STATISTICAL DISEASE DETECTION WITH RESTING STATE FUNCTIONAL MAGNETIC RESONANCE IMAGING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

EBRU ÖZTÜRK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

SEPTEMBER 2017

Approval of the thesis:

STATISTICAL DISEASE DETECTION WITH RESTING STATE FUNCTIONAL MAGNETIC RESONANCE IMAGING

submitted by **EBRU ÖZTÜRK** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Ayşen D. Akkaya _____
Head of Department, **Statistics**

Assoc. Prof. Dr. Özlem İlk Dağ _____
Supervisor, **Department of Statistics, METU**

Examining Committee Members:

Assoc. Prof. Dr. Ceylan Yozgatlıgil
_____

Assoc. Prof. Dr. Özlem İlk Dağ
_____

Prof. Dr. Ergun Karaağaoğlu
_____

Prof. Dr. Serpil Aktaş Altunay
_____

Assoc. Prof. Dr. Sevtap Selçuk-Kestel
_____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: EBRU ÖZTÜRK

Signature :

ABSTRACT

STATISTICAL DISEASE DETECTION WITH RESTING STATE FUNCTIONAL MAGNETIC RESONANCE IMAGING

ÖZTÜRK, Ebru
M.S., Department of Statistics
Supervisor : Assoc. Prof. Dr. Özlem İlk Dağ

September 2017, 49 pages

Most of the functional magnetic resonance imaging (fMRI) data are based on a particular task. The fMRI data are obtained while the subject performs a task. Yet, it's obvious that the brain is active even when the subject is not performing a task. Resting state fMRI (R-fMRI) is a comparatively new and popular technique for assessing regional interactions when a subject is not performing a task. This study focuses on classifying subjects as healthy or diseased with the diagnosis of schizophrenia by analyzing R-fMRI data. The resting state situation in the dataset of "UCLA Consortium for Neuropsychiatric Phonemics LA5c Study" is used to extract brain signals in the Region of Interest (ROI) analysis. The default mode network (DMN) ROIs were selected since the DMN is a perception depending on an interconnected set of areas displaying higher activity during rest than task related activity (Raichle and Snyder, 2007). Pre-processing of fMRI images is achieved with toolbox of Statistical Parametric Mapping version 8 (SPM8). ROI-based on brain signals are obtained from Functional Connectivity (CONN). After brain signals are obtained, the disease status is predicted by adjusting for the magnitude of brain signals, the demographic information's of subjects such as gender and age. Logistic regression model, marginal model, random effect model and k-means clustering, hierarchical clustering and clustering genes with replications (CGR) followed by logistic regression approaches are conducted to classify the subjects in the UCLA data set by using R-Studio. Marginal model with smoking status and k-means clustering algorithm followed with logistic

regression model excluding smoking status give best results.

Keywords: Clustering, R-fMRI, Statistical Models, Statistical Disease Detection

ÖZ

DİNLENİM DURUMU FONKSİYONEL MAGNETİK REZONANS GÖRÜNTÜLEME İLE HASTALIK DURUMUNUN İSTATİSTİKSEL OLARAK BELİRLENMESİ

ÖZTÜRK, Ebru
Yüksek Lisans, İstatistik Bölümü
Tez Yöneticisi : Doç. Dr. Özlem İlk Dağ

Eylül 2017 , 49 sayfa

Fonksiyonel manyetik rezonans görüntüleme (fMRI) verilerinin çoğu belirli bir görevi temel alır. FMRI verileri, denek bir görevi yerine getirirken elde edilir. Ancak, denek bir görevi yerine getirmediği zamanlarda bile beynin aktif olduğu açıktır. Dinlenim durumu fMRI (R-fMRI), bir denek görev yapmadığında bölgesel etkileşimleri değerlendirmek için kullanılan yeni ve popüler bir tekniktir. Bu çalışma, R-fMRI verilerini analiz ederek, çalışmaya katılan bireyleri sağlıklı veya hasta olarak sınıflandırmaya odaklanmaktadır. "UCLA Consortium for Neuropsychiatric Phonemics LA5c Study" veri setindeki dinlenim durumu, işlevsel bağlantısallık (ROI) beyin sinyallerinin çıkarılması için kullanılmıştır. Varsayılan mod modeli ağları (DMN) dinlenim süresince görevle ilişkili etkinlikten daha yüksek aktivite göstermektedir (Raichle ve Snyder, 2007). Bu nedenle bu tezde DMN ağları üzerine çalıştık. fMRI görüntülerinin ön işleme tabi tutulması, Statistical Parametric Mapping sürüm 8 araç kutusu (SPM8) ile gerçekleştirilmiştir. Beyin sinyalleri Functional Connectivity araç kutusu kullanılarak (CONN) elde edilmiştir. Beyin sinyallerinin elde edilmesinden sonra, beyin sinyalleri, cinsiyet ve yaş gibi kişilerin demografik bilgileri ile birlikte kullanılarak hastalık durumu sınıflandırılmıştır. UCLA veri setinde bulunan denekleri R-Studio kullanarak sınıflandırmak için lojistik regresyon modeli, marjinal model, rastgele etki modeli ve k-ortalama, hiyerarşik ve CGR kümelemesi yöntemleriyle birlikte lojis-

tik regresyon yaklaşımı uygulanmıştır. Sigara içme durumu olan marjinal model ve k-ortalama kümeleme algoritmasını takip eden sigara içme durumunu içermeyen lojistik regresyon modeli en iyi sonucu vermektedir.

Anahtar Kelimeler: R-fMRI, İstatistiksel Modelleme, İstatistiksel Hastalık Durumu Belirleme, Kümeleme

To my lovely mommy and daddy

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my advisor Assoc. Prof. Dr. Özlem İlk Dağ for support, guidance, motivation and patience during thesis study. Without her guidance and feedbacks, this study would not have been completed. From my undergraduate studies. It has always been an honour for me to be her student.

I would like to present my grateful thanks to my examining committee members: Prof. Dr. Ergun Karağaoğlu Prof. Dr. Serpil Aktaş Altunay, Assoc. Prof. Dr. Ceylan Yozgatlıgil and Assoc. Prof. Dr. Sevtap Selçuk-Kestel for their valuable time to review my study.

I would also like to thank to my dear friend and college Elif Akça for her nice friendship and kind supports from the begining of this study. Moreover, I would thank to all members of METU Department of Statistics.

My special thanks go to my dear friend Öznur Demirel for her warm friendship, inspirational soul, supporting my dreams and being with me for twelve years.

Finally, I would like to represent to my deepest thanks to my dear family, Hacer Öztürk, İsmail Öztürk, Burcu Öztürk, Hatice Öztürk and Arif Ferah, for their unconditional love, support and patience. Without them, I could not success to finish this thesis.

This data was obtained from the OpenfMRI database. Its accession number is ds000030.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ACC	Accuracy Rate
ADHD	Attention Deficit Hyperactivity Disorder
AIC	Akaike Information Criteria
BA	Broadmann Area
BIDS	Brain Imaging Data Structure
BOLD	Blood-Oxygen-Level-Dependent
CGR	Clustering Genes with Replicates
CONN	Connectivity Toolbox
CT	Computed Tomography
DMN	Default Mode Network
fMRI	Functional Magnetic Resonance Imaging
FN	False Negative
FP	False Positive
LLP	Left Inferior Parietal Lobe
MPFC	Medial Prefrontal Cortex
MR	Magnetic Resonance
MR	Magnetic Resonance Imaging
MTL	Medial Temporal Lobe
NIH	National Institute of Health
NPV	Negative Predictive Value
PCC	Posterior Cingulate Cortex
PET	Position Emission Tomography
R-fMRI	Resting State Functional Magnetic Resonance Imaging
RLP	Right Inferior Parietal Lobe
ROI	Region of Interest
PPV	Positive Predictive Value
SPM8	Statistical Parametric Mapping version 8
Std. Error	Standard Error

TN	True Negative
TP	True Positive
TR	Time Response

CHAPTER 1

INTRODUCTION

1.1 Functional Magnetic Resonance Imaging

In spite of the fact that alterations in the neural activity in the human brain occur in near real time, its non-invasive observation is enabled by an exhilarating magnetic resonance imaging (MRI) technology, called Functional Magnetic Resonance Imaging (fMRI) that leads researchers to study the structure and function of the brain. Its impacts not only have led to the evolution in most of the neuroscience like cognitive psychology and perception, the presence of the fields (social neuroscience, developmental neuroscience, neuroeconomics, neuromarketing and the like) also proves that its influences have catalyzed the emergence of sub-disciplines within this branch of science (Ashby, 2011).

