Likelihood Paradigm on Multiple Subjects with Task-Induced fMRI Data

By

Cassandra Hennessy

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

August 31, 2020

Nashville, Tennessee

Approved:

Hakmook Kang, Ph.D.

Simon Vandekar, Ph.D.

TABLE OF CONTENTS

Page

CHAPTERS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Functional Magnetic Resonance Imaging

Functional magnetic resonance imaging (fMRI) is used to identity activated regions of the brain. This is done using information on how blood flow changes happen with neural activity, which cannot be done with a regular MRI. FMRI has high spatial resolution, helping to locate the region of the brain that is activated [1]. FMRI most commonly uses Blood Oxygenation Level-Dependent (BOLD) contrast to study local changes in deoxyhemoglobin concentration in the brain [1]. Neuronal activity causes changes in the brain hemodynamics. Cerebral blood flow refreshes areas of the brain that are active during a mental task with oxygenated blood, which causes changes in the MR signal in the active brain region [1]. BOLD contrast is used to estimate the Hemodynamic Response Function (HRF), which is the response to a neural event [1]. The HRF describes changes in BOLD signal over time. Looking at different brain volumes, changes in hemodynamics over time can be used to infer when and where activity is taking place.

Since neuronal activity occurs both in space and time, spatial and temporal resolution of fMRI studies will limit conclusions that can be made [1]. It is not possible to increase both spatial and temporal resolution simultaneously [1]. For spatial resolution, it is common to spatially smooth fMRI data prior to analysis [1]. Brains of varying sizes and shapes are needed for population inference, but in order to compare across subjects, the data needs to be on a standard template [1]. This introduces some spatial imprecision and blurring in group data [1].

Temporal resolution depends on the time between acquisition of each image. The time between each image is not the only factor, though. The speed of the hemodynamic response from a neural event needs to be considered for temporal resolution. Since the neuronal signal that is seen is from the hemodynamic response and not the actual neuronal activity, the absolute timing of brain activation cannot be determined, but the relative timing can be [1].

FMRI data consist of BOLD signal, nuisance parameters, and noise. It is common to assume a linear relationship between neuronal activity and BOLD responses, implying that the magnitude and shape of the HRF does not depend on preceding stimuli [1]. FMRI signal is corrupted by random noise and nuisance components that come from hardware reasons and from the subjects (e.g. the subject moving, breathing, and heartbeat all add noise in space and time). Certain regions of the brain have higher amounts of variability, so the noise component is spatially dependent [1].

1.2 Conventional Approach

The conventional approach for this data analysis is to run a general linear regression at each voxel. From this regression, a t-statistic is obtained at each voxel to test the hypothesis of voxel-level activation. Using a predetermined threshold, the voxel is active if the t-statistic reaches this threshold. From the active and inactive voxels, a brain activation map can be created. Running the statistical analysis on all of the voxels creates a multiple comparisons problem which leads to inflation of the family-wise error rate (FWER). Increasing the threshold for significance will control family wise error rate but will also increase the Type II error rate [2].

Standard approaches to controlling multiple comparisons include controlling the false discovery rate (FDR) [3] or finding a trade-off between Type I and Type II error rates. Both Bonferroni corrections and random field theory (RFT) control the FWER by setting a predetermined level to balance Type I and II error rates [4]. Spatial correlation in fMRI data can be problematic for Bonferroni corrections by making the FWER too conservative [5], which is why random field theory is preferred. RFT can be used to find the threshold necessary for the FWER from smooth statistical maps [5]. In both of these approaches, the Type I error rate does not converge to zero as sample size increases, so there will always be some false positive findings [2].

1.3 Likelihood Paradigm

In the conventional approach, a p-value is obtained at each voxel, used as a measure of the strength of evidence. Another approach, used by Kang et al. (2015), is to measure how often misleading evidence will be observed with a likelihood ratio. Likelihood ratios are derived at each voxel to measure the strength of evidence in the data about the level of voxel activation [2]. Using this approach, the family-wise error rate is controllable even with a large number of tests. This approach has multiple advantages over the conventional approaches like Random Field Theory and controlling the FDR. The likelihood ratio is not adjusted for the number of comparisons. The false positive and negative error rates both converge to zero as the number of images and the sample size increases [2]. Global error rates typically inflate with multiple comparisons because global rates are the accumulation of rates over each comparison. In the likelihood paradigm approach, the global error rates converge to zero [2].

CHAPTER 2

METHODS

2.1 Data

Albert et al. (2016) studied amygdala activity and functional connectivity with attention bias in women with depression. FMRI images were collected for 74 postmenopausal women with a history of major depression, using an emotion dot probe task. Event-related design was used for the fMRI. In an event-related design, the stimulus is short discrete events, whose timing can be randomized. This design allows for estimation of key features of the hemodynamic response function, such as onset and width [1]. These features can be used to make inference about relative timing of activation across different conditions [1], in order to test the hypotheses of interest. By using event-related designs, effects from boredom and systematic patterns do not have an effect because of the short events [1]. However, this design typically has lower power to detect activation than block designs.

2.2 Preprocessing

An important step in analyzing fMRI data is preprocessing. Preprocessing helps reduce the influence from acquiring data and physiological artifacts, such as muscle movement and breathing [1]. It is assumed that voxels in a brain volume were acquired simultaneously, that all brains are registered (important for group analysis), and that the data come from a single location, meaning the subjects don't move between scans. Preprocessing also standardizes locations of brain regions across subjects. This increases the validity and sensitivity in group analysis [1]. The main steps in preprocessing are motion correction, slice-timing correction, coregistration and normalization, spatial smoothing, and temporal filtering. These steps can be reordered, some can be combined, and some can be omitted depending on the study.

The first step is motion correction, which helps reduce motion caused by the subjects moving during the fMRI. The assumption that is made is that each data point in a voxel's timeseries only consists of a signal from that voxel (i.e. the participant did not move between scans) [1]. When movement does happen, the signal from the neighboring voxels will mess up the signal [1]. To correct for this, the functional data is realigned to a common reference. FSL's MCFLIRT (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MCFLIRT) is a motion correction tool based on FLIRT. MCFLIRT uses the middle volume as the reference image [7]. Each fMRI volume is registered to the reference image separately using rigid body transformation with 6 degrees of freedom [8]. This transformation matches the input image to the target image by minimizing a cost function [1]. The image is then resampled using interpolation to create new motion corrected voxel values, and then repeated for each brain volume [1]. Motion parameters are estimated during the rigid body realignment stage and included as nuisance parameters in the analysis [9].

FMRI data are collected sequentially, imaging one slice at a time to create a full brain volume. It is assumed that that all slices are acquired simultaneously, so slice-timing correction is done during preprocessing to compensate the time delay between slices. The timeseries for each slice is shifted back in time by the duration it took to acquire that slice. Sladky et al. (2011) showed that slice-timing correction can lead to significant increases in statistical power for studies with longer TR values (e.g. TRs 2 seconds or longer) [10].

The next step in preprocessing is coregistration and normalization. The assumption that is made is that all individual brains are registered so that each voxel is located in the same anatomical region for all subjects for group analysis [1]. To correct for this assumption, first the structural and functional data need to be reoriented to the same orientation of the standard image. MNI-152 standard was used as the standard space in this analysis. Also, the structural data needs to have the brain extracted from the skull to use. This can be done using the "bet" tool from FSL. FSL's FLIRT tool is used for coregistration and normalization. FLIRT performs coregistration by using affine transformation to register the input image to the reference image [11]. Normalization registers each subject's anatomy to a standardized space using a template brain [1]. This aligns brains of different subjects so that a given voxel represents the "same" location. This results in a normalized image that can be compared with similarly normalized images from other subjects. Normalization offers the benefits that results can be generalized, it allows comparison across subjects, and spatial locations can be interpreted consistently. A drawback though, is that normalization reduces spatial resolution [1].

Depending on the analysis being done, spatial smoothing might be a step in the preprocessing. In order to use random field theory (RFT) in the data analysis, spatial smoothing is necessary because it ensures that the assumptions of random field theory are valid. It can help reduce random noise in individual voxels and increase the signal-to-noise ratio within a region [1]. From FSL, fslmaths uses a Gaussian kernel and mean filtering to smooth the data [11]. Spatial smoothing is done by replacing a voxel value with a weighted average of nearby voxels using Gaussian weighting [1]. The full width half maximum (FWHM) value equal to 6mm is used for spatial smoothing in this analysis. Spatial smoothing can improve inter-subject registrations and overcome limitations in spatial normalization [1]. Spatial smoothing is not a necessary step for the likelihood paradigm approach.

A final step in preprocessing is temporal filtering. There are multiple sources of noise in fMRI data (i.e. physiological, subject movement, the fMRI machine). Temporal filtering is used in the preprocessing in order to remove some of this noise and improve the signal-to-noise ratio [12]. A band pass filter can be used for temporal filtering. It uses both high-pass and low-pass filters to remove frequencies outside of the specified band. The goal is to preserve the signal while also eliminating noise outside of the band. Going forward, noise inside of the band can be handled in the model. For this analysis, a low-pass cutoff of .08 Hz and a high-pass cutoff of .01 Hz are used with FSL math's bptf option. This option uses a nonlinear high-pass and a Gaussian linear low-pass [13].

Preprocessing is done on the whole brain following these steps to create the timeseries at each voxel for each subject. In order to implement the likelihood paradigm approach, spatial smoothing is not done in the preprocessing. For the likelihood approach, the cerebrospinal fluid (CSF) region of the brain is used for computing the alternatives. The CSF region is extracted during preprocessing to create a timeseries at each voxel, for each subject, with only the CSF region. This gives us two sets of timeseries: one for the whole brain and one just for the CSF region.

2.3 Data Analysis

After the data is preprocessed, the timeseries are used to model the fMRI signal. General linear model methods were used to model the timeseries for a single subject at each voxel. The goal is to test whether the activity in a brain region is systematically related to any input functions, such as the convolution between an external stimulus function and the hemodynamic response function (HRF). The HRF was created using the spm_hrf function from SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). Assuming the HRF is known, the data can be modeled with a multiple linear regression model. The task-related BOLD response is summarized into the design matrix, with a column for each predictor. The design matrix is 270 x 12 for the 270 scans and 12 predictors, including 6 experimental conditions and 6 motion parameters. FMRI activation is modeled in a single voxel for a single subject:

$$Y_i = X_i\beta_i + \epsilon_i$$

$$i = 1, ..., N(\text{where N is the number of voxels})$$

X = the design matrix at each subject and voxel

$$\epsilon_i \text{ is assumed to follow AR(1) process}$$

At each voxel, data from neighboring voxels were used for spatial correlation and temporal correlation was modeled as an autoregressive process of order 1 (AR(1)). Regression parameters and variance components were estimated using generalized least squares at each voxel. At each voxel, we have:

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$$
$$\hat{\Omega} = (X'\Sigma^{-1}X)^{-1}$$
$$\text{where } \Sigma = Cov(\epsilon_i) \text{ and } \Omega = Cov(\beta)$$

Once the parameters were estimated, they were used to test for the effect between the two parameters of interest, $\beta_1 - \beta_2$, where β_1 and β_2 are the effects of the first and second stimuli. It was noticed that one participant's values of this difference were extremely different than the other subjects. With no explanation for this, this subject was removed from the rest of the analysis.

2.4 Conventional (Traditional) Approach

For the conventional approach for fMRI data analysis, spatial smoothing was done during preprocessing using a FWHM value of 6mm. The smoothed timeseries data were used in the GLM approach described above to estimate the model parameters. To test the hypothesis, $H_0 : \beta_1 - \beta_2 = 0$, the difference between $\hat{\beta}_1$ and $\hat{\beta}_2$ was taken at each voxel. Using MATLAB's one sample t-test, t-tests were run at every voxel across the subjects. A sample consisted of data from each subject at one voxel (i.e. for N = 14 subjects with ~200,000 voxels, 200,000 t-tests were run with 14 data points in each one). For each t-test, a decision was made to either reject or fail to reject the null hypothesis. The p-values and test statistics for each t-test were collected to be used for further steps in the analysis.

