Linear Model Fall 2018 - Final Project 2

Pei-Shan Yen, Yanli Gao and Ran Gao

ARTICLE HISTORY
Compiled December 19, 2018

1. Introduction

fMRI (functional magnetic resonance imaging) data is generally used for the detection of many neurological discovers, such as depression, Alzheimers disease, and autism. Statistical modeling and multiple comparison are essential while analyzing fMRI data in neural connectivity research. Bhaumik (2018) [1] proposes a specific mix-effects model to address the importance of within-subject correlations and group-variation. In order to solve the problems that occur when multiple comparisons are performed, the method of False Discovery Rate (FDR) is employed to locate the significant connectivity links. This method is introduced by Benjamn and Hochberg (1995) [2] and exhibits two major advantages. First, the method achieves the smallest fraction of false positives. Second, the method shows more power comparing to the Bonferroni method.

Our primary purpose in this report is to demonstrate the whole FDR procedure for the detection of the disrupted links. The fMRI dataset applied in the report is from one of the research projects hold by Dr. Bhaumik. Seven subjects are captured among the healthy control group and the depression group, respectively. In each subject, fMRI data is presented as an 87 by 87 symmetric functional connectivity matrix. For model fitting, we derive $\frac{87\times 86}{2} = 3741$ dimensional vectors from the lower triangle elements of the mode matrix. The results of estimators for fix effects ($\widehat{\boldsymbol{\beta}}_{0i}$ and $\widehat{\boldsymbol{\beta}}_{1i}$) and the corresponding standard errors are already given for this project. In the data analysis section of this report, we will examine the characteristics of the dataset and then perform the FDR procedure to identify significant links under different FDR-levels (q-values).

2. Methodology

2.1. Model Specification

For each link, the total variation among different subjects can be partitioned into two components: (1) the difference between two groups: diseased patient and control, and (2) the difference between subjects. As the subject effect is not of interest in the work, we treat each subject as a random sample from the study population. We adopt a mixed-effects model to analyze fMRI data,

$$y_{ijk} = \beta_{ki} + \gamma_{jk} + \epsilon_{ijk} \tag{1}$$

where y_{ijk} is the fMRI measurement for the ith link from the jth subject in the kth group, $i = 1...m$ and $j = 1...n$, k=0 for the control group, and k=1 for the diseased group. γ_{jk} is the random effect term for the jth subject in the kth group and ϵ_{ijk} is the error term. We further assume that $\gamma_{jk} \sim N(0, \sigma_\gamma^2)$, $\epsilon_{ijk} \sim N(0, \sigma_{ki}^2)$. We assume γ_{jk} is independent of ϵ_{ijk} .

Group effects β_{0i} and β_{1i} are treated as fixed effects, while the subject effect γ_{jk} is a random effect. This model allows each link to have its own mean and its own variance and for those to vary between groups.

$$y_{ijk}|\gamma_{jk} \sim N(\beta_{ki} + \gamma_{jk}, \sigma_{ki}^2) \tag{2}$$

The dataset contains fMRI measurements from 14 subjects. We work with 3741 link-based measurements from each subject.

2.1.1. Parameter Estimation:

The model specified in (1) can be written following the general expression, so that the mixed-effects model is expressed as

$$\boldsymbol{y}_{jk} = \boldsymbol{X}_{jk}\boldsymbol{\beta}_k + \boldsymbol{Z}_{jk}\gamma_{jk} + \boldsymbol{\varepsilon}_{jk}, \tag{3}$$

where $\boldsymbol{y}_{jk}$ denotes the vector of fMRI measurements for j^{th} subject in the k^{th} group. The $m \times 1$ vector of $\boldsymbol{y}_{jk}$ is modeled with both fixed-effect parameter vector $\boldsymbol{\beta}_k$ and random effect γ_{jk} . $\boldsymbol{\beta}_k$ is a $m \times 1$ vector and γ_{jk} is a scalar. $\boldsymbol{X}_{jk}$ is a $m \times m$ binary design matrix for the fixed effects and $\boldsymbol{Z}_{jk}$ is a $m \times 1$ design matrix for the random effect. $\boldsymbol{\varepsilon}_{jk}$ is a $m \times 1$ error vector. This is equivalent to (1) when $\boldsymbol{X}$ is the identity matrix and $\boldsymbol{Z}$ is a column of 1s, for every combination of j and k.

The fixed-effect vectors $\boldsymbol{\beta}_k$ and random effect terms γ_{jk} can be estimated via an Expectation-Maximization (EM) algorithm. Briefly, the random effect γ is estimated in the E step via Empirical Bayes (EB) estimation, whereas the fixed-effect vector $\boldsymbol{\beta}_k$, its variance and variance components are estimated in the M step via maximum likelihood estimation (MLE).

