doi:10.1093/scan/nss059 SCAN (2012) 7, 738–742

A power calculation guide for fMRI studies

Jeanette A. Mumford
Department of Psychology, University of Texas at Austin, University Station A8000, Austin, TX 7812-0187, USA

In the past, power analyses were not that common for fMRI studies, but recent advances in power calculation techniques and software development are making power analyses much more accessible. As a result, power analyses are more commonly expected in grant applications proposing fMRI studies. Even though the software is somewhat automated, there are important decisions to be made when setting up and carrying out a power analysis. This guide provides tips on carrying out power analyses, including obtaining pilot data, defining a region of interest and other choices to help create reliable power calculations.

Keywords: functional magnetic resonance imaging; classification analysis; MVPA; beta series estimation; rapid event-related design

INTRODUCTION

When running a functional magnetic resonance imaging (fMRI) experiment, we hope that our data have signal related to our task of interest and, more importantly, that we have collected enough data to detect this signal. The ability to detect an effect, when present, is referred to as statistical power and we typically aim for power of 80% or higher. The interpretation of 80% power is if we were to repeat our study 100 times, and the signal truly existed, it would be detected in 80 of the studies. Unlike other statistical analyses, a power analysis needs to be performed prior to collecting data and is used as a study planning tool. Most commonly, power analyses are included in a grant proposals. Although different strategies for fMRI power analyses have existed since the early 2000s, the earlier approaches either required lengthy simulations (Desmond and Glover, 2002) or were too complicated for a non-statistician to apply (Hayasaka *et al.*, 2007; Mumford and Nichols, 2008). The recently developed software package, fMRIPower (fmripower.org), accomplishes region of interest (ROI) power calculations based on a simplified version of the methods in Mumford and Nichols (2008). fMRIPower and the upcoming power analysis software described in Joyce and Hayasaka (2011) make it possible for any investigator to perform power analyses. This is of great use as power analyses are more commonly required in fMRI-based grant applications and should help reduce the number of underpowered fMRI studies in the future.

The power analysis model described in Mumford and Nichols (2008) has the flexibility to calculate power according to both how many subjects you include and how many stimuli (or blocks of stimuli) are presented in a run. Changing the number of stimuli requires the generation of new first-level design matrices, which is difficult to automate in a software package. Therefore, fMRIPower only allows for power calculation as a function of the overall sample size. The power analysis assumes that the future data will use the same stimulus presentation and number of runs as the pilot data used to drive the power analysis. This Matlab-based software uses pilot analysis output from the FSL (www.fmrib.ox.ac.uk/fsl/) or SPM (www.fil.ion.ucl.ac.uk/spm/) software packages and automatically extracts the data necessary for a power calculation. fMRIPower can calculate power for one-sample, two-sample and paired *t*-tests. The software will automatically detect what type of analysis was originally run on your pilot data and you will only need to specify the number of subjects for which you would like to calculate power and the ROI for the power analysis. Although it could not be automated for general analysis of variance (ANOVA) models, most of the contrasts of interest from an ANOVA model can be put into the two-sample *t*-test framework and then used in fMRIPower. Of course, it could be the case that your future data will not exactly match the study design used for the pilot data and in this case, your power analysis may not be accurate. For example, you may expect your future study to have higher variability. By using the output from fMRIPower you can alter values of the mean or variance of the activation and study how this impacts power.

Sometimes it is difficult to know how to get started with a power analysis and the purpose of this work is to provide some guidelines for the process and tips for producing a power calculation that more closely represents your future data. This information can be used to perform power calculations with fMRIPower and, when released, the power calculation tool described in Joyce and Hayasaka (2011). Next is a brief overview of how power is calculated. Then, the two most important pieces of information, the ROI and pilot data, are discussed in detail. The pitfalls of a power analysis as well as the lesser realized benefits of running a power analysis are discussed. In addition, it is important to realize that power analyses are only appropriate when predicting the power of a *future* study, which differs from the misguided idea of *post hoc* power. This is typically done in an attempt to estimate power for a study that has already occurred, perhaps to build a case for the null or explain that a result was not found due to a lack in power. Lastly, a brief overview of the steps you should take when performing a power analysis is discussed.

