsiveness. Neuroimage 49: 885-893.

89. Voon V, Gao J, Brezing C, Symmonds M, Ekanayake V, et al. (2011) Dopamine agonists and risk: Impulse control disorders in parkinson's disease. Brain 134: 1438-1446.

90. Wagner DD, Dal Cin S, Sargent JD, Kelley WM, Heatherton TF. (2011) Spontaneous action representation in smokers when watching movie characters smoke. J Neurosci 31: 894-898.

91. Wang L, Metzak PD, Honer WG, Woodward TS. (2010) Impaired efficiency of functional networks underlying episodic memory-for-context in schizophrenia. J Neurosci 30: 13171-13179.

92. Wang WC, Lazzara MM, Ranganath C, Knight RT, Yonelinas AP. (2010) The medial temporal lobe supports conceptual implicit memory. Neuron 68: 835-842.

93. Wildenberg JC, Tyler ME, Danilov YP, Kaczmarek KA, Meyerand ME. (2011) High-resolution fMRI detects neuromodulation of individual brainstem nuclei by electrical tongue stimulation in balance-impaired individuals. Neuroimage 56: 2129-2137.

94. Wilson SM, Dronkers NF, Ogar JM, Jang J, Growdon ME, et al. (2010) Neural correlates of syntactic processing in the nonfluent variant of primary progressive aphasia. J Neurosci 30: 16845-16854. 10.1523/JNEUROSCI.2547-10.2010; 10.1523/JNEUROSCI.2547-10.2010.

95. Wolbers T, Zahorik P, Giudice NA. (2011) Decoding the direction of auditory motion in blind humans. Neuroimage 56: 681-687. 10.1016/j.neuroimage.2010.04.266; 10.1016/j.neuroimage.2010.04.266.

96. Wu T, Chan P, Hallett M. (2010) Effective connectivity of neural networks in automatic movements in parkinson's disease. Neuroimage 49: 2581-2587.

97. Wu T, Wang L, Hallett M, Li K, Chan P. (2010) Neural correlates of bimanual anti-phase and in-phase movements in parkinson's disease. Brain 133: 2394-2409.

98. Yassa MA, Stark SM, Bakker A, Albert MS, Gallagher M, et al. (2010) High-resolution structural and functional MRI of hippocampal CA3 and dentate gyrus in patients with amnestic mild cognitive impairment. Neuroimage 51: 1242-1252.

99. You H, Gaab N, Wei N, Cheng-Lai A, Wang Z, et al. (2011) Neural deficits in second language reading: FMRI evidence from chinese children with english reading impairment. Neuroimage 57: 760-770. 10.1016/j.neuroimage.2010.12.003; 10.1016/j.neuroimage.2010.12.003.

100. Zempleni MZ, Michels L, Mehnert U, Schurch B, Kollias S. (2010) Cortical substrate of bladder control in SCI and the effect of peripheral pudendal stimulation. Neuroimage 49: 2983-2994.

CHAPTER FIVE

CONCLUSIONS

With a growing literature in functional MRI studies, there are some issues around sample size and power calculations. Specifically, a subset of issues includes: (1) documenting reported input parameters necessary to a sample size calculation in the results section, (2) developing an effective strategy to deal with uncertainty in input parameters and to set a sample size with maximum returns per unit cost, and (3) suggesting a strategy to help improve overall quality of reporting and hence to reduce bias in estimated input parameters leading to an effective future sample size calculation in the long run.

These issues have been investigated in a manuscript-based thesis, with each paper addressing one of the three issues. The three papers make contributions as follows: (1) identifying the inadequate reporting of observed input data in the results section as a potential impediment to implement sample size and power calculations, (2) illustrating how cost-efficiency can be used in conjunction with conventional methods as a short-term strategy to calculate an optimal sample size, and (3) providing empirical evidence to highlight the room for improvement in the current reporting of observational clinical fMRI studies and making recommendations on how to improve the overall quality of reporting as a long term strategy to facilitate a future sample size calculation. In this chapter, the key study findings, practical implications, and limitations for each paper are summarized.

Chapter 2 provides empirical evidence of the lack of reported sample size calculations in the literature, points to important impediments to conducting these calculations, and recommends strategies to reduce the potential bias associated with the reported results. Of the 100 studies included in our systematic review, one reported a sample size calculation. Furthermore, implementing these calculations requires estimates of effect sizes and variance components. However, we found that reporting of these data is uncommon. Therefore, routine reporting of observed input estimates in the results section would facilitate sample size calculations for future studies. Moreover, to ensure that the calculations are valid and reliable, bias associated with the reported data (e.g., using reported peak *t*, *z*, or *F* statistic in statically significant regions, regions of interest (ROIs) defined in a dependent fashion, selective reporting, and publication bias) should be diminished.

This is the first systematic review to evaluate the reporting of necessary information in the results section for investigators to carry out a power analysis for new studies. The practical implications of this work are as follows. First, there is a need for routine reporting of observed effect sizes and variance components either in the results section or online supplement if publication space is limited, in order to facilitate well-powered sample size calculations for future studies. Second, we have made recommendations to minimize bias in the reported results, including but not limited to, defining ROIs either anatomically using an anatomical atlas or functionally using independent data such as pilot data to avoid biased selection of ROIs, and reporting results from both negative and

positive regions to reduce reporting bias. Third, investigators who are conducting observational fMRI studies can consider using standard guidelines such as the STROBE statement (von Elm et al., 2007) to guide their reporting for observational studies, especially in terms of how sample size was arrived at, methods used to control for confounding variables, precise and detailed description of sampling strategy (e.g., convenience or random sample) and participant recruiting process. Including these components in their reports helps check representativeness of the study sample, validity or generalizability of findings, identify bias and enhance interpretation and replication. Fourth, editorial policies can be updated to encourage reporting observed input parameters in the manuscript, endorsing and using the guidelines in journals.

