Facilitating open-science with realistic fMRI simulation: validation and application

Cameron T. Ellis[1*], Christopher Baldassano[2,3], Anna C. Schapiro[4],

Ming Bo Cai[2], & Jonathan D. Cohen[2]

[1]Department of Psychology, Yale University, New Haven, CT, USA

[2]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA

[3]Department of Psychology, Columbia University, New York, NY, USA

[4]Department of Psychiatry, Harvard Medical School, Harvard University, Boston, MA, USA

Article type: Research article

Abstract: 202 / 250 words

Main text: 6964 words (incl. captions)

Display elements: 5

References: 46

Corresponding Author: Cameron Ellis, cameron.ellis@yale.edu

Abstract

Background:
With advances in methods for collecting and analyzing fMRI data, there is a concurrent need to understand how to reliably evaluate and optimally use these methods. Simulations of fMRI data can aid in both the evaluation of complex designs and the analysis of data.

New Method:
We present fmrisim, a new Python package for standardized, realistic simulation of fMRI data. This package is part of BrainIAK: a recently released open-source Python toolbox for advanced neuroimaging analyses. We describe how to use fmrisim to extract noise properties from real fMRI data and then create a synthetic dataset with matched noise properties and a user-specified signal.

Results:
We validate the noise generated by fmrisim to show that it can approximate the noise properties of real data. We further show how fmrisim can help researchers find the optimal design in terms of power.

Comparison with other methods:
fmrisim ports the functionality of other packages to the Python platform while extending what is available in order to make it seamless to simulate realistic fMRI data.

Conclusions:
The fmrisim package holds promise for improving the design of fMRI experiments, which may facilitate both the pre-registration of such experiments as well as the analysis of fMRI data.

Keywords: fMRI, simulation, multivariate design, reproducibility

Highlights:

- fmrisim can simulate fMRI data matched to the noise properties of real fMRI.
- This can help researchers investigate the power of their fMRI designs.
- This also facilitates open science by making it easy to pre-register analysis pipelines.

1. Introduction

Over the past two decades, the use of functional Magnetic Resonance Imaging (fMRI) has exploded as a method for studying human brain function. In the early years of fMRI research, linear models were used to describe the voxel by voxel differences in activity between conditions (Friston et al., 1995). However, over time there has been a progressive shift toward methods that assess multivariate activity in order to probe representations and dynamics that are distributed throughout the brain (McIntosh, Bookstein, Haxby, & Grady, 1996; Norman, Polyn, Detre, & Haxby, 2006). These new methods have taken advantage of high performance computing and advanced software packages to answer questions that would otherwise be difficult to test with univariate procedures alone (Cohen et al., 2017). However, this shift has not been accompanied by commensurate advances in our understanding of how fMRI noise influences these analyses. Here we describe fmrisim, a new tool to address this need that can help researchers evaluate and anticipate nuances that advanced neuroimaging analyses may introduce, and that can facilitate open science.

In the past, simulation has been a useful approach to addressing important issues regarding fMRI data. For example, simulations have been used to characterize the hemodynamic response function (HRF) in the analysis of event-related designs (Burock, Buckner, Woldorff, Rosen, & Dale, 1998), as well as the spatial noise properties of fMRI data in cluster thresholding (Foreman et al., 1995; Ward, 2000) and assessing inter-subject variability (Erhardt, Allen, Wei, Eichele, & Calhoun, 2012). However, others have noted (Welvaert & Rosseel, 2014) that the procedures for constructing these simulations are often not standard across studies and/or not clearly described. One remedy to this situation is neuRosim — an extensively developed toolbox in R that can be used to accurately simulate task-based activity, especially for univariate designs, in a standardized way across experiments (Welvaert, Durnez, Moerkerke, Verdoolaege, & Rosseel, 2011). Another tool is STANCE, a MATLAB package that takes real data as a baseline and adds user-specified signal on top of it (Hill, Liu, Nutter, & Mitra, 2017). These tools provide

different approaches to simulating univariate brain activity, however they have been focused largely on univariate analyses. Similar tools are needed that address the effect of noise in multivariate analyses.

Synergistic with their use in data analysis, simulations can also facilitate open science and reproducibility (Simonsohn, Nelson, & Simmons, 2014). Reproducibility is especially problematic in fMRI research, where researchers have innumerable preprocessing choices that can have important consequences for their results, but are not always carried out in a standard and/or clearly described manner. Although many remedies to this problem have been proposed, such as standardized preprocessing pipelines (Esteban et al., 2018), one potential solution is pre-registration (Munafò et al., 2017). Using services such as those offered by the Open Science Framework, researchers can specify in advance their design procedures and analysis pipeline and embargo it until review. Current guidelines for reproducible science encourage the pre-registration of both experimental design and analysis. This can be difficult for fMRI analyses: because of their scope and complexity it is sometimes difficult to anticipate all of the possible concerns that otherwise may arise in analysis — especially those caused by noise — which often invites post-hoc analyses. However, the failure to provide algorithmically precise analysis plans in advance can inflate the risk of false discovery due to post-hoc exploration of different analysis pipelines.

We propose that simulations could facilitate pre-registration of fMRI data. Before any data is collected, researchers can simulate expected experiment effects embedded within a realistic model of fMRI noise, and construct a pipeline for analyzing the results that takes this into account. This would not only ensure explicit specification of the analysis procedure, but could provide more accurate and reliable estimates of effect sizes that can be used for power calculations. The result would be that interested readers could observe for themselves exactly what the researcher hypothesized and what analyses were planned in advance. This is not to discourage post-hoc, exploratory analyses; rather, it would simply assist in demarcating what was

predicted and what should be considered post-hoc analysis, so that the proper analyses and interpretations can be applied to each.

To be most effective, a simulator should be integrated with the tools used for experimental design and data analysis. While there are a number of packages used for standard fMRI analysis (e.g. FSL [Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012]), and others for multivariate analysis (e.g. the Princeton MVPA toolbox [Detre, et al., 2006]), there has recently been a large scale migration to Python for scientific computing, where tools such as Nilearn (Abraham et al., 2014) and Nipype (Gorgolewski et al., 2011) make it an increasingly attractive environment for fMRI data analysis. Among these is BrainIAK (Brain Imaging Analysis Kit, http://brainiak.org), a newly released open-source Python toolbox that supports advanced neuroimaging analyses, such as multivariate pattern analysis (Norman, et al., 2006), full-correlation matrix analysis (Wang, Cohen, Li, & Turk-Browne, 2015), Inter-Subject Functional Connectivity (Simony et al., 2016), Bayesian Representational Similarity Analysis (Cai, Schuck, Pillow, & Niv, 2016), Hierarchical Topographical Factor Analysis (Manning, et al., 2018) and Shared Response Modeling (Chen et al., 2015). These methods are computationally intensive but have been optimized to exploit modern advances in high performance cluster computing. Thus, this represents a potentially valuable environment for a simulation package, and in particular one that addresses the impact of noise in these advanced, multivariate analysis methods. Ideally, this should be customizable, have appropriate defaults, and integrate seamlessly with existing tools for experiment design and analysis. Critically, it must be able to generate realistic, 'brain-like' data in order for it to work in a pre-registration pipeline.