REVIEW OF POWER ANALYSIS

The definition of statistical power is the probability of rejecting the null hypothesis, given that the alternative hypothesis is true. So, in order to calculate power, we need to know our statistic threshold for rejecting the null hypothesis, based on the null distribution, and then we would calculate the probability of being larger than this threshold according to the alternative distribution. This generally requires four pieces of information: the mean of the activation, its variance, the Type I error rate and sample size. Furthermore, for fMRI, we also need a brain ROI to focus on for the power analysis. Figure 1 illustrates power calculations for three different alternative distributions. In each case, we must first specify the null distribution, which is shown in red. Typically, this is centered at 0 and the variance is based on a previously obtained estimate and the proposed sample size. If your statistic follows the normal distribution, your null would be $N(0,\sigma^2/n)$ and the alternative would be $N(\mu,\sigma^2/n)$, where n is the proposed sample size, μ and σ are the mean and s.d. of the future data. Without loss of

Received 13 April 2012; Accepted 22 May 2012
Advance Access publication 28 May 2012

Correspondence should be addressed to Jeanette A. Mumford, Department of Psychology, University of Texas at Austin 1 University Station A8000 Austin, TX, 78712-0187, USA. E-mail: mumford@mail.utexas.edu

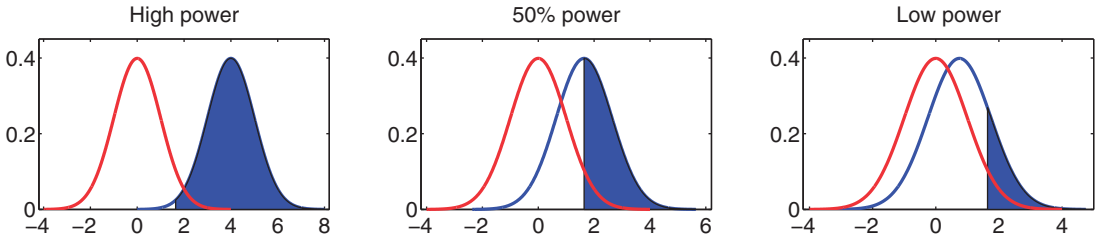

Fig. 1 Examples of three different levels of power, assuming a normal distribution. In each case, the null distribution is centered at 0 and the alternative is centered at $\sqrt{n}\mu/\sigma$, both with a variance of 1. A statistic threshold of 1.64 controls the Type I error rate at 5% for a one-sided hypothesis. In all cases, power is the area under the alternative distribution for statistic values larger than the threshold of 1.64. The left panel shows a high power example. The middle panel illustrates that if the mean of the distribution is exactly the threshold, the power is 50%. Lastly, the right panel shows a case where power is very low.

generality, we can standardize these statistics by dividing by the s.d. and these are the distributions shown in Figure 1, where the null is $N(0,1)$ and the alternative is $N(\sqrt{n}\mu/\sigma, 1)$. Then, according to the null distribution and Type I error rate, the statistic threshold that defines significance is specified, in this case, a threshold of 1.64 is used, as it corresponds to a *P*-value of 0.05 for a one-sided test based on the standard normal distribution. The area under the alternative distribution to the right of this threshold is the power. Since the mean of the alternative is defined as $\sqrt{n}\mu/\sigma$, increasing n or μ, or decreasing σ will increase the mean of the alternative distribution and, therefore, the power. In addition, one could increase power by decreasing the statistic threshold used to define significance, but this comes at the cost of an increase in the Type I error rate, which is not recommended.

PILOT DATA

The above examples of power analyses assumed that we knew the true mean and variance, μ and σ^2, but in reality, we must estimate these values. Commonly, funds are not available to run a pilot study to supply data for a power analysis, but, if you can do so, it is highly recommended. In addition to acquiring data for your power calculation, you may possibly uncover important improvements that can be made in your study design and modeling strategy. How many subjects should be included in your pilot study is not exactly clear, perhaps somewhere between 6 and 10 subjects for a one-sample *t*-test. If you are running a task for which you expect the magnitude of the signal to be small or the variance to be large, possibly due to working with a patient population, you will need more subjects in your pilot data. A two-sample *t*-test generally requires more subjects per group for a pilot study since the effects are typically small and the variance is larger than a one-sample *t*-test. You basically need an estimate of the size and variance of your effect and small sample sizes will produce highly variable estimates of both. Still, it will be closer than what you would have had without pilot data and you can always test how much the power changes if you decrease the effect size a little or increase the variance. Hopefully, for the sample size you choose, the power will not change too much with these small changes. If, however, the power changes greatly you may want to increase the proposed sample size of your future study.

