Title
Analysis of Biomarkers of Symptoms in Patients with Schizophrenia

Author
Zhang, Haorui

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Analysis of Biomarkers of Symptoms in Patients with Schizophrenia

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Applied Statistics

by

Haorui Zhang

2020

ABSTRACT OF THE THESIS

Analysis of Biomarkers of Symptoms in Patients with Schizophrenia

by

Haorui Zhang

Master of Applied Statistics

University of California, Los Angeles, 2020

Professor Frederic R. Paik Schoenberg, Chair

This paper aims to apply machine learning methods to analyze the biomarkers of symptoms in patients with schizophrenia. By reducing the dimension of brain patterns via random forest models and mapping brain patterns to symptoms of schizophrenia using multivariate regression models, we will explore the relationship between brain patterns and symptoms of schizophrenia and the association between different types of antipsychotic medication and brain patterns and symptoms.

The thesis of Haorui Zhang is approved.

Ariana Anderson

Mahtash Esfandiari

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

The functional magnetic resonance imaging (fMRI) is a dynamic imaging approach to measure the brain activity by detecting changes associated with blood flow. The introduction of fMRI into neuroscience has instigated a revolution in the magnitude and type of research relating brain function to behavior. For example, fMRI is widely used for detecting neural network problems in patients with schizophrenia. Schizophrenia is a mental illness characterized by relapsing episodes of psychosis [1]. Hallmark symptoms include hallucinations, delusions, and emotional withdrawal. This study, by examining the imaging measures of how 15 individual brain networks behave individually and how they correlate with other networks, explains the relationship among brain patterns, symptoms of schizophrenia, and brain patterns and symptoms of schizophrenia and the role different types of antipsychotic medication play in this. These measures were obtained from resting-state fMRI scans for patients with schizophrenia from two institutions, Centers of Biomedical Research Excellence (COBRE) at University of New Mexico (UNM) and Consortium for Neuropsychiatric Phenomics (CNP) at University of California, Los Angeles (UCLA).

CHAPTER 2

Problem Defining

The Positive and Negative Syndrome Scale (PANSS) in COBRE dataset, and Scale for the Assessment of Negative Symptoms (SANS) and Scale for the Assessment of Positive Symptoms (SAPS) in CNP dataset are well-established scales for evaluating symptom severity in Schizophrenia [2]. To examine the relationship between brain patterns and symptoms of schizophrenia, we apply machine learning tools to explore if the fMRI patterns map to the PANSS symptoms in COBRE dataset, and SANS and SAPS in CNP dataset. Besides, we notice that the patients in the COBRE study are taking different medications, so we would also like to know if there is an interaction between fMRI and medication in predicting symptoms of schizophrenia. To achieve this, we first research on whether the medications taken by each patient belong to first or second generation antipsychotic drugs, and then train some statistical models to study the effect of interactions between fMRI patterns and types of medications on the PANSS variables.

CHAPTER 3

Data

3.1 Variables of the Study

COBRE Dataset

Predictors (All Numerical)		Outcomes (All Numerical)
91 Between-Network Variables	➜	Negative Dimension of PANSS
70 Within-Network Variables		Positive Dimension of PANSS
		Disorganized Dimension of PANSS
		Excited Dimension of PANSS
		Anxiety Dimension of PANSS

Table 3.1 Structure of COBRE study

Clarification of Variables:

- The PANSS variables include a total of 30 symptoms rated for severity on a 7-point scale (min = 1, max = 7). They are grouped into 5 dimensions according to the 30-item Oblimin-rotated model appeared in prior analyses of the PANSS [3].

1. Negative = Blunted Affect + Emotional Withdrawal + Poor Rapport + Passive Apathetic Social Withdrawal + Lack of Spontaneity and Flow of Conversation + Motor Retardation + Active Social Avoidance + Disturbance of Volition
2. Positive = Delusions + Hallucinatory Behavior + Grandiosity + Suspiciousness Persecution + Unusual Thought Content + Preoccupation
3. Disorganized = Stereotyped Thinking + Lack of Judgment and Insight + Conceptual Disorganization + Difficulty in Abstract Thinking + Mannerisms and Posturing + Poor Attention + Disturbance of Volition + Preoccupation + Disorientation
4. Excited = Poor Impulse Control + Excitement + Hostility + Uncooperativeness
5. Anxiety = Anxiety + Depression + Tension + Guilt Feelings + Somatic Concern

The 5 variables of PANSS are all numerical variables with integer values. Higher PANSS scores always reveal more severe symptoms, regardless of the dimensions.

- Between-network variables

The 91 between-network measurements are all numerical, informing us how individual brain networks correlate with other networks.

- Within-network variables

The 70 within-network measurements are all numerical. They inform us how individual brain network work within itself. Within-network variables include fMRI measurements, such as small worlds and global efficiency.

CNP Dataset

Predictors (All Numerical)		Outcomes (All Numerical)
91 Between-Network Variables	➜	Sum of SANS
70 Within-Network Variables		Sum of SAPS

Table 3.2 Structure of CNP study

Clarification of Variables:

- Sum of SANS

  This is the sum of 24 scales for the assessment of negative symptoms (SANS) for each individual. It is a numeric variable with a minimum value 0 and a maximum value 24. A patient with a higher value of sum of SANS has more severe symptoms of schizophrenia.

- Sum of SAPS

  This is the sum of 35 scales for the assessment of positive symptoms (SAPS) for each individual. It is a numeric variable with a minimum value 0 and a maximum value 27. A patient with a higher value of sum of SAPS has more severe symptoms of schizophrenia.

- Between-Network and Within-Network Variables are the same as defined above.

COBRE_Med Dataset

Predictors		Outcomes (All Numerical)
Small Worlds fMRI (14 Numerical Variables)	➜	Negative Dimension of PANSS
Medication Type (1 Categorical Variable)		Positive Dimension of PANSS
Medication * fMRI (14 Interactions)		Disorganized Dimension of PANSS
		Excited Dimension of PANSS
		Anxiety Dimension of PANSS

Table 3.3 Structure of COBRE_Med study

Clarification of Variables:

- Small-World fMRI Measurements

  Small-world network is a highly clustered system but with small mean path length between networks which allow the information transferred with high efficiency. The human brain can be considered as a sparse, complex network modeled by the small-world properties. Once the brain network was disrupted by disease, the small-world properties would be altered to manifest that the information integration was inefficiency and the network was loosely organized [4]. Here we include the fMRI measurements for 14 small-world networks such as structural cerebellar network, dorsal attention network, fronto-parietal control network, cingulo-opercular network, etc. All of these measurements are numeric variables, with values ranging from 0.004167 to 3.9325.

- Medication Type

  This variable indicates the type of medication the patient was taking. It is a factor with three levels, (1.0), (0.1) and (1.1) where (1.0) means the patient was taking 1st-generation antipsychotics, (0.1) means the patient was taking 2nd-generation antipsychotics, and (1.1) means the patient was taking both 1st-and 2nd-generation antipsychotics. 1st-generation antipsychotics refer to those developed in the 1950s, and the 2nd-generation antipsychotics refer to those developed since the 1980s. The medications are classified according to prior comparative effectiveness review (CER) of Food and Drug Administration (FDA)-approved first-generation and second-generation antipsychotic medications [5].
- Interaction between Small Worlds fMRI and Medication Type

  14 interaction terms are added in the multivariate model showing the interaction between different medications and the 14 small world fMRI measurements.
- 5-dimension PANSS variable is the same as defined for COBRE dataset.

3.2 Exploratory Data Analysis

3.2.1 Exploratory Data Analysis for COBRE

COBRE data has a total of 69 observations, 91 between-network predictors, 70 within-network predictors, and 5 outcome variables – 5 dimensions of PANSS.

- Exploring the predictors

  Summary statistics of the 161 between and within-network predictors show that most of these predictors are numerical variables ranging from 0 to 1. The minimum value of between-network measurements is -0.2396064, and the minimum of within-

network measurements is 0. The maximum value of between-network measurements is 0.7352335, and the maximum of within-network measurements is 172.7878, which is an outlier. We noticed that the variable “Memory_retrieval.char_path_length” containing the outlier 172.7878 is not included in the final predictors selected by the random forest method. The second largest value of within-network measurements was 6.69197, and the variable containing it – "Uncertain.char_path_length" is also excluded from the final predictors.