In what follows we describe such a package — called fmrisim — that is a set of open-source Python functions integrated into the BrainIAK toolbox. This package builds on the innovations of previous fMRI simulators, but focuses on its application to advanced, multivariate neuroimaging methods. The goal is to make realistic simulations of brain data that can be inserted into a standard fMRI analysis stream that makes use of multivariate methods. Below we describe

the steps necessary to generate a dataset appropriate for multivariate analysis. We then describe analyses to assess the quality of the simulation. Finally, we discuss applications of the simulator for evaluating the efficacy of experimental design and facilitating reproducibility.

2. fmrisim

The process of simulating participant data with fmrisim can be described in three steps: (1) specifying parameters, (2) generating noise, and (3) generating signal to add to the noise. Here we describe at a conceptual level what is needed to simulate a multivariate dataset with noise properties matched to a raw dataset. We encourage interested readers to go through the Jupyter Notebook in BrainIAK that demonstrates the relevant code to perform these steps (github.com/brainiak/brainiak/blob/master/examples/utils/fmrisim_multivariate_example.ipynb).

2.1. Specify parameters

In order to simulate fMRI data, the signal, noise, and acquisition parameters must be described. fmrisim also needs a template that represents an approximate baseline of the fMRI signal to which noise can be added. This spatial volume is generated in fmrisim by averaging each voxel across time (Welvaert, et al., 2011). This template is also used to create a binary mask of brain and non-brain voxels (Abraham, et al., 2014).

2.2. Generate noise

fmrisim can generate various types of realistic fMRI noise. Much of the noise modeling was inspired by neuRosim (Welvaert, et al., 2011). In fmrisim, a single function receives the specification of noise parameters and simulates whole brain data with noise properties approximating those parameters. Figure 1 shows examples of the noise types generated by fmrisim.

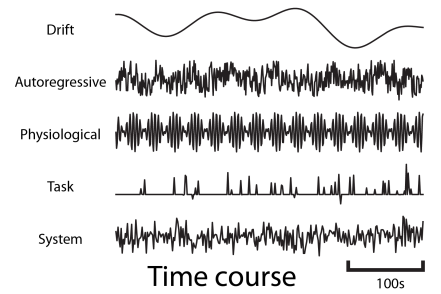

Figure 1. Example time courses of the different types of noise that are generated by fmrisim. Each plot represents a voxel's activity for each type of noise.

The noise of fMRI data is comprised of multiple components: drift; auto-regressive/moving-average (ARMA); physiological; task-related noise and system noise. Drift and system noise are assumed to reflect machine-related noise and thus affect the entire field of view of the acquisition. The remaining components are assumed to be specific to the brain and thus have a smoothness component related to the smoothness of functional data. Using Gaussian random fields of a certain Full-Width Half-Max (FWHM), we can determine how smooth the temporal noise is across voxels (Chumbley & Friston, 2009). This volume of spatial noise is masked to only include brain voxels. Equivalent drift and system noise is added to all voxels in a volume.

To simulate drift, cosine basis functions of different periods and phases are combined, with longer runs being comprised of more basis functions (Welvaert, et al., 2011). ARMA noise is estimated by creating a time course of Gaussian noise values that are weighted by previous values of the time course (Purdon & Weisskoff, 1998). Physiological noise is modeled by sine waves comprised of heart rate (1.17Hz) and respiration rate (0.2Hz) (Biswal, Deyoe, & Hyde, 1996) with random phase. Finally, task-related noise is simulated by adding noise to time points where there are events (determined by the design of the experiment). These four noise components are mixed together according to user-defined parameters and then set to the appropriate magnitude of temporal variance, derived from the Signal-to-Fluctuation-Noise Ratio (SFNR). The SFNR, also known as temporal signal-to-noise, reflects how much temporal variation there is in brain voxels

relative to their mean activity (Friedman & Glover, 2006). If the SFNR is high then the amount of temporal variation will be low and thus so will these temporal noise components.

The volume created by the combination of these noise components is then combined with system noise. System noise results from heat-related motion in MRI scans (Bodurka, Ye, Petridou, Murphy, & Bandettini, 2007). System noise is Rician (Gudbjartsson & Patz, 1995); however, we have determined that the distribution of voxel intensity in non-brain regions (which includes the skull and eyes which are sensitive to T2* measurements) is also approximately a Rician distribution (e.g. the eyes have high MR values but the vast majority of voxels are near zero). As stated above, fmrisim creates a template of voxel activity (both brain and non-brain voxels) that noise is built on top of. This template also has non-brain voxels with a Rician distribution so adding Rician noise to voxels with an already Rician distribution leads to inappropriately 'spiky' data. Instead, we have found that system noise is better approximated by Gaussian noise added to the template of average voxel activity. The magnitude of system noise is determined by both a spatial noise component, which depends on the Signal to Noise Ratio (SNR) value, and a temporal component of system noise, that depends on the SFNR.

All the noise parameters, like SNR, can be manually specified; however, it is possible to instead extract some of these noise parameters directly from an fMRI dataset using a single automated function in fmrisim. Specifically, fmrisim utilizes tools for estimating the SNR, SFNR, FWHM and auto-regressive noise properties of real data. To calculate the SNR, fmrisim compares brain signal with non-brain spatial variance. To do this the mean activity in brain voxels for the middle time point is divided by the standard deviation in activity across demeaned non-brain voxels for that time point (Triantafyllou et al., 2005). To estimate SFNR, fmrisim divides each voxel's mean activity by the standard deviation of its activity over time, after it has been detrended with a second order polynomial. This is done for every brain voxel and then averaged to give an estimate of the SFNR. The FWHM of a volume is calculated by fmrisim first in the X, Y, and Z dimension and then averaged both within a volume and across a sample of

time points. Finally, fmrisim estimates ARMA using the statsmodels Python package by assuming an AR order of 1 and an MA order of 1. This is calculated for many voxels and then averaged.

The noise estimates of real fMRI data provided by the fmrisim tools described above can then be used to simulate data with equivalent noise parameters; however, there are some noise properties that fmrisim does not estimate. This is because they are otherwise trivially dealt with in preprocessing (e.g. drift), or they are impossible to retrieve from the raw data without knowing the latent signal that generated the brain data (e.g. task noise) or without greater temporal precision (e.g. physiological noise). These types of noise must be specified manually in order for them to be used by fmrisim.