If you do not have any pilot data and you are thinking of your power analysis well in advance of your grant submission deadline, you can try to obtain data from other research groups. Power analyses are a good motivation for why our data should be made public when we are done running our own analyses. Examples of fMRI databases are the Open fMRI Project (openfmri.org) and the open fMRI Data Center (http://fmridc.org/f/fmridc). Open fMRI is a newer database, which currently supplies 12 datasets (220 subjects total) online and is expected to add more in the near future. The fMRI Data Center has 107 datasets which may be obtained by request. These two resources supply whole brain data, which is necessary for most power analyses, whereas other databases, such as the Brain Map database (www.brainmap.org), only supply coordinates indicating the active voxels. Granted, these coordinates could be used to form ROIs for the power analysis using independent data.

The last resort is obtaining effect sizes and variance estimates from results published in articles. This is, by far, is the least desirable approach. Primarily because most articles only report significant activation and obviously these will have high power estimates according to your calculation. A power analysis is not supplying any new information in this case and you can simply state that with a similar sample size you would hope to find a similar effect without running an actual power analysis. Of course, if the effect sizes reported came from maximum statistic voxels or biased ROIs, they will not be useful at all in a power analysis. The only time this would be acceptable is if a non-significant result was reported using an unbiased ROI.

A very important consideration, if you have run a power analysis using pilot data, is that these data can only be used for the purposes of your pilot study and cannot be combined with your future data to perform the final analysis. If you use your pilot data to estimate the mean and variance of the effect, or even just the variance (a mean is assumed a priori) combining your pilot data with more data will increase your Type I error rate. This is because your sample size, which is normally a fixed value, becomes a random variable that is dependent on the distribution of your pilot data. Since this newly introduced sample size variability is not accounted for in our standard analyses, the Type I error rate increases. Consider the case of a one-sided, one-sample *t*-test where the true mean is 1, true s.d. is 2.5 and our goal Type I error rate is 0.05. For 80% power, this study would require 41 subjects. In reality, we do not know the true values. Let us assume our guess of the mean is correct ($\mu = 1$) and we collect data from five subjects to obtain a variance estimate and run a power analysis to determine how many more subjects we need. Lastly, we collect only as many more subjects as we would need to meet this sample size and analyze all of the data, *including* the five pilot subjects. Based on 5000 simulated datasets, the estimated true Type I error rate, when thresholding your statistics based on a 0.05 threshold, is actually 0.0614. In other words, although we were thresholding our statistic such that we would have a Type I error of 5%, our true Type I error is actually 6.14%. Although our power is higher, this comes at the cost of an increase in the number of false positives that will result in the group analysis. The inflation in the Type I error rate will vary according to how large your pilot study is with respect to the final total sample size, the type of test you are running and other factors. In the simulation work by Wittes *et al.* (1999) for a two-sample *t*-test, the Type I error ranged up to 0.08 and they recommended that this practice only be used in very large studies (hundreds of subjects) as the bias is negligible in this case. Not only is the Type I error bias unpredictable, no study has been conducted to characterize this bias in fMRI studies where hundreds of thousands of tests are performed within a single brain.

It is unknown whether the bias would be larger or small when multiple comparison correction is used. Thus, this reuse of pilot data should not be practiced for our fMRI studies.

CHOOSING A ROI

The ROI should be chosen carefully for a power analysis. However, the ROI is chosen, we need to ensure that it was not done in a manner that would bias the effect size. Most often this occurs because the ROI was selected based on significant activations in the pilot data. This topic has been thoroughly discussed from the standpoint of Type I error rate inflation by Kriegeskorte *et al.* (2009) and Vul *et al.* (2009), who illustrated that defining ROIs in a biased fashion can lead to overestimates of effect sizes that are driven by noise in the data. Another important point made by Yarkoni (2009) is, due to the small sample sizes we tend to use in fMRI studies, the effects found to be significant are necessarily large. For example, for a one-sample *t*-test with 10 subjects, the effect size (μ/σ) must be at least 0.58 to reject the null, whereas for a sample size of 20, an effect size of 0.39 is required to reject the null. Hence, the effect sizes from significant findings in studies with small samples run the risk of being much larger than the true mean of the alternative distribution. In other words, this sample likely came from the upper tail of the alternative distribution. Yarkoni (2009) suggests that sample sizes of 20 subjects used in many current fMRI studies are much too small and sample sizes of 50 or larger are most likely more appropriate to detect the effects in fMRI. Therefore, if you select an unbiased ROI, do not be surprised if a power analysis suggests you need 50 or more subjects as this could be a more realistic sample size for imaging studies.