- Exploring the outcomes

Variable	Min	1st Quant	Median	Mean	3rd Quant	Max
Negative	8	11	14	15.62	19	32
Positive	8	14	16	17.42	21	38
Disorganized	9	12	15	15.46	18	25
Excited	4	4	4	5.426	6	12
Anxiety	5	8	12	11.36	15	28

Table 3.4 Summary Statistics of Outcome Variable (5-Dimensional PANSS Scores)

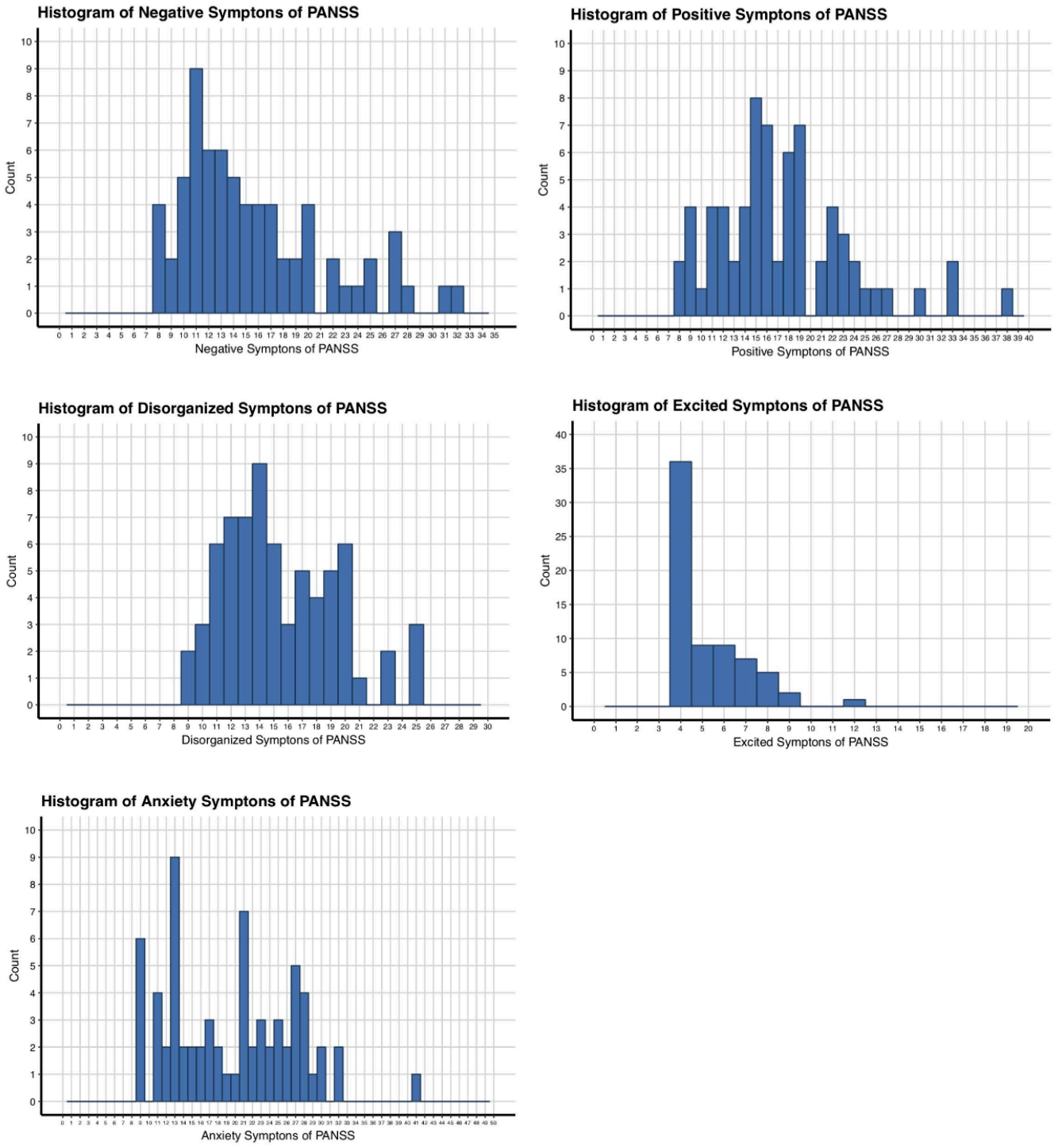

Figure 3.1 Histogram of Outcome Variable (5-Dimensional PANSS Scores)

- Exploring the relationship between predictors and outcomes

Between-Network vs. Negative PANSS	Within-Network vs. Negative PANSS
Min = -0.354, Max = 0.033, Mean of abs = 0.149	Min = -0.388, Max = 0.345, Mean of abs = 0.125
Between-Network vs. Positive PANSS	Within-Network vs. Positive PANSS
Min = -0.261, Max = 0.247, Mean of abs = 0.095	Min = -0.349, Max = 0.166, Mean of abs = 0.081
Between-Network vs. Disorganized PANSS	Within-Network vs. Disorganized PANSS
Min = -0.287, Max = 0.274, Mean of abs = 0.086	Min = -0.391, Max = 0.337, Mean of abs = 0.117
Between-Network vs. Excited PANSS	Within-Network vs. Excited PANSS
Min = -0.240, Max = 0.031, Mean of abs = 0.080	Min = -0.359, Max = 0.326, Mean of abs = 0.127
Between-Network vs. Anxiety PANSS	Within-Network vs. Anxiety PANSS
Min = -0.288, Max = 0.242, Mean of abs = 0.085	Min = -0.371, Max = 0.211, Mean of abs = 0.093

Table 3.5 Summary Statistics of Correlation Coefficients between Predictors and Outcomes

The table above shows that, on average, the correlation coefficients are quite low. Between-network measurements have the highest correlation coefficient with negative dimension of PANSS in the COBRE data.

3.2.2 Exploratory Data Analysis for CNP

CNP data has a total of 42 observations, 91 between-network predictors, 70 within-network predictors, and 2 outcome variables – sum of SANS and sum of SAPS.

- Exploring the predictors

Summary statistics of the 161 between and within-network predictors shows that most these predictors are numerical variables ranging from 0 to 1. The minimum value

of between-network measurements is -0.2258408, and the minimum of within-network measurements is 0. The maximum value of between-network measurements is 0.7976123, and the maximum of within-network measurements is 5.67975, which may be an outlier. We noticed that the variable “Cerebellar.char_path_length” containing the value 5.67975 is not included in the final predictors selected by the random forest model.

- Exploring the outcomes

Variable	Min	1st Quant	Median	Mean	3rd Quant	Max
SANS	0	11	14.5	13.9	18.75	24
SAPS	0	8	12	11.74	14.75	27

Table 3.6 Summary Statistics of Outcome Variable (Sum of SANS and SAPS Scores)

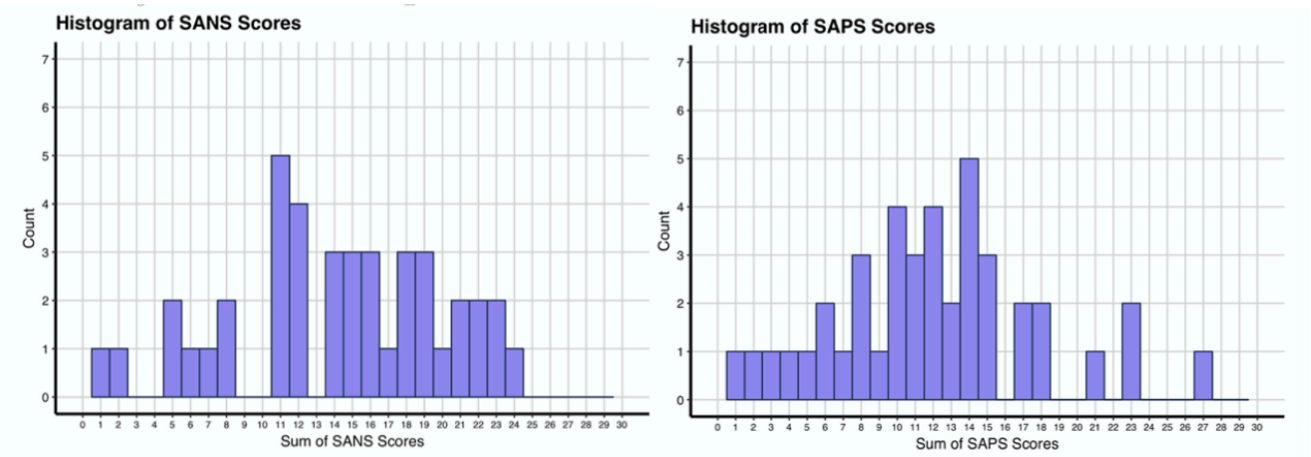

Figure 3.2 Histogram of Outcome Variable (Sum of SANS and SAPS Scores)

- Exploring the relationship between predictors and outcomes

Between-Network vs. SANS	Within-Network vs. SANS
Min = -0.220, Max = 0.139, Mean of abs = 0.070	Min = -0.241, Max = 0.210, Mean of abs = 0.087
Between-Network vs. SAPS	Within-Network vs. SAPS
Min = 0.044, Max = 0.494, Mean of abs = 0.269	Min = -0.319, Max = 0.392, Mean of abs = 0.134

Table 3.7 Summary Statistics of Correlation Coefficients between Predictors and Outcomes

The table above shows that, on average, the correlation coefficients are quite low. Between-network measurements have the highest correlation coefficient with SAPS in the CNP data.

