NeuroImage 86 (2014) 172–181

Contents lists available at ScienceDirect

NeuroImage

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

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

ARTICLE INFO

Article history:
Accepted 5 August 2013
Available online 15 August 2013

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

ABSTRACT

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. 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. 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* statistics. The inter-rater agreement was very good, with a prevalence-adjusted bias-adjusted kappa (PABAK) value greater than 0.88. We concluded that 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.

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 (Carp, 2012a; Simmons et al., 2011). 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, and 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 the i) size of effect to be detected (e.g., mean activation or percent signal

* Corresponding author at: Department of Epidemiology and Biostatistics, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada.
E-mail address: guoq@mcmaster.ca (Q. Guo).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.08.012

change between two conditions), ii) between-subject variance, iii) within-subject variance, and iv) temporal auto-correlation variance–covariance matrix. In other fields, input parameters for sample size calculations are often estimated from previously published studies (Lenth, 2001; Wilkinson and Task Force Stat Inference, 1999; 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 (Glahn et al., 2005; Monk et al., 2008; Sheline et al., 2001; Siegle et al., 2002; Snitz et al., 2005; 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.

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.

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.

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.

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).

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), $\sigma_w{}^2$ (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.

Hayasaka et al. (2007)

Hayasaka et al. (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

Table 1
Sample size approaches and parameters required to report.

Approach	Parameters
Mumford & Nichols (*Neuroimage*. 2008; 39 (1): 261–268)	Δ, $\sigma_w{}^2$, σ_b^2, *V*
Desmond & Glover (*J. Neurosci. Methods*. 2002; 118 (2): 115–128)	Δ, $\sigma_w{}^2$, 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); $\sigma_w{}^2$ (within-subject variance); *V* (temporal autocorrelation matrix); σ_b^2 (between-subject variance).