Importantly, the noise parameter estimates depend on assumptions about the appropriate noise properties of fMRI data and how they interact. If the data is already preprocessed then this can interfere with the way noise properties ought to be inferred. For instance, if all non-brain voxels have been masked, then these zeroed-out voxels will be an inappropriate baseline for non-brain variability. In addition, because of the non-linearity and stochasticity of the simulation, the estimation is not fully invertible: a brain simulated with a set of noise parameters will have similar but not necessarily the same noise parameters as what were specified. To address this, and generate simulations with noise characteristics as close to the ones specified as possible, fmrisim iteratively simulates and checks the noise parameters until they fall within a specified tolerance of what was specified. This is useful for matching noise parameters to a specific participant, not just to typical human fMRI data. The fitting tolerance is by default set to within 5% (i.e. the difference between each real and simulated noise parameter is less than 5%). This could be made more strict by the user if desired; however there is inherent stochasticity in the noise generation process, especially for ARMA and system noise, that makes it less likely to converge. Once the noise parameters are estimated, they are then used to generate time series data to which one or more hypothesized neural signals are added.

2.3. Generate signal and add it to noise

fmrisim has tools to simulate many different types of neural signal. These tools are useful for generating signals intended to represent patterns of activation in specific regions of interest (ROIs). However, custom scripts may be necessary for more complex signals, such as activity coupling between regions. To design a signal it is necessary to establish the predicted effect size, the ROI/spatial extent of the signal, and its time course. For instance, the notebook describes a case in which two conditions evoke different patterns of activity in the same set of voxels in the brain. This pattern does not manifest as a uniform change in voxel activity across the voxels containing signal (i.e. the mean evoked activity is not different between conditions). Instead, each trial of a condition evokes an independent pattern across voxels. This activity can then be convolved with the hemodynamic response in order to estimate the predicted pattern of measured activity. fmrisim also makes it easy to calibrate the magnitude of a signal using various metrics, such as percent signal change or contrast to noise ratio (Poldrack, Mumford, & Nichols, 2011; Welvaert & Rosseel, 2013).

The final step is to combine the noise volume with the signal volume to create a brain that is ready for preprocessing and analysis.

2.4. Discussion

Above we described how fmrisim generates realistic fMRI data. For more detail we point readers to the documentation, and to the online notebook that demonstrates the implementation of a simulation to see how the tool is actually used. fmrisim is intended to be flexible, allowing for the generation and specification of many different types of fMRI noise and signal. Of course, the simulator is not perfect: a complete simulation would require that we understand all the latent causal interactions in the brain and how these manifest in fMRI data — a goal that remains the focus of the research enterprise itself. There are likely many sources of noise that remain poorly understood or undetected, that contribute to the fMRI signal. Moreover, there are types of noise that fmrisim does not attempt to model (e.g., higher order structure

inherent in baseline or resting state activity; Raichle et al., 2001). Nevertheless, the use of fmrisim to simulate known properties of fMRI data holds promise for improving experimental design and data analysis, as illustrated in the examples below.

3. Experiment 1: Validation of fmrisim

To evaluate the validity of the simulator, we test whether various noise parameters of the simulated data match the noise parameters of real data. All the code for these analyses is available online (https://github.com/CameronTEllis/fmrisim_validation_application.git).

3.1.1. Materials and methods

For these analyses we used the publicly available Corr_MVPA dataset (Bejjanki, Da Silveira, Cohen, and Turk-Browne (2017): http://arks.princeton.edu/ark:/88435/dsp01dn39x4181). This dataset contains two task runs, two rest runs and an anatomical image for each of 17 participants. This is appropriate for our analyses because it is a typical sized dataset with no preprocessing. For this data set we only considered the rest runs. These were collected on a 3T scanner (Siemens Skyra) with a 16 channel head coil and a T2* gradient-echo planar imaging sequence (TR=1.5s, TE=28ms, flip angle=64^0, matrix=64x64, slices=27, resolution=3x3x3.5mm).

We compared the noise properties of the real and simulated data to test the efficacy of fmrisim. To do this, the noise properties of the participant data were extracted and then used to create simulated data. We then estimated the noise properties of the simulated data, including SNR, SFNR, auto-regression, and FWHM. The noise properties of the original data were then compared with those measured in the simulated data in order to determine the fidelity of the simulation. In other words, we tested whether the known noise properties utilized by the simulator were similar to the noise properties estimated from the data generated by the simulator.

We used the default fitting procedure and created 10 simulations for the rest runs of each participant in order to measure the reliability of each simulation.

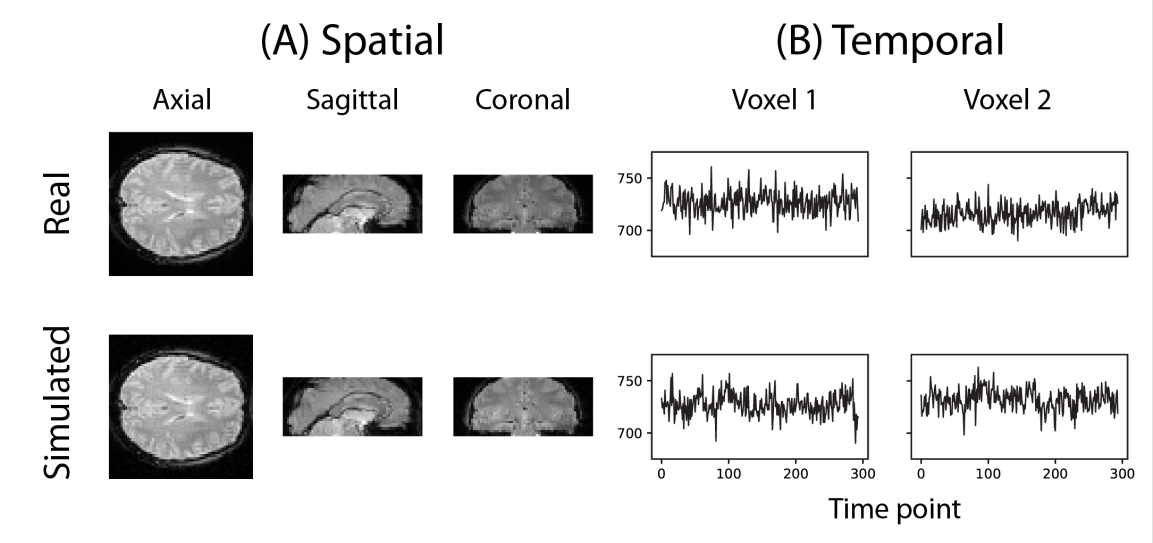

Figure 2. Example of the spatial and temporal structure of real and simulated data. (A) depicts the spatial structure of real data (top) and fitted simulated data (bottom). (B) shows the time course of sample voxels.