Running all of these t-test can create multiple comparisons problems. One way to account for multiple comparisons is to control the false discovery rate (FDR). The FDR is the expected proportion of falsely rejected null hypotheses [3]. Controlling the FDR controls the proportion of false positives among the rejected tests [1]. If all null hypotheses are true, then the FDR equals the FWER, but otherwise the FDR is a smaller error rate [3]. This means that the method of controlling the FDR is less conservative than a method that controls the FWER.

To control for the FDR, the fdh_bh function was used in MATLAB. This function controls the FDR of a family of hypothesis tests using the Benjamini and Hochberg procedure [14]. Using the p-values from the one sample t-tests, each test rejected or failed to reject the null hypothesis (that all the data came from a normal distribution with mean = 0) at a false discovery rate of .05. For voxels where the test failed to reject the null hypothesis, these voxels are considered inactive, and the rest were considered active voxels. This information was used to map the active voxels onto MNI-152 standard space.

Controlling the FDR can be a useful method to account for multiple comparisons in fMRI analysis, but it still has some issues. Using this method to control the false discovery rate can result in many false negatives [2]. Also, the Benjamini and Hochberg approach assumes all tests are independent of each other. However, this is not that realistic in fMRI data. It is more realistic to assume that tests are dependent because neighboring voxels will likely have similar p-values [1].

Since multiple comparisons inflate the family-wise error rate, another way to account for the multiple comparisons is to control the FWER. One way this can be done in fMRI data analysis is by using random field theory (RFT). Random field theory uses a height threshold to find voxels that have a value greater than the threshold, in which it is concluded that there is an effect at these voxels [5]. This process can help localize which voxels are activated. In order to control the FWER with random field theory, the number of tests needs to be considered [5].

To use random field theory, the spatial correlation of the statistical maps is estimated using a Gaussian kernel in the preprocessing stage [5]. The probability of a family wise error is approximately equivalent to the expected Euler Characteristic [5]. The expected Euler characteristic is found at different thresholds based on the smoothness values. This is then used to calculate the threshold where it is expected that 5% of statistical maps from the null to have at least one area above the threshold [5]. The expected EC is calculated using the equation:

$$E[EC] = resel(4\log(2))(2\pi)^{-\frac{3}{2}}Z_t e^{-\frac{1}{2}Z_t^2}$$

In this equation, the number of ‘resels’ are calculated based on Worsley’s procedure [4] using the FWHM value (in this analysis, FWHM = 6mm). The calculated threshold is compared to the test statistics from the voxel-level one sample t-tests to determine active and inactive voxels. These voxels were also mapped back onto MNI-152 standard space to help visualize the brain activation across methods.

2.5 Likelihood Paradigm

Instead of using the traditional approaches of controlling the family-wise error rate or controlling the false discovery rate, a different approach to control the error rates in fMRI data analysis is to use the likelihood paradigm. With the traditional approaches, there will always be some false positives [2]. The likelihood paradigm approach uses likelihood ratios to measure the strength of evidence in the data about each voxel's activation. The likelihood ratio is used as the measure of strength of evidence instead of the p-value [2]. Unlike the p-value, the likelihood ratio is unaffected by the number of examinations of the data because it is a measure of observed evidence [15]. The effects of multiple comparisons are from the hypothesis testing process. In hypothesis testing, we look at the tail area probability. But in the likelihood paradigm, the likelihood ratio itself measures the strength of statistical evidence.

Using the likelihood paradigm has advantages over the traditional approaches, as the family-wise error rate stays small even with a large number of tests and will converge to 0 [2]. The likelihood ratio does not need to be adjusted for the number of comparisons because it is only a measure of the observed evidence [2]. Also, the rates from the likelihood paradigm approach that are similar to the traditional false positive and false negative error rates both converge to 0 as the amount of information increases [2]. In this approach, instead of fixing one error rate and minimizing the other, the law of likelihood minimizes the average error rate at the value of k [2]. "k" is the cutoff point used as the indication of sufficient evidence for one hypothesis over the other in the likelihood ratio.

In the likelihood paradigm, the likelihood ratios are used as measures of statistical evidence for one hypothesis over another [15]. The law of likelihood says that the data favor the hypothesis that better predicts the observed data or gives a higher likelihood [2]. The strength of evidence for one hypothesis over another is measure by the likelihood ratio:

$$\mathrm{LR} = \frac{L(\theta_1)}{L(\theta_0)}$$

Cutoff points can be used to determine "strong" evidence for one hypothesis over the other. Using k = 20 as the cutoff controls the probability of misleading evidence, bounded at 1/20 = .05. The actual probability of observing misleading evidence is often much less than its bound (1/20), and eventually converges to 0 [2]. This gives a similar comparison to the Type I error rate in the traditional approach [2]. Blume outlined three quantities for the likelihood framework [2]. First is the likelihood ratio: a measure of the strength of evidence. This can be used as a descriptive tool for what the data say and can answer the question, "How strong is the statistical evidence for a particular voxel that is claimed as activated?" [2]. Next is the probability of observing misleading evidence. Misleading evidence is strong evidence in favor of the incorrect hypothesis over the correct hypothesis [15]. The probability of observing misleading evidence can be written as:

$$\mathrm{mis}_0 = P(\mathrm{LR} > k | H_0)$$
$$\mathrm{mis}_1 = P(\mathrm{LR} < \frac{1}{k} | H_1)$$

These quantities are similar to the traditional Type I and Type II error rates, although their approaches have different goals [2]. The third quantity is the probability that observed evidence is misleading, which can be measured by:

$$P(H_0 | \mathrm{LR} > k)$$
$$P(H_1 | \mathrm{LR} < \frac{1}{k})$$

2.6 Likelihood Paradigm Analysis

The likelihood paradigm approach was used in the data analysis following the preprocessing and estimation steps. For this approach, spatial smoothing was not used in the preprocessing stage. Spatial smoothing was accounted for by using data from neighboring voxels in the estimation step. At each voxel, the difference between $\hat{\beta}_1$ and $\hat{\beta}_2$ was taken to test the hypothesis of interest. This difference was treated as the parameter of interest used in the likelihood ratio. To account for multiple subjects, the mean of this difference was taken at each voxel, across subjects. In order to take this mean, voxels were removed if any of the patients did not have data at a specific voxel, only keeping voxels that all subjects had in common.

The covariance estimates from the generalized least squares were also used in the likelihood ratio formula. Since the difference of $\hat{\beta}_1$ and $\hat{\beta}_2$ is the parameter of interest, the variance of the difference was calculated using:

$$\hat{\sigma}^2 = \hat{\Omega}(1,1) + \hat{\Omega}(2,2) - 2\hat{\Omega}(1,2)$$

With the variance at each voxel, the mean of $\mathrm{Var}(\widehat{\beta_1 - \beta_2})$ was taken at each voxel across each subject.

In order to create the alternatives, the timeseries from the CSF region of the brain were used. Through the same estimation process, regression parameters and variance components were estimated for each subject. Again, the difference between $\hat{\beta}_1$ and $\hat{\beta}_2$ was taken at each voxel. The mean, median, variance and interquartile range were each taken across all voxels and all subjects to result in one mean value, one median value, one variance value, and one IQR value. These values were used in the four alternatives of interest:

$$\delta_1 = \text{Mean} + 2\sqrt{\text{Var}}$$
$$\delta_2 = \text{Median} + 2\text{IQR}$$
$$\delta_3 = \text{Mean} + 3\sqrt{\text{Var}}$$
$$\delta_4 = \text{Median} + 3\text{IQR}$$

Likelihood ratios were computed for each alternative. The null values used for all four likelihood ratios were 0. The likelihood ratio (derivation in Appendix A.3) was defined as:

$$LR = \frac{L_n(\mu_1)}{L_n(\mu_0)} = \exp\{\frac{n(\mu_1 - \mu_0)}{\sigma^2}[\bar{X}_n - \frac{(\mu_1 + \mu_0)}{2}]\}$$

$$\begin{aligned} \mu_1 &= \text{alternative} \\ \mu_0 &= \text{null} \\ \bar{X}_n &= \text{average difference in } \widehat{\beta_1 - \beta_2} \\ \sigma^2 &= \text{average variance for } \widehat{\beta_1 - \beta_2} \end{aligned}$$

This process was of taking the mean of $(\widehat{\beta_1 - \beta_2})$ and the mean of Var $(\widehat{\beta_1 - \beta_2})$ was repeated for the different number of subjects: N = 14, 29, 44, 59, and 73. All four likelihood ratios were computed for each number of subjects.

For each likelihood ratio, two sets of outputs were computed: binary outputs and outputs with an inconclusive zone. Using a cutoff value of k = 20, the binary results were computed as:

$$\text{If } \log(\text{LR}) \geq \log(20) \rightarrow \text{active voxel}$$

$$\text{If } \log(\text{LR}) < \log(20) \rightarrow \text{inactive voxel}$$

The resulting matrix of 0's and 1's was mapped onto MNI-152 standard space to visualize the brain activation. To also include an inconclusive zone, the results were computed as:

$$\begin{gathered} \text{If } \log(\text{LR}) \geq \log(20) \rightarrow \text{active voxel} \\ \text{If } \log(\frac{1}{20}) < \log(\text{LR}) < \log(20) \rightarrow \text{inconclusive zone} \\ \text{If } \log(\text{LR}) < \log(\frac{1}{20}) \rightarrow \text{inactive voxel} \end{gathered}$$

The results with the inconclusive zone were also used to create a brain activation map.

2.7 Error Rates:

These different approaches to analyzing fMRI data have different ways of accounting for multiple comparisons. A potential argument about the comparison of these approaches is that the likelihood approach will perform better because the weak evidence is ignored in the likelihood approach [2]. However, in hypothesis testing, weak evidence is evidence for the null. To make these comparisons with a little less bias, the likelihood ratio results were dichotomized so all approaches have two regions.

For each method, false positive and false negative rates were computed, using the activation results from all subjects (N = 73) as the "truth". To compare across the different approaches, only the binary outputs from the likelihood paradigm approach were used. The false positive and false negative rates are calculated as:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}$$

In the above error rates, FP = false positive, FN = false negative, TN = true negative, and TP = true positive. The FPR and FNR were calculated for N = 14, 29, 44, and 59. In the likelihood paradigm, these rates were calculated for each of the four likelihood ratios. The same formulas for false positive and negative error rates were used for all the three methods: likelihood paradigm, RFT, and controlling the FDR.

2.8 Multi-Subject Estimation Process

Another approach for multiple subject analysis in fMRI data is to consider integrating subject-level data to estimate the regression parameters and variance components, instead of first estimating parameters for each subject. This approach was considered as a secondary analysis.

To compare this approach to the main analysis approach, five random subjects were selected using MATLAB's RandStream and datasample functions. With these five subjects, the same analysis process was used as described in Section 2.6, Likelihood Paradigm Analysis. The same five subjects were used in the new approach. Each subject still had their own design matrix and timeseries data. Only voxels that all five subjects had in common were considered in this analysis.

The same estimation process was used in this method with a few modifications to integrate subject-level data. Then, the model parameters were estimated at each voxel.

$$\hat{\beta} = [\sum_{i=1}^{n} X_i^T \Sigma_i^{-1} X_i]^{-1} [\sum_{i=1}^{n} X_i^T \Sigma_i^{-1} y_i] \quad \text{(where n = 5)}$$

$$\hat{\Omega} = [\sum_{i=1}^{n} X_i^T \Sigma_i^{-1} X_i]^{-1} \quad \text{(where n = 5)}$$

In order to use the Likelihood Paradigm approach, parameters from the CSF region were also estimated using the two approaches. For both approaches described here, the likelihood ratios were calculated as done before. Brain activation maps of all likelihood ratios were constructed to compare the two approaches.