2.1.2. Empirical Bayes Estimation:

Observed fMRI measurements $\boldsymbol{y}_{jk}$ and random effect γ_{jk} have the following joint distribution:

$$\begin{bmatrix} \boldsymbol{y}_{jk} \\ \gamma_{jk} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{X}_{jk}\boldsymbol{\beta}_k \\ 0 \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{V}_{jk} & \boldsymbol{Z}_{jk}\sigma_\gamma^2 \\ \sigma_\gamma^2 \boldsymbol{Z}_{jk}^T & \sigma_\gamma^2 \end{bmatrix} \right), \tag{4}$$

where $\boldsymbol{V}_{jk} = \sigma_\gamma^2 \boldsymbol{Z}_{jk}\boldsymbol{Z}_{jk}^T + \Sigma_k$, σ_γ^2 is the variance of the random effect, and Σ_k is an $m \times m$ diagonal error covariance matrix for the k^{th} group. As above, $\boldsymbol{X}_{jk} = \boldsymbol{I}_m$ and

$\boldsymbol{Z}_{jk} = \mathbf{1}$.

Given values for $\hat{\boldsymbol{\beta}}_k$, $\hat{\sigma}^2_\gamma$, and $\hat{\Sigma}_k$, we can estimate $\hat{\boldsymbol{V}}_{jk}$ and use the following equations to calculate conditional (EB) estimates of the mean and variance of γ_{jk}, denoted as $\tilde{\gamma}_{jk}$ and $\tilde{\Sigma}_{\gamma|\boldsymbol{y}_{jk}}$, respectively.

$$\begin{aligned}\tilde{\gamma}_{jk} =& \widehat{\sigma}^2_\gamma \boldsymbol{Z}^T_{jk} \widehat{\boldsymbol{V}}^{-1}_{jk} \left(\boldsymbol{y}_{jk} - \boldsymbol{X}_{jk} \widehat{\boldsymbol{\beta}}_k \right) \\ \tilde{\Sigma}_{\gamma|\boldsymbol{y}_{jk}} =& \widehat{\sigma}^2_\gamma \ (1 - \boldsymbol{Z}^T_{jk} \widehat{\boldsymbol{V}}^{-1}_{jk} \boldsymbol{Z}_{jk})\end{aligned} \tag{5}$$

2.1.3. Maximum marginal likelihood estimation:

As noted in the previous section, in order to obtain EB estimates of the mean and variance of random effect γ, $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\beta}_k$ in (5) need to be estimated from the marginal distribution of $\boldsymbol{Y}$. There are several approaches to obtain estimates of $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\beta}_k$, including the maximum likelihood estimation used here.

The marginal distribution of the fMRI measurements for the j^{th} subject in the k^{th} group, $\boldsymbol{y}_{jk}$, can be derived as a conditional distribution from the joint distribution of $\boldsymbol{y}_{jk}$ and γ_{jk}. For ease of notation, we stack all observations in the k^{th} group together. Then $\boldsymbol{V}_k = \widehat{\sigma}^2_\gamma \boldsymbol{Z}_k \boldsymbol{Z}^T_k + \widehat{\Sigma}_k \otimes \boldsymbol{I}_{N_k}$. Generalized least square theory yields the following results, where $\tilde{\gamma}_{jk}$ and $\tilde{\Sigma}_{\gamma|\boldsymbol{y}_{jk}}$ come from the previous step and $e_{ijk} = y_{ijk} - x_{ijk}\widehat{\beta}_k - \tilde{\gamma}_{jk}$:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_k =& \left(\boldsymbol{X}^T_k \widehat{\boldsymbol{V}}^{-1}_k \boldsymbol{X}_k \right)^{-1} \boldsymbol{X}^T_k \widehat{\boldsymbol{V}}^{-1}_k (\boldsymbol{y}_k - \tilde{\boldsymbol{\gamma}}_k) \\ Cov(\widehat{\boldsymbol{\beta}}_k) =& \left(\boldsymbol{X}^T_k \widehat{\boldsymbol{V}}^{-1}_k \boldsymbol{X}_k \right)^{-1} \\ \widehat{\sigma}^2_\gamma =& \frac{1}{N} \sum_{k=0}^{1} \sum_{j=1}^{N_k} \left[\tilde{\gamma}_{jk} \tilde{\gamma}^T_{jk} + \tilde{\Sigma}_{\gamma|\boldsymbol{y}_{jk}} \right]\end{aligned} \tag{6}$$

$$\widehat{\sigma}^2_{ki} = \frac{1}{N_k} \sum_{j=1}^{N_k} \left[\tilde{\Sigma}_{\gamma|\boldsymbol{y}jk} + e^2_{ijk} \right], \quad \widehat{\boldsymbol{\Sigma}}_k = diag(\widehat{\sigma}^2_{ki}). \tag{7}$$