3.2.3 Exploratory Data Analysis for COBRE_Med

The COBRE_Med has 69 observations. The minimum value of the 14 small world fMRI variables is 0.004167 and the maximum value is 3.9325.

Medication	Number of Patients	Mean Age	Minimum Age	Maximum Age	Percentage of Female
Generation 1	3	42.3	35	49	0%
Generation 2	41	35.5	18	62	22%
Generation 1 & Generation 2	23	45.5	22	65	21.7%
No Medication	2	27	26	28	0%

Table 3.8 Summary Table of Medications

The table above suggests that there are 3 patients taking 1st-generation medications, 41 patients taking 2nd-generation medications, and 23 patients taking both generations, while 2 patients are not taking any kind of medications. This table also includes some demographic features of each group of patients.

CHAPTER 4

Methodology

4.1 Random Forest Feature Selection

Often in data science we have hundreds or even millions of features and we want a way to create a model that only includes the most important features. The process of identifying only the most relevant features is called “feature selection” [6]. In this study, because we have 161 predictors (91 between-network and 70 within-network measurements) in both COBRE and CNP datasets, feature selection appear to be necessary.

Random forest, as one the most popular machine learning algorithms, is so successful because it provides in general a good predictive performance, low overfitting, and high interpretability. This interpretability is given by the fact that it is straightforward to derive the importance of each variable on the tree decision. In other words, it is easy to compute how much each variable is contributing to the decision. Feature selection using random forest comes under the category of embedded methods which have the benefits of high accuracy, good generalization, and high interpretability [7].

Random forests consist of 4 –12 hundred decision trees, each of them built over a random extraction of the observations from the dataset and a random extraction of the features. Not every tree sees all the features or all the observations, and this guarantees that the trees are de-

correlated and therefore less prone to over-fitting. Each tree is also a sequence of yes-no questions based on a single or combination of features. At each node, the tree divides the dataset into 2 buckets, each of them hosting observations that are more similar among themselves and different from the ones in the other bucket. Therefore, the importance of each feature is derived from how "pure" each of the buckets is [7].

In this study, the feature selection process was completed using the package "*randomForestSRC*" in R which allows us to build multivariate random forest trees. We built 4 random forest trees, mapping COBRE between-network to 5-dimensional PANSS, COBRE within-network to 5-dimensional PANSS, CNP between-network to SANS and SAPS, and CNP within-network to SANS and SAPS.

Due to the random splitting of trees, each time we would get a different set of important predictors from the variable selection process. We repeat this process 5 times for each tree model. Common predictors which appear at least twice among the five trials, are selected as final predictors. Other predictors that appear once among five attempts are considered as potential predictors.

4.2 Multivariate Regression

Multivariate regression is one of the simplest algorithm. It comes under the class of supervised learning algorithms i.e, when we are provided with training dataset. Multivariate analysis is used to address the situations where multiple measurements are made on each experimental unit and the relations among these measurements and their structures are important

[8]. It attempts to determine a formula that can describe how elements in a vector of variables respond simultaneously to changes in others [9].

The multivariate regression model has the form:

$$Y_{ik} = \beta_{0k} + \sum_{j=1}^{p} \beta_{jk} x_{ij} + \varepsilon_{ik}$$

where Y_{ik} is the *k*-th real-valued response for the *i*-th observation, β_{0k} is the regression intercept for the *k*-th response, β_{jk} is the *j*-th predictor's regression slope for the *k*-th response, x_{ij} is the *j*-th predictor for the *i*-th observation, and $(\varepsilon_{i1}, ..., \varepsilon_{im}) \overset{iid}{\sim} N(\mathbf{0_m}, \mathbf{\Sigma})$ is a multivariate Gaussian error vector.

4.2.1 Brain Patterns vs. Symptoms of Schizophrenia

After selecting important features from the random forest models, we built a multivariate regression model to map significant predictors of brain patterns to 5-dimensional PANSS in COBRE dataset and SANS and SAPS in CNP dataset, respectively. A multivariate regression model was chosen because the variables we would like to predict, PANSS in COBRE dataset and SANS and SAPS in CNP dataset are both multi-dimensional, and all of our predictors and outcomes are numeric.

Four model building schemes, as shown below, were used. The model with the smallest MSE is selected.

1) Common Predictor (Frequency >= 2)

2) Age + Gender + Common Predictor (Frequency >= 2)

3) Age + Gender + Common Predictor (Frequency >= 2) + Potential Predictors

4) Common Predictor(Frequency >= 2) + Potential Predictors

4.2.2 PANSS vs. Interactions between fMRI and Medication

In statistics, multiple linear regression is a linear approach to model the relationship between one dependent variable and multiple independent variables, while multivariate regression pertains to multiple dependent variables and multiple independent variables. Here, again, a multivariate regression model was chosen because the variable we would like to predict, PANSS, has 5 dimensions. It also helps us examine the interaction effects between fMRI and the medication types. We dropped all 3 observations with medication (1,0) — patients taking only 1st-generation antipsychotics, because these observations have high collinearity.

4.3 Cross-Study Validation

Cross-validation and related resampling methods are de facto standard for ranking supervised learning algorithms. They allow estimation of prediction accuracy using subsets of data that have not been used to train the algorithms. This avoids over-optimistic accuracy estimates caused by ‘re-substitution’. It is common to evaluate algorithms by estimating prediction accuracy via cross-validation for several datasets. This approach recognizes possible variations in the relative performances of learning algorithms across studies or fields of application. However, it is not fully consistent with the ultimate goal, in the development of models with applications of statistics in medical research, of providing accurate predictions for fully independent samples, originating from institutions and processed by laboratories that did not generate the training datasets. Therefore, another perspective was promoted: a good learning

algorithm should be a generalist, in the sense that it yields models that may be suboptimal for the training population, or not fully representative of the dataset at hand, but that perform reasonably well across different populations and laboratories employing comparable but not identical methods [10], which can be called cross-study validation.

In this study, performing a cross-study validation makes sense because SANS in CNP dataset is essentially measuring the same symptoms with the negative dimension of PANSS in COBRE dataset and SAPS in CNP dataset is equivalent to the positive dimension of PANSS in COBRE dataset. Moreover, all the brain pattern variables, including 91 between-network and 70 within-network, are measured on the same scale. We train brain patterns, negative and positive dimensions of PANSS in the COBRE data, and test the model using brain patterns in the CNP data. We expect that the COBRE model could explain a good amount of variations in SANS and SAPS appeared in the CNP dataset.

Nevertheless, PANSS and SANS/SAPS are not on the same scale (as shown in the data exploration part). Therefore, we need to scale these response variable in both COBRE and CNP data before conducting cross-study model validation. After scaling the response variables in both datasets to have means of 0 and standard deviations of 1, the PANSS in COBRE and SANS/ SAPS in CNP are now comparable.