This project has several limitations: First, this study is conducted on the basis of six top journals. The findings may not apply to studies published in a wider range of journals. Second, it does not rule out the possibility that power-based sample size calculations were conducted but simply not reported. However, it is difficult to make such judgment in this study.

Chapter 3 demonstrates through an fMRI case study how the cost-efficiency method supplements conventional methods in the face of substantial uncertainty in input parameters. The innovative nature of the case study and limited reporting of input estimates (i.e., effect sizes, within- and between-subject variances, and temporal autocorrelation structures) result in a wide range of candidate sample sizes (e.g., 50 to 800 in this case study) using a sensitivity analysis and a conventional sample size method

(Mumford and Nichols, 2008) at a desired statistical power of 80% and a significance level of 5%. Cost-efficiency provides a way of narrowing down the range and choosing a sample size among the range on the basis of maximum return upon investment. This work provides a new direction for investigators in designing studies to be cost efficient and methodologically sound; that is, cost-efficiency methodology accommodates cost into calculating a sample size so as to maximize the information gained from the study per dollar spent, while ensuring the new study reaches a desired statistical power to detect at least one of the proposed effect sizes. The implications of this work are as follows. First, the cost-efficiency approach gives a sample size which is more cost efficient than any larger choice, meaning that cost-efficiency provides an upper bound of the sample size for investigators to consider. Second, rather than a stand-alone method, cost-efficiency supplements traditional power analyses to narrow down the range of candidate sample sizes and target a sample size among them by considering both cost and statistical power.

This work has the following limitations. First, the financial cost must be estimated precisely in order to obtain a valid sample size on the basis of cost-efficiency calculations. Other costs such as social cost and opportunity cost beyond financial cost are not considered. Second, the cost efficiency approach does not consider whether study value is less than study cost, a situation under which a new study should not be undertaken. Finally, cost efficiency could yield a sample size beyond the budget, a challenge which is shared by many other methods for power calculations.

Chapter 4 evaluates the overall reporting of observational clinical fMRI studies through a systematic review, highlights key components that are insufficiently reported, and suggests creating a shortened list of core items. Of the 83 items adapted from the guidelines for fMRI studies (Poldrack et al., 2008) and used for data abstraction in this review, around one third of the items are uncommonly reported. These items cover domains of experimental design, imaging properties, data preprocessing, statistical modeling, statistical methods to correct for multiplicity, and statistical inferences on ROI analysis. Incomplete reporting of these methodological details would hinder evaluating methodological rigor, understanding and replicating results. Sample size calculation is one among many key methodologies in reporting fMRI studies. Increasing the overall quality of reporting helps implement a rigorous sample size calculation in the long run.

This is the first attempt to evaluate the quality of reporting in observational fMRI studies involving clinical participants. The challenge of reporting clinical fMRI studies points to a need for appropriate guidelines to aid a better reporting and to allow comparisons between studies. Adhering to the guidelines proposed by Poldrack and his colleagues could help meet the need. However, reporting guidelines are evolving and need continual updates. The practical concerns relating to lengthy guidelines and strict word limits for publication suggest constructing a shortened version of Poldrack's checklist. We suggest removing seven items from Poldrack's checklist given these items involve much subjectivity. We hope that our suggestion could help lead to an updated reporting guideline and to discussion of what is indeed essential to report under tight

space limitations. There are two main limitations for this study. First, findings in this study reflect the quality of reporting in six top journals, results which may not be generalized to journals with a low profile. Second, several items evaluated in this systematic review involve subjectivity. Using duplicate review and solving disagreement through consensus helps increase interpretability and understandability of the items.

Once a better quality of reporting practice becomes routine, reliable sample size calculations can be performed to yield adequately powered studies. The next step is to disseminate the knowledge from this evidence-based thesis research to as wide an audience as possible in fMRI community and end users (e.g., health practitioners, patients, editors, and policy makers) including but not limited to the following activities and media. First and foremost, we will present this work in conferences, especially in neuroimaging or fMRI-specific symposia or colloquia, and get them published in peer-reviewed journals to make these findings accessible to the fMRI community. Secondly, I myself or my colleagues could, in a collaborative context, to use the cost-efficiency methodology to calculate sample sizes for future studies. Through collaborations, we hope to introduce the methodology into practice. Third, I can make my written R codes for calculating sample sizes using cost-efficiency available and free for all interested users; moreover this could eventually be built into existing power analysis software packages that are either free (e.g., *G*Power*) or paid (e.g., *PASS*, *nQuery advisor*, *UnifyPow*, and *Power and Precision*). Fourth, I may inform journal editors, e.g., by addressing them a letter, about the empirical evidence and findings of research and

suggest updating the editorial policy, e.g., endorsing and adhering to the guidelines developed by Poldrack et al. and requiring authors to provide observed input estimates in the results section.

References

Mumford, J.A., Nichols, T.E., 2008. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. Neuroimage 39, 261-268.

Poldrack, R.A., Fletcher, P.C., Henson, R.N., Worsley, K.J., Brett, M., Nichols, T.E., 2008. Guidelines for reporting an fMRI study. Neuroimage 40, 409-414.

von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gotzsche, P.C., Vandenbroucke, J.P., STROBE, I., 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann. Intern. Med. 147, 573-577.