3.1.2. Results

Figure 2 illustrates that the simulator can at least superficially approximate both the spatial and temporal properties of real fMRI data. To evaluate whether fmrisim can accurately simulate brain data with specific noise parameters, we compared the noise parameters estimated from the real and simulated data. The ideal pattern of results is that there is no difference between the real and simulated data. Figure 3 shows boxplots of the noise parameters of the real and simulated data across participants. Notably, the majority of simulated brains (across all participants, runs and resamples) are within 5% of the target parameter (proportion within 5% for each noise component: SNR=98.5%, SFNR=100%, Auto-Regression=92.1%, FWHM=100%). Although advantageous, fitting is not perfect because of the sequential nature of the fitting process (e.g. SNR is fit before autoregressive noise) and the inherent randomness in the simulation.

Not depicted in Figure 3 is the variance in the estimated noise parameters on each simulation. Although each simulation has inherent randomness, the variance in estimates is small relative to the variance between individuals with different noise parameters (mean variance of noise parameters extracted from simulations: SNR: M_{Var}=0.008, SFNR: M_{Var}=1.428, Auto-Regression: M_{Var}=0.0006, FWHM: M_{Var}=0.0002). Execution times for this data set were reasonable: the simulation with fitting took 278.3s (SD=39.8s, max=381.9s) on average to run for each participant on a single core of an Intel Xeon E5 processor (8 cores, 2.6GHz).

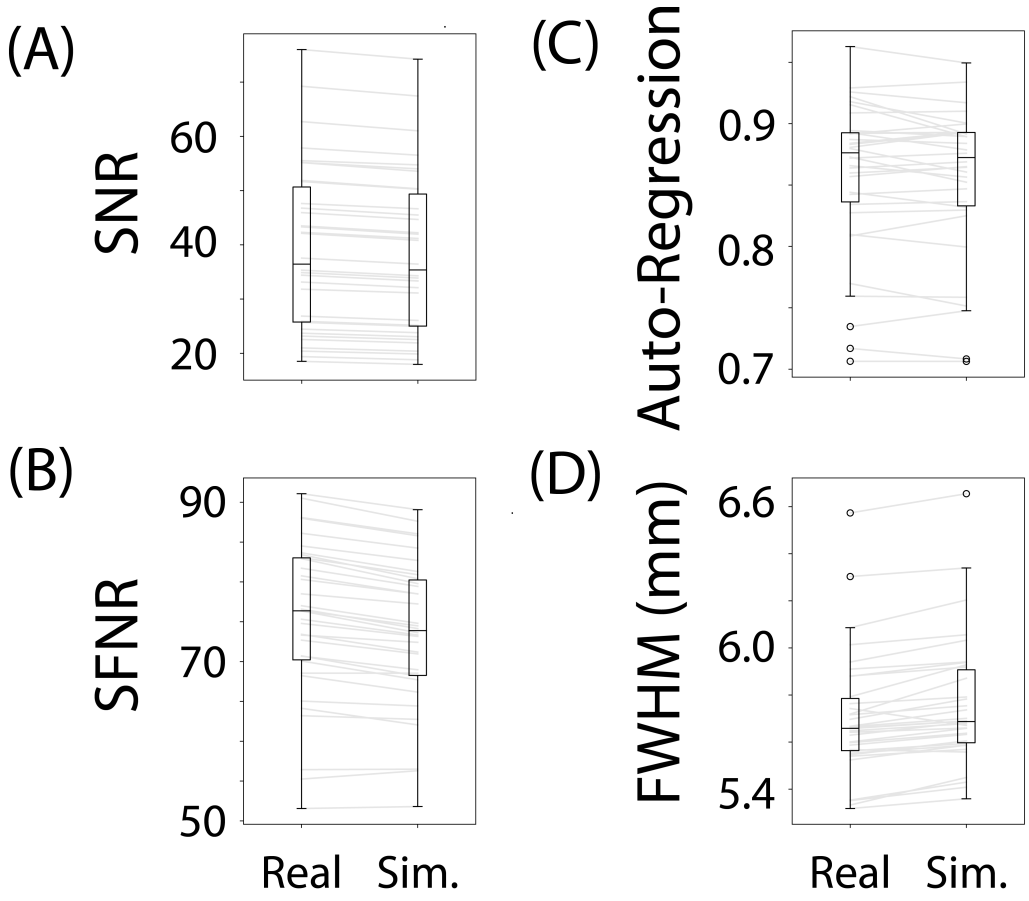

Figure 3. Boxplots of noise parameters from real data and fitted simulated data. For each participant and run, 10 simulations of their real data were generated and their noise parameters estimated. These noise parameters were then averaged within participant/run and plotted, and compared with the noise parameters from their corresponding real data. The grey lines connect the participant/run parameter value, the ellipses refer to outliers. The noise parameters tested were (A) Signal-to-Noise Ratio, (B) Signal-to-Fluctuation-Noise Ratio, (C) Auto-Regression, (D) Full-Width Half-Max.

3.1.3. Discussion

We evaluated how well fmrisim is able to recreate the noise properties of fMRI data and showed that it can do so adequately. Critically, the noise parameter estimates were participant/run specific, which means that fmrisim can simulate data with noise properties matched to individual participants.

Capturing the multifaceted nature of fMRI noise is not trivial — the different sources of noise are combined in fMRI such that retrieving and isolating those sources of noise requires that some assumptions are made. These assumptions are numerous and described explicitly in fmrisim (e.g. it is assumed that the properties of system noise are different in space versus time, which is evident in real data). Yet, when these assumptions are violated or inappropriate, the simulator will not accurately recreate fMRI data. The simulations described above are necessarily incomplete, since potentially important sources of noise, such as physiological noise (Raj, Anderson, & Gore, 2001), could not be estimated from the data. Nevertheless, the success of fmrisim in capturing some of the noise properties of real data suggests that at least some of these assumptions are appropriate.

Importantly, the tests outlined above reflect a lower bound of what an fMRI simulator is able to do. Future improvements to fmrisim or other simulators should capture more than just the descriptive statistics of noise in fMRI data but should also reflect the latent noise generation properties of the brain. Nevertheless, fmrisim provides a useful advance in the ability to evaluate how fMRI noise and signal interact, that can be used for experimental design and analysis. We demonstrate such potential in the example application described below.

4. Experiment 2: Application of fmrisim

The interactions between experiment design, noise characteristics of the data, and their consequences on statistical power can be complex and difficult to anticipate without appropriate

tools. Here, we describe how fmrisim can be used to optimally design an experiment for maximizing statistical power.