For negative symptoms of schizophrenia, the R-squared value is calculated as:

$$R^2 = \frac{SSreg}{SStot} = \frac{\sum_{i=1}^{n} (Predicted\ Negative\ Symptoms\ of\ CNP_i - Actual\ SANS\ of\ CNP_i)^2}{\sum_{i=1}^{n} (Actual\ SANS\ of\ CNP_i - Mean\ of\ Actual\ SANS\ of\ CNP)^2}$$

For positive symptoms of schizophrenia, the R-squared value is calculated as:

$$R^2 = \frac{SSreg}{SStot} = \frac{\sum_{i=1}^{n} (Predicted\ Positive\ Symptoms\ of\ CNP_i - Actual\ SAPS\ of\ CNP_i)^2}{\sum_{i=1}^{n} (Actual\ SAPS\ of\ CNP_i - Mean\ of\ Actual\ SAPS\ of\ CNP)^2}$$

CHAPTER 5

Summary of Results

5.1 Results of Multivariate Regression Model (Brain Patterns vs. Symptoms of Schizophrenia)

The final model is shown in the table below:

	COBRE Model	CNP Model
Between-Network Predictors	Sensory.Hand..Fronto.parietal Visual.Auditory Visual.Sensory.Hand. Salience.Dorsal_Attention Sensory.Mouth..Fronto.parietal Sensory.Mouth..Auditory Salience.Auditory Sensory.Hand..Dorsal_Attention Sensory.Mouth..Dorsal_Attention Visual.Dorsal_Attention	Sensory.Hand..Default_Mode Default_Mode.Auditory Fronto.parietal.Cerebellar Visual.Cerebellar Uncertain.Default_Mode Salience.Cerebellar Default_Mode.Cerebellar
Within-Network Predictors	Sensory.somatomotor_Hand.small_world Sensory.somatomotor_Hand.char_path_length Dorsal_attention.clust_coef Cingulo.opercular_Task_Control.global_eff	Dorsal_attention.small_world Sensory.somatomotor_Hand.small_world Sensory.somatomotor_Hand.global_eff Auditory.mod
	⬇	⬇
Outcome	Negative Dimension of PANSS Positive Dimension of PANSS Disorganized Dimension of PANSS Excited Dimension of PANSS Anxiety Dimension of PANSS	Sum of SANS Sum of SAPS

Table 5.1 Final Models (Brain Patterns vs. Symptoms of Schizophrenia)

In this final model, the number of between-network variables exceeds the number of within-network variables in both COBRE and CNP, which suggests that between-network brain patterns may play a more important role in explaining the symptoms of schizophrenia than within-network brain patterns. This model explains 27.24% of variation in negative dimension of PANSS, 42.44% of variation in positive dimension of PANSS, 44.99% of variation in disorganized dimension of PANSS, 19.76% of variation in excited dimension of PANSS, and 26.07% variation in anxiety dimension of PANSS. Here the adjusted R-square value is preferable because the number of variables in our study is larger than the sample size.

Data	Outcome	Adjusted R-squared
COBRE	Negative	27.24 %
	Positive	42.44%
	Disorganized	44.99%
	Excited	19.76%
	Anxiety	25.54%
CNP	SANS_sum	30.02%
	SAPS_sum	32.24%

Table 5.2 Table of Adjusted R-squares

Some significant brain patterns mapped to 5 dimensions of PANSS are listed in the table below:

Red: High values of the brain pattern = More severe symptoms of schizophrenia

Blue: Low values of the brain pattern = Less severe symptoms of schizophrenia

Data	Outcome	Significant Predictor	Coefficient Estimate	P-value
COBRE	Negative	None	/	/
	Positive	Sensory.Hand..Fronto.parietal	- 43.176	0.000679
		Sensory.Mouth..Fronto.parietal	39.195	0.001315
		Sensory.Mouth..Auditory	26.517	0.002078
		Sensory.Hand..Dorsal_Attention	53.030	0.003089
		Sensory.Mouth..Dorsal_Attention	-35.692	0.002186
		Visual.Dorsal_Attention	-32.674	0.006364
	Disorganized	Sensory.Hand..Fronto.parietal	-15.5873	0.04594
		Sensory.Mouth..Fronto.parietal	18.3486	0.01589
		Sensory.Mouth..Auditory	15.9711	0.00347
		Sensory.Hand..Dorsal_Attention	32.0675	0.00482
		Sensory.Mouth..Dorsal_Attention	-20.2867	0.00585
		Visual.Dorsal_Attention	-21.5572	0.00483
		Cingulo.opercular_Task_Control.global_eff	-21.0985	0.04379
	Excited	Sensory.Hand..Fronto.parietal	-7.8761	0.0490
	Anxiety	Sensory.Hand..Fronto.parietal	-19.930	0.0497
		Visual.Dorsal_Attention	-22.055	0.0247
CNP	SANS_sum	None	/	/
	SAPS_sum	None	/	/

Table 5.3 Significant Predictors

We would like to know whether our model contains at least one useful predictor in predicting the 5-dimensional PANSS and the 2-dimensional SANS/SAPS.

- Null Hypothesis: All slope coefficients of the brain pattern predictors are zero.
- Alternative Hypothesis: At least one of the slope coefficients of the brain pattern predictors is not zero.

To predict the negative dimension of PANSS, the p-values associated with all predictors are greater than 0.05. Therefore, we fail to reject the null hypothesis, and conclude that none of the brain patterns is useful predictor of negative dimension of PANSS.

To predict the positive dimension of PANSS, the p-values associated with sensory hand fronto-parietal, sensory mouth fronto-parietal, sensory mouth auditory, sensory hand dorsal attention, sensory mouth dorsal attention, and visual dorsal attention are smaller than 0.05. We reject the null hypothesis. Keeping all other constant, if sensory mouth fronto-parietal/sensory mouth auditory/sensory hand dorsal attention increases, the patient will score higher on the positive dimension of PANSS. If sensory hand fronto-parietal/sensory mouth dorsal attention/ visual dorsal attention increases, the patient will score lower on positive dimension of PANSS.

To predict the disorganized dimension of PANSS, the p-values associated with sensory hand fronto-parietal, sensory mouth fronto-parietal, sensory hand dorsal attention, sensory mouth dorsal attention, visual dorsal attention, and cingulo-opercular task control global efficiency are smaller than 0.05. We reject the null hypothesis. Keeping all other constant, if sensory mouth fronto-parietal/ sensory mouth auditory/ sensory hand dorsal attention increases, the patient will score higher on the disorganized dimension of PANSS. If sensory hand fronto-parietal/sensory mouth dorsal attention/visual dorsal attention/cingulo-opercular task control global efficiency increases, the patient will score lower on disorganized dimension of PANSS.

To predict the excited dimension of PANSS, the p-value associated with sensory hand fronto-parietal is smaller than 0.05. We reject the null hypothesis. Keeping all other constant, if sensory hand fronto-parietal increases, the patient will score lower on excited dimension of PANSS.

To predict the anxiety dimension of PANSS, the p-values associated with sensory hand fronto-parietal and visual dorsal attention are smaller than 0.05. We reject the null hypothesis. Keeping all other constant, if sensory hand fronto-parietal or visual dorsal attention increases, the patient will score lower on the anxiety dimension of PANSS.

To predict the sum of SANS, the p-values associated with all predictors are greater than 0.05. Therefore, we fail to reject the null hypothesis, and conclude that none of the brain patterns is useful predictor of sum of SANS.

To predict sum of SAPS, the p-values associated with all predictors are greater than 0.05. Therefore, we fail to reject the null hypothesis, and conclude that none of the brain patterns is useful predictor of sum of SAPS.

A cross-study model validation using COBRE as the training dataset and CNP as the testing dataset was also conducted to check the performance of the model we trained as well as guard against the possibility of overfitting. For negative symptoms of Schizophrenia, the R-squared value is:

$$R^2 = \frac{SSreg}{SStot} = \frac{\sum_{i=1}^{n} (Predicted\ Negative\ Symptoms\ of\ CNP_i - Actual\ SANS\ of\ CNP_i)^2}{\sum_{i=1}^{n} (Actual\ SANS\ of\ CNP_i - Mean\ of\ Actual\ SANS\ of\ CNP)^2} = \frac{16.42962}{41} = 40.072\ \%$$

For positive symptoms of schizophrenia, the R-squared value is:

$$R^2 = \frac{SSreg}{SStot} = \frac{\sum_{i=1}^{n} (Predicted\ Positive\ Symptoms\ of\ CNP_i - Actual\ SAPS\ of\ CNP_i)^2}{\sum_{i=1}^{n} (Actual\ SAPS\ of\ CNP_i - Mean\ of\ Actual\ SAPS\ of\ CNP)^2} = \frac{22.31222}{41} = 54.420\ \%$$

The COBRE multivariate regression model explains 40.072% of variation in SANS of the CNP data, and explains 52.420% of variation in SAPS in the CNP data.

5.2 Results of Multivariate Regression Model (PANSS vs. Interactions between fMRI and Medication)

Again, we use the adjusted R-square value to evaluate our model because the number of variables in our study is larger than the sample size. The multivariate regression model examining the interaction between medication and fMRI explains 41.17 % of variation in negative dimension of PANSS, 62.82 % of variation in positive dimension of PANSS, 58.05 % of variation in disorganized dimension of PANSS, 56.81 % of variation in excited dimension of PANSS, and 64.30 % variation in anxiety dimension of PANSS.