To obtain estimates of $\boldsymbol{\beta}_k$, $\boldsymbol{cov}(\widehat{\beta}_k)$, γ_{jk}, σ^2_γ, $\boldsymbol{\Sigma}_k$, we iterate through the following steps of our EM algorithm:

(1) Initialize the parameters $\boldsymbol{\beta}_k$, $\boldsymbol{cov}(\widehat{\beta}_k)$, γ_{jk}, σ^2_γ, $\boldsymbol{\Sigma}_k$ to some random values.

(2) Compute MLE for $\boldsymbol{\beta}_k$, $\boldsymbol{cov}(\widehat{\beta}_k)$, $\boldsymbol{\Sigma}_k$, using initial values.

(3) Obtain EB estimates of γ_{jk}, σ^2_γ by plugging in estimated $\boldsymbol{\beta}_k$, $\boldsymbol{cov}(\widehat{\beta}_k)$, $\boldsymbol{\Sigma}_k$.

(4) Re-estimate MLE given updated γ_{jk}, σ^2_γ.

(5) Steps 3-4 are repeated until relative change of the estimated values is smaller than some limit of tolerance (i.e. 10^{-5}).

2.2. Hypothesis testing

The goal of our project is to identify those links whose correlations are significantly changed in depressed subjects as compared with healthy subjects. The comparison will be based on the difference between fixed effects of $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$, , which are the functional connectivity estimated for the ith link in healthy and depressed subjects , respectively. For all links, we will test the following hypothesis simultaneously, where $i = 1...3741$:

$$H_{0i} : \hat{\beta}_{0i} \text{ - } \hat{\beta}_{1i} = 0 \text{ versus } H_{1i} : \hat{\beta}_{0i} \text{ - } \hat{\beta}_{1i} \neq 0$$

Individually for each link, our t test statistic is:

$$t_i = \frac{\hat{\beta}_{0i} - \hat{\beta}_{1i}}{\sqrt{\frac{SE(\hat{\beta}_{0i})^2 + SE(\hat{\beta}_{1i})^2}{n_i}}} \tag{9}$$

where n_i is the number of subjects in each group, which is 7 in our data within each link per group. The t-value above compares the strength of the difference between two groups about the same link with 6 degrees of freedom.

2.3. False Discovery Rate

We here have thousands of hypothesis tests conducted simultaneously. A type I error happens when we choose to reject a null hypothesis that is actually true . In order to be able to identify as many significant comparisons as possible while still maintaining a low false positive rate, the False Discovery Rate (FDR) and its analog the q-value are utilized.

FDR is the rate that features called significant are truly null. FDR = expected (# false predictions/# total predictions). Just as we set alpha as a threshold for the p-value to control the false positive rate (FPR), we can also set a threshold for the q-value, which is the FDR analog of the p-value. The q-value is the expected proportion of false positives among all features as or more extreme than the observed one. For example, a q-value threshold of 0.05 yields a FDR of 5% among all features called significant.

The FDR at a certain threshold t, is FDR(t). FDR(t) = E[F(t)]/E[S(t)] , which can be estimated as the expected# of false positives at that threshold divided by the expected# of features called significant at that threshold.

E[S(t)] is simply S(t), the number of observed p-values t (i.e. the number of features we call significant at the chosen threshold). How do we estimate E[F(t)], which is the expected number of false positives for a given threshold t, and it follows that $E[F(t)] = m_0 t$, which is the number of truly null features times the probability a null feature will be called significant. However, the true value of m_0 is unknown. Instead, we can estimate the proportion of features that are truly null, $\pi_0 = m_0/m$, where m is total # of hypothesis tests.

We assume that p-values of null features will be uniformly distributed (have a flat distribution) between [0,1]. The height of the flat distribution gives a conservative estimate of the overall proportion of null p-values, π_0, which is quantified as $\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i=1,...,m\}}{m(1-\lambda)}$, where λ is the tuning parameter. The proportion of truly null features equals the number of p-values greater than λ divided by m(1-λ). As λ approaches 0 (when most of the distribution is flat), the denominator will be approximately m, as will the numerator since the majority of the p-values will be greater than lambda, and π_0 will be approximately 1 (all features are null). The choice of λ is automated by qvalue function from R software. $\hat{\pi}_0(\lambda)$ is calculated in such function as well.