Statistical power, in the context of null hypothesis testing, reflects the likelihood of getting a significant result from a sample, assuming that there is a real effect to be found. Designing experiments with sufficient power is critical for conducting reliable research (Cohen, 1992). Power tests are important at the design stage for estimating how the number of participants or trials impacts the likelihood of reaching significance (Poldrack, et al., 2011). However, they can also be useful after data collection for determining the likelihood of a given result (Button et al., 2013).

Power tests are particularly important for fMRI research: data acquisition is expensive, sample sizes are small, and effect sizes can be weak. Statistical power in fMRI research is affected by the size and character of the signal to be measured (which can vary widely across brain region; Gläscher, 2009; Desmond & Glover, 2002), the number of participants that will be studied, and the threshold for significance. Typically, power has been assessed in fMRI studies with reference to previous studies with similar designs, using these to predict the effect size for that design. However, in cases for which such information is not available (e.g., in novel designs and/or populations), simulations provide a valuable alternative approach: by simulating a hypothesized signal in the brain, it is possible to estimate how many trials, participants, etc., are needed to pass a significance threshold for the analysis of interest.

One aspect of fMRI design for which power issues have been explored is in selecting the inter-stimulus interval (ISI). To maximize power in task-based fMRI, it is considered best to have slow designs in which events are spaced out to mitigate the overlap of the hemodynamic response (Friston, Zarahn, Josephs, Henson, & Dale, 1999). However, sometimes the only way to evaluate a research question is to use faster event-related designs, in which events occur within 10s of each other, either because a slower design would be impractically long or because the cognitive process requires faster paced stimuli and/or responses. Although fast designs mean the brain's

slow hemodynamic response overlaps more across events, a shorter ISI can allow for more trials (i.e. the efficiency is increased), which can increase the statistical power for both univariate (Dale, 1999) and multivariate (Kriegeskorte, Mur, & Bandettini, 2008) analyses. This is especially true when the time between events is randomly jittered (Burock, Buckner, Woldorff, Rosen, & Dale, 1998), as different combinations of overlap can be sampled to improve deconvolution. Moreover, if conditions are presented multiple times then it is considered a good design choice to present them in a randomized order, since it will result in different overlaps between evoked representations (Burock, et al., 1998). Thus, ISI can be a critical factor in determining the effect size and power of the design.

Although longer ISIs with randomized event sequences are generally considered to optimize power, most demonstrations of this have been limited to univariate analyses (although see Kriegeskorte, et al. (2008)). However, several factors could complicate, and potentially invalidate generalization of these findings to multivariate analyses. For instance, calculating efficiency and covariance are important and tractable within univariate analyses, in which different voxels are treated as representing different stimulus conditions; however, doing so becomes much more complex if an effect is multivariate — that is, a single voxel is involved in some or all of the stimulus conditions.

To evaluate the effects of ISI and some related design choices on power, we used fmrisim with data from a study by Schapiro, Rogers, Cordova, Turk-Browne, and Botvinick (2013). Below, we first describe use of fmrisim to observe what signal magnitude is necessary to match their result. We then show how different design choices can influence the expected effect size. This provides an example not only of how fmrisim can be used to explore the power of experimental designs, but also how this facilitates open science by making it easy for researchers to pre-register experiment analyses and clearly demarcate planned versus post-hoc analyses. All of the analyses described in what follows, as well as the necessary data to perform them, are available online (https://github.com/CameronTEllis/fmrisim_validation_application.git).

4.1. Materials

Schapiro, et al. (2013) examined how event segmentation is achieved in the brain. Figure 4 shows a graph representing the 15 stimuli participants saw in this study and the possible transitions (grey) between stimuli during the learning phase of the experiment. In their fMRI experiment, the stimuli were fractal-like images presented for 1s with a 1, 3, or 5s ISI. Although the assignment of images to graph nodes was random, the structured temporal transitions between these stimuli were expected to change the image representations to reflect the graph community structure (Fortunato, 2010). As participants observed a random walk through the possible transitions, the graph structure tends to produce long sequences of stimuli from within one community before transitioning to another community. Transitions between communities were only possible by passing through certain nodes on the graph. Behavioral evidence suggested that over the course of exposure, participants come to represent the transitions between communities as event boundaries.

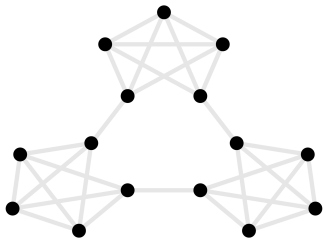

Figure 4. Graph of the community structure from Schapiro, et al. (2013). Black dots represent the 15 abstract visual stimuli participants were shown. Grey lines represent possible transitions between the dots.

Participants were scanned while observing the sequences of stimuli, and the imaging data was used to examine the similarity structure in the neural representations of the stimuli. fMRI data were collected in a 3T scanner (Siemens Allegra) with a 16 channel head coil and a T2* gradient-echo planar imaging sequence (TR=2s, TE=30ms, flip angle=90°, matrix=64x64, slices=34, resolution=3x3x3mm, gap=1mm). After participants had 35 minutes of exposure to

random walks on the graph outside the scanner, they viewed alternations between random walks and Hamiltonian paths (i.e. every node visited exactly once) in the scanner. One particular Hamiltonian path was chosen for each subject (e.g. the perimeter around the dots in Figure 4) and was presented in clockwise and counterclockwise directions. In order to avoid potentially confounding item repetition effects in the random walks, the first three items of each Hamiltonian path were ignored and analyses were performed on the remaining 12 items in the Hamiltonian paths. Over five runs, 20 participants were presented with 25 sets of these Hamiltonian walks. Each run was Z scored in time and the TRs corresponding to 4s after the stimulus onset were extracted and averaged to represent the brain's response to the stimulus. A "searchlight" approach was used to test which, if any, voxels in the brain represented the community structure. This involves iterating the same computation over a subset ("searchlight") of voxels centered on different voxels in the brain. The searchlight was a tensor of 3 x 3 x 3 x 15 voxels centered on every voxel in the brain. In each searchlight the correlation of patterns of voxel activity was computed for all pairs of stimuli, and these were tested to determine whether the correlation for stimuli belonging to the same community was higher than that for stimuli belonging to different communities. A permutation test evaluated the robustness of these metrics across participants. Some regions of the brain, including the inferior frontal gyrus and superior temporal gyrus, reliably showed a greater correlation between stimuli within a community compared to stimuli in different communities.

4.2. Methods

Five steps were performed to simulate the data from Schapiro, et al. (2013): generate a template of average voxel activity; extract the noise parameters using fmrisim; represent the timing of each stimulus onset; implement a region of interest mask of the significant voxels; and characterize the signal. All of these were performed using the aforementioned dataset, and then the nature and magnitude of the signal was manipulated as described below.