Outcome	Adjusted R-squared
Negative	41.17 %
Positive	62.82 %
Disorganized	58.05 %
Excited	56.81 %
Anxiety	64.30 %

Table 5.4 Table of Adjusted R-squares

Outcome	Significant Predictor	Coefficient Estimate	P-value
Negative	Sensory.somatomotor_Hand.small_world (Not an interaction; none of the interactions are significant)	-13.12387	0.0394
Positive	Cerebellar.small_world (Not an interaction)	2.8123	0.033399
	Medication 1.1 * Auditory.small_world	27.5664	0.000456
	Medication 1.1 * Fronto.parietal_Task_Control.small_world	-31.3821	0.022279
	Medication 1.1 * Memory_retrieval.small_world	6.4381	0.006562
	Medication 1.1 * Salience.small_world	23.3722	0.035478
Disorganized	Medication 1.1 * Auditory.small_world	11.8079	0.029385
Excited	Medication 1.1 * Salience.small_world	-7.994422	0.02127
Anxiety	Ventral_attention.small_world (Not an interaction)	9.1749	0.003292
	Medication 1.1 * Auditory.small_world	20.2296	0.000691
	Medication 1.1 * Fronto.parietal_Task_Control.small_world	-20.5424	0.047610
	Medication 1.1 * Salience.small_world	17.2762	0.041185
	Medication 1.1 * Uncertain.small_world	36.5656	0.010055

Table 5.5 Significant Predictors

We would like to know whether our model contains at least one useful predictor in predicting the 5-dimensional PANSS.

- Null Hypothesis: All slope coefficients of the brain pattern predictors are zero.
- Alternative Hypothesis: At least one slope coefficient of the brain pattern predictors is not zero.

We reject the null hypothesis because the p-values associated with multiple interaction effects are smaller than 0.05. Further interpretations of the results will be provided along with the interaction plots in the following section.

- Interpretation of Interaction Plots of Significant Interaction Effects

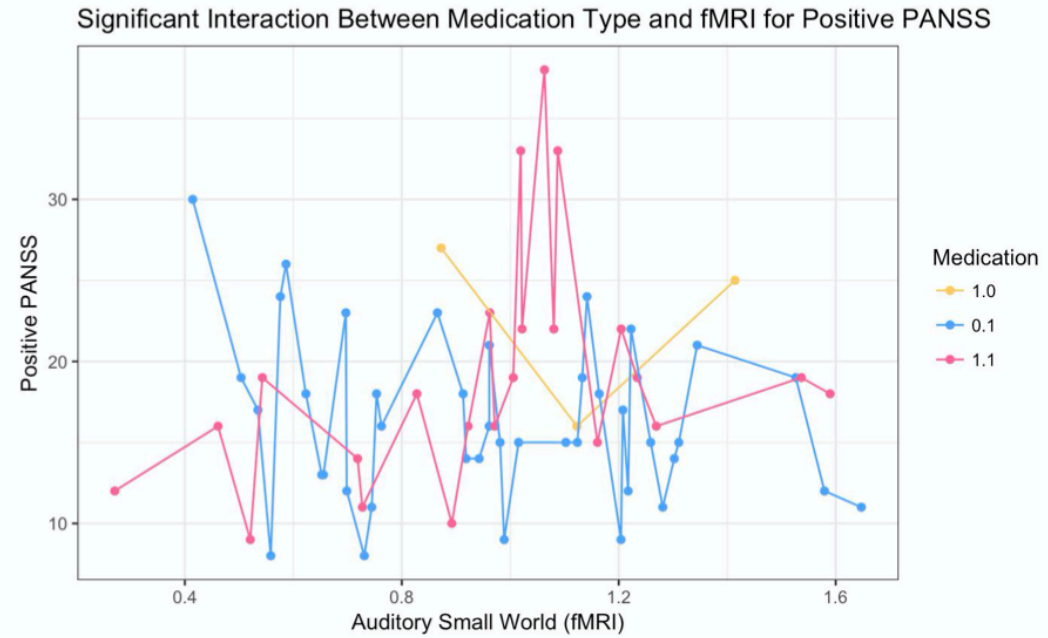

Figure 5.1 Interaction Between Medication and Auditory Small World for Positive PANSS

The effect of auditory small world on positive PANSS is not similar for patients taking different medication types. When auditory small world is smaller than 0.9, patients taking the 2nd-generation medications score higher on positive PANSS than patients taking both-generation medications. When auditory small world is larger than 0.9, patients taking both-generation medications score higher on positive PANSS with only a few exceptions.

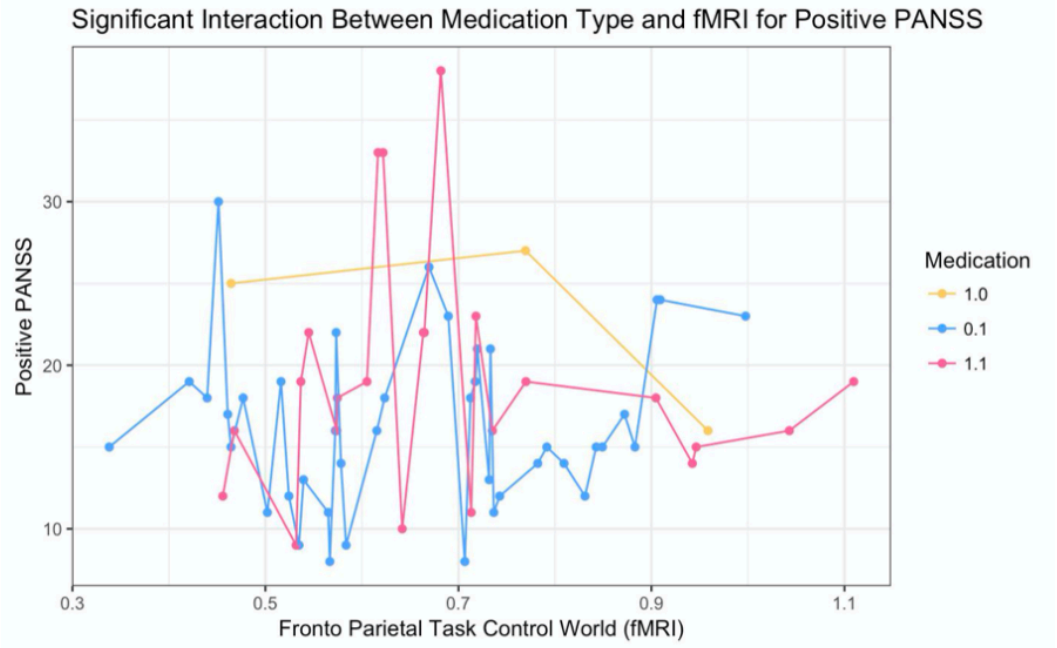

Figure 5.2 Interaction Between Medication and Fronto-Parietal Task Control World for Positive PANSS

The effect of fronto-parietal task control world on positive PANSS is not similar for patients taking different medication types. When fronto-parietal task control world is smaller than around 0.53, patients taking 2nd-generation medications score higher on positive PANSS than patients taking both-generation medications. When fronto-parietal task control world is approximately between 0.53 and 0.9, patients taking both-generation medications score higher on positive PANSS. When parietal task control world is higher than 0.9, patients taking 2nd-generation medications score higher again.