This multiple subject approach has the advantage to be able to perform valid population-level inference. However, it has extreme computational demands due to the large amount of data being analyzed [1].

CHAPTER 3

RESULTS

3.1 Error Rates

Table 3.1 shows the false positive rates for each method at each sample size. The four likelihood ratios are all 0. This is because there was no activation found in our “true” results from N=73. Both the random field theory and false discovery rate methods have increasing false positive rates as the sample size increases. Although, the error rates are still relatively small.

Table 3.1 False Positive Rates

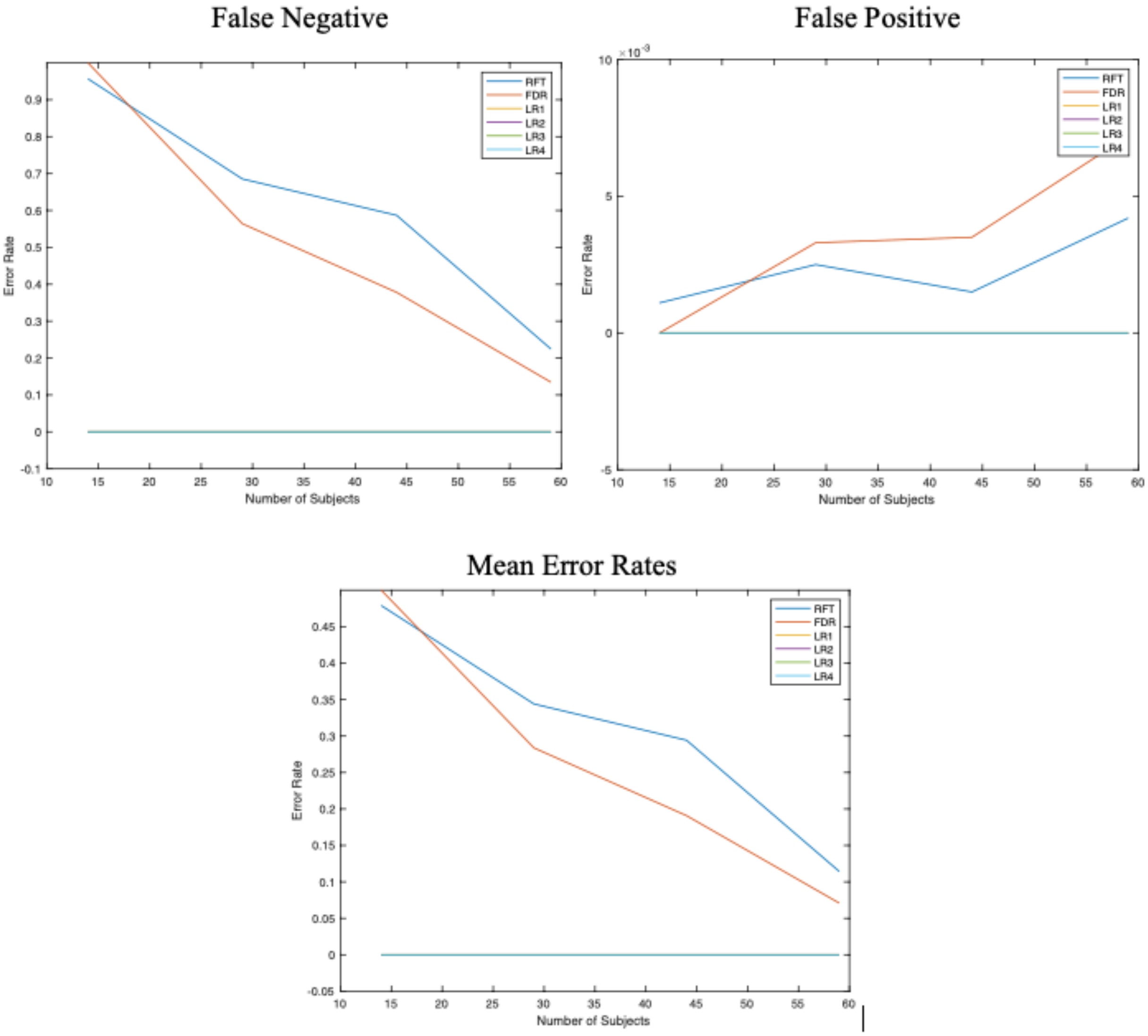
Sample Size	RFT	FDR	LR1	LR2	LR3	LR4
14	.0011	0	0	0	0	0
29	.0025	.0033	0	0	0	0
44	.0015	.0035	0	0	0	0
59	.0042	.0072	0	0	0	0

Table 3.2 shows the false negative rates for all the methods at each sample size. For the likelihood ratio, there are no true activations, so all the rates are 0. For the random field theory method, the false negative rate is high at the smallest sample size, N = 14, and decreases as the sample size increases. However, at N = 59, the false negative rate is still relatively high at .2245. For the method controlling the false discovery rate, the false negative rate is 1 for N = 14, meaning there were no true negatives. This is because there were no activated voxels in the FDR method at N = 14. As the sample size increased, the false negative rate decreased down to .1346 for N = 59 subjects. The false negative rates were significantly higher for the two traditional methods than for the likelihood ratio methods.

Table 3.2 False Negative Rates

Sample Size	RFT	FDR	LR1	LR2	LR3	LR4
14	.9567	1	0	0	0	0
29	.6855	.5642	0	0	0	0
44	.5869	.3781	0	0	0	0
59	.2245	.1346	0	0	0	0

Figure 3.1 Error rates for all methods across subjects



False Negative

False Positive

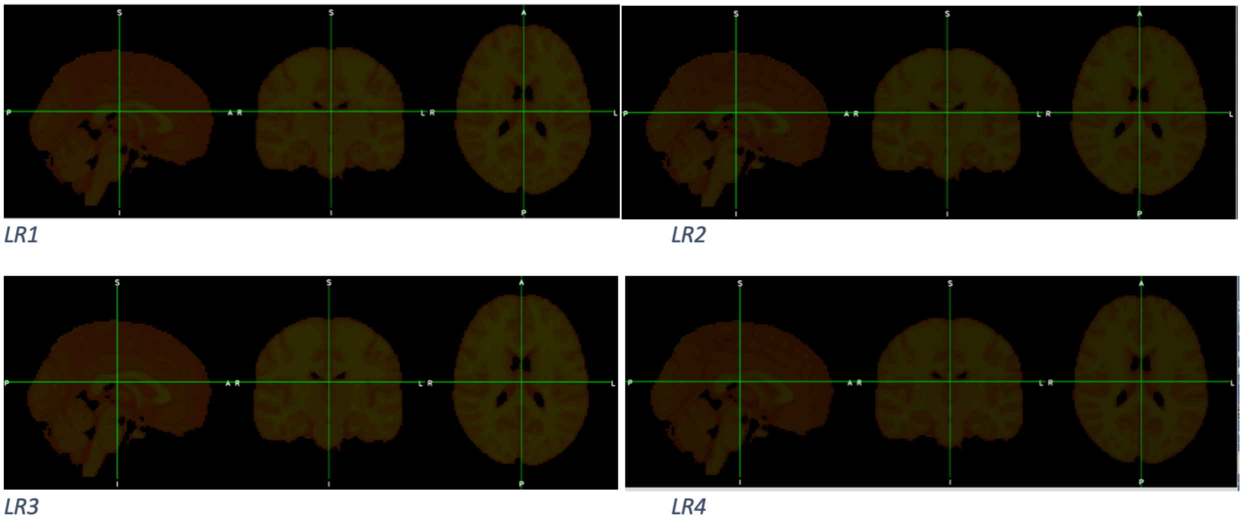
Mean Error Rates

The false positive and negative error rates from Tables 3.1 and 3.2 were summarized into Figure 3.1. The likelihood ratio false positive rates all stay fixed at 0, while the RFT and FDR false positive rates increase. For the false negative rates, the likelihood ratios are all 0 because there was no activation detected in the likelihood ratio method, while the RFT and FDR rates decrease as the number of subjects increases.

3.2 Brain Activation Maps

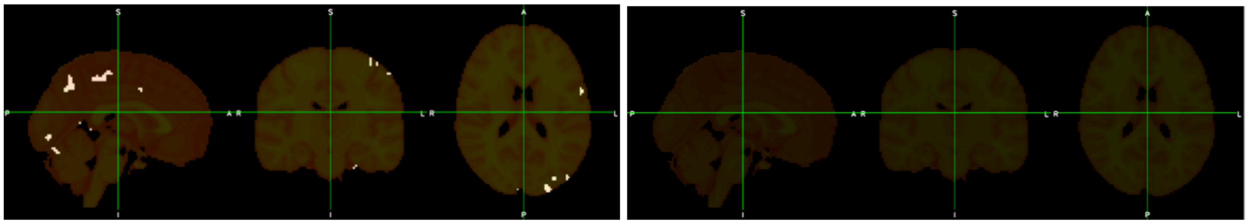
Figure 3.2 shows the activated voxels in the brain for the four likelihood ratios at N = 73. The likelihood ratio images are based on the binary results. Images for the brain activation maps from the likelihood ratio binary outputs for the other sample sizes (N = 14, 29, 44, and 59) all show no activation as well. These brain activation maps are included in Appendix A.1. To compare the likelihood ratio approach to the RFT and FDR methods, the focus was on the binary results. The brain activation maps for the likelihood ratio results with the inconclusive zone are also shown in Appendix A.1.

Figure 3.2 Brain activation maps for N = 73 (LR binary results)



The brain maps for random field theory and controlling the false discovery rate methods do show some activation. Figures 3.3 – 3.7 show the brain activation maps for these two methods at each sample size. At the smallest sample size, N = 14, there is little activation in the RFT method, and no activation in the FDR methods. For both RFT and FDR, the number of activated voxels increase as the sample size increases. The activated regions are consistent with the different sample sizes.

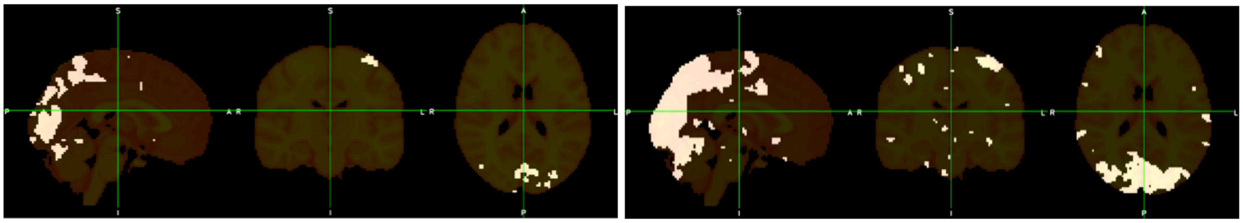
Figure 3.3 Brain activation maps for random field theory and false discovery rate (N=14)



RFT

FDR

Figure 3.4 Brain activation maps for random field theory and false discovery rate (N=29)



RFT

FDR

Figure 3.5 Brain activation maps for random field theory and false discovery rate (N=44)

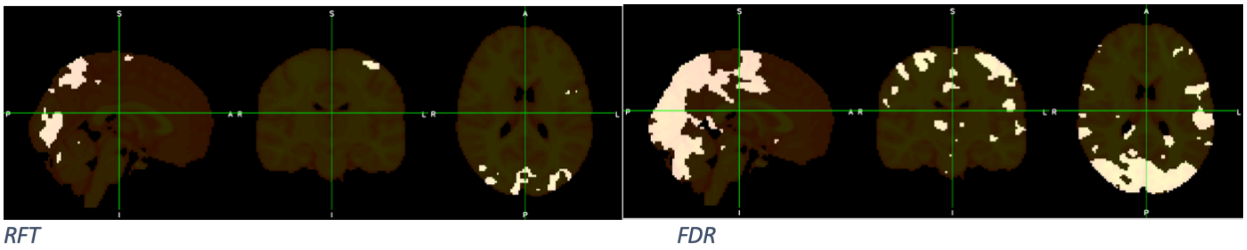


Figure 3.6 Brain activation maps for random field theory and false discovery rate (N=59)