We embedded this signal in a region of interest (ROI), identified by Schapiro et al., (2013) spanning the left superior temporal gyrus, as well as the anterior temporal lobe and inferior frontal gyrus (totaling 442 voxels). The simulated signal was generated with an event-related design in which the different stimuli on the community structure graph (Figure 4) occurred one after the other. Each stimulus evoked a positive pattern of activity across all the voxels in the ROI. This pattern was determined by that voxel's position in the high-dimensional embedding of the community structure. In particular, we generated a 2 dimensional graph of the community structure, then an orthonormal transformation was used to embed these points in 442 dimensional space (matching the number of ROI voxels) in a way that preserved Euclidean distance. This meant that each of these voxels' activity represented a dimension along which each event could vary and the pattern across the voxels determined the event's representation. The specified response to every event for each voxel was convolved with the double gamma HRF (Friston, et al., 1998). The time course of the events was determined by the stimulus presentation timing during the Hamiltonian test phases.

There were two critical parameters determining the representation of the community structure: magnitude and density. Magnitude reflects the neural response that each stimulus presentation evokes and this determines the overall range of the graph values in the simulation. Specifically, percent signal change was used to determine the evoked response. Percent signal change was calculated for each voxel by taking the mean activity of that voxel's noise activity, dividing it by one hundred to make it a proportion and then multiplying it by the evoked response. Critically, the percent signal change is set for the voxel with the highest value in this embedding (i.e. the coordinate furthest from 0 after convolution) and all other voxels are scaled in proportion to that peak voxel.

Density is the amount of clustering within a community. A high-density representation means that all of the stimuli within a community are similar to one another, reflecting strong event segmentation. In the simulation, a density of 1 means that points within a community are

completely overlapping, whereas a density of 0 means that all transitions between stimuli result in equal length edges on the graph (Figure 5A). Importantly, magnitude and density are independent: a participant's representation of the communities can be perfectly separated (high density) but the neural response evoked by each stimulus might be minimal (low magnitude), meaning that the community structure is not measurable due to neural noise. Hence, the combination of both magnitude and density jointly contribute to the degree of community structure that can be measured neutrally.

We simulated fMRI data for pairs of magnitude and density parameter values to observe how they compared to the real data. These were processed using the same analysis pipeline, including the metric for calculating within vs. between community distance, used for the real data to get a summary statistic of the degree of community structure in the ROI where a known signal was added. We took the difference in the permuted t-test statistic for ROI voxels in the real and simulated participants to get an estimate of simulated signal strength relative to the real data.

In addition to re-creating the design of the experiment, we used fmrisim to manipulate design parameters and examine their impact on effect size. As discussed above, a critical parameter in event-related design is the ISI. We used fmrisim to explore the impact that this may have had on the results. To do this, time was added between events while preserving the trial jittering, ordering, and trial number (e.g. if the original participant experienced a 3s ISI between events then in this simulation more time was inserted to space out these events, but otherwise nothing else was changed); allowing new data to be simulated based on this time course. A simulation with 0s added to the ISI was a direct simulation of Schapiro, et al. (2013). This analysis was run 10 times (i.e. 10 sets of participants were generated for each level of ISI and signal parameter) to estimate the variability of these summary statistics.

We also used fmrisim to examine how the stimulus sequence may have impacted the results; that is, what would have happened if stimuli were not presented in the Hamilitonian path as was done by Schapiro, et al. (2013), but rather presented in a random order, irrespective of

their community structure (note that a randomized order was undesirable in the experiment because it may have interfered with the learned transition probabilities). For this analysis, the order of stimuli from each Hamiltonian walk was shuffled (such that the same stimuli could never occur consecutively). According to evidence from randomized univariate designs (Dale, 1999) this should boost signal greatly.

Finally, although increasing the ISI may boost signal, it also increases the duration of the experiment. To investigate this, we also examined the effect of limiting the number of trials per run. For this analysis, we increased the ISI between events in a Hamiltonian walk, as before, but cut-off the sequence of stimuli when the run ended for the real participants, and analyzed the truncated dataset to determine the influence of this manipulation on power.

4.3. Results

The first aim of these analyses was to determine the range of percent signal change required to replicate the results from Schapiro, et al. (2013). To do so, we compared the summary statistic for real and simulated data for different pairs of signal magnitude and density parameters. Figure 5B shows the absolute value of this difference, with low difference scores indicating the simulation more accurately approximated the real test statistic compared to high difference scores. When the simulated community structure is low density, the signal magnitude required to approximate the test statistic of Schapiro, et al. (2013) is higher than the signal magnitude required when the density is greater. This analysis shows that the necessary magnitude is within the plausible range, regardless of community density. Hence, if this simulation were done before data collection (using default noise parameters from other participants), this analysis would show that this design, in terms of the specified participant number, trial number and ISI, is sufficiently powered to identify community structure.

We chose three pairs of magnitude and density parameters that lie along the trough of Figure 5B (i.e. parameters that minimize the difference in test statistic between the real and simulated data) and varied the ISI to observe the effect it has on the test statistic, as shown in

Figure 5C. The peak test statistic is greatest with a minimum ISI of 5s. Critically, the test statistic plateaus after this ISI, suggesting that a slower event-related design would confer no additional benefit, especially given the cost of additional experimental time. This pattern is present regardless of the signal density, although a denser signal structure has a higher test statistic.

Surprisingly, and counter to dogma conquering univariate analyses, randomizing the order of test stimuli did not greatly change the test statistic. Figure 5D shows that when the minimum ISI is 1s, as in Schapiro, et al. (2013), and the stimulus sequences are randomized, then the test statistic is lower than when they are presented in a hamilitonian order especially with low density representations. In other words, randomizing the stimulus sequence would decrease the test statistic. However, if the ISI for these randomized sequences is increased then the test statistics are only slightly greater than the test statistic for a Hamiltonian sequence. Hence, randomizing would only be beneficial for this experiment design if the ISI were long.

Although increasing ISI increases the signal when the experiment duration is potentially limitless (Figure 5C and 5D), experiments are practically constrained in terms of how long a session can run. Figure 5E shows that when the experiment duration is constrained to be the same as Schapiro, et al. (2013), the test statistic in general decreases. Interestingly, if the density is above zero then there is a slight gain by adding an ISI of 2s but no such benefit exists when the density is 0. This suggests that given the time constraints of fMRI, the design chosen by Schapiro, et al. (2013) was efficient and nearly optimal. However, such foresight is not always forthcoming as noted in the discussion below.