Now that we have estimated π_0, we can estimate FDR(t) as $F\hat{D}R(t) = \frac{\hat{\pi}_0 mt}{S(t)} = \frac{\hat{\pi}_0 mt}{\#\{p_i \leq t\}}$. The numerator for this equation is just the expected number of false positives, since $\pi_0 m$ is the estimated number of truly null hypotheses and t is the probability of a truly null feature being called significant (being below the threshold t). The denominator, as we said above, is simply the number of features called significant.

A software for computing q-values based on a list of p-values can be found at http://genomine.org/qvalue/. After getting p-values from multiple t-tests, the qvalue package was used to calculate q-values for each link.

3. Data Analysis And Results

3.1. Descriptive Statistics: The difference of Connectivity Links

First, we measure the difference of the connectivity links between the healthy group and the depression group. As shown in Table 1, the estimated values of mean correlation for the healthy group and the depression group are 0.1378 and 0.1392, respectively.

Table 1. The descriptive statistics of the correlation among two group

		Min	1st Qu.	Median	Mean	3rd Qu.	Max
Healthy Group	$\widehat{\beta_0}$	-0.2500	0.0324	0.1055	0.1378	0.2039	1.3981
Depression Group	$\widehat{\beta_1}$	-0.2236	0.0207	0.1046	0.1392	0.2139	1.4647
Difference	$\widehat{\beta_0} - \widehat{\beta_1}$	-0.0264	0.0117	0.0009	-0.0014	-0.0100	-0.0666

Despite the small difference in the mean correlation, this does not, however, mean that the number of significant links is low. From Fig. 1 which shows the histogram with the bin size of 0.02, it indicates that there are 682 links out of 3,741 links (18.23%) around the point where the correlations for the two groups (healthy vs. depression) is the same, i.e., $x = 0$ in Fig. 1, which implies that the numbers of significant links should be substantial.

3.2. P-values Performance

The histogram of log_{10} of p-values is shown in Fig. 2. It shows two findings. The first is that there are 793 p-values which are smaller than 0.05, indicating that up to 21.20%$(= 793/3,741)$ of links are significant. The second is that there are 2,541 p-values above 0.1, representing that 67.92% of links could be considered as the true null hypotheses. Both findings lead to the same conclusion as previously indicated by the examination of the connectivity links, that is, the number of significant links should not be small.