Note that the proposed fMRI power analysis project of Joyce and Hayasaka (2011) is a voxelwise approach that uses voxelwise multiple comparison correction through random field theory. In this case, it may be tempting to skip ROI selection, as the power maps are corrected for multiple comparisons, but, especially with small sample sizes, it is possible for large effects to be present in the data that are driven by noise. To help prevent noise from driving your power analysis, choose an ROI independent of your data prior to looking at the power map.

If you have multiple ROIs you will want to adjust your Type I error accordingly. A Bonferroni correction according to the number of ROIs you are investigating should be adequate (use $0.05/n$, where n is the number of ROIs). Bonferroni tends to be conservative, but in the case of power, it is better to be slightly conservative in your estimates.

If, prior to collecting the pilot data, you had ROIs, use those for the power analysis. Or, another option is to look through similar published studies and use the regions, or a combination of regions, from the work of others for your power analysis. Do not limit yourself to a single ROI, but perhaps a couple of reasonable regions and run multiple power analyses to obtain a range of sample sizes. If you have enough data, I recommend randomly splitting the data into two halves, using one half to define the ROI and the other half for the power analysis.

WHEN A POWER ANALYSIS IS NOT APPROPRIATE

More often than we would like, the results of our data analysis are not what we would expect and we are left wondering whether there is no effect (the null hypothesis is true) or if the power was too low to detect the effect. Due to the relationship between the Type I error rate and power, running a power calculation on a dataset to predict the power of *that* study is not informative (Hoenig and Heisey, 2001; Levine and Ensom, 2001). These types of power analyses are often referred to as *posthoc* power analyses or *observed* power. Unfortunately, popular statistical software packages, such as SPSS (http://www-01.ibm.com/software/analytics/spss/), often report observed power estimates and journal reviewers will sometimes request them. Power is the probability of rejecting the null hypothesis given the true activation is of a particular magnitude. There is no way to determine whether our observed effect size resulted from a sample from a null distribution or an alternative distribution with a non-zero mean. It is analogous to being told that the number 2 was randomly sampled from the null or alternative distribution and you now must specify what distribution this single observation came from (null or alternative). Clearly, you would need more data sampled from that same distribution to make this sort of conclusion.

In addition, the *P*-value and observed power have a relationship such that if you fail to reject the null hypothesis your observed power, based on your estimated effect size, is guaranteed to be low and hence is not informative. This is shown in the middle and right panels of Figure 1. If you fail to reject the null and then use this effect size to define the mean of your alternative distribution, assuming symmetric distributions the power will be 50% when the statistic equals the threshold (middle panel) or smaller (right panel). This does not give us any information about our true alternative distribution. In this sense, a *posthoc* power analysis is not informative at all, high *P*-values always imply low observed power. Note for non-symmetric distributions (such as the noncentral *t*) and for two-sided hypothesis, the power may be > 50% when the null is not rejected. All possible arguments for *post hoc* power are clearly refuted in Hoenig and Heisey (2001).

What can you do if you do not detect anything significant for one of your hypotheses and reviewers request a *posthoc* power analysis? Since it is not informative and there are plenty of references explaining why (Goodman and Berlin, 1994; Hoenig and Heisey, 2001; Levine and Ensom, 2001), you can respectfully decline to run a *posthoc* power analysis. One alternative, which is often suggested, is to supply a confidence interval for the estimate (Goodman and Berlin, 1994). The reason behind this is that the confidence interval provides intuition for the range of values supported by the data that we do have.

BENEFITS OF A POWER ANALYSIS

Power analyses are not an exact science, meaning following the suggested sample size from a power analysis does not guarantee you will have a significant result in your study. One of two things can prevent this. First, our estimate of, $\hat{\mu}/\hat{\sigma}$, may be much larger than the truth, causing our power to be lower than we had hoped (true alternative is shifted to the left). In addition, even if our estimates are perfect, 80% power means that one out of five replications of our study will fail to find a result (your study's result falls in the lower tail of the alternative distribution). Due to this, some are skeptical about whether or not power analyses are useful. If we are merely making an educated guess that may be wrong, why bother? Surprisingly, even though by this point, the investigator has thought through their study in depth and is usually almost finished writing a grant, the act of going through the power analysis often results in an even deeper understanding of the study, including the exact models that will be used to analyze the data, the effects of interest that will be studied and exactly where in the brain the activation is expected. The aims of the study may change slightly, so for this and other reasons, the power analysis should be performed early on in the grant writing process.