4.4. Discussion

The aim of the present analyses was to show how fmrisim can help researchers make informed design choices, regarding factors such as ISI and stimulus ordering, to attain the highest statistical power in their design as possible. Across the different signal parameters, the same trends in effect size were found for different ISIs: the effect size increases as the ISI increases up

to 5s, after which it plateaus; however, this gain in effect size was greatest when the signal was dense. Counter to typical assumptions in fast event-related designs (Burock, et al., 1998), a randomized order of stimuli slightly reduced the effect size when the ISI was short and only slightly helped when the ISI was longer. Of course any increase in ISI needs to be weighed against how this affects experiment duration and thus the number of allowable trials. Others have shown that longer ISIs can help the effect size despite such limits (Friston, et al., 1999), whereas for this design we observed that increasing ISI resulted in a diminished effect size (Kriegeskorte, et al., 2008), although this was non-monotonic for some signal densities.

The analyses described above provide an example of how fmrisim can validate the plausibility of an experimental analysis; however, fmrisim can also be used to show the implausibility of an experiment/analysis. We have taken this approach in other work using fmrisim to suggest that a particular analysis procedure — topological data analysis — has limited usefulness for analyzing multivariate representations in fMRI because of the noise properties of fMRI data (Ellis, Lesnick, Henselman-Petrusek, Keller, & Cohen, 2018). Hence we believe that fmrisim can be used to demarcate what experiments and analyses are and are not plausible.

The example outlined above also suggests ways in which simulation can be used for pre-registration. For example, consider designing a study similar to that of Schapiro, et al. (2013), but that required a change in some parameter(s), such as ISI. The data generation and analysis procedures described above could be used to evaluate the influence of this manipulation on the hypothesized effect, which can then be pre-registered and made openly accessible both during review and after publication. The hypothesized results and final analyses could also be compared to what was pre-registered and be interpreted accordingly.

In sum, the kinds of simulations supported by fmrisim can be used to improve the power of designs and facilitate open science practices in research.

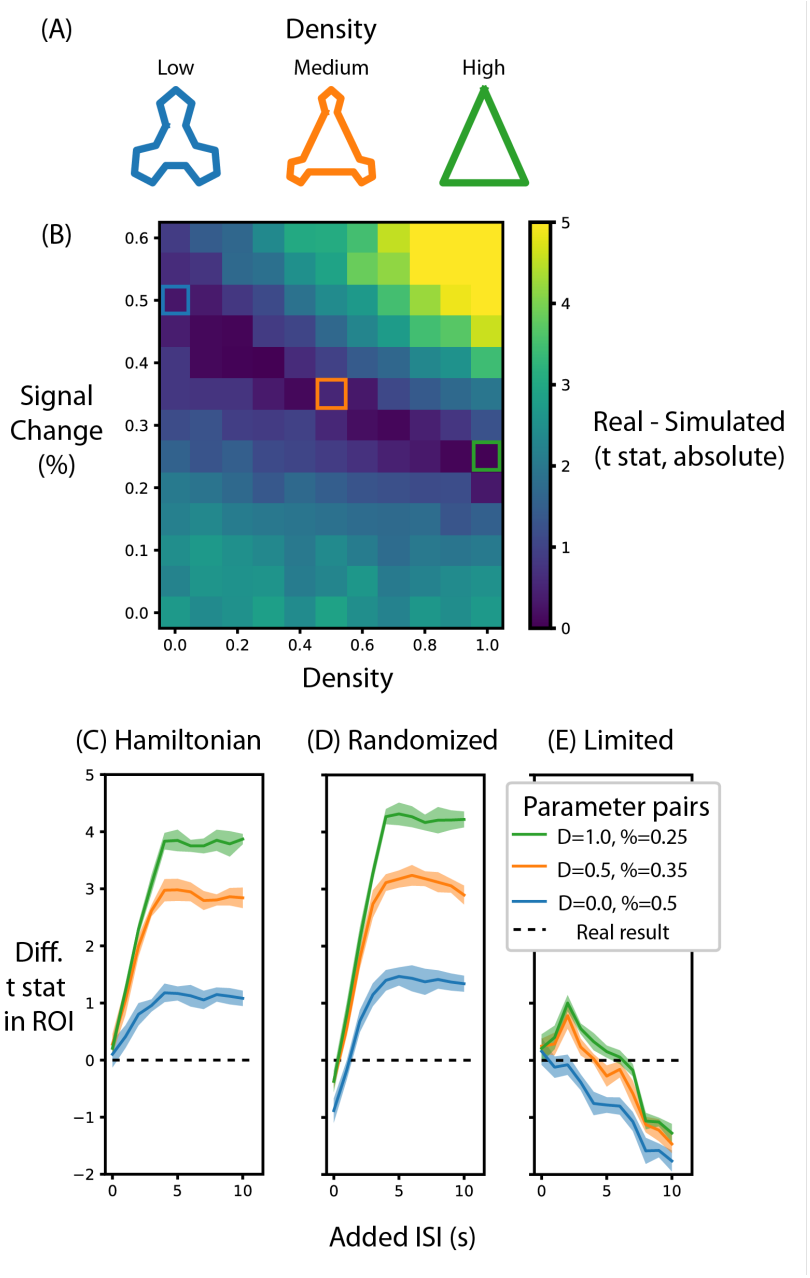

Figure 5. Alternative designs for Schapiro, et al. (2013) and their relative test statistic. (A) Depictions of representations corresponding to different densities. (B) A heat map of the difference between the real data and simulated data for different parameters of signal magnitude and density. Three pairs of parameters were chosen that minimize the difference between the simulated and real data. (C) For these three pairs of parameters, the average t-statistic in the ROI that was significant in Schapiro, et al. (2013) was subtracted from the real data and compared at different ISIs when the stimuli were presented in a Hamiltonian sequence. The black line represents the real data. (D) is the same as (C) except that the sequence of stimuli was randomized in the simulation. (E) is the same as (C) except that the maximum duration of the run is limited, regardless of ISI. Shaded lines represent the standard deviation across the 10 permutations of this condition. 'D' in the legend refers to the density of the representation, '%' refers to the percent signal change.

5. Concluding remarks

We have presented a new Python package, fmrisim, for simulation of realistic fMRI data. This package is capable of taking in raw fMRI data and generating synthetic fMRI data with approximately matched noise properties. fmrisim can be profitably used to evaluate the effects of design and analysis parameters and facilitate pre-registration of fMRI data. With the accelerating development of advanced neuroimaging methods, fmrisim can help evaluate and compare these methods on a standardized and openly accessible software platform that can facilitate both data analysis and further methods development.

Acknowledgements

Thank you to N. Malek for comments on an earlier draft.

Contributions

C.T.E., C.B. & J.D.C. designed the project. C.T.E., C.B. and M.C. wrote fmrisim. C.T.E., A.C.S. performed the analyses. C.T.E. wrote the initial draft of the manuscript.