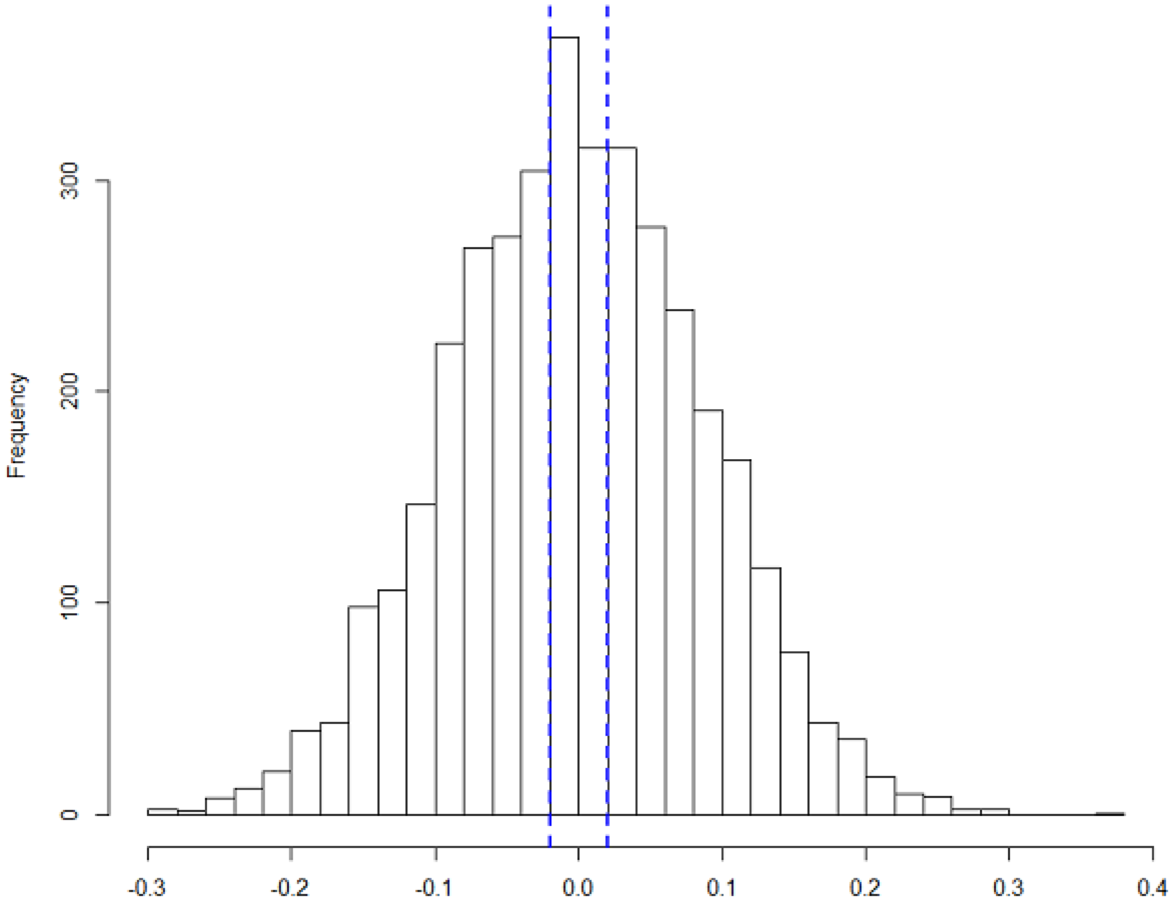


Figure 1. Histogram (across links) of difference between healthy group and depression group

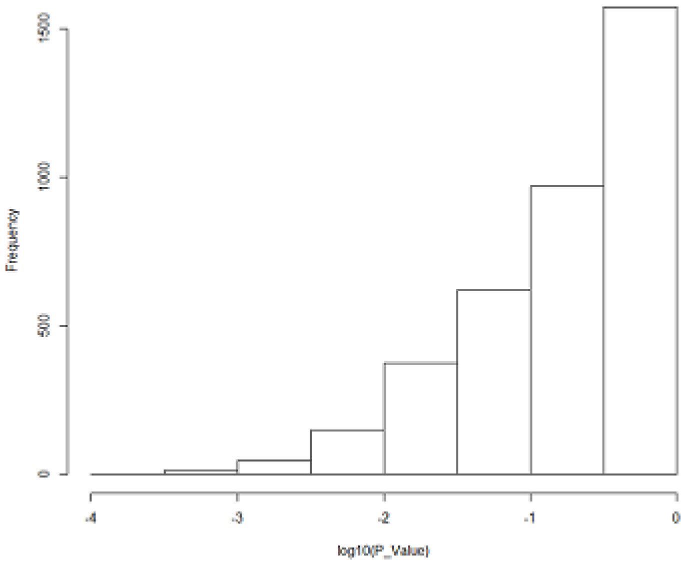


Figure 2. Histogram of log_{10} of p-values form mixed-effects analysis

3.3. The Result of FDR procedure

Now we proceed to the formal analysis of multiple comparison. We apply the method of false discovery rate to evaluate the number of significant links. First, we explore the p-value distribution in our 3,741 links (Fig. 3). This distribution of large p-values converges to a flat Uniform distribution and the estimated proportion of true null hypothesis $\widehat{\pi}_0$ is 0.541.

Consistently as shown in Fig. 4(a), $\widehat{\pi}_0$ is inversely proportional to $\hat{\lambda}$, and $\widehat{\pi}_0$ asymptotically reaches the local minimum which is 0.541 as $\hat{\lambda}$ is close to one, i.e., $\lim_{\hat{\lambda}\to 0} \widehat{\pi}_0 = 0.541$.

Second, we use FDR procedure to calculate q-values through the information of p-values, and the result is shown in Fig 4(b). After deriving the q-values, we can plot the number of the significant links as a function of q-values, as shown in Fig 4(c), and the selective cutoffs are given in Table 2. With the FDR-level of $q = 0.0910$, the number of significant links is 185.

The detailed information with these 185 links is described in Table 3. Finally, Fig 4(d) displays the expected false positive as a function of significant links. The whole FDR procedure shows how to detect the number of significant links under the different cutoffs of the q-values.

4. Conclusions

fMRI (functional magnetic resonance imaging) is an essential tool for the detection of aberrant activities in human brain. In this report, our purpose is to identify significant neural connectivity links which caused the major depressive disorder. A mix-effected model gives information of estimated correlation between healthy groups and depres-