The use of fMRI as a technique for noninvasive mapping and analysis of cortical activity in human brain due to its effectiveness to track the local alterations in cerebral blood volume, blood flow and blood oxygenation depending on the increased neuronal activity have brought about a way to probe numerous matters that are once assumed as unreachable. On the basis of measuring differences in the magnetic properties of certain molecules, MRI operates to comprehend the structure and function of brain. In 1977, first human MRI has appeared, and its clinical use has been approved by The Food and Drug Administration in 1985. In the following ten years, the number of the MRI instruments installed across the United States has reached the thousands. As it is completely noninvasive and poses less health risk, MRI have become one of the routine medical procedures for diagnostic and scientific purposes. In comparison with Computed Tomography (CT) that produces images by using x-rays and Positron

Emission Tomography (PET) scanning that is applied through the injection of a radioactive drug, MRI can said to be a substantial innovation in neuroimaging (Ashby, 2011).

The observation of how human brain functions in real time with high spatial resolution would be the ideal way of how fMRI works. However, the fact that a typical fMRI experiment that reports a slow and indirect measurement with a temporal resolution of 1–3 seconds and a spatial resolution of 3–5 mm^3 is the evidence that the methodology that is currently used does not fulfill its ideal requirements. As stated by the numerous publications on fMRI, in spite of its ineffectuality to reach its theoretical objectives, the great impacts of fMRI on the study of mind and brain cannot be denied (Ashby, 2011).

It is postulated that the oxygen consumption in active areas of brain is more than in inactive areas. Therefore, most of the fMRI studies measure blood oxygen level-dependent (BOLD) signal. The reason behind is that oxygenated hemoglobin is transported to the area in which the increased neural activity ascending the metabolic demand observed. The low BOLD signal level that is observed due to the increased metabolic demand, rises after the oxygenated hemoglobin molecules are transferred to the active brain area (Ashby, 2011).

1.2 Resting State

Recently, resting state fMRI (R-fMRI) method has been developed, which enables the observation of regional interactions among different brain regions to be performed without any involvement of the subject in external task. At state of rest, significant correlations between distant grey matter regions are observed through low-frequency (<0.1 Hz) BOLD fluctuations (Lee et. al 2013). Although dynamics behind the neural fluctuations in brain is still ambiguous, it is assumed that it is BOLD fluctuations which cause the fluctuations in spontaneous neural activity. The spatial patterns of R-fMRI correlations show resemblance with the correlations observed in the states such as eyes-open, eyes-closed and fixation. As it is not task dependent, R-fMRI facilitates experimental design, subject compliance, and training demands and is preferable for studies of development and clinical populations(Components of the Human Connec-

tome Project,2017).

1.3 Default Mode Network

Brain's Default Mode Network (DMN) is deemed to be a group of brain structure that display their highest level activity during resting state (Raichle et al., 2001). In the absence of attention demanding tasks, brain adjusts itself to a default mode of stimulus-independent thought in which subject's brain is stimulated by internally focused tasks consisting of autobiographical memories, future envisioning, personal introspection and the like (Buckner et al., 2008). Regarding that it supports self-referential mental activities, Default Mode Network introduces a new aspect towards the spontaneous human thoughts and feelings, and how they are formed in mental disorders such as Alzheimer's disease, autism, schizophrenia.

The Default Mode Network consists of interconnected brain regions functioning with a strong temporal synchrony which are identified as the medial prefrontal cortex (MPFC) [Brodmann area (BA) 10, 24, 32], posterior cingulate cortex (PCC) (BA 29/30, 23/31) and left and right inferior parietal lobules (LLP and RLP) (BA 39, 40). Brodmann areas are illustrated in Figure 1.1. In addition to that core regions, there are additional ones that can be considered as a part of the DMN, such as the medial temporal lobe (MTL) (Buckner et al., 2008).

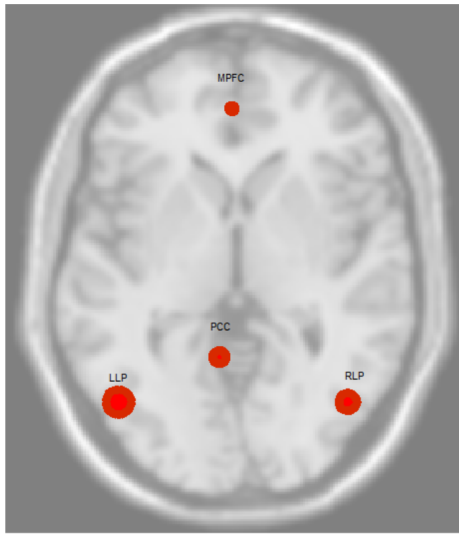


Figure 1.1: Default Mode Networks Regions

1.4 The Focus of Thesis

The motivation of this study is diagnosing subjects as patient or healthy by using statistical methods. Moreover, we shorten time for modeling data set and diagnosing subjects by keeping them less in the MRI device. This gives time and financial efficiency to researchers and patients. The main focus of this thesis is to compare several statistical methods for classifying patients diagnosed as schizophrenia and healthy subjects. Some of these methods just use covariates, while others include longitudinal BOLD signals as well. We aim to compare and show the contribution of BOLD signals with covariates against modeling with only covariates.

In order to achieve the aim of this thesis,

- Logistic regression model is conducted with the covariates (such as age, gender, religion...), but without BOLD signals,
- Marginal models are conducted with the covariates together with BOLD signals,
- Random effects model is conducted with the covariates and BOLD signals,
- Clustering approaches are conducted on BOLD signals; then by using the cluster information of each subject, logistic regression model is conducted.

The plan of this thesis is as follows: In Chapter 2, information about the data set is given. In Chapter 3,the methods and performance measures are discussed. In Chapter 4, the results are demonstrated. Finally, the discussion and conclusion follow in Chapter 5.

CHAPTER 2

DATA SET

2.1 Data Set

Both healthy individuals (138 subjects) and individuals with neuropsychiatric disorders including schizophrenia (58 subjects), bipolar disorder (49 subjects), and attention deficit/hyperactivity disorder (45 subjects) have been selected as shared neuroimaging dataset by UCLA Consortium for Neuropsychiatric Phenomics, which is a multidisciplinary team concentrating on the dimensional structure of memory and cognitive control (response inhibition) functions.
A set of task-based fMRI assessments, resting fMRI, structural MRI, and high angular resolution diffusion MRI are included in the dataset which is shared on OpenfMRI project, and put in a format based on the Brain Imaging Data Structure (BIDS) standard (Poldrack, et al., 2016).
In this thesis, we study only with healthy subjects and schizophrenia subjects in the resting situation status (Poldrack, et al., 2016).

2.2 Participants

For the healthy adults, ages between 21-50, community advertisements from the Los Angeles area; for the adults with ADHD, Bipolar and Schizophrenia, a patient-oriented strategy including local clinics and online portals have been used to determine subjects. Each candidate have participated in a telephone interview prior to an in-person interview. They were asked to identify themselves in either one of the

NIH ethnic and racial categories: White but not Hispanic or Latino; or Hispanic or Latino. Participants were also asked whether they fulfil the following inclusion criteria: Primary language (as determined by verbal-fluency tests in both languages) either English or Spanish; completed at least 8 years of formal education; no significant medical illness by self-report; sufficiently willing to finish assessments; and visual acuity not worse than 20/60. Moreover, urinalysis was checked to detect possible drug abuse. The subjects determined as drug abuse were excluded from the study (Poldrack, et al., 2016).

2.3 MRI Data Acquisition

In this part, we give the technical features of the MR devices used in the experiment for the preprocessing of the fMRI data since these features gain importance. For MRI data acquisition, one of two 3T Siemens Trio scanners from Ahmanson Lovelace Brain Mapping Center (Siemens version syngo MR B15) and the Staglin Center for Cognitive Neuroscience (Siemens version syngo MR B17) at UCLA are. Data are obtained with the use of a $T2^*$weighted echoplanar imaging (EPI) sequence with the following parameters: slice thickness = 4 mm, 34 slices, TR = 2 s, TE = 30 ms, flip angle = 90°, matrix 64 x 64, FOV = 192 mm, oblique slice orientation, in addition to a $T2^*$ weighted matched bandwidth high resolution anatomical scan with the parameters of 4mm slices, TR/TE=5000/34 ms, 4 averages, matrix = 128x128, 90 ° flip angle (with the same slice prescription as the fMRI scan) and MPRAGE with the parameters of TR = 1.9 s, TE = 2.26 ms, FOV = 250 mm, matrix = 256 x 256, sagittal plane, slice thickness = 1 mm, 176 slices. To collect the diffusion weighted imaging (DWI) data, an echoplanar sequence with the parameters of 64 directions, 2mm slices, TR/TE=9000/93 ms, 1 average, 96x96 matrix, 90° flip angle, axial slices, b=1000 s/mm^2, is used (Poldrack, et al., 2016).

In the resting fMRI, it was required from the participants to try to stay relaxed with open eyes, not show any stimulation or respond for 304 seconds. To get the BOLD signals, we used the Statistical Parametric Mapping software version 8 (SPM8). First of all, we conducted the pre-process of fMRI data, including slice timing correction, motion correction, coregistration and normalization and smoothing steps. Then, we

used in the analysis of Voxel-based and ROI-based correlation the Functional Connectivity (CONN) toolbox of SPM8.