Competing interests

No conflicts are declared.

Funding

This work was made possible by support from Intel Corporation and the John Templeton Foundation. The opinions expressed in this publication do not necessarily reflect the views of these agencies.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., . . . Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics, 8*, 1-14.

Bejjanki, V. R., Da Silveira, R. A., Cohen, J. D., & Turk-Browne, N. B. (2017). Noise correlations in the human brain and their impact on pattern classification. *PLoS computational biology, 13*(8), e1005674.

Biswal, B., Deyoe, E. A., & Hyde, J. S. (1996). Reduction of physiological fluctuations in fMRI using digital filters. *Magnetic Resonance in Medicine, 35*(1), 107-113.

Bodurka, J., Ye, F., Petridou, N., Murphy, K., & Bandettini, P. (2007). Mapping the MRI voxel volume in which thermal noise matches physiological noise—implications for fMRI. *Neuroimage, 34*(2), 542-549.

Burock, M. A., Buckner, R. L., Woldorff, M. G., Rosen, B. R., & Dale, A. M. (1998). Randomized event-related experimental designs allow for extremely rapid presentation rates using functional MRI. *Neuroreport, 9*(16), 3735-3739.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376.

Cai, M. B., Schuck, N. W., Pillow, J. W., & Niv, Y. (2016). A Bayesian method for reducing bias in neural representational similarity analysis. In *Advances in Neural Information Processing Systems* (pp. 4951-4959).

Chen, P.-H. C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., & Ramadge, P. J. (2015). *A reduced-dimension fMRI shared response model.* Paper presented at the Advances in Neural Information Processing Systems.

Chumbley, J. R., & Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *Neuroimage, 44*(1), 62-70.

Cohen, J. (1992). Statistical power analysis. *Current directions in psychological science, 1*(3), 98-101.

Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., . . . Turk-Browne, N. B. (2017). Computational approaches to fMRI analysis. *Nature Neuroscience, 20*(3), 304-313.

Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human brain mapping, 8*(2-3), 109-114.

Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *Journal of Neuroscience Methods*, *118*(2), 115-128.

Detre, G., Polyn, S. M., Moore, C. D., Natu, V. S., Singer, B., Cohen, J. D., ... & Norman, K. A. (2006, June). The multi-voxel pattern analysis (MVPA) toolbox. In *Poster presented at the annual Meeting of the organization for human brain mapping.*

Ellis, C. T., Lesnick, M., Henselman-Petrusek, G., Keller, B., & Cohen, J. D. (2018). Limitations of Topological Data Analysis for event-related fMRI. *bioRxiv,* 457747.

Erhardt, E. B., Allen, E. A., Wei, Y., Eichele, T., & Calhoun, V. D. (2012). SimTB, a simulation toolbox for fMRI data under a model of spatiotemporal separability. *Neuroimage, 59*(4), 4160-4167.

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Aliaga, A. E., . . . Snyder, M. (2018). FMRIPrep: a robust preprocessing pipeline for functional MRI. *bioRxiv*, 306951.

Fortunato, S. (2010). Community detection in graphs. *Physics reports, 486*(3-5), 75-174.

Friedman, L., & Glover, G. H. (2006). Report on a multicenter fMRI quality assurance protocol. *Journal of Magnetic Resonance Imaging, 23*(6), 827-839.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M., & Turner, R. (1998). Event-related fMRI: characterizing differential responses. *Neuroimage, 7*(1), 30-40.

Friston, K. J., Holmes, A. P., Poline, J., Grasby, P., Williams, S., Frackowiak, R. S., & Turner, R. (1995). Analysis of fMRI time-series revisited. *Neuroimage, 2*(1), 45-53.

Friston, K. J., Zarahn, E., Josephs, O., Henson, R., & Dale, A. M. (1999). Stochastic designs in event-related fMRI. *Neuroimage, 10*(5), 607-619.

Gläscher, J. (2009). Visualization of group inference data in functional neuroimaging. *Neuroinformatics, 7*(1), 73-82.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics, 5*, 13.

Gudbjartsson, H., & Patz, S. (1995). The Rician distribution of noisy MRI data. *Magnetic resonance in medicine, 34*(6), 910-914.

Hill, J. E., Liu, X., Nutter, B., & Mitra, S. (2017). *A task-related and resting state realistic fMRI simulator for fMRI data validation.* Paper presented at the Medical Imaging 2017: Image Processing.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). FSL. *Neuroimage*, *62*(2), 782-790.

Kay, K. N., Winawer, J., Mezer, A., & Wandell, B. A. (2013). Compressive spatial summation in human visual cortex. *Journal of neurophysiology, 110*(2), 481-494.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience, 2*.

Manning, J. R., Zhu, X., Willke, T. L., Ranganath, R., Stachenfeld, K., Hasson, U., ... & Norman, K. A. (2018). A probabilistic approach to discovering dynamic full-brain functional connectivity patterns. *NeuroImage. 180*: 243-252.

McIntosh, A., Bookstein, F., Haxby, J. V., & Grady, C. (1996). Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage, 3*(3), 143-157.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., . . . Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences, 10*(9), 424-430.

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. Cambridge University Press.

Purdon, P. L., & Weisskoff, R. M. (1998). Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. *Human brain mapping, 6*(4), 239-249.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences, 98*(2), 676-682.

Raj, D., Anderson, A. W., & Gore, J. C. (2001). Respiratory effects in human functional magnetic resonance imaging due to bulk susceptibility changes. *Physics in medicine and biology, 46*(12), 3331.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature neuroscience, 16*(4), 486-492.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534.

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature communications, 7*, 12141.

Triantafyllou, C., Hoge, R., Krueger, G., Wiggins, C., Potthast, A., Wiggins, G., & Wald, L. (2005). Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters. *Neuroimage, 26*(1), 243-250.

Wang, Y., Cohen, J. D., Li, K., & Turk-Browne, N. B. (2015). Full correlation matrix analysis (FCMA): An unbiased method for task-related functional connectivity. *Journal of Neuroscience Methods, 251*, 108-119.

Ward, B.D. (2000). Simultaneous inference for fMRI data. Retrieved December 15, 2018, from http://homepage.usask.ca/~ges125/fMRI/AFNIdoc/AlphaSim.pdf

Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G., & Rosseel, Y. (2011). neuRosim: An R package for generating fMRI data. *Journal of Statistical Software, 44*(10), 1-18.

Welvaert, M., & Rosseel, Y. (2013). On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data. *PloS one, 8*(11), e77089.

Welvaert, M., & Rosseel, Y. (2014). A review of fMRI simulation studies. *PloS one, 9*(7), e101953.