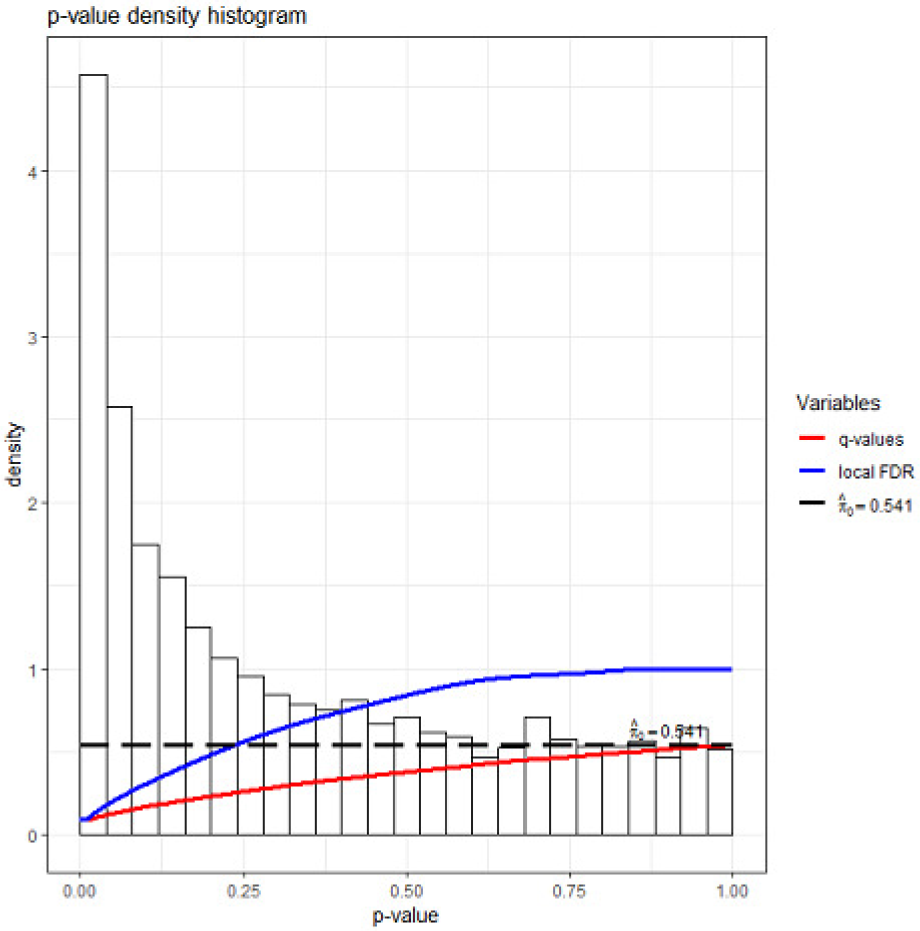


Figure 3. Density histogram of p values across all links

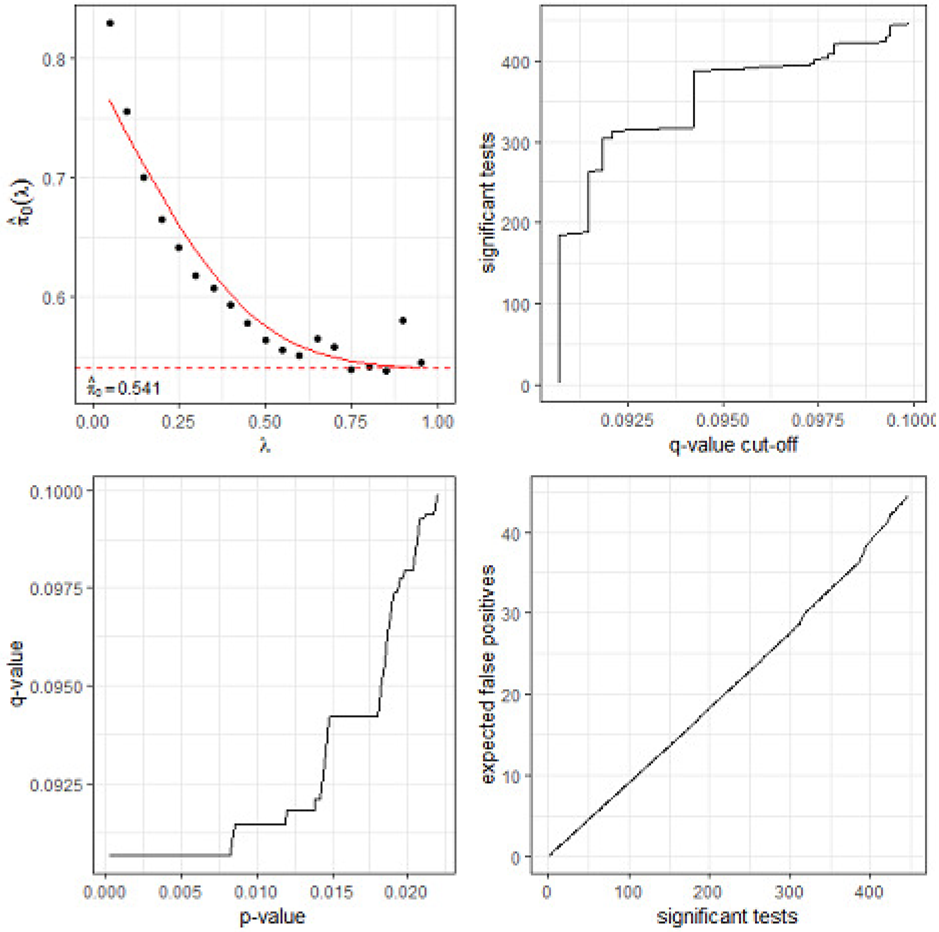


Figure 4. Plots produced by qvalue package. (a) $\widehat{\pi}_0$ versus the tuning parameter λ (b) The q-values versus their respective p-values (c) The number of significant links as a function of q-value cutoff (d) The expected number of false positive links as a function of the total number of significant links.

Table 2. The number of significant links under different q-values cutoff

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
q values	0.0907	0.1379	0.2461	0.2732	0.3973	0.5407
Number of significant links	185	751	938	149	783	935

Table 3. The corresponding significant link with FDR=0.0910

83	355	996	1476	1809	2022	2666	2896	3251	3688
93	359	1000	1560	1810	2023	2675	2905	3257	3690
99	408	1005	1612	1812	2103	2689	2924	3263	3706
103	570	1010	1620	1816	2241	2698	2944	3264	3713