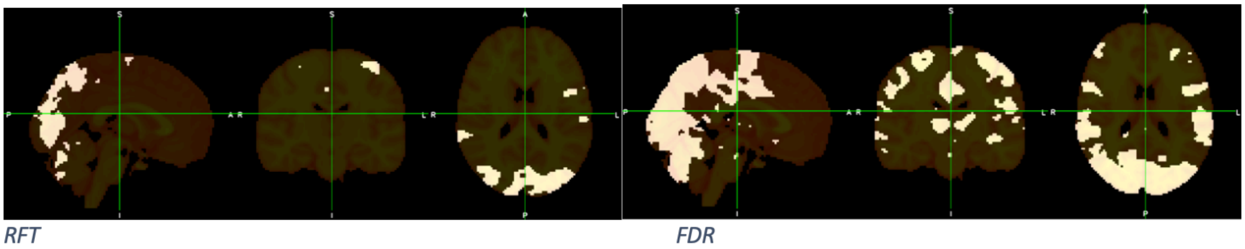
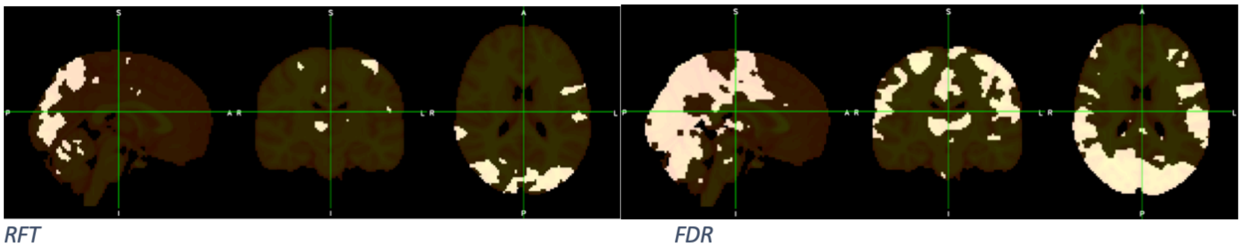


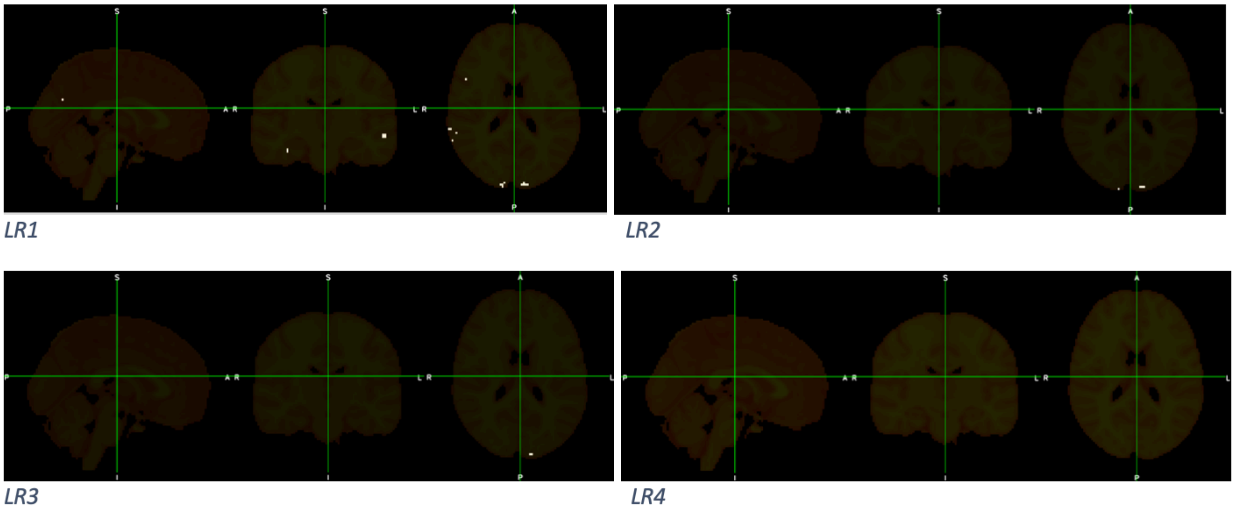
Figure 3.7 Brain activation maps for random field theory and false discovery rate (N=73)



3.3 Five Subject Test

Brain activation maps for the four likelihood ratios from the five random subjects following the main analysis approach are shown in Figure 3.8. There are very few activated voxels in this analysis. There are a few activated voxels in the first, second and third likelihood ratios, but there were no activated voxels in the fourth likelihood ratio.

Figure 3.8 Binary LR activation maps for single subject 5 random subjects


LR1

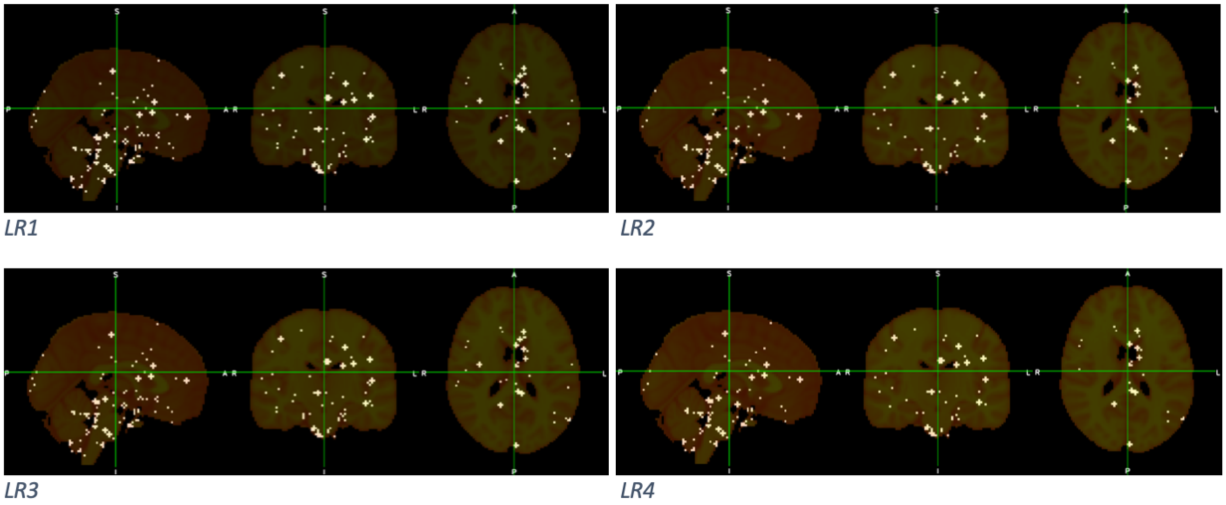
LR2

LR3

LR4

The secondary analysis approach, integrating subject-level data in the estimation process, resulted in brain activation maps seen in Figure 3.9. These activation maps show some scattered activation throughout the brain for all four likelihood ratios. Compared to the analysis with single subject estimation and then group analysis, the multi-subject estimation did result in some activated voxels, unlike the almost empty activation maps in Figure 3.8.

Figure 3.9 Binary LR activation maps for multi-subject 5 random subjects


LR1

LR2

LR3

LR4

CHAPTER 4

SIMULATION

A simulation study was ran to validate the results of the data analysis using spatially and temporally correlated timeseries for each voxel. A general linear model was fit at each voxel and then similar approaches of the data analysis were used to calculate the false positive and false negative rates in each approach: Likelihood Ratio, Random Field Theory, and False Discovery Rate.

4.1 Data Generation:

The timeseries were generated with a length of T = 128 and spatial dimension 32x32 voxels. Three beta coefficients were used for simplicity. In order to create the design matrix, two boxcar stimuli were convolved with the canonical hemodynamic response function (HRF). The HRF was created using the spm_hrf function from SPM8. It is assumed that there are two active, square blocks: one is 6x6 voxels and the other is 8x8 voxels.

Spatial correlation was considered by using an exponential covariance function with a decaying parameter = 2 and variance = 2.5. An AR(1) process was used for temporal correlation, with an AR parameter = .4.

The effect size at each voxel was created by fixing the β's and adding random noise. Three effect sizes were created, a small (about .2), medium (about .4), and large (about .6). Data were generated with a varying number of subjects (N=10, 20, 30, 40) with each effect size to compare across sample size. For the likelihood ratio analysis, the CSF region of the brain was simulated using an effect size of 0. At a later step, data were generated with a larger effect size (about 1) for 10 subjects.

4.2 Methods:

With the simulated timeseries, the same process was used as in the data analysis. A general linear regression was fit at each at voxel, while considering temporal and spatial correlation. Temporal correlation was modeled as an AR(1) process and spatial correlation was accounted for using data from neighboring voxels. $\hat{\beta}'\mathrm{s}$ and the covariance matrix were estimated at each voxel for each subject.

In order to test the null hypothesis, $H_0 : \beta_1 - \beta_2 = 0$, the difference was taken at each voxel for all subjects in each of the simulations. Within each simulation, the mean was taken at each voxel across the subjects, resulting in an 32x32 matrix. The estimated covariance matrices also had the mean taken across subjects at each voxel. The mean difference in $\hat{\beta}'\mathrm{s}$ and mean covariances were used to estimate the likelihood ratio. The same four alternatives were used in the simulation as in the data analysis for the likelihood ratios. The likelihood ratios were calculated with the formula:

$$LR = \frac{L_n(\mu_1)}{L_n(\mu_0)} = \exp\{\frac{n(\mu_1 - \mu_0)}{\sigma^2}[\bar{X}_n - \frac{(\mu_1 + \mu_0)}{2}]\}$$

$$\mu_1 = \text{alternative}$$

$$\mu_0 = \text{null}$$

$$\bar{X}_n = \text{average difference in } \widehat{\beta_1 - \beta_2}$$

$$\sigma^2 = \text{average variance for } \widehat{\beta_1 - \beta_2}$$

Using the cutoff k = 20, voxels were dichotomized into “active” and “inactive” for the binary outputs results. The 100 voxels in the two active squares in the generated data were used as the truth compared to the binary outputs in calculating the false positive and false negative rates. Once the error rates were calculated, the error rates for each simulation were averaged to result in one error rate for each scenario (each effect size for the different sample sizes).

The same generated data was used for the other methods: controlling for the false discovery rate and random field theory. For these two methods, the data was smoothed using a FWHM = 6mm. As in the data analysis, the difference was calculated for the hypothesis, $H_0 : \beta_1 - \beta_2 = 0$, at each voxel for each subject. A t-test was performed at each voxel across the subjects. To control for the false discovery rate, the Benjamini and Hochberg procedure was used [1]. The procedure results in voxels that are significant (reject the null) and voxels that are not significant (fail to reject the null). If the null is rejected, that voxel is considered “active” and the other voxels are “inactive”. Using the same assumed true active voxels as in the LR analysis, the false positive and false negative rates were calculated for each simulation. Then, the error rates were averaged over the 300 simulations to result in an average false positive rate and average false negative rate for each sample size and effect size.

For Random Field Theory, the same methods were used as in the data analysis. The FWHM was set to 6 in order to calculate the resels. Then from the number of resels, the expected Euler characteristic (EC) is calculated and then the Z threshold is found. This threshold was used to determine “active” and “inactive” voxels. The false positive and false negative rates were calculated the same way as the LR and FDR methods, using the 100 active voxels as the truth. Again, the error rates were averaged over the simulations, resulting in average rates for each sample size and effect size.