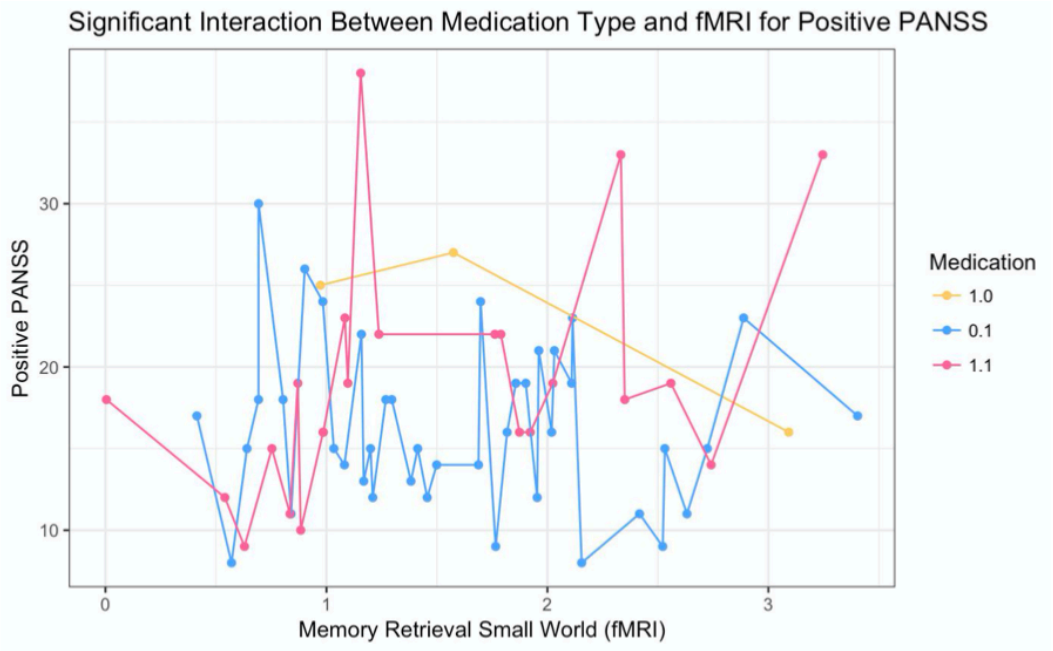

Figure 5.3 Interaction Between Medication and Memory Retrieval Small World for Positive PANSS

The effect of memory retrieval small world on positive PANSS is not similar for patients taking different medication types. When memory retrieval small world is smaller than 1, patients taking the 2nd-generation medications generally score higher on positive PANSS. When memory retrieval small world is above 1, patients taking both-generation medications score higher on positive PANSS, with some exceptions when memory retrieval small world is around 1.8 and 2.8.

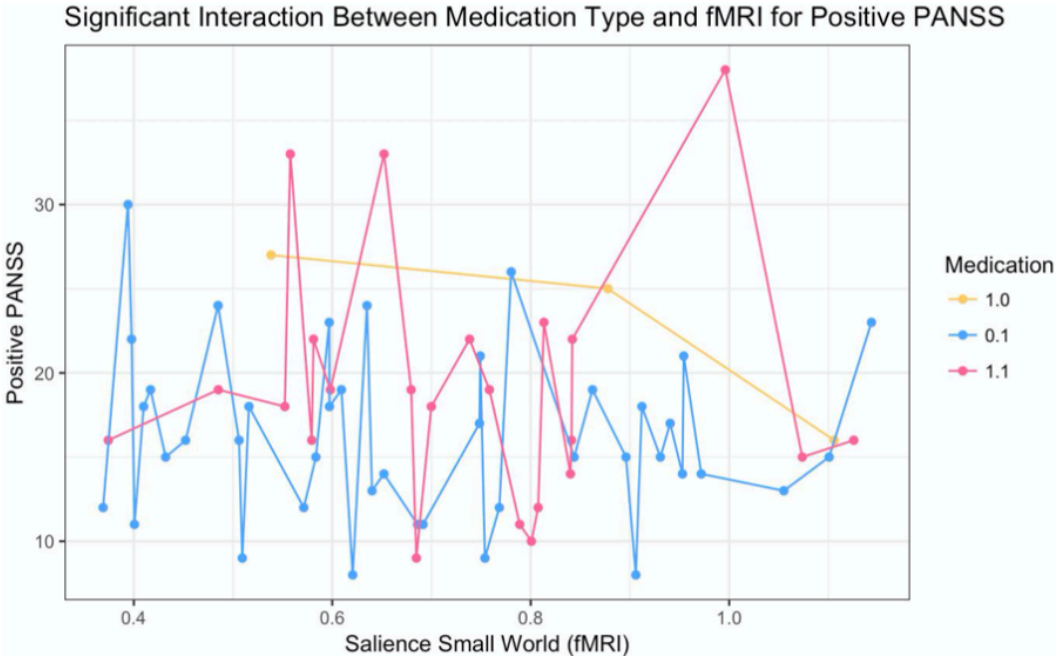

Figure 5.4 Interaction Between Medication and Salience Small World for Positive PANSS

The effect of salience small world on positive PANSS is not similar for patients taking different medication types. When salience small world is smaller than 0.55, patients taking 2nd-generation medications score higher on positive PANSS. When salience small world is above 0.55, most patients taking both-generation medications score higher on positive PANSS.

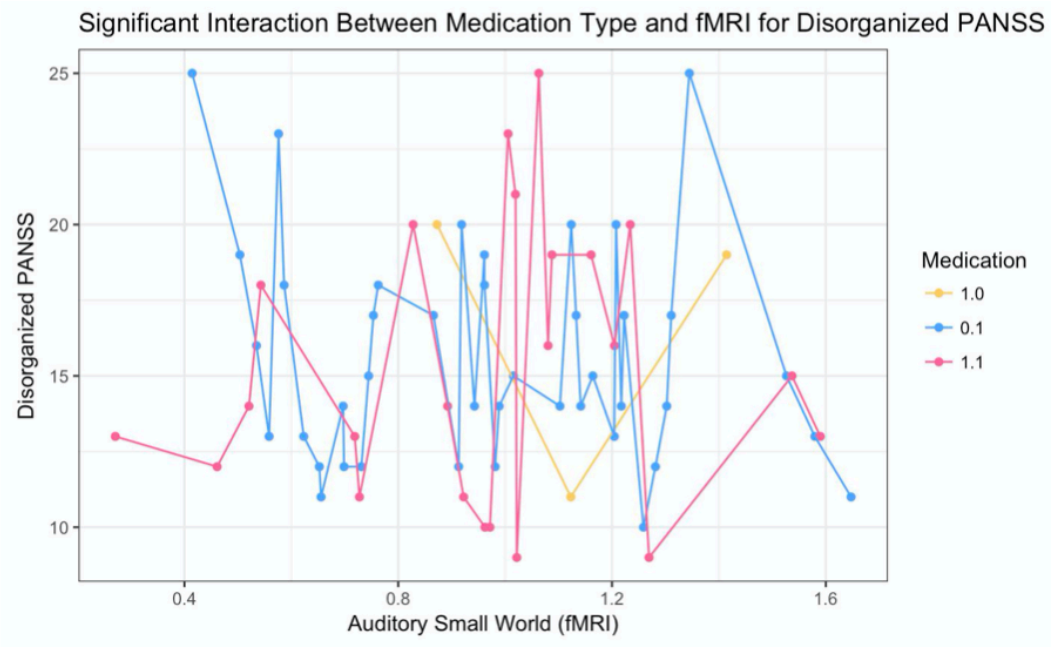

Figure 5.5 Interaction Between Medication and Auditory Small World for Disorganized PANSS

The effect of auditory small world on disorganized PANSS is not similar for patients taking different medication types. When auditory small world is smaller than around 0.97, patients taking 2nd-generation medications tend to score higher on disorganized PANSS, with three exceptions in which patients taking both-generation medications score higher. When auditory small world is between 0.97 to 1.25, patients taking both-generation medications generally score higher on disorganized PANSS. When auditory small world is larger than 1.25, patients taking 2nd-generation medications score higher on disorganized PANSS again.

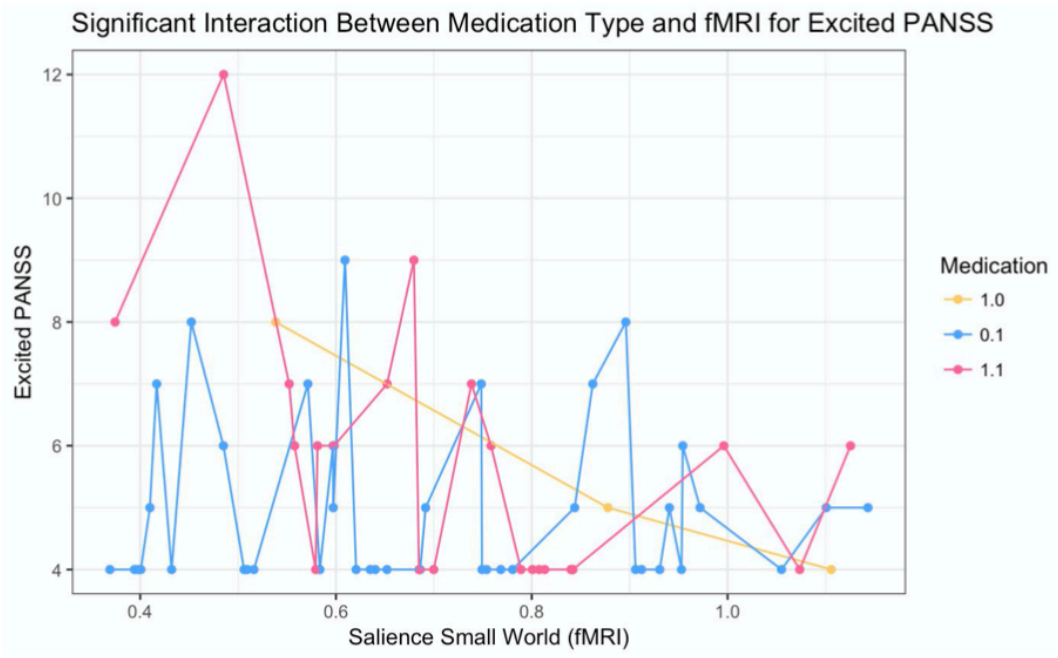

Figure 5.6 Interaction Between Medication and Salience Small World for Excited PANSS