2.4 Covariates

In the data set, the information recorded from all subjects are race, gender, smoking status, civil status, education level of subject's mother (in years), religion, education level of subject's father (in years) , education level of subject. The number of observations is low on some levels of the categorical variables. Therefore, some levels are merged before the analysis. In order to clarify the data set, we share the coding system of categoric variables in Table 2.1. The detailed information is in the Appendix.

Table 2.1: Labels of Categoric Variables

Variable	Value	Label
Race	1	White
	2	Other
Smoking Status	1	No
	2	Yes, current
	3	Yes, past
Civil Status	1	Married
	2	Separated/Divorced
	3	Never Married
Religion	1	Christian
	2	Not Affiliated
	3	Other
Ethnicity	1	Hispanic
	2	Other
Gender	1	Female
	2	Male

In Table 2.2 we give a portion of the data set which combined BOLD signals and demographic covariates. In Table 2.2, smoking status, civil status, education level of subject's mother, education level of subject's father, education level of subject, children number and diagnosis are represented with smoke, civil, sc_mother, sc_father, sc, child_number and diag, respectively.

Table 2.2: Portion of the Data Set

id	Time	race	gender	smoke	civil	sc_mother	religion	sc_father	sc	age	child_num	ethnicity	diag	MPFC	PCC	LLP	RLP
1	1	2	2	1	3	19	3	20	16	30	0	2	Control	-0.042	0.172	0.023	-0.267
1	2	2	2	1	3	19	3	20	16	30	0	2	Control	-0.107	0.071	-0.211	-0.068
1	3	2	2	1	3	19	3	20	16	30	0	2	Control	-0.132	0.007	-0.422	0.037
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	144	2	2	1	3	19	3	20	16	30	0	2	Control	-0.516	-0.038	-0.134	-0.171
1	145	2	2	1	3	19	3	20	16	30	0	2	Control	-0.737	-0.042	0.404	-0.645
1	146	2	2	1	3	19	3	20	16	30	0	2	Control	-0.903	-0.030	0.823	0.297
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
171	1	1	1	1	3	13	1	12.5	12	25	0	1	Scz	-0.079	-0.024	-0.045	-0.075
171	2	1	1	1	3	13	1	12.5	12	25	0	1	Scz	-0.149	0.007	-0.050	-0.019
171	3	1	1	1	3	13	1	12.5	12	25	0	1	Scz	-0.207	0.0359	-0.069	0.009
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
171	144	1	1	1	3	13	1	12.5	12	25	0	1	Scz	0.144	0.150	-0.086	0.143
171	145	1	1	1	3	13	1	12.5	12	25	0	1	Scz	-0.169	0.375	0.307	0.313
171	146	1	1	1	3	13	1	12.5	12	25	0	1	Scz	-0.426	0.538	0.553	0.384

CHAPTER 3

METHODOLOGY

In this chapter the methods that are used in this thesis namely logistic regression, marginal model, random effects model and clustering algorithms are introduced. Next, we touch briefly on the performance measures.

3.1 Preprocessing of fMRI Data

It is significant to reduce the influence of data acquisition and physiological artifacts to a minimum, to verify statistical assumptions, and to determine the standardized locations of brain regions, which are obtained from different subjects to reach increased validity and sensitivity in statistical analysis of fMRI data.To satisfy these objectives, basically for the elimination of artifacts and validation of the following model assumptions, fMRI data has to be preprocessed. Analysis of fMRI data essentially is based on the assumptions of that all voxels in a specific brain volume are concurrently procured, that each data point in a specific voxel's time series only consists of a signal from that voxel (i.e., that the participant did not move in between measurements), and that all individual brains are registered to locate each voxel in the same anatomical region (Lindquist,2008).

3.1.1 Slice-Time Correction

For the fMRI data analysis, it is important to acquire 3D fMRI data by eliminating the temporal offset occurred between slices while measuring the whole brain. Although

it is assumed that the brain slices are concurrently measured in the data acquisition, they are sampled sequentially at different time points. Therefore, the time difference has to be compensated by temporally shifting the similar time courses from different slices. Basically, slice timing correction is one of preprocessing fMRI data steps to correct the temporal offsets between slices which uses either interpolation or Fourier shift theorem (Lindquist,2008).

3.1.2 Motion Correction

For any fMRI study, it is essential to tackle subject movement, which can occur in data acquisition process and result in errors in imaging, in the most proper way possible. In the case of movement, the image that is rendered from a signal obtained from a particular voxel will be disturbed due to the signal coming from neighboring voxel. Moreover, an accurate estimation of the degree of motion and correction of the images are critical. The first action to be taken regarding the motion correction is to provide the best match between the input image and some target image (e.g., the first image or the mean image), which aims at matching the input image to target image. We used mean image as target image in this research. Through the use of a rigid transformation that involves six parameters (three translations, three rotations) , the input image can be translated (shifted in the x, y and z directions) and rotated (altered roll, pitch and yaw). The minimization of the cost functions (e.g., sums of squared differences), as a tool to evaluate the resemblance between the two images, enables the matching process to be practiced properly. For the acquisition of the corrected voxel values, the interpolation is used to resample the image when the parameters reach their optimal realignment. To complete the study, motion correction procedure has to be iterated for each brain volume. If the degree of motion is high, the subject should be removed from study (Lindquist, 2008).

3.1.3 Coregistiration and Normalization

Since fMRI data has low spatial resolution that results in illustrating less anatomical detail, it is practical to project the results of functional data on high resolution struc-

tural MR image. Coregistration as one of the steps for the preprocessing fMRI data is used for providing the alignment between structural and functional images, which uses either a rigid body (6 parameters) or an affine (12 parameters) transformation. Although each individual brain differs from the other in shape and feature, it is critical for a group analysis to consider each voxel within the same brain structure. Therefore, normalization in the preprocessing of data analysis is a way to register the anatomy of each individual to a standardized stereotaxic space defined by a template brain (e.g., the Talairach or Montreal Neurological Institute (MNI) brain). Since each individual brain has an inherent structure, the use of a rigid body transformation would be unsuitable for normalization procedure. Hence, to match the local features, nonlinear transformations are commonly used. Normalization procedure is based on a high resolution image warped onto a template image. To that end, a smooth continuous map is constructed between the points in an input image and the ones in the target image, which provides a normalized image comparing to the similarly normalized images from other subjects. Due to the reduction in spatial resolution of the images, errors can be encountered in normalization procedure. Yet, it enables to report and interpret the spatial locations in a consistent manner, and to generalize the results for a population with a greater number (Lindquist, 2008).

3.1.4 Spatial Smoothing

In addition to the preprocessing steps for fMRI data given above, spatial smoothing is a widely used process in which a Gaussian kernel is employed to convolve the functional images. Gaussian kernel is defined by the full width of the kernel at half its maximum height (FWHM) with the common values varying between 4–12 mm FWHM.

The common use of spatial smoothing of fMRI data is based on various reasons as enabling to improve inter-subject registration and to overcome the limitations observed in the spatial normalization, of ensuring that the random field theory assumptions are validated, which are facilitated for the correction of the multiple comparisons and require a FWHM 3 times the voxel size for their validation (e.g., 9 mm for 3 mm voxels). Furthermore, it provides a reduction in the noise in individual voxels and

incline in the signal-to-noise ratio within the region (Lindquist, 2008). In this study, we used 6mm Gaussian kernels for spatial smoothing.

3.1.5 Functional Connectivity Toolbox

The functional connectivity is used to demonstrate the correlations between spatially different brain regions both in resting state and task dependent experiments. In this experiment, the subjects were needed to stay in MR for 304 seconds. TR=2 implies that the image and signal was record for every 2 seconds. Therefore, 152 images were recorded in this fMRI experiment. Due to the irregularity of the MR signals in the initial of the experiment, the first 6 TR were deleted. After the connectivity step was implemented, consequently, in the data set, BOLD signals were recorded for 146 time points.

3.2 Models

In the subsection of 3.2 the models are introduced shortly. The formulations and the logic behind these models are given. For obtaining BOLD signals, we use MATLAB (2012) version R2012b, SPM8 and CONN (Whitfield-Gabrieli and Nieto-Castanon,2012) toolboxes. For modeling part, we use RStudio (2017) version 1.0.136.