4.3 Results:

False negative and false positive rates were averaged across all 300 simulations to get the average rates shown in Table 4.1. For the small and medium effect sizes, the average false negative rates for the likelihood ratio methods were all approximately equal to 1. For the large effect size, the false negative rates were smaller than 1, especially for the first likelihood ratio, but increased with sample size, converging to 1. For all three effect sizes, the average false negative rates decrease as the sample size increases for both RFT and FDR methods. For the medium and large effect sizes, the rates start to converge to 0 as the sample size increases.

For all three effect sizes and all sample sizes, the false positive rates for the likelihood ratios are zero. For RFT and FDR methods, the false positive rate increases as the sample size increases. This is the same trend that occurred in the data analysis, although the rates get relatively larger in the medium and large effect sizes.

Table 4.1 False negative and false positive error rates across the simulations.

Average False Negative Rates for Small Effect Size

Sample Size	RFT	FDR	LR1	LR2	LR3	LR4
10	.7207	.9742	1	1	1	1
20	.5290	.5255	1	1	1	1
30	.1955	.1104	1	1	1	1
40	.1525	.0744	1	1	1	1

Average False Positive Rates for Small Effect Size

Sample Size	RFT	FDR	LR1	LR2	LR3	LR4
10	.0102	.0002	0	0	0	0
20	.0128	.0140	0	0	0	0
30	.0264	.0493	0	0	0	0
40	.0316	.0620	0	0	0	0

Average False Negative Rates for Medium Effect Size

Sample Size	RFT	FDR	LR1	LR2	LR3	LR4
10	.1609	.1912	.9887	.9960	.9982	1
20	.0168	.0058	.9985	.9997	1	1
30	.0026	.0004	.9999	1	1	1
40	.0003	0	1	1	1	1

Average False Positive Rates for Medium Effect Size

Sample Size	RFT	FDR	LR1	LR2	LR3	LR4
10	.0429	.0376	0	0	0	0
20	.0760	.1106	0	0	0	0
30	.1041	.1539	0	0	0	0
40	.1231	.1787	0	0	0	0

Average False Negative Rates for Large Effect Size

Sample Size	RFT	FDR	LR1	LR2	LR3	LR4
10	.0161	.0126	.6748	.7687	.8242	.9687
20	.0002	0	.8061	.8883	.9264	.9960
30	0	0	.7823	.8816	.9284	.9981
40	0	0	.7460	.8637	.9195	.9989

Average False Positive Rates for Large Effect Size

Sample Size	RFT	FDR	LR1	LR2	LR3	LR4
10	.0972	.1071	0	0	0	0
20	.1303	.1740	0	0	0	0
30	.1686	.2283	0	0	0	0
40	.1941	.2559	0	0	0	0

Figure 4.1 Simulation false positive and negative error rates

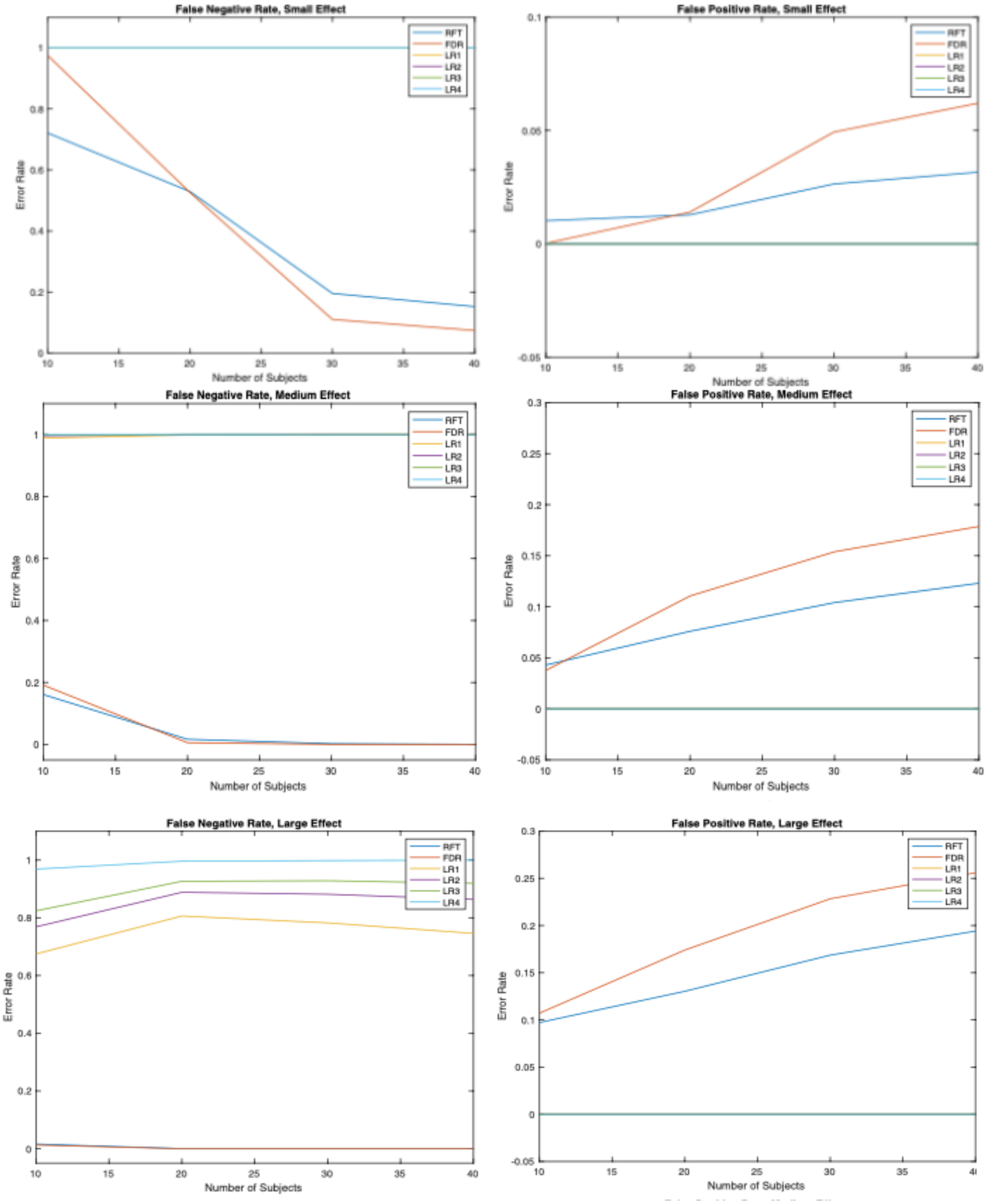


Figure 4.1 shows the error rates visually for the small, medium and large effect. All four likelihood ratios are included in the graphs, although most of them have the same value (i.e. all false positive rates are 0 and many of the false negative rates are 1).

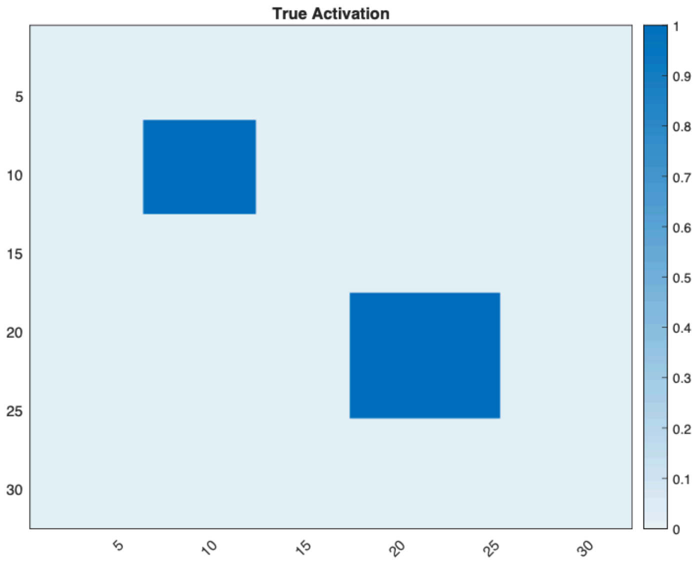
Table 4.2 shows the error rates for the larger effect size added in later. This effect size was only simulated with 10 subjects. The false positive rates stay around 0, as seen in Table 4.1 with the smaller effect sizes. The false negative rates are much smaller now with this larger effect size. In the original three effect sizes, the false negative rates were all 1 or close to 1. Here, the false negative rates are no longer close to 1, but they are still relatively large compared to the RFT and FDR methods with the smaller effect sizes, seen in Table 4.1.

Table 4.2: Error rates for effect size = 1; N = 10

	False Negative Rate	False Positive Rate
LR1	.1358	.0006
LR2	.1792	.0003
LR3	.2090	.0001
LR4	.3830	0

Heatmaps were generated for all the methods, effect sizes, and sample sizes in the simulation. Heatmaps for each method were generated with the average voxel activation for each simulation. One likelihood ratio method was used for the visual comparison. The likelihood ratio methods produced very similar results. Figure 4.2 shows the voxels that are truly activated in the simulation. For simplicity, only the small and large effect sizes were compared, as there did not appear to be much difference in the small and medium effects (medium effect sizes shown in Appendix A.2). Also, for simplicity, only sample sizes of 10 and 30 were compared visually (sample sizes 20 and 40 are shown in Appendix A.2).

Figure 4.2 Heatmap of the truly activated voxels in the simulation.



The heatmaps in Figures 4.3 and 4.4 show the average voxel activation over all the simulations for the small effect size for both sample sizes. For both of the likelihood ratio maps, there seems to be no activation (or a very small amount). There is a little bit of voxel activation for the FDR method with 10 subjects, and a little more activation for the RFT method. With the sample size of 10, the RFT method looks to have less activation than the truth, but in the same regions as the truth. For the sample size of 30, both the FDR and RFT methods have similar amount and regions of activations as the truth (Figure 4.2).

Figure 4.3 Heatmaps for sample size 10, small effect size. From left to right, the methods used are the first likelihood ratio, FDR, and RFT.

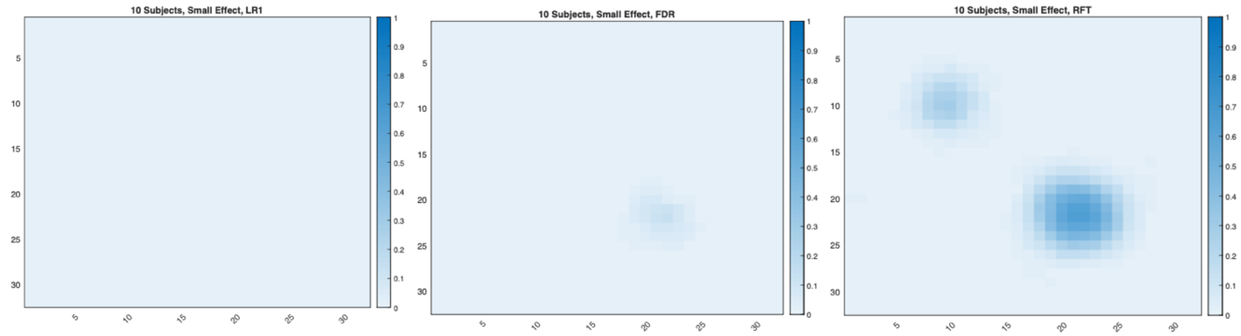
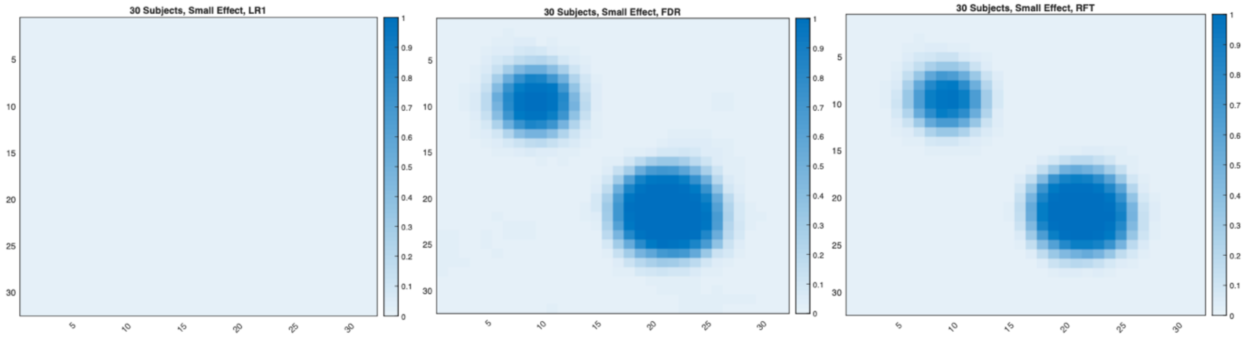


Figure 4.4 Heatmaps for sample size 30, small effect size. From left to right, the methods used are the first likelihood ratio, FDR, and RFT.