105	698	1024	1678	1817	2271	2700	2949	3269	3741
125	734	1026	1688	1830	2282	2702	2955	3303	
139	749	1030	1690	1844	2326	2738	2957	3378	
140	783	1039	1708	1846	2417	2740	3021	3384	
159	784	1062	1721	1886	2424	2743	3065	3411	
197	803	1120	1722	1932	2494	2758	3066	3412	
207	831	1124	1724	1937	2501	2764	3079	3434	
215	892	1134	1725	1947	2508	2787	3095	3453	
217	897	1154	1750	1951	2510	2823	3111	3458	
222	901	1158	1755	1955	2523	2849	3116	3560	
225	935	1227	1762	1966	2601	2853	3119	3593	
287	968	1254	1784	1973	2613	2863	3172	3601	
288	976	1336	1796	1989	2615	2864	3214	3627	
289	982	1432	1799	1999	2623	2865	3218	3669	
323	984	1435	1801	2003	2630	2890	3228	3671	
326	992	1439	1806	2009	2657	2894	3237	3678	

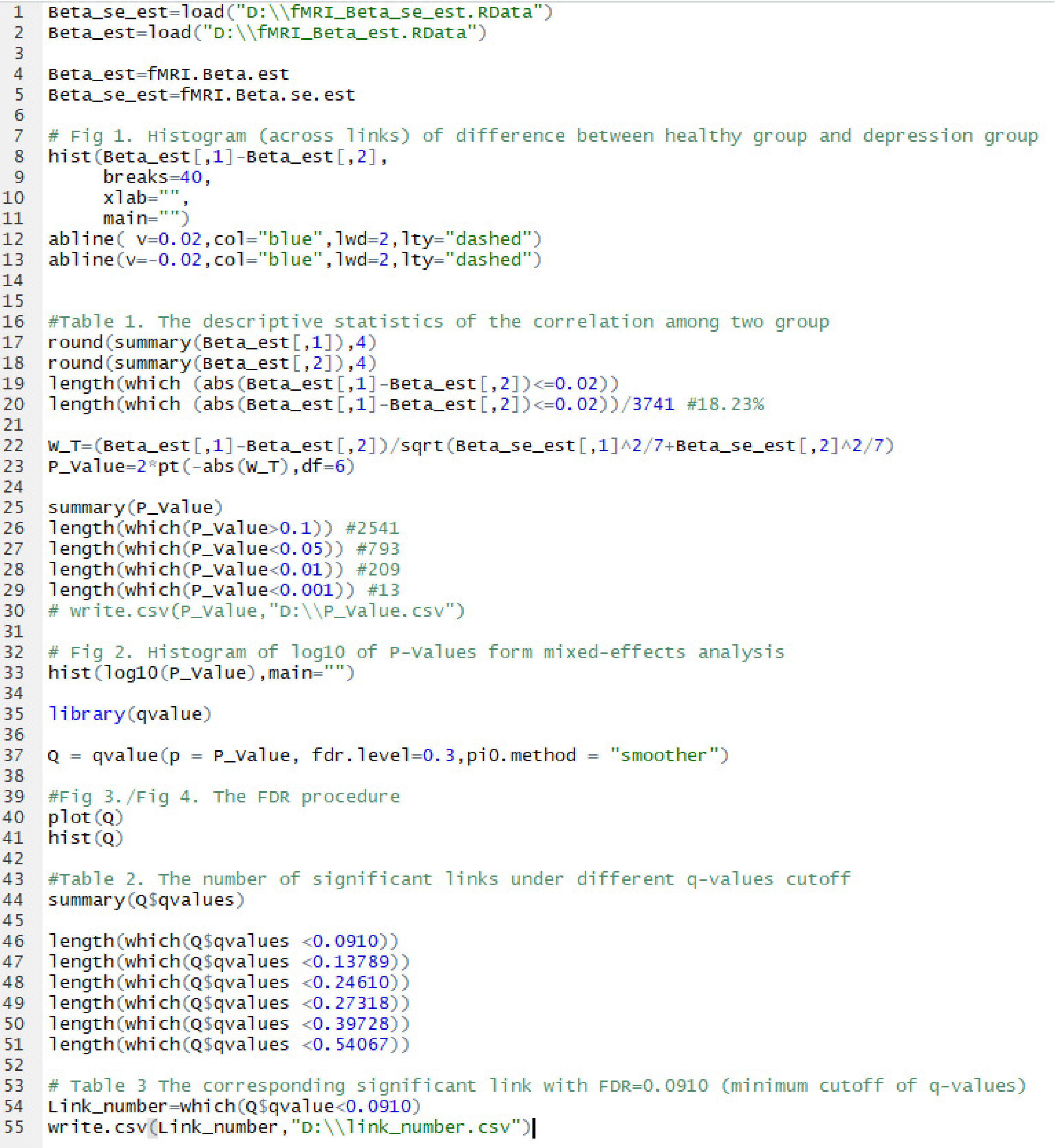
sion groups. We perform the method of False Discover Rate to deal with multiple comparison issues. Through controlling the q-value level=0.0910, we successfully identify 183 disrupted links for the future therapeutic benefit.