3.2.1 Logistic Regression

Regression analysis is used to examine and model the relationship between a response variable and explanatory variables. For the logistic regression analysis, the response variable is categorical; it can take two or more possible values. In this thesis, the response variable is the disease status, whether the subject is a healthy subject or diagnosed with schizophrenia. Therefore, the response variable can take either 0 or 1 values,which indicates response variable Y has a Bernoulli distribution with parameter π. Generally, the logit function is used as a link function in logistic regression. Logit function is based on natural logarithm of odds. Odds represent the probability of something happening to not happening. Mathematically, odds is the ratio of

$P(Y=1|X)$ to $P(Y=0|X)$. The general form of binary logistic regression can be written as the following:

$$log(\frac{P(Y_i=1|X_i)}{P(Y_i=0|X_i)}) = \beta_0+\beta_1 x_{i1}+...+\beta_k x_{ik}, \quad i=1,2,3,...,n \quad \text{and} \quad k=1,2,3,...,p. \tag{3.1}$$

The Equation 3.1 leads to the following probability:

$$P(Y_i=1|X_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+...+\beta_k x_{ik}}}{1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_k x_{ik}}}, \quad i=1,2,3,...,n \quad \text{and} \quad k=1,2,3,...,p. \tag{3.2}$$

The most frequent approach for the parameter estimation is based on Maximum Likelihood Estimation. The interpretation of the coefficients rely on estimated odds ratio. For instance, in Equation 3.3, the odds ratio formulation is written for the x covariate which have two categories such as 1 and 2. Equation 3.3 gives the change in odds when the covariate x changes by one unit:

$$\hat{OR} = \frac{\frac{P(Y=1|X=2)}{P(Y=0|X=2)}}{\frac{P(Y=1|X=1)}{P(Y=0|X=1)}} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1(x=2)}}{e^{\hat{\beta}_0+\hat{\beta}_1(x=1)}} = e^{\hat{\beta}_1((x=2)-(x=1))} = e^{\hat{\beta}_1}. \tag{3.3}$$

The logistic regression is commonly used in biological sciences especially in epidemiology, for instance to find the relationship between disease and the influencing factors of the disease. In logistic regression by using the cut off point for the probability, one can construct a classification table. As a result, the success of models can be decided via comparing with model selection criteria. The details about logistic regression can be found in Montgomery et al. (2012). For logistic regression, we use "stats" package (R Core Team, 2016) in RStudio.

3.2.2 Marginal Models

These models are also known as "population-average models". Suppose Y_{ij} be the binary response (whether the subject is diagnosed with schizophrenia or healthy) at time *t*, for the i^{th} subject. In the marginal models, the marginal expectation, E(Y_{ij}), of the response, which can be continuous or binary in general, is modelled via the explanatory variables. Marginal expectation implies the average response over the subpopulation that has common value of explanatory variables. The coefficients of the marginal models, β, are interpreted like coefficients in cross-sectional analysis.

However, the within subject correlation is taken into account in the marginal models. Moreover, just like in logistic regression, in the marginal models, the link functions are used such as logit function for binary response. The formulation of the marginal model is given below.

$$\begin{aligned} log(\frac{P(Y_i=1|X_i)}{P(Y_i=0|X_i)}) &= \beta_0 + \beta_1 x_{it1} + ... + \beta_k x_{itk}, \\ i = 1,2,3,...,n, \quad t = 1,2,3,...,T \quad &\text{and} \quad k = 1,2,3,...,p. \end{aligned} \tag{3.4}$$

Equation 3.4 leads to

$$P(Y_i = 1|X_i) = \frac{e^{\beta_0+\beta_1 x_{it1}+...+\beta_k x_{itk}}}{1 + e^{\beta_0+\beta_1 x_{it1}+...+\beta_k x_{itk}}}. \tag{3.5}$$

In this study, some of these covariates involve demographic information, others are longitudinal BOLD signals. In Equation 3.6, we present marginal model formulation adapted to the data set that we used.

$$\begin{aligned} log(\frac{P(Y_i=1|X_i)}{P(Y_i=0|X_i)}) &= \beta_0 + \beta_1 Age_i + ... + \beta_k BOLD_{it}, \\ i = 1,2,3,...,n, \quad &\text{and} \quad t = 1,2,3,...,T \end{aligned} \tag{3.6}$$

Since data for subject i is collected in a very short time period, age variable, for instance, is not changing over time. Hence, we drop the index t for this variable in Equation 3.6. The detailed information about marginal models can be obtained from Diggle et al. (2013). For marginal models, we use "gee" (Carey et al., 2015) package in RStudio.

3.2.3 Random Effects Models

They are also known as "subject-specific" models. In the random effects models, the regression coefficients might vary from one subject to another. Therefore, in these models, the individual heterogeneity is taken into account.

$$\begin{aligned} log(\frac{P(Y_i=1|X_i)}{P(Y_i=0|X_i)}) &= \beta_0 + \beta_1 x_{it} + ... + \beta_k x_{itk} + U_i, \\ i = 1,2,3,...,n, \quad t = 1,2,3,...,T \quad &\text{and} \quad k = 1,2,3,...,p. \end{aligned} \tag{3.7}$$

Equation 3.7 leads to

$$P(Y_i = 1|X_i) = \frac{e^{\beta_0+\beta_1 x_{it}+...+\beta_k x_{itk}+U_i}}{1 + e^{\beta_0+\beta_1 x_{it}+...+\beta_k x_{itk}+U_i}}. \tag{3.8}$$

In Equations 3.7 and 3.8, U_i is known as the random intercept which is allowed to be different for each subject. These intercepts are assumed to be independent and follow normal distribution with mean 0 and variance D. This implies that the expected response is just like the one in the marginal model. However, the variance is different for each subject. This variance assures the individual heterogeneity for each subject. The details of random effects model can be found in Diggle et al. (2013). For random effects model, we use "lme4" package (Bates et al., 2015) in RStudio.

3.2.4 Logistic Regression with Clustering Algorithms

In this method, first, the BOLD signals are clustered by using clustering algorithms, namely k-means, hierarchical and CGR (Cinar et al., 2017). After the cluster of each subject is identified, by adding the cluster information as categorical variable to the demographic information, the logistic regression model is conducted again.

3.2.4.1 k-means Algorithm

The k-means algorithm is one of the most popular and widely used partitioning methods. In k-means algorithm, first input parameter k is specified. Then, the objects in the data set, say n, is divided to k clusters with respect to high similarity within clusters and low similarity between clusters. Cluster similarity is evaluated according to the mean value of the objects in a cluster that can be considered as the center of the cluster. The k-means algorithm works as follows:

1. Choose randomly k of the objects, which primarily presents a cluster mean
2. Assign remaining objects to the most similar cluster based on distance between the object and the cluster mean
3. Calculate new mean of each cluster
4. Return Step 1 and repeat each step until convergence is satisfied

The detailed information about clustering can be found in Han and Kamber (2006). For k-means clustering, we use "stats" package (R Core Team, 2016) in RStudio.

3.2.4.2 Hierarchical Clustering

A hierarchical method generates a hierarchical subdivision of given data set. Agglomerative and divisive are types of hierarchical clustering. Agglomerative hierarchical clustering is based on bottom-up strategy. It begins by laying its own object to its own cluster and then combines these small clusters into larger and larger clusters. It stops when all objects are in one cluster or final termination condition is satisfied. Divisive hierarchical clustering is based on top-down strategy. It does the opposite of agglomerative hierarchical clustering. First, all objects are in one cluster then divided into smaller and smaller clusters until final termination condition is fulfilled. In this study we use agglomerative strategy. Moreover, the detailed information is in the Han and Kamber(2006). We also use "stats" package (R Core Team, 2016) in RStudio for hierarchical clustering.

3.2.4.3 CGR Algorithm

In the CGR algorithm, we aim to cluster BOLD signals considering similarities between their behavior through time. Due to time and behavior, two distance measures are used. The first one is to detect magnitude differences in signals. The second one is to capture the differences in trends. Moreover, CGR algorithm is also use hierarchical clustering with agglomerative strategy. The details about this algorithm can be found in Cinar at al. (2017). For clustering, we use "cgr" package (Cinar et al., 2015) in R-Studio.

3.3 Model Selection Criteria

In the subsection of 3.2, the performance measures are given to compare models. The cross-validation is applied to investigate how the models give results when a new and independent data is given. Therefore, by cross-validation, we observe how the models classify a new data set. In this thesis, 80% of the data set is allocated for model building and 20% of the data is allocated for model prediction. The accuracy, sensitivity, specificity, positive and negative predictive values are evaluated. For per-

formance measures, we use "caret" (Max, 2008) package in RStudio. In Table 3.1 the classification table is drawn due to the response with two categories. Here, 0 and 1 indicate healthy and patient diagnosed with schizophrenia, respectively.

Table 3.1: Classification Table

		Predicted	
		0 (-)	1 (+)
Observed	0 (-)	n_{11} (TN)	n_{12} (FP)
	1 (+)	n_{21} (FN)	n_{22} (TP)

According to Table 3.1, n_{11} represents the number of subjects who are healthy and classified as healthy also by the model; n_{12} represents the number of subjects who are healthy but classified as diseased by the model; n_{21} represents the number of subjects who are diseased but classified as healthy by the model; n_{22} represents the number of subjects who are diseased and classified also as diseased by the model. The cells in Table 3.1, n_{11}, n_{12}, n_{21} and n_{22} are also known as "True Negative (TN)", "False Positive (FP)", "False Negative (FN)" and "True Positive (TP)", respectively.