The effect of salience small world on excited PANSS is not similar for patients taking different medication types. When salience small world is smaller than 0.78, patients taking both-generation medications score higher on excited PANSS, with two exceptions around 0.6. When salience small world is larger than 0.78, patients taking 2nd-generation medications score higher on excited PANSS, with two exceptions around 1 and 1.13.

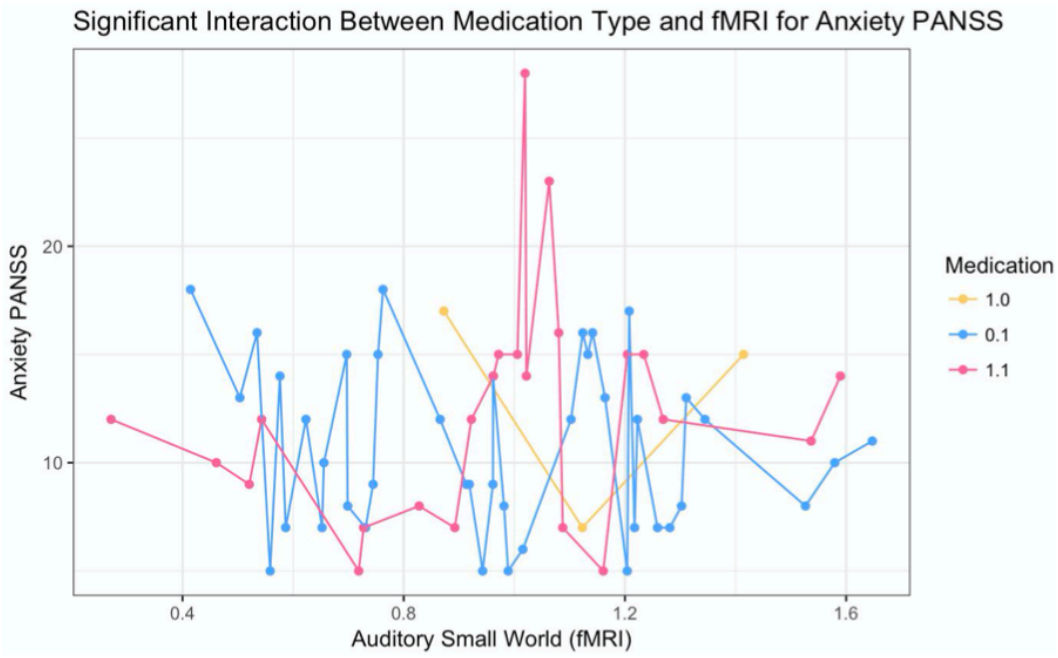

Figure 5.7 Interaction Between Medication and Auditory Small World for Anxiety PANSS

The effect of auditory small world on anxiety PANSS is not similar for patients taking different medication types. When auditory small world is smaller than 0.9, patients taking 2nd-generation medications score higher on anxiety PANSS. When auditory small world is between 0.9 and 1.1, patients taking both-generation medications score higher on anxiety PANSS. Patients taking 2nd-generation medications score higher on anxiety PANSS when auditory small world is between 1.1 and 1.2, but they score lower when auditory small world is above 1.2.

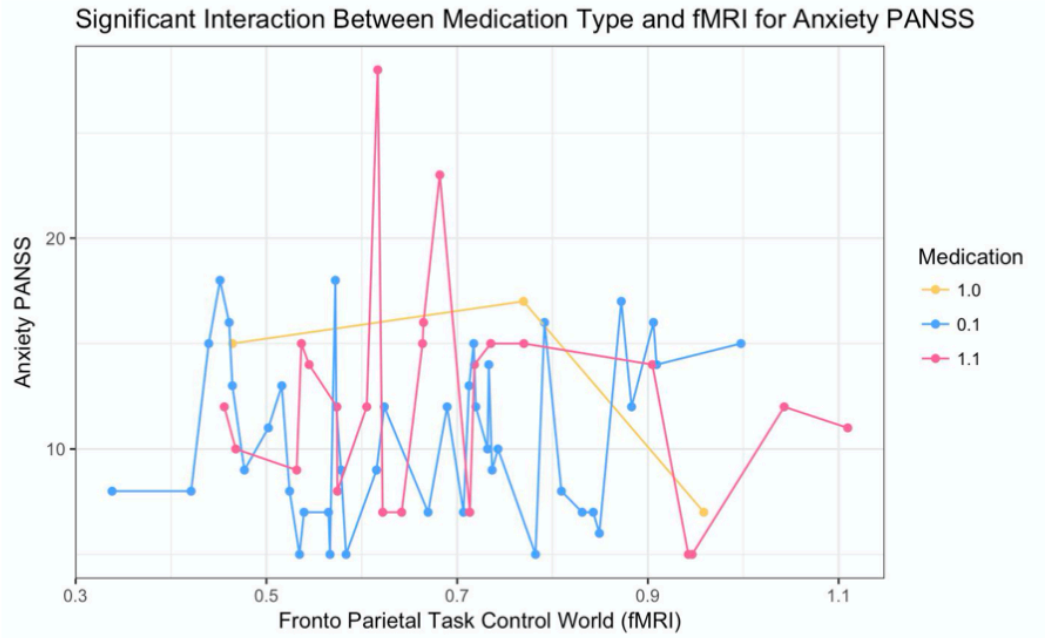

Figure 5.8 Interaction Between Medication and Fronto-Parietal Task Control World for Anxiety PANSS

The effect of fronto-parietal task control world on anxiety PANSS is not similar for patients taking different medication types. When fronto-parietal task control world is smaller than 0.53, patients taking 2nd-generation medications score higher on anxiety PANSS. When fronto-parietal task control world is between 0.53 and 0.77, patients taking both-generation medications score higher on anxiety PANSS. When fronto-parietal task control world is above 0.77, patients taking 2nd-generation medications score higher again.

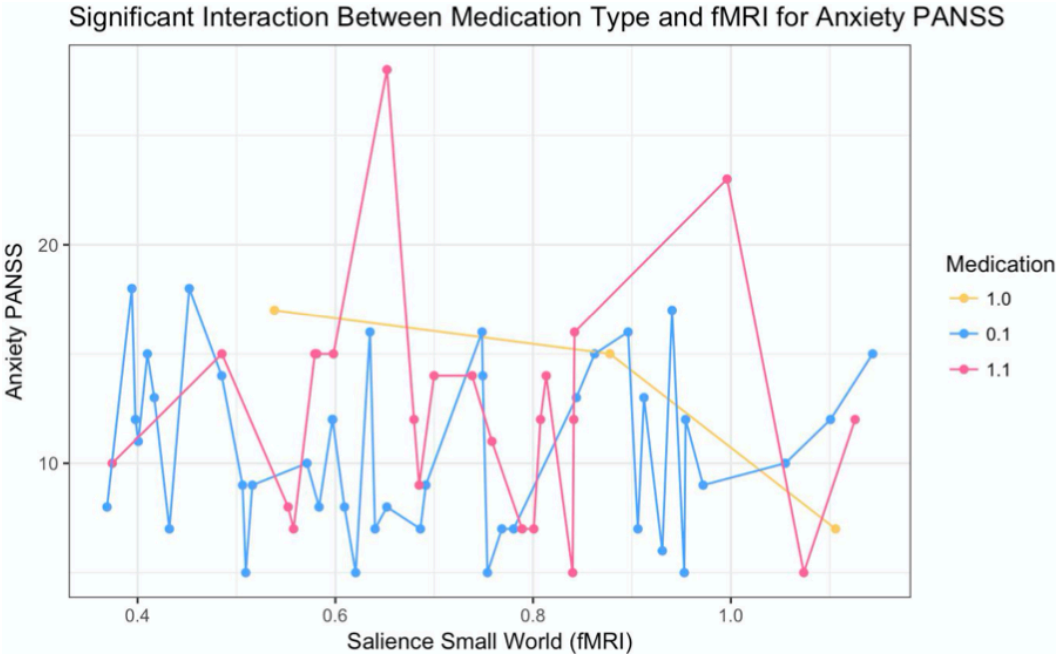

Figure 5.9 Interaction Between Medication and Salience Small World for Anxiety PANSS

The effect of salience small world on anxiety PANSS is not similar for patients taking different medication types. When salience small world is smaller than 0.48, patients taking 2nd-generation medications score higher on anxiety PANSS. When salience small world is larger than 0.48, patients taking both-generation medications tend to score higher on anxiety PANSS, although one exception occurs when salience small world is around 0.75 and two exceptions occur when salience small around is larger than 1.05.