The next set of heatmaps shown in Figures 4.5 and 4.6 look at the sample sizes N=10 and N=30 for large effect size. With the large effect, voxel activation is apparent now for the likelihood ratio method. Although the activation is less than the truth, the regions are the same. For both sample sizes, the FDR and RFT methods have activation in the same areas as the true activation. For sample size = 30, the areas appear larger than for sample size = 10 for both of these methods. The regions of activation also are larger than the truth. This corresponds to the larger false positive rates we saw in Table 4.1 for the larger sample sizes.

Figure 4.5 Heatmaps for sample size 10, large effect size. From left to right, the methods used are the first likelihood ratio, FDR, and RFT.

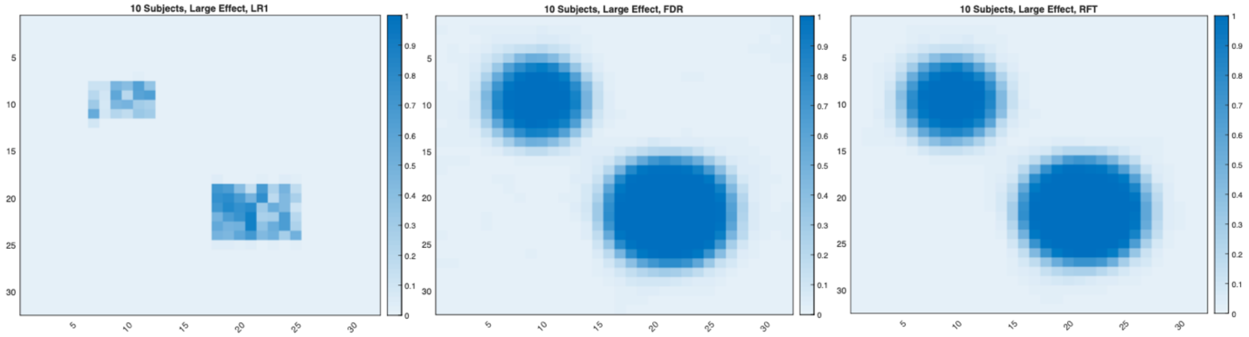
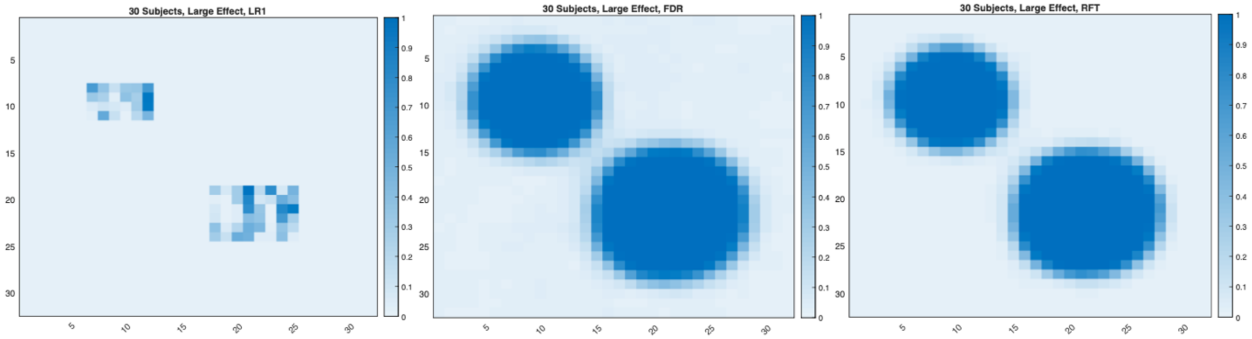


Figure 4.6 Heatmaps for sample size 30, large effect size. From left to right, the methods used are the first likelihood ratio, FDR, and RFT.



CHAPTER 5

CONCLUSION AND FUTURE WORK

The likelihood results showed that there is strong evidence for the null hypothesis over the alternative hypothesis, as we found no active voxels based on our binary cutoff of k = 20. The likelihood ratio brain maps show no activation. Because there is no activation, the likelihood ratios all had false negative rates equal to 0. If the "truth" were to have any activation, then we would see false negative rates equal to 1 instead of 0. For random field theory and false discovery methods, the false negative rate decreased as the sample size increased. Due to no activation in all sample sizes, the false positive rates were also 0. The RFT and FDR false positive rates both increased as the sample size increased.

In the main analyses described in this paper, subjects were not randomized. When doing analysis on each set of subject numbers (i.e. N = 14, 29, 44, 59, & 73), the first 14 subjects were selected, then the first 29 subjects, and so on. The only time subjects were randomly selected was in the secondary analysis, multiple subject estimation approach described in Section 2.8. For future work, subjects for all of the analyses could be randomized. Future work could also redefine the alternatives or the nulls.

The five random subject test shows that there are different results for the two methods. The single subject estimation method (like the main analysis) resulted in very little activation. Integrating multiple-subject data into the estimation process produced brain maps with more activation, though the activation was very scattered. For future work, the multiple subject estimation approach could be used on the entire data set. This approach will take a lot of computing power but would be useful in population inference.

The simulation yielded similar results as the main data analysis for the RFT and FDR methods. The false positive rates increased as the sample size increased, and the false negative rates decreased as the sample size increased. Results for the likelihood ratios were also similar to those in the data analysis. All false positive rates were 0 in the simulation, just like we saw in the data analysis. This is also because of the little activation in all of the likelihood ratio methods.

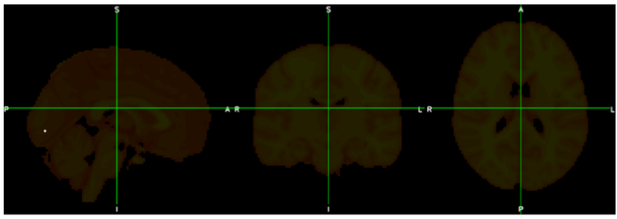
The false negative rates, however, were mostly all about 1, which is different than the data analysis results because here we do know 100 truly active voxels. In the data analysis, there are no false negative values because there is not any true activation. In the simulation, there are false negative values, but no (or very little) true positive values, yielding false negative rates = 1. There is some activation in the large effect size, suggesting larger effect sizes should be looked into.

Adding in one more effect size that was larger than the three original effect sizes, a decrease in false negative rates was observed. Now there are true positive values, and less false negative values so we see false negative rates around .13 - .38 for the different likelihood ratios. The first three effect sizes may have been too small. For future work, larger effect sizes should be considered.

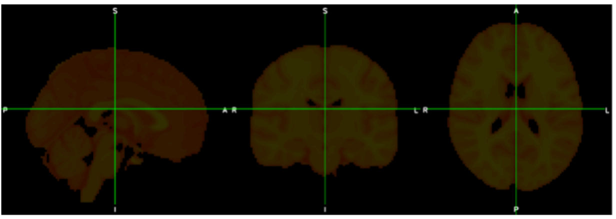
APPENDIX

A.1 Brain Activation Maps

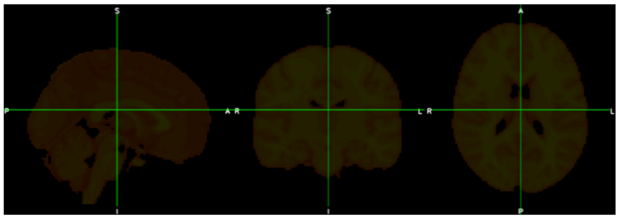
Brain Activation Maps, Binary Outputs, N = 14



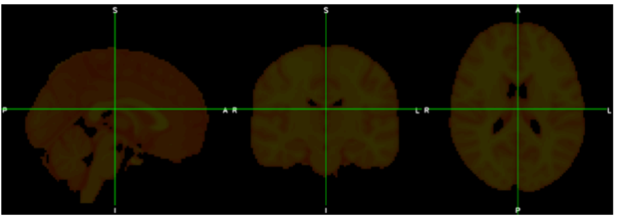
LR1



LR2

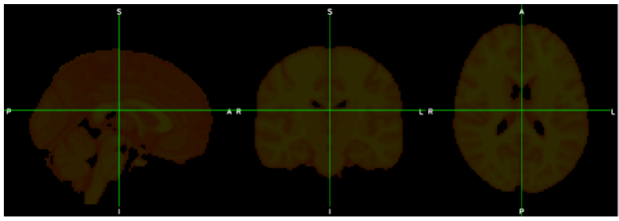


LR3

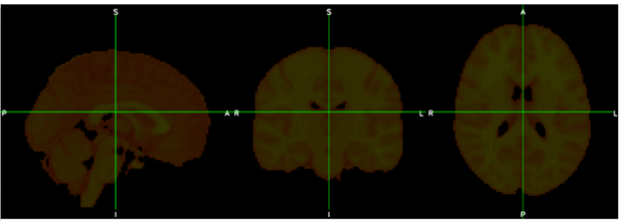


LR4

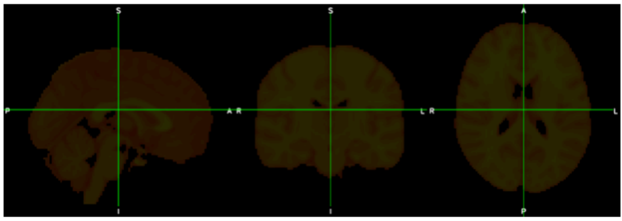
Brain Activation Maps, Binary Outputs, N = 29



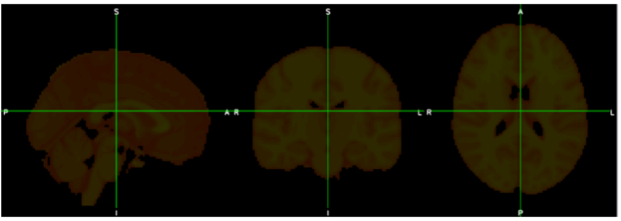
LR1



LR2

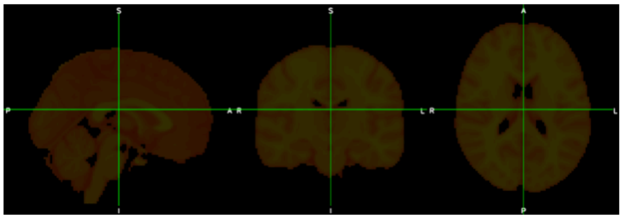


LR3

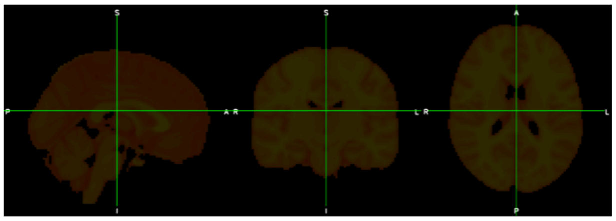


LR4

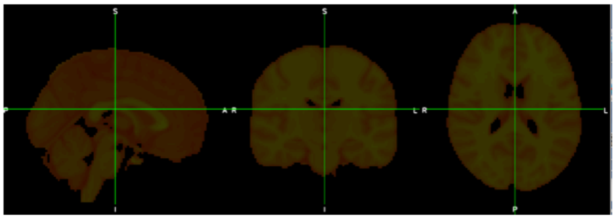
Brain Activation Maps, Binary Outputs, N = 44