3.3.1 Accuracy Rate

Accuracy rate (ACC) gives the true classification rate. The formulation of accuracy rate based on Table 3.1 is given in Equation 3.9. According to the this thesis, accuracy rate gives the proportion of correctly classified healthy subjects and patients.

$$ACC = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} = \frac{TP + TN}{TN + FP + FN + TP}. \tag{3.9}$$

3.3.2 Sensitivity

Sensitivity is known as true positive rate, which is basically the rate of number of true positive predictions to the total number of diseased subjects. In this thesis, sensitivity is the proportion of the number of correctly classified diseased subjects diagnosed with schizophrenia to the total number of those diseased subjects. The formulation can be represented as in Equation 3.10.

$$Sensitivity = \frac{n_{22}}{n_{21} + n_{22}} = \frac{TP}{TP + FN}. \tag{3.10}$$

3.3.3 Specificity

Specificity is known as true negative rate, which means the rate of number of true negative predictions to the total number of non-diseased subjects. Specificity is the proportion of the number of correctly classified healthy subjects to the total number of subjects observed as healthy. The formulation can be represented as in Equation 3.11.

$$Specificity = \frac{n_{11}}{n_{11} + n_{12}} = \frac{TN}{TN + FP}. \tag{3.11}$$

3.3.4 Positive Predictive Value

Positive predictive value (PPV) is the proportion of diseased subjects among all of those predicted as diseased. In other words, PPV is the proportion of diseased subjects to the total number of diseased subjects predicted by the model. The formulation can be represented as in Equation 3.12.

$$PPV = \frac{n_{22}}{n_{22} + n_{12}} = \frac{TP}{TP + FP}. \tag{3.12}$$

3.3.5 Negative Predictive Value

Negative predictive value (NPV) is the proportion of healthy subjects to the total number of healthy subjects predicted by the model. The formulation can be represented as in Equation 3.13.

$$NPV = \frac{n_{11}}{n_{11} + n_{21}} = \frac{TN}{TN + FN}. \tag{3.13}$$

3.3.6 Akaike Information Criteria (AIC)

The most common model selection criteria is AIC. AIC is based on likelihood theory to estimate Kullback-Leibler (KL) distance between the real and candidate model. The aim of AIC is minimizing this distance as well as holding with simple model. In AIC, if the number of parameters to estimate increases, then bias will also increase as it can be seen in Equation 3.14.

$$AIC = -2log(L) + 2p. \tag{3.14}$$

According to Equation 3.14, p and L represent number of estimated parameters and likelihood, respectively. The lower AIC offers better model. The more theoretical detail about AIC can be found in Bozdagan (1987).

CHAPTER 4

RESULTS

In this chapter,we first present the descriptive statistics for covariates and the plots of BOLD signals in order to show the structure of the data set. After that, we discuss the results of the logistic regression, marginal model, random effects model and clustering approach. Finally, the results of the performance measures are shared.

4.1 Descriptive Statistics

As we mentioned in Chapter 2.1, there are 138 healthy subjects and 58 patients diagnosed with schizophrenia in the study. However, in this thesis, we could only include 121 healthy subjects and 50 patients due to the lack of anatomical images or resting state of some subjects. There exist missing values in 4 covariates, which are race, education level of subject's father and mother and number of children that subject has. The percentages of the missing values change between 1.17 % to 7.60 %. Since these percentages are small, we applied mode or median imputations to those variables depending on whether they are categorical or continuous variables. In Table 4.1, one can see the characteristics of the sample with respect to health condition.

Table 4.1: Descriptive Statistics of Covariates*

	Healthy Controls (n=121)	Patient Group Diagnosed with Schizophrenia (n=50)
Education level of mother (years)	14 (12-16)	12 (12-14)
Education level of father (years)	13 (12-16)	12.25 (12-16)
Education level (years)	16 (14-16)	12 (12-13.75)
Age	28 (24-39)	37.5 (29-43.75)
Number of children	0 (0-0)	0 (0-0)
Race, white (%)	76.86	70.00
Gender, female (%)	53.72	76.00
Religion (%)		
Christian	73.56	78.00
Other (Jewish, Muslim, Other)	9.08	16.00
Not affiliated	17.36	6.00
Civil Status (%)		
Never Married	72.73	90.00
Married	14.05	2.00
Separated/Divorced	13.22	8.00
Ethnicity, Hispanic (%)	35.54	58.00
Smoking		
No	74.38	40.00
Yes (current)	9.09	40.00
Yes (past)	16.53	20.00

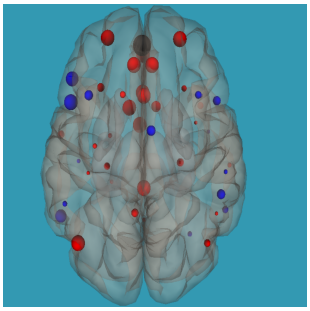
*Data represented are median (25^{th} and 75^{th} percentiles) or percentage of frequency

According to descriptive statistics, average education level of schizophrenia patients seem to be lower than those of healthy subjects. Education level of subject's father and mother also show the same pattern with education level of subject. Average age of schizophrenia patients is higher than healthy controls. Average number of children are same for schizophrenia patients and healthy controls. For both groups, the percentage of white subjects and the percentage of Christian subjects are similar. However, the percentage of female subjects, Hispanic subjects and subjects who have never been married are higher in schizophrenia patients than healthy controls. For smoking status, the percentage of subjects who never smoked is higher in healthy controls.

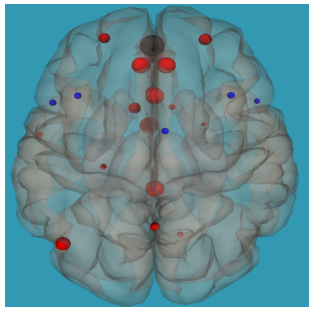
4.2 BOLD Signals

In this part we give the connectivity results of 4 DMN regions for healthy subjects and patients separately.

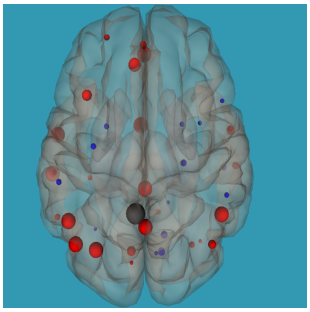
In Figure 4.1, the black points represent specific DMN regions, for instance, MPFC in panels (a) and (b) . While the red points represent positive relation of specific DMN region with the other regions, the blue points represent negative relationship. Moreover, the size of red and blue points are related to the degree of connectivity. This means the bigger the size is the higher the degree.



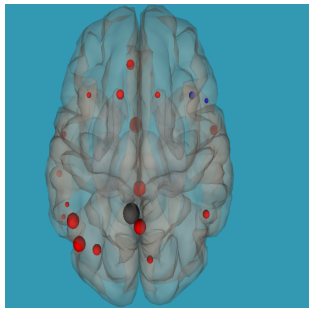
(a) Connectivity of MPFC for control group



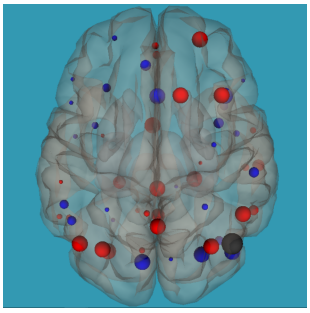
(b) Connectivity of MPFC for schizophrenia patients



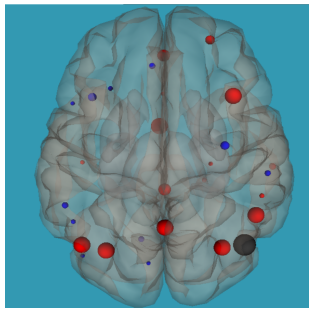
(c) Connectivity of PCC for control group



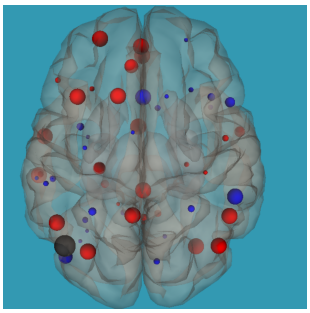
(d) Connectivity of PCC for schizophrenia patients



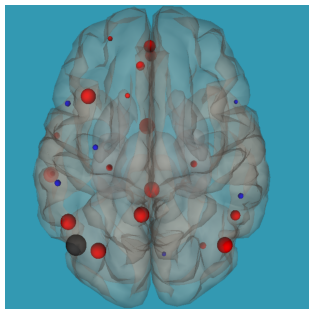
(e) Connectivity of RLP for control group