If care and thought are put into a power calculation, it is likely that you will end up revising parts of your study design. You may change how many trials you are including, based on what other people have reported. In order to calculate power you must know the specific contrasts that you will be estimating and how they will be estimated, which may alter your modeling strategy or other features of the study such as the baseline task. For example, if you were planning on a

longitudinal fMRI study with five imaging sessions, it may sound like an impressive plan at first, but once you start thinking about this from the modeling perspective, you will quickly discover that running whole-brain repeated measures analyses like this is not easily done. It is a complicated model and currently most standard fMRI software packages cannot handle this type of data, especially in the case of missing data. In realizing this, a more simple design may be proposed or an ROI-based approach may be considered instead of a whole-brain analysis.

LIMITATIONS OF A POWER ANALYSIS

The data used to create the mean and variance estimates for the power calculations may not be good representations of the future data. If the mean is over-estimated or variance is under-estimated, the predicted sample size will be too small. Likewise if the mean is too small or variance too large, the predicted sample size will be too large. Within reason, a conservative power calculation is preferred as it increases the chances that the effect will be seen. It is likely that this can occur, especially in an fMRI study where the pilot data you are using most likely will not be exactly the same as what will be used in the future study and even the types of subjects may change. Obviously, if you already had data that could test your hypothesis you would not be proposing a new study to collect exactly the same data. This is not unique to the field of fMRI, but occurs with almost all power analyses.

One way to help avoid overly optimistic estimates of effect size is to tweak the different variables involved in the power analysis to see how much the predicted sample size fluctuates. Try a couple of ROIs and a range of effect sizes within reason for your experiment. Based on all of these findings be sure not to rely on the best case scenario. So, if you ran three variations of your power analysis and obtained sample sizes of 20, 35 and 38, be honest with yourself that although the sample size of 20 seems really tempting, you are probably safer with a sample size of 35 or 40. In addition to looking at a variety of ROIs and effect sizes, it is highly recommend that you obtain some pilot data. This will have many benefits, including supplying better estimates of the mean and variance estimates.

GETTING ORGANIZED FOR YOUR POWER ANALYSIS

If you know you have a grant deadline approaching, start working on your power analysis as soon as possible. It takes time to find pilot data and almost every time you run a power analysis you will discover that you need more subjects than you thought and so your budget and possibly your primary hypotheses may change as a result. Most likely you will not perform a power analysis for every task you are proposing, so try to perform power analyses on the important aims of your proposal. If you are using fMRIPower for your power analysis, it requires a pilot data analysis that was performed in either FSL or SPM and you must also input an ROI. Thorough, step-by-step instructions for using the GUI in fMRIpower can be found at http://fmripower.org/instructions.pdf. Since this is an automated program, it can only perform one-sample, two-sample and paired *t*-tests. Although by default it will calculate power for each of the regions in the automated anatomical labeling (AAL) atlas (Tzourio-Mazoyer *et al.*, 2002), for reasons described in the 'Choosing a ROI' section you should think carefully about what region you will use, as opposed to simply running it for all regions of the AAL atlas and choosing the region with the most satisfying level of power. Since anatomical atlases have such large ROIs, you typically will not have very high power, so it is recommended that you form your own ROI based on evidence from literature and not from what you found in your pilot data. Last you will enter a single or range of sample sizes and a Type I error rate. I would simply