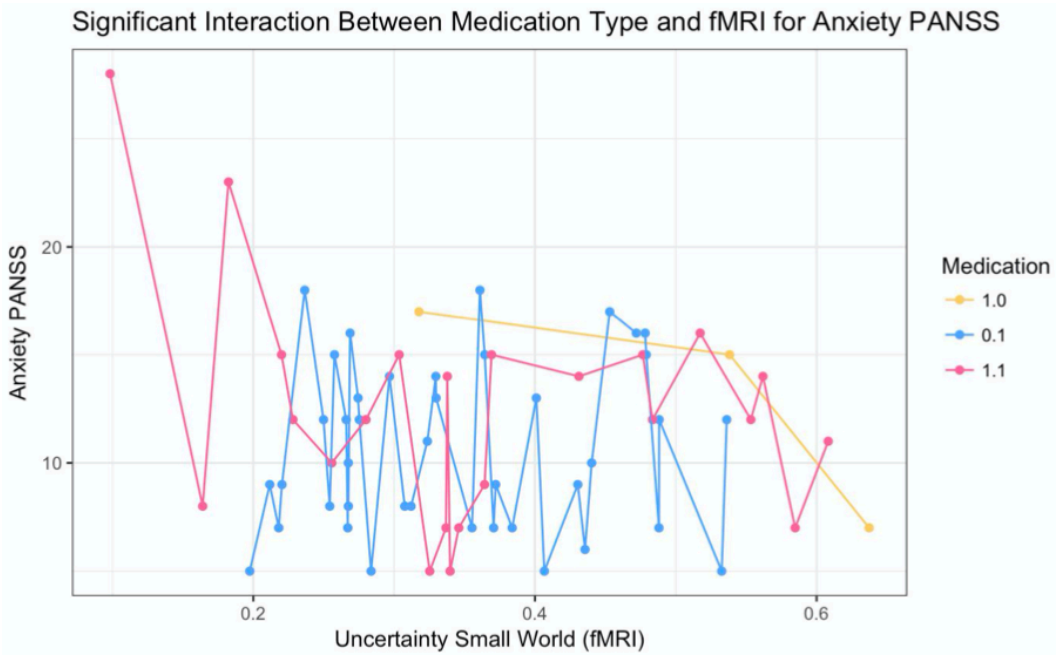

Figure 5.10 Interaction Between Medication and Uncertainty Small World for Anxiety PANSS

The effect of uncertainty small world on anxiety PANSS is not similar for patients taking different medication types. When uncertainty small world is smaller than 0.25, patients taking both-generation medications score higher on anxiety PANSS. When uncertainty small world is between 0.25 to 0.48, patients taking 2nd-generation medications generally score higher on anxiety PANSS, although there are three exceptions. When uncertainty small world is above 0.48, patients taking both-generation medications score higher again.

CHAPTER 6

Conclusion and Recommendation

6.1 Conclusion

The objective of this paper is to train multiple supervised learning algorithms to explore the relationship between brain patterns and symptoms of schizophrenia and the role different types of antipsychotic medication play in this.

For the COBRE study, we found that sensory hand fronto-parietal, sensory mouth front parietal, sensory mouth auditory, sensory mouth dorsal attention, sensory hand dorsal attention, and cingulo-opercular task control global efficiency have significant effects on PANSS scores. Specifically, keeping all else constant, higher values of sensory mouth fronto-parietal, or sensory mouth auditory, or sensory hand dorsal attention will result in more severe symptoms of schizophrenia. On the other hand, keeping all else constant, higher values of sensory hand fronto-parietal or sensory mouth dorsal attention or cingulo-opercular task control global efficiency will result in less severe symptoms of schizophrenia.

For the CNP study, none of the brain patterns has significant effect on sums of SANS/ SAPS scores. Testing the COBRE model with CNP data shows that the COBRE model explains a good amount of variation in sums of SANS/SAPS, suggesting that our model is generalized enough to provide accurate predictions for fully independent samples.

Besides, we examined the interaction effects between fMRI and medication types (whether it is 1st generation or 2nd generation) on the five dimensions of PANSS. It turns out that the interaction effects between medication types and auditory small world, fronto-parietal task control small world, salience small world were significant. Specifically, when auditory small words or fronto-parietal task control or salience small world has small values, patients taking 2nd-generation medications tend to have more severe symptoms of schizophrenia than patients taking both-generation medications. When auditory small words or fronto-parietal task control or salience small world has large values, patients taking both-generation medications tend to have more severe symptoms of schizophrenia than patients taking 2nd-generation medications.

6.2 Recommendation

Real data with high quality usually improves accuracy because we can go for more complex models without worrying about overfitting. Therefore, the next step of analysis should first include acquiring more data, especially data on patients who are not taking any medications as there are only 2 patients without medication in our dataset.

Besides, the interaction plots in Section 5.3 showed the comparison of different medication groups on their effect on small-world fMRI features for each symptom domain. Though we could observe clear difference from most of these plots, the results were merely observational and have not been tested using statistical methods. To get more powerful result, further statistical tests should be run.

REFERENCES

[1] Schizophrenia. (2020). *En.Wikipedia.Org.* https://en.wikipedia.org/wiki/Schizophrenia.

[2] S.R. Kay, A. Fiszbein, L.A. Opler. (1987). The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophrenia Bulletin, 13.* p. 261.

[3] Anderson, A. E., Mansolf, M., Reise, S. P., Savitz, A., Salvadore, G., Li, Q., and Bilder, R. M. (2017). Measuring pathology using the PANSS across diagnoses: Inconsistency of the positive symptom domain across schizophrenia, schizoaffective, and bipolar disorder. *Psychiatry Research, 258.* 207-216. https://doi.org/10.1016/j.psychres.2017.08.009.

[4] S. Teng, P. S. Wang, Y. L. Liao, T.-C. Yeh, T.-P. Su, J. C. Hsieh, Y. T. Wu (2009). Small-world Network for Investigating Functional Connectivity in Bipolar Disorder: A Functional Magnetic Images (fMRI) Study. *13th International Conference on Biomedical Engineering.* IFMBE Proceedings, vol 23. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-92841-6_178.

[5] Abou-Setta, A. M., Mousavi, S., Spooner, C., Schouten, J. R., Pasichnyk, D., Armijo-Olivo, S., ... Hartling, L. (2012). First-generation versus second-generation antipsychotics in adults: comparative effectiveness. *Comparative Effectiveness Review, 63.* https://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0049165/.

[6] Albon, Chris. (2017). Feature Selection Using Random Forest. *Chrisalbon.com.* https://chrisalbon.com/machine_learning/trees_and_forests/feature_selection_using_random_forest/.

[7] Dubey, Akash. (2018). Feature Selection Using Random Forest. *Towardsdatascience.com.* https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f.

[8] Olkin, I.; Sampson, A. R. (2001-01-01), Smelser, Neil J.; Baltes, Paul B. (eds.). Multivariate Analysis: Overview. *International Encyclopedia of the Social & Behavioral Sciences.* Pergamon, pp. 10240-10247, ISBN 9780080430768.

[9] Multivariate Statistics. (2020). *En.Wikipedia.Org.* https://en.wikipedia.org/wiki/Multivariate_statistics.

[10] Bernau, C., Riester, M., Boulesteix, A., Parmigiani, G., Huttenhower, C., Waldron, L., and Trippa L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics, Volume 30.* https://doi.org/10.1093/bioinformatics/btu279.