(f) Connectivity of RLP for schizophrenia patients



(g) Connectivity of LLP for control group



(h) Connectivity of LLP for schizophrenia patients

Figure 4.1: Connectivity of DMN Regions for Healthy Subjects and Schizophrenia Patients

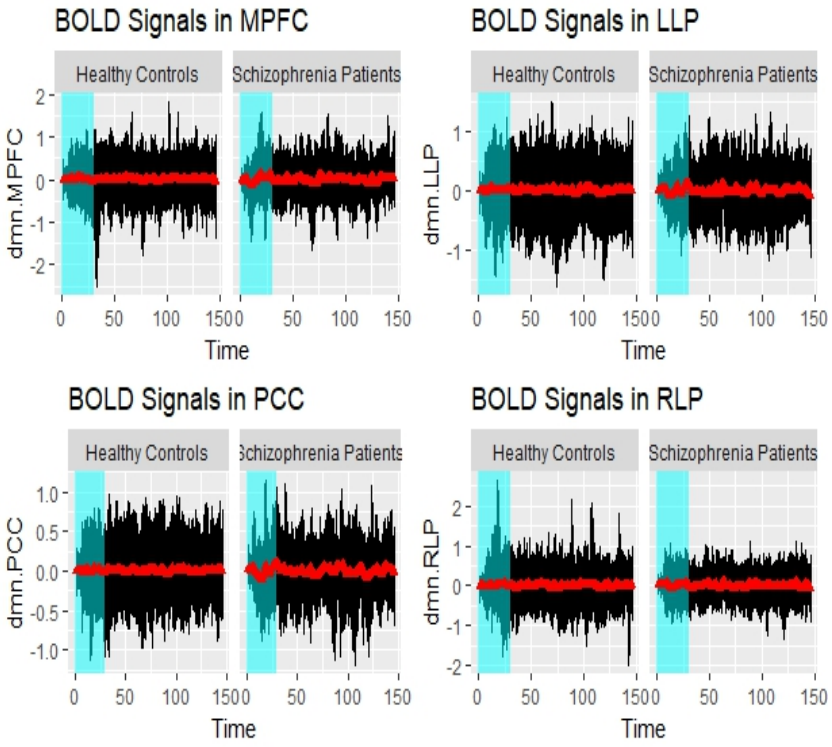


Figure 4.3: BOLD Signals for 4 DMN Regions with respect to Health Condition

In Figure 4.3, we show the BOLD signals for 4 DMN regions with respect to health conditions. We draw the plots in Figures 4.3, 4.6 and 4.7 by using "ggplot2" package (Wickham,2009) in RStudio. In these figures, red line represents the mean of the signals over time. The values of BOLD signals mostly change between -2 to 2 for healthy controls and schizophrenia patients. Although the range is narrow, when the time is taken account, the peak points and trends over time show some differences. In the analysis part, we take into account first 30 second, which is highlighted with blue in Figure 4.3 since when we model data set with 146 seconds, we face with some convergence problems. Moreover, one of the motivation of this thesis is modeling data set and diagnosing subjects by keeping them less in the MRI device. This satisfies time and financial efficiency to researchers and patients.

4.3 Results of the Models

In this subsection, we give the results of the parameter estimation from train data set. After that we demonstrate and discuss the performance measures for both train and test data sets.

4.3.1 Logistic Regression Model

In Chapter 3.2.1, we touched the logistic regression modeling briefly. For logistic regression, the response is the health condition, whether the subject is healthy or patient. The covariates of the logistic regression model is demonstrated in Table 4.1.We attempted different models starting from the full model and then applying the backward elimination and forward selection approaches. When we represent the results, we show only the best models. The only exception is about the smoking status. Smoking status is statistically significant only in the logistic regression model. For other models, the smoking status is not statistically significant. In order to see the contribution of the smoking status and to compare the models, all of the models are conducted with and without smoking.

In Table 4.1, it can be seen that race, gender, religion, civil status, ethnicity and smoking status are nominal variables. To include these variables in the models, we create dummy variables. In Table 4.2, the reference category of the nominal variables are demonstrated.

Table 4.2: List of Reference Categories

Variables	Reference Category
Race	White
Gender	Female
Religion	Christian
Civil Status	Never Married
Ethnicity	Hispanic
Smoking	No

Table 4.3: Results of Logistic Regression with Smoking Status

Variables	Estimate (β)	Std. Error	exp(β)	95% of Confidence Interval		z value	p-value
				Lower Limit	Upper Limit		
(Intercept)	8.395	2.640				3.179	0.001
smoking(yes current)	1.347	0.669	3.846	1.037	14.269	2.014	0.044
smoking(yes past)	-0.255	0.821	0.775	0.155	3.874	-0.311	0.756
civil(married)	-1.620	1.299	0.198	0.016	2.523	-1.247	0.212
civil(divorced/separated)	-4.994	1.345	0.007	0.000	0.095	-3.712	<0.001
school year	-0.875	0.197	0.417	0.283	0.613	-4.440	<0.001
ethnicity	-1.182	0.581	0.307	0.098	0.958	-2.033	0.042
age	0.111	0.034	1.117	1.045	1.194	3.246	0.001

The results of logistic regression with smoking status are presented in Table 4.3. The interpretations of parameters follow:

- Subjects who are currently smoking have 3.85(=$e^{1.347}$) times higher odds of being diagnosed as patient compared to subjects who never smoked. Note that in Table4.1, the percentage of subjects who are currently smoking is higher in patient group compared to healthy group. In other words, the coefficient for currently smoking variable is in the direction of what we expect.

- The odds of observing schizophrenia are 147 times higher for subjects who have never been married as compared with subjects who are separated or divorced. The percentage of subjects who have never been married is higher in patient group as in descriptive statistics in Table 4.1.

- Subjects who are Hispanic have 3.26 times higher odds of being diagnosed as schizophrenia with regard to subjects who have other ethnicities. This interpretation is consistent with descriptive statistics in Table 4.1.

- When the education level of subject increases by one unit, the odds of being diagnosed as patient have decreased by 58%. In Table 4.1, education level of healthy subjects is higher than patients.

- When age is increased by one year, the odds of subjects being diagnosed as patient have increased 1.12 times. Mean of age is higher in patient group as stated in Table 4.1

- Confidence intervals for odds ratios (e^{β}) are also included in Table 3. One can check whether confidence interval does not include 1 or p-value is smaller than

0.05 to check if the corresponding variable is significant. For instance, one of the dummy variables for smoking is insignificant with a p-value of 0.756 in Table 3, while the other one is barely significant (p-value=0.044). Therefore, we also fit a model by excluding the smoking status.

Table 4.4: Results of Logistic Regression without Smoking Status

Variables	Estimate (β)	Std. Error	exp(β)	95% of Confidence Interval		z value	p-value
				Lower Limit	Upper Limit		
Intercept	9.261	2.384				3.885	<0.001
civil(married)	-1.635	1.322	0.195	0.015	2.601	-1.237	0.216
civil(divorced/separated)	-4.473	1.218	0.011	0.001	0.124	-3.672	<0.001
school year	-0.920	0.181	0.398	0.280	0.568	-5.096	<0.001
ethnicity	-1.095	0.557	0.334	0.112	0.995	-1.968	0.049
age	0.111	0.033	1.117	1.047	1.192	3.348	0.001

The interpretations of the parameters have the same tendency in Table 4.4.

4.3.2 Marginal Model

In Chapter 3.2.2, we mentioned marginal models shortly. The response and the covariates are the same ones that we used in the logistic regression. Additionally, in marginal models, the BOLD signals in DMN regions for 30 time points are included. Furthermore, as a separate attempt, we model the response by only BOLD signals to see its marginal effect. In this study, we only present the results of MPFC region as the rest of three regions do not show the significant results. In Table 4.5, co-

Table 4.5: Results of Marginal Model with Smoking Status

Variables	Estimate ($beta$)	Robust Std. Error	exp(β)	95% of Confidence Interval		Robust z value
				Lower Limit	Upper Limit	
(Intercept)	8.395	2.466				3.404*
smoking(yes current)	1.347	0.761	3.846	0.865	17.091	1.770
smoking(yes past)	-0.255	0.745	0.775	0.180	3.337	-0.342
civil(married)	-1.620	0.980	0.198	0.029	1.351	-1.653
civil(divorced/separated)	-4.994	1.350	0.007	0.0005	0.096	-3.700*
school year	-0.875	0.162	0.417	0.303	0.573	-5.392*
ethnicity	-1.182	0.556	0.307	0.103	0.912	-2.124*
age	0.111	0.033	1.117	1.047	1.192	3.372*
MPFC	0.000016	0.000005	1.00002	1.00001	1.00003	2.967*

variates that are marked with "*" are statistically significant in 95% confidence level (i.e. $|z \quad value| \geq z_{0.05/2} = 1.96$). Although the smoking status is not statistically significant, we decide to keep this variable to compare with the results of logistic

regression model. BOLD signals in MPFC region are statistically significant. The covariates about demographic information have the same interpretations with logistic regression. Additionally, 1 unit increase in MPFC signals increases the odds of being diagnosed as patient by less than 1%.