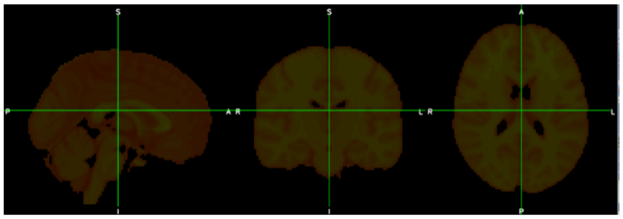
LR1



LR2

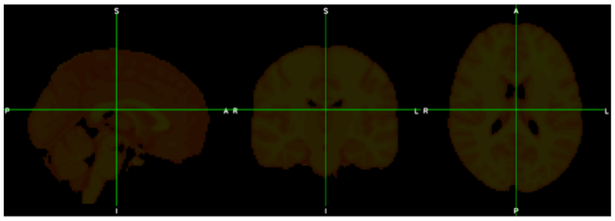


LR3

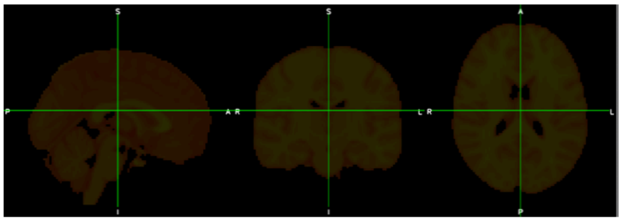


LR4

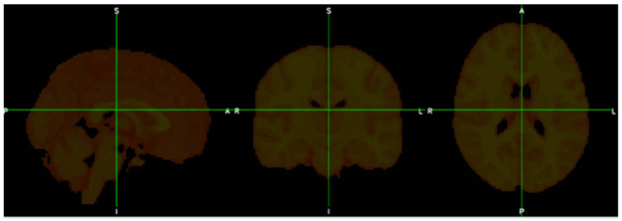
Brain Activation Maps, Binary Outputs, N = 59



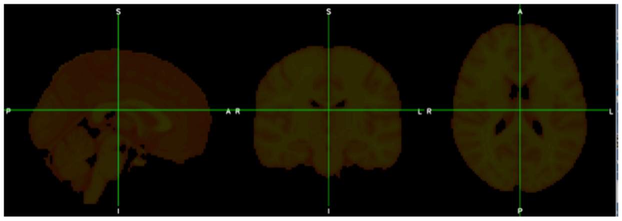
LR1



LR2

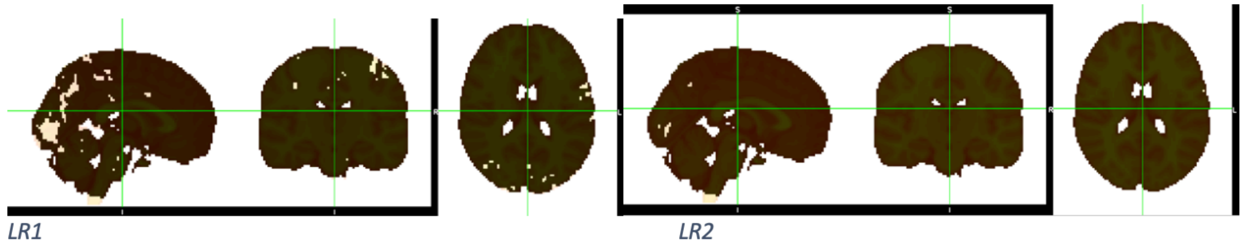


LR3

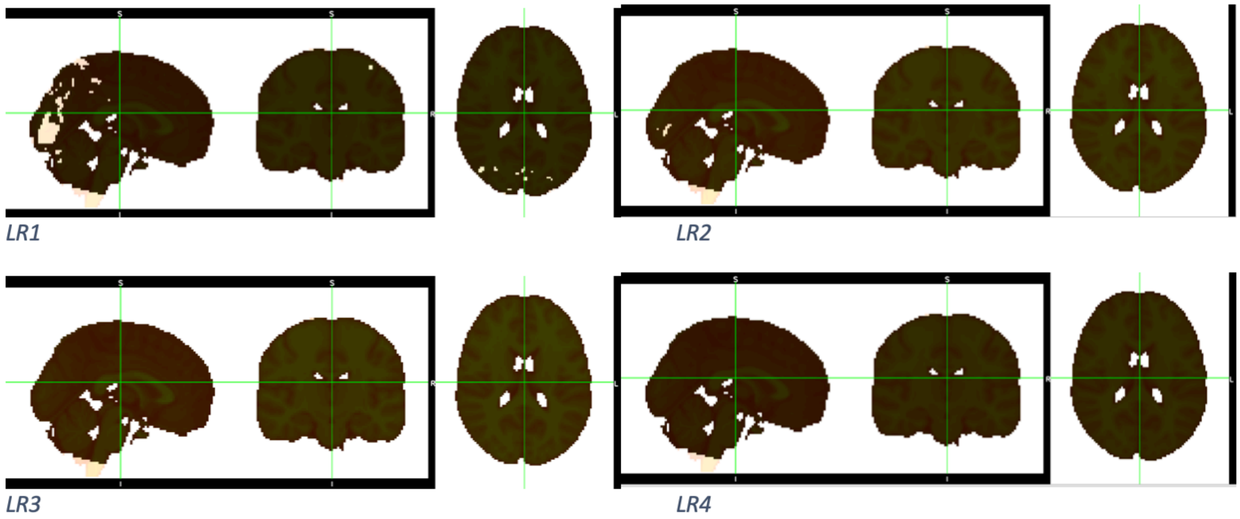


LR4

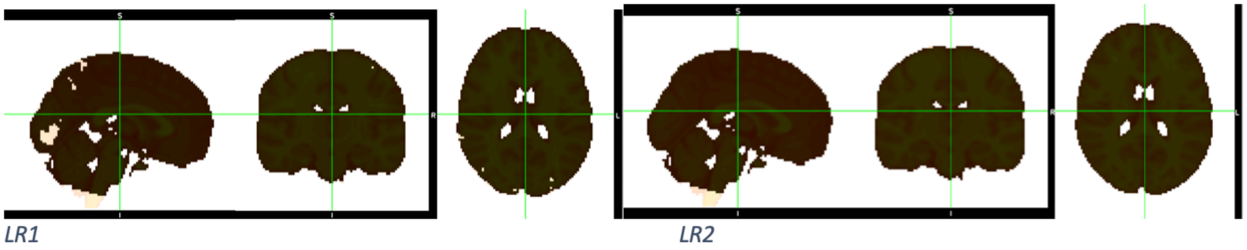
Brain Activation Maps with Inconclusive Zone, N = 14



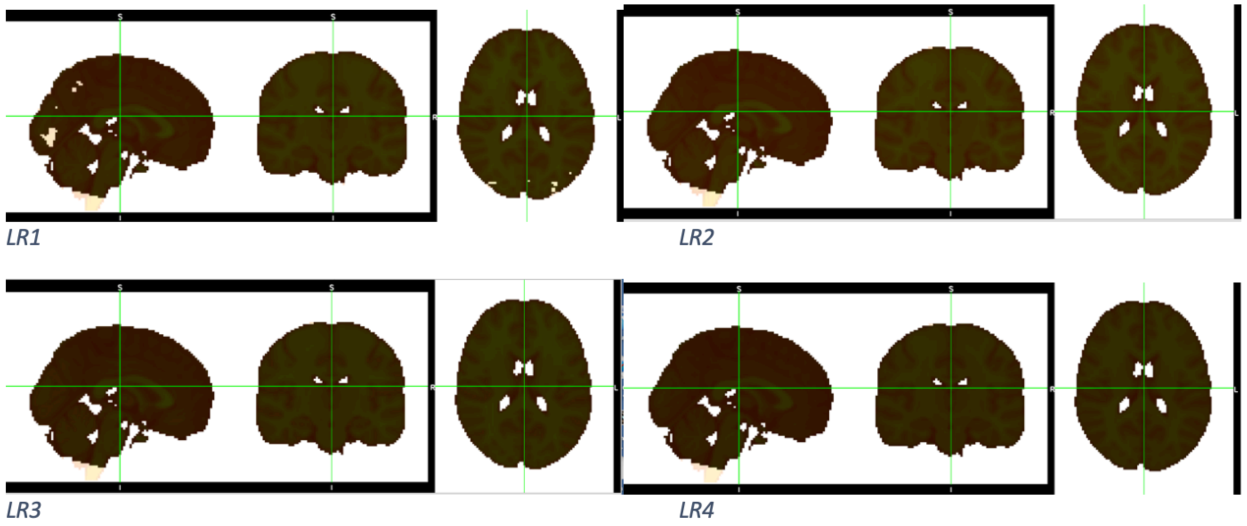
Brain Activation Maps with Inconclusive Zone, N = 29



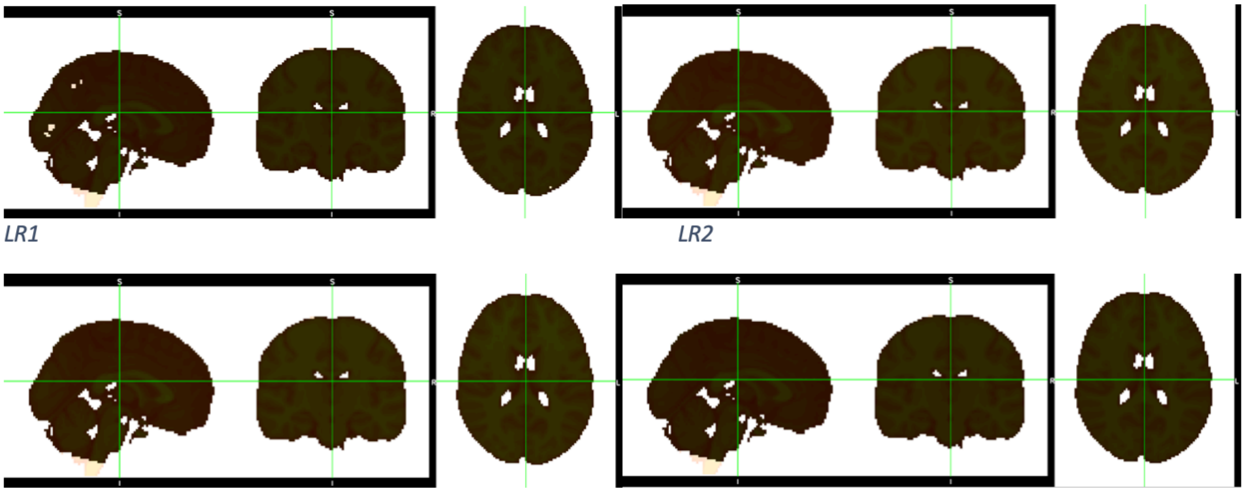
Brain Activation Maps with Inconclusive Zone, N = 44