References

[1] D. Bhaumik, F. Jie, R. Nordgren, R. Bhaunik, and B. Sinha, *A Mixed-Effects Model for Detecting Disrupted Connectivities in Heterogeneous Data*, IEEE Transactions on Medical Imaging, vol:37, Issue:11, Nov. 2018.

[2] Y. Benjamini and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1 (1995), pp. 289-300.

Appendix: R code for FDR analysis



```r
Beta_se_est=load("D:\\fMRI_Beta_se_est.RData")
Beta_est=load("D:\\fMRI_Beta_est.RData")

Beta_est=fMRI.Beta.est
Beta_se_est=fMRI.Beta.se.est

# Fig 1. Histogram (across links) of difference between healthy group and depression group
hist(Beta_est[,1]-Beta_est[,2],
     breaks=40,
     xlab="",
     main="")
abline( v=0.02,col="blue",lwd=2,lty="dashed")
abline(v=-0.02,col="blue",lwd=2,lty="dashed")


#Table 1. The descriptive statistics of the correlation among two group
round(summary(Beta_est[,1]),4)
round(summary(Beta_est[,2]),4)
length(which (abs(Beta_est[,1]-Beta_est[,2])<=0.02))
length(which (abs(Beta_est[,1]-Beta_est[,2])<=0.02))/3741 #18.23%

W_T=(Beta_est[,1]-Beta_est[,2])/sqrt(Beta_se_est[,1]^2/7+Beta_se_est[,2]^2/7)
P_Value=2*pt(-abs(W_T),df=6)

summary(P_Value)
length(which(P_Value>0.1)) #2541
length(which(P_Value<0.05)) #793
length(which(P_Value<0.01)) #209
length(which(P_Value<0.001)) #13
# write.csv(P_Value,"D:\\P_Value.csv")

# Fig 2. Histogram of log10 of P-Values form mixed-effects analysis
hist(log10(P_Value),main="")

library(qvalue)

Q = qvalue(p = P_Value, fdr.level=0.3,pi0.method = "smoother")

#Fig 3./Fig 4. The FDR procedure
plot(Q)
hist(Q)

#Table 2. The number of significant links under different q-values cutoff
summary(Q$qvalues)

length(which(Q$qvalues <0.0910))
length(which(Q$qvalues <0.13789))
length(which(Q$qvalues <0.24610))
length(which(Q$qvalues <0.27318))
length(which(Q$qvalues <0.39728))
length(which(Q$qvalues <0.54067))

# Table 3 The corresponding significant link with FDR=0.0910 (minimum cutoff of q-values)
Link_number=which(Q$qvalue<0.0910)
write.csv(Link_number,"D:\\link_number.csv")
```