Table 4.6: Results of Marginal Model without Smoking Status

Variables	Estimate (β)	Robust Std. Error	exp(β)	95% of Confidence Interval		Robust z
				Lower Limit	Upper Limit	
(Intercept)	9.261	2.403				3.854*
civil(married)	-1.635	0.963	0.195	0.030	1.287	-1.698
civil(divorced/separated)	-4.473	0.968	0.011	0.002	0.076	-4.623*
school year	-0.920	0.158	0.399	0.292	0.543	-5.827*
ethnicity	-1.095	0.532	0.335	0.118	0.949	-2.058*
age	0.111	0.033	1.117	1.047	1.192	3.351*
MPFC	0.000013	0.000004	1.000013	1.000005	1.000021	2.916*

Removal of smoking status from the model causes slight changes in the coefficients as presented in Table 4.6. BOLD signals in MPFC region are still statistically significant. The interpretations of covariates have the same tendency with the former marginal model.

Table 4.7: Results of Marginal Model with only BOLD Signals

Variables	Estimate (β)	Robust Std.Error	exp(β)	95% of Confidence Interval		Robust z
				Lower Limit	Upper Limit	
Intercept	-0.886	0.188				-4.714*
MPFC	0.000004	0.000001	1.000004	1.000002	1.000006	3.847*

BOLD signals in MPFC region are still statistically significant when the other covariates are ignored in Table 4.7. Note that, the estimate of the coefficients are very small in all of the marginal models. However, the effect of these small coefficients would be much more clearer in performance measures.

4.3.3 Random Effects Model

In Chapter 3.2.3, we touch random effects model briefly. Here, we apply same procedure as in the marginal models. When we attempt to model data set via random effects, we face with some convergence problems. To overcome convergence problems, first, we standardize numeric covariates which are education level and age. However, just standardizing variables does not solve the convergence problem. Therefore, we

decide to change the optimizer method via "optimx" package (Nash and Varadhan, 2011) in RStudio. Changing optimizer method as "bobyqa" solve our problem for most of the models. Even though we try all optimizers, we cannot solve convergence problem which including smoking status in the model. Consequently, random effects model with smoking status is not represented in this subsection. After solving convergence problem, when we conduct the models with all covariates (except for smoking status), some estimated standard errors of covariates are high, which lead to suspicious results. To handle with these suspicious results, we also exclude some covariates. In this subsection, the results we present are best results that solve convergence and high standard error problems.

Table 4.8: Results of Random Effects Model with Demographic Covariates

Variables	Estimate (β)	Std. Error	exp(β)	95% of Confidence Interval		z value	p-value
				Lower Limit	Upper Limit		
Intercept	-15.397	1.981				-7.774	<0.001
ethnicity	-1.520	2.773	0.219	0.001	50.153	-0.548	0.584
age	0.672	1.325	1.958	0.146	26.285	0.507	0.612
MPFC	0.116	2.717	1.123	0.005	230.747	0.043	0.966

In Table 4.8, all of the covariates including BOLD signals in MPFC region are not statistically significant. For comparison, we apply random effects model only with BOLD signals in Table 4.9.

Table 4.9: Results of Random Effects Model with only BOLD Signals

Variables	Estimate (β)	Std. Error	exp(β)	95% of Confidence Interval		z value	p-value
				Lower Limit	Upper Limit		
Intercept	-16.254	1.486				-10.937	<0.002
MPFC	0.129	2.844	1.138	0.004	299.837	0.045	0.964

Since we standardize continuous variables, interpretation of continuous variables has changed. For instance, one standard deviation unit increase in age leads to 1.325 tines unit increase in the odds of disease diagnosed as patient according to Table 4.8.

In both of the random effects models, BOLD signals are not statistically significant. However, we still decide to examine results of random effects models. Furthermore, we compare the performance measures of the random effects model with other modeling. One of the reasons of examining results of random effects models is that the individual heterogeneity is very high. As mentioned in Chapter 3.2.3, in random ef-

fects modeling, $U_i \sim N(0, D)$, D represents the variance of individual heterogeneity. When we explore the results of the these two models, we observe that the standard deviation of random effects is high (Table 4.10).

Table 4.10: Standard Deviation of Random Effects

Random Effects Models	Standard Deviation of Random Effects
With Demographic Covariates	125.300
Only BOLD Signals	129.600

4.3.4 Clustering Approach

We use in order of k-means, hierarchical and CGR algorithms to cluster the BOLD signals of subjects. Then, by using cluster information of the subjects and other covariates, we conduct logistic regression model.

4.3.4.1 k-means Algorithm

One of the main problems in cluster analysis is predicting how many clusters of data set the researchers want to separate. We use "NbClust" (Malika et al., 2014) package in R to decide on the number of clusters. Moreover, in NbCLust package, we choose "ch" index which is proposed in 1974 since Milligan and Cooper (1985) showed "ch" index is the best one based on their simulation studies. According to "ch" index result, the number of cluster for our data set is found to be 3. Therefore, in k-means, hierarchical and CGR algorithms, we decide to use 3 clusters.

In Table 4.11 and Table 4.12, we show the results of logistic regression with 3 clusters we obtained from k-means algorithm included as covariate. Clusters are included into the model as nominal variables. Therefore, we create 2 dummy variables and we choose the Cluster 1 as a reference.

Table 4.11: Results of Logistic Regression with Smoking Status with k-means Algorithm Cluster Information

Variables	Estimate (β)	Std. Error	exp(β)	95% of Confidence Interval		z value	p-value
				Lower Limit	Upper Limit		
(Intercept)	9.492	3.098				3.064	0.002
smoking(yes current)	1.358	0.744	3.888	0.905	16.714	1.827	0.068
smoking(yes past)	0.311	0.982	1.365	0.199	9.353	0.317	0.751
civil(married)	-1.896	1.401	0.150	0.010	2.340	-1.353	0.176
civil(divorced/separated)	-6.130	1.577	0.002	0.000	0.048	-3.887	<0.001
school year	-1.039	0.249	0.354	0.217	0.576	-4.169	<0.001
ethnicity	-1.727	0.701	0.178	0.045	0.703	-2.465	0.014
age	0.133	0.040	1.142	1.056	1.235	3.355	0.001
Cluster 2	1.986	0.759	7.286	1.646	32.254	2.615	0.009
Cluster 3	-0.861	0.891	0.423	0.074	2.424	-0.967	0.333

Table 4.12: Results of Logistic Regression without Smoking Status with k-means Algorithm Cluster Information

Variables	Estimate (β)	Std. Error	exp(β)	95% of Confidence Interval		z value	p-value
				Lower Limit	Upper Limit		
Intercept	10.887	2.872				3.791	<0.001
civil(married)	-2.161	1.436	0.115	0.007	1.922	-1.505	0.132
civil(divorced/separated)	-6.019	1.593	0.002	0.000	0.055	-3.778	<0.001
school year	-1.137	0.239	0.321	0.201	0.512	-4.756	<0.001
ethnicity	-1.584	0.653	0.205	0.057	0.738	-2.427	0.015
age	0.14	0.039	1.150	1.066	1.242	3.569	<0.001
Cluster 2	2.034	0.73	7.645	1.828	31.970	2.786	0.005
Cluster 3	-0.81	0.855	0.445	0.083	2.377	-0.947	0.344

The results of this approach show similar results with logistic regression models. Moreover, some levels of variables related with cluster are statistically significant.

4.3.4.2 Hierarchical Algorithm

As we mentioned in k-means algorithm, subjects are divided into 3 cluster according to BOLD signals in MPFC region. In Figure 4.4, cluster dendrogram is presented. We choose depth cutoff point as 14.

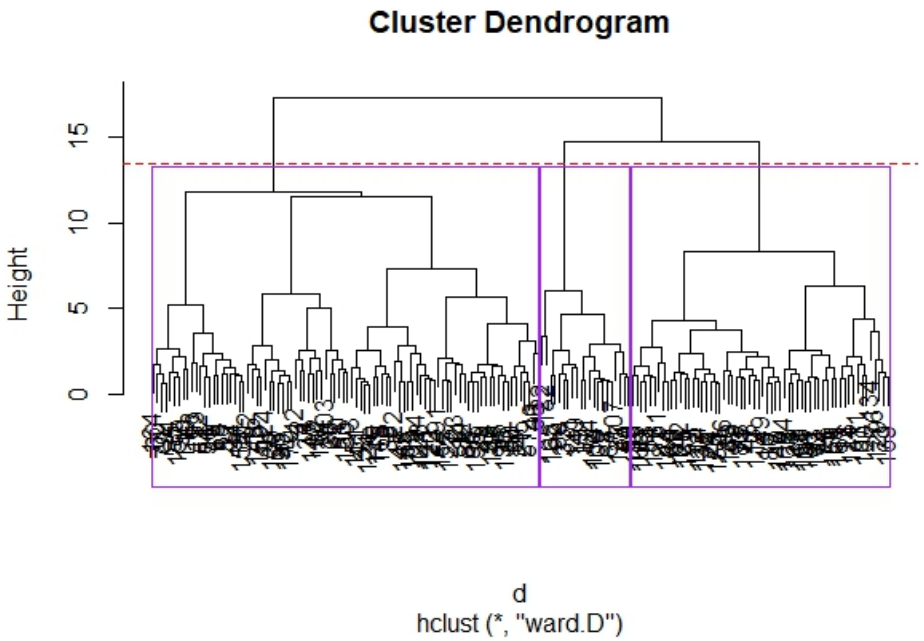


Figure 4.4: Dendrogram of Hierarchical Clustering

In this part, we follow the same path in k-means algorithm. We choose Cluster 1 as reference category again.

Table 4.13: Results of Logistic Regression with Smoking Status with Hierarchical Algorithm Cluster Information

Variables	Estimate (β)	Std. Error	exp(β)	95% of Confidence Interval		z value	p-value
				Lower Limit	Upper Limit		
(Intercept)	8.787	2.863				3.069	0.002
smoking(yes current)	1.452	0.726	4.272	1.029	17.725	1.999	0.046
smoking(yes past)	0.29	0.885	1.336	0.236	7.573	0.328	0.743
civil(married)	-1.565	1.316	0.209	0.016	2.758	-1.189	0.235
civil(divorced/separated)	-5.257	1.394	0.005	0.000	0.080	-3.771	<0.001
school year	-0.851	0.207	0.427	0.285	0.641	-4.117	<0.001
ethnicity	-1.325	0.633	0.266	0.077	0.919	-2.092	0.036
age	0.118	0.037	1.125	1.047	1.210	3.197	0.001
Cluster 2	-1.757	0.925	0.173	0.028	1.058	-1.9	0.057
Cluster 3	-0.393	0.911	0.675	0.113	4.025	-0.431	0.666

Table 4.14: Results of Logistic Regression without Smoking Status with k-means Algorithm Cluster Information

Variables	Estimate (β)	Std. Error	exp(β)	95% of Confidence Interval		z value	p-value
				Lower Limit	Upper Limit		
Intercept	10.119	2.653				3.814	<0.001
civil(married)	-1.65	1.353	0.192	0.014	2.723	-1.22	0.223
civil(divorced/separated)	-4.866	1.296	0.008	0.001	0.098	-3.753	<0.001
school year	-0.943	0.194	0.389	0.266	0.570	-4.858	<0.001
ethnicity	-1.221	0.602	0.295	0.091	0.960	-2.028	0.043
age	0.121	0.036	1.129	1.052	1.211	3.351	0.001
Cluster 2	-1.625	0.908	0.197	0.033	1.167	-1.79	0.074
Cluster 3	-0.174	0.899	0.840	0.144	4.894	-0.194	0.846

The results of this approach show similar results with logistic regression models. Moreover, variables related with cluster are not statistically significant but cluster 2 variable is significant at 10% significance level.

4.3.4.3 CGR Algorithm

The number of clusters for BOLD signals is unknown in advance. Algorithm CGR suggests two different validation methods to decide on the number of clusters. Both methods make use of within and between distances among clusters, and take the ratio of the descriptive statistics of these distances. Figure 4.5 presents the graphs of these 2 validation methods for our data set. We aim to choose the number of clusters which provide the minimum ratio or the one which gives a significant decrease. In Figure 4.5, we plot the validation scores for a maximum number of clusters as 20.

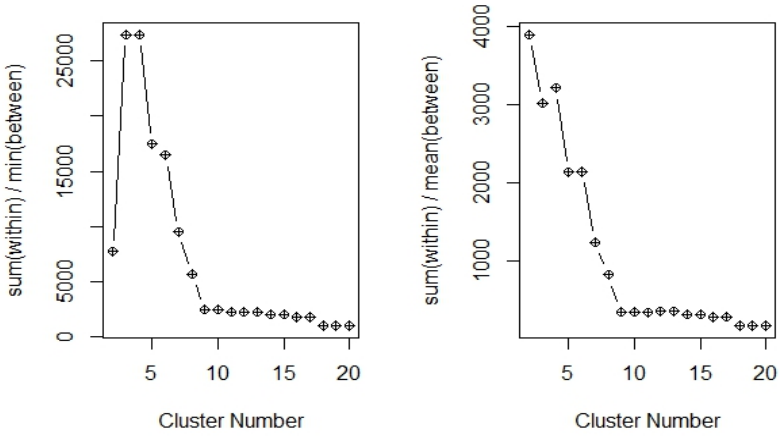


Figure 4.5: Two Validation Score Graphs

As a result, the minimum score is satisfied with 18 clusters. However, this number of cluster is high for modeling.Therefore, we aim to decrease the number of clusters by checking BOLD signals.

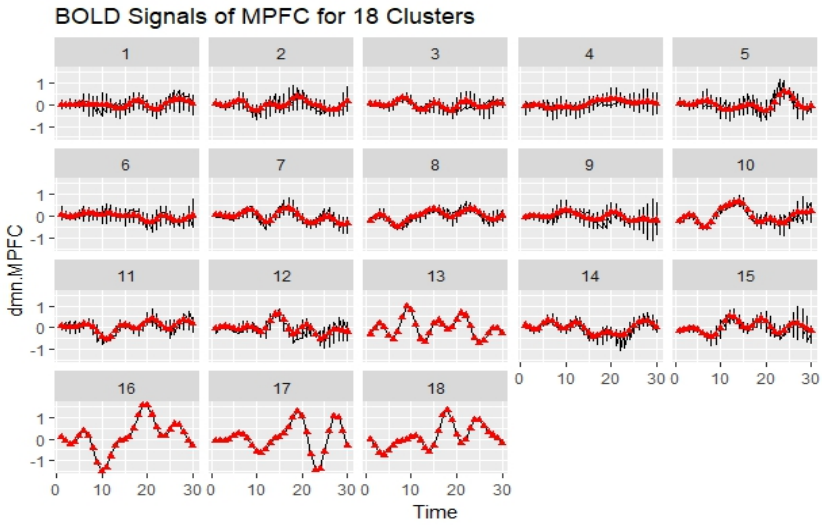


Figure 4.6: BOLD Signals for 18 Clusters

We merge clusters which have similar trends according to Figure 4.6:

- Cluster 1: Cluster 1, Cluster 11, Cluster 12, Cluster 14, Cluster 16 and Cluster 17
- Cluster 2: Cluster 3, Cluster 4, Cluster 7, Cluster 9, Cluster 13 and Cluster 19

- Cluster 3: Cluster 2, Cluster 5, Cluster 6,Cluster 8, Cluster 10 and Cluster 15

Figure 4.7 presents the plot of BOLD signals with respect to 3 clusters.

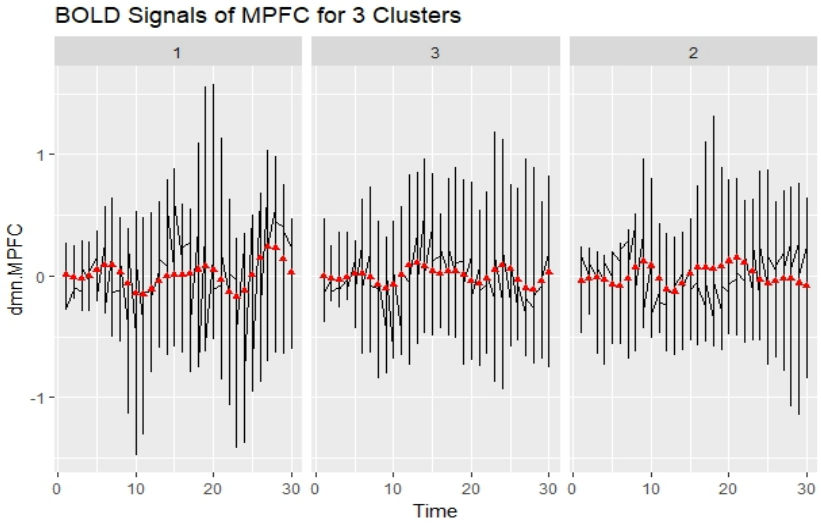


Figure 4.7: BOLD Signals for 3 Clusters

Table 4.15 and 4.16 show the results of logistic regression models with 3 clusters we obtained from CGR algorithm included as covariate. Clusters are included into the model as nominal variables. Therefore, we create 2 dummy variables and we choose the Cluster 1 as reference category.

Table 4.15: Results of Logistic Regression with Smoking Status with CGR algorithm Cluster Information

Variables	Estimate (β)	Std. Error	exp(β	95% of Confidence Interval		z value	p-value
				Lower Limit	Upper Limit		