Brain Activation Maps with Inconclusive Zone, N = 59



Brain Activation Maps with Inconclusive Zone, N = 73



LR1

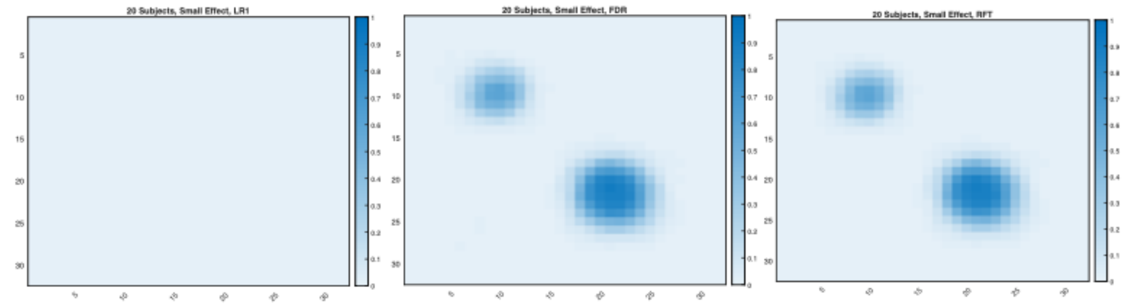
LR2

LR3

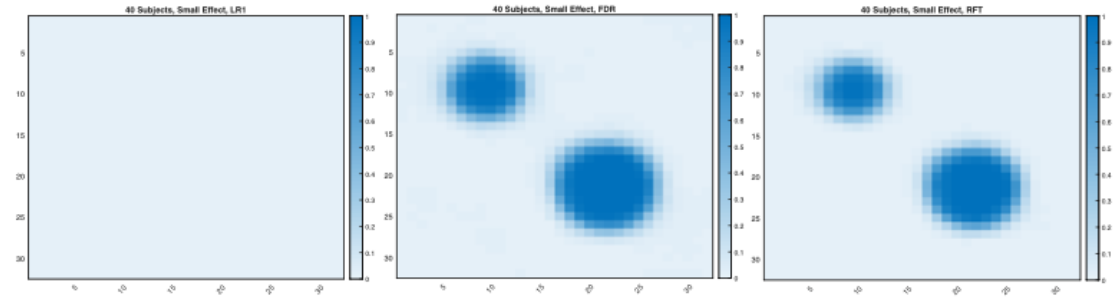
LR4

A.2 Heatmaps from Simulation not included in Section 4

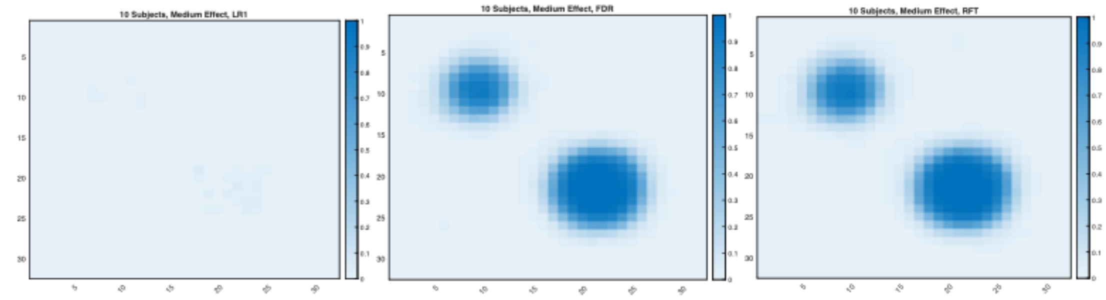
Small Effect Size, 20 Subjects



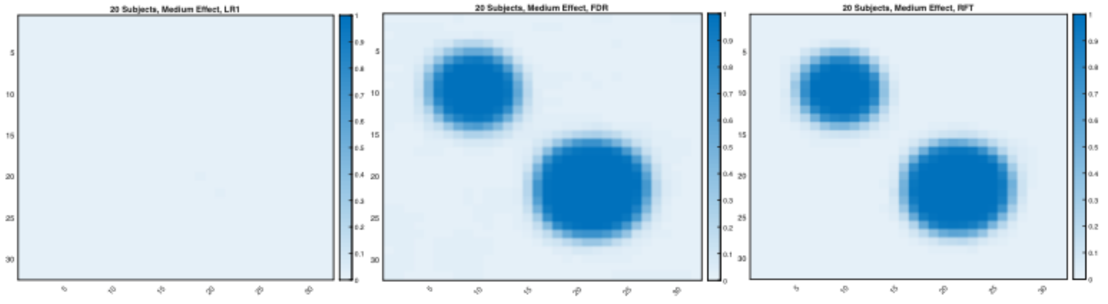
Small Effect Size, 40 Subjects



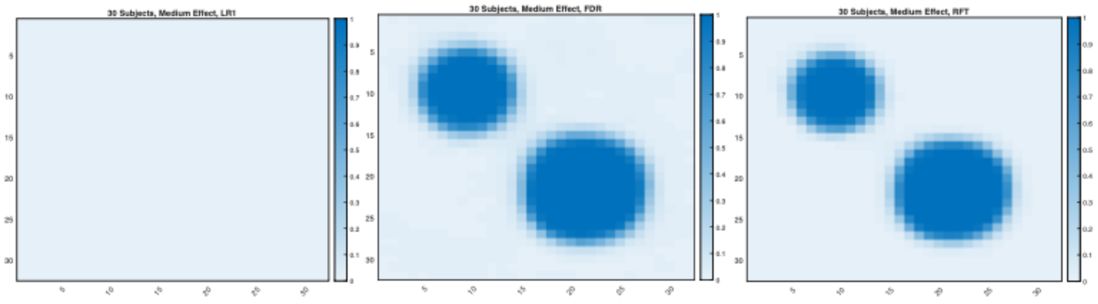
Medium Effect Size, 10 Subjects



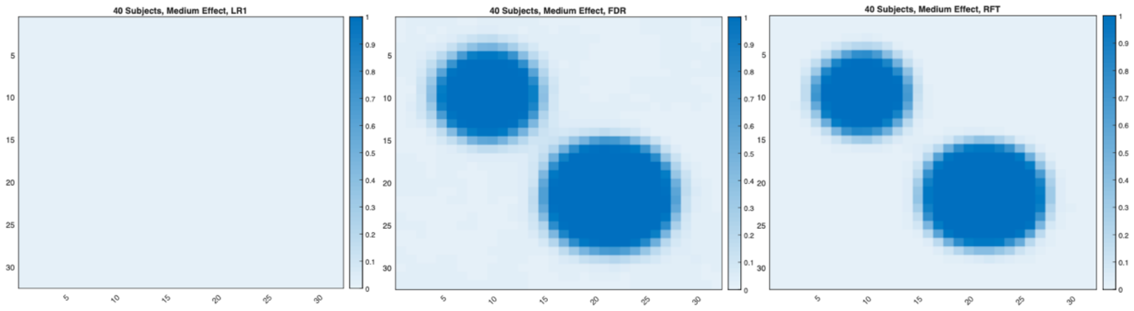
Medium Effect Size, 20 Subjects



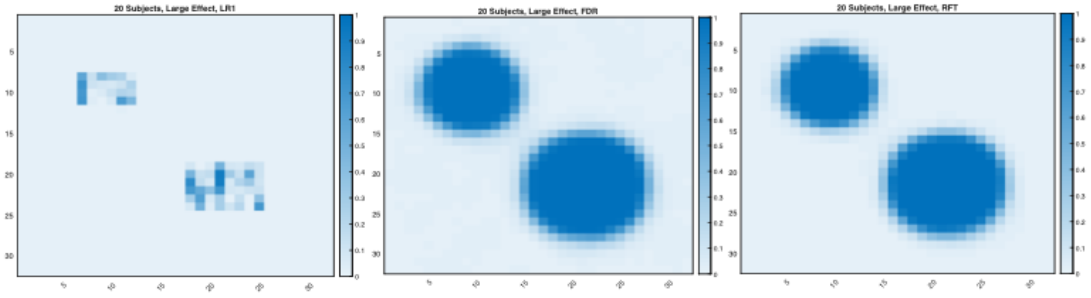
Medium Effect Size, 30 Subjects



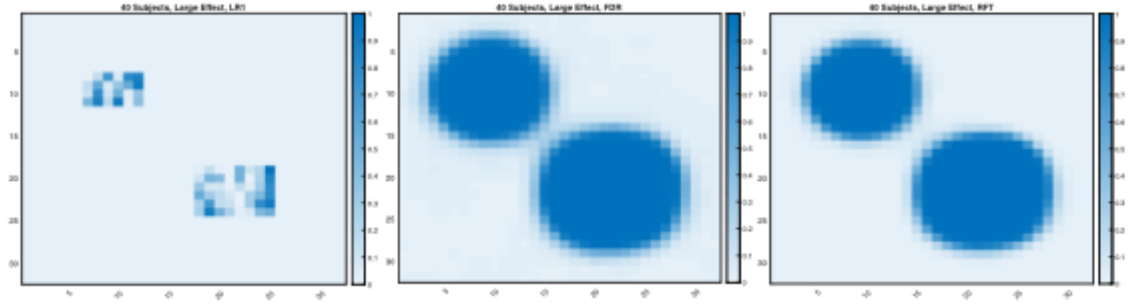
Medium Effect Size, 40 Subjects



Large Effect Size, 20 Subjects



Large Effect Size, 40 Subjects



A.3 Likelihood Ratio Derivation

$$\mathrm{LR} = \frac{L_n(\mu_1)}{L_n(\mu_0)} = \frac{\exp[-\frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu_1)^2]}{\exp[-\frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu_0)^2]}$$

$$= \exp\{-\frac{1}{\sigma^2}[n(\mu_1 + \mu_0)(\mu_1 - \mu_0) - 2(\mu_1 - \mu_0)\sum_{i=1}^{n} x_i]\}$$

$$= \exp\{\frac{n(\mu_1 - \mu_0)}{\sigma^2}[\bar{X}_n - \frac{(\mu_1 + \mu_2)}{2}]\}$$

REFERENCES

1. Lindquist, Martin. (2008). The Statistical Analysis of fMRI Data. *Statistical Science,* 23(4), 439–464.

2. Kang, H., Blume, J., Ombao, H., & Badre, D. (2015). Simultaneous control of error rates in fMRI data analysis. *NeuroImage*, *123*, 102–113.

3. Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.

4. Worsley, K. J., Evans, A. C., Marrett, S., & Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of cerebral blood flow and metabolism: official journal of the International Society of Cerebral Blood Flow and Metabolism*, *12*(6), 900–918.

5. Brett, M., Penny, W.D., & Kiebel, S.J. (2003). An Introduction to Random Field Theory.

6. Albert, K., Gau, V., Taylor, W. D., & Newhouse, P. A. (2017). Attention bias in older women with remitted depression is associated with enhanced amygdala activity and functional connectivity. *Journal of affective disorders*, 210, 49–56.

7. Jenkinson, M., Bannister, P., Brady, J. M. and Smith, S. M. (2002). Improved Optimisation for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, 17(2), 825-841.

8. Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M. (2012). FSL. *NeuroImage*, 62:782-90.

9. Parker, D. B., & Razlighi, Q. R. (2019). The Benefit of Slice Timing Correction in Common fMRI Preprocessing Pipelines. *Frontiers in neuroscience*, 13, 821.

10. Sladky, R., Friston, K.J., Tröstl, J., Cunnigton, R., Moser, E., Windischberger, C. (2011). Slice-timing effects and their correction in functional MRI. *NeuroImage*, 58(2), 588-594.

11. Smith, S. M, Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., and Matthews, P.M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(S1):208-19.

12. Ngan, S & Laconte, Stephen & Hu, Xiaoping. (2000). Temporal Filtering of Event-Related fMRI Data Using Cross-Validation. *NeuroImage*. 11. 797-804.

13. Smith, S., Flitney, D., Jenkinson, M., Clare, S., Nichols, T., and Webster, M. (2008). *fslmaths.cc Image processing routines.* FMRIB's Software Library. http://ftp.nmr.mgh.harvard.edu/pub/dist/freesurfer/tutorial_packages/OSX/fsl_501/src/avwutils/fslmaths.cc

14. David Groppe (2020). fdr_bh (https://www.mathworks.com/matlabcentral/fileexchange/27418-fdr_bh), MATLAB Central File Exchange.

15. Blume, Jeffrey. (2002). Likelihood methods for measuring statistical evidence. Statistics in medicine. 21. 2563-99.