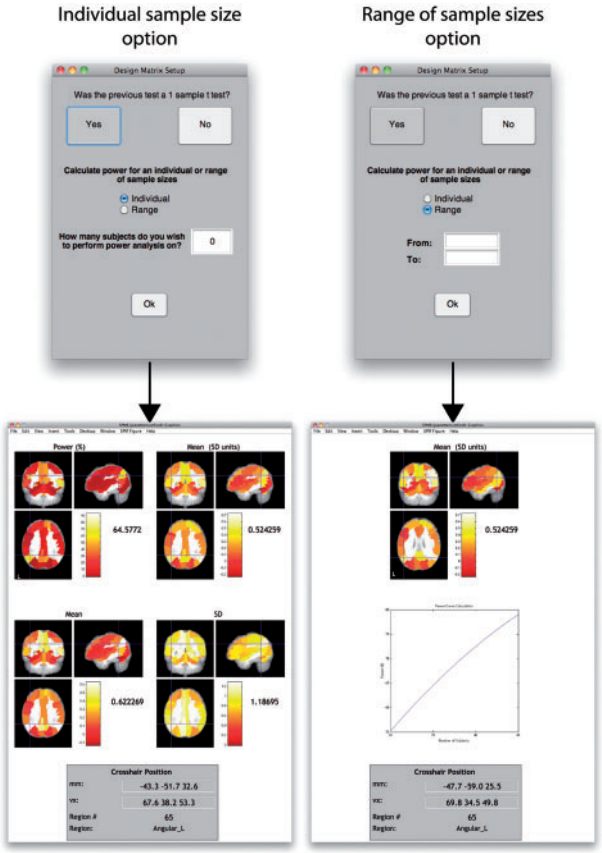

Fig. 2 Two options for sample size specification. If an individual sample size is selected you will obtain power estimates for that sample size only (left), whereas a range of sample sizes supplies a power curve (right).

Bonferroni correct the Type I error rate according to how many ROIs your are using.

fMRIPower uses the average mean and s.d. over the ROI in the power analysis. If you have V voxels, where the data in each voxel, x_v, are distrubuted, $N(\mu,\sigma^2)$, the variance of the mean across the voxels is $\mathrm{Var}\left(\frac{1}{V}\Sigma_v x_v\right) = \frac{1}{V}\sigma^2$. Hence, if the averaged data over the ROI were first calculated and then the mean and s.d. of the averaged data were used for the power analysis, the size of the ROI would influence the size of the variance and the power such that larger ROIs would have higher power. Thus, the power analysis produced by fMRIPower applies to the average voxel in the ROI. Furthermore, the power analysis assumes the same number of runs, trials per run and stimulus presentation as the pilot data, so any possible differences should be noted. Often, if I suspect my future data will more variable than the pilot data, I will also test power for slight variations in the s.d. If a single sample size is used for the power analysis, fMRIPower will display the power, mean in s.d. units, mean and s.d. images as shown in the left panel of Figure 2. You can obtain $\hat{\mu}$ and $\hat{\sigma}$ either from this image or the Matlab structure that is saved. If you run power for a range of sample sizes (right panel of Figure 2), the mean and s.d. images are instead replaced by a power curve, showing power as a function of sample size. If you want to run the power analysis for a different s.d., just use your favorite power calculation software for your test and the values.

Once the mean and s.d. have been calculated, the power calculation is actually quite simple if you are fixing the length of the run and study design. This makes it easy to test slight variations in the variance.

I find that a concise way to report power calculations for a range of mean or variance estimates is to plot a few power curves on the same axis.

CONCLUSION

Recent research has not only supplied the field of fMRI with approaches for calculating power, but tools for doing so. Although some are skeptical about the utility of a power analysis, following the advice given here will help improve the quality of sample size estimates. This will have a positive impact on the quality of grant applications and will cut down on the number of underpowered studies. In addition, careful thought about the details of your study that are necessary when running a power analysis will often impact your study positively by improving the task or study design.

FUNDING

This work was supported by National Institute of Health (NIH) (grant R03 EB008675-01A1).

REFERENCES

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

Goodman, S.N., Berlin, J.A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, *121*(3), 200–6.

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*, 721–30.

Hoenig, J.M., Heisey, D.M. (2001). The abuse of power. *The American Statistician*, *55*(1), 19–24.

Joyce, K., Hayasaka, S. (2011). Development of PowerMap: a software package for power analysis in neuroimaging studies. Poster presented at Organization for Human Brain Mapping, Quebec City, Quebec.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, *12*, 535–40.

Levine, M., Ensom, M.H. (2001). Post hoc power analysis: an idea whose time has passed? *Pharmacotherapy*, *21*, 405–9.

Mumford, J.A., Nichols, T.E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage*, *39*, 261–8.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, *15*, 273–89.

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*(3), 274–90.

Wittes, J., Schabenberger, O., Zucker, D., Brittain, E., Proschan, M. (1999). Internal pilot studies I: type I error rate of the naive *t*-test. *Statistics in Medicine*, *18*, 3481–91.

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*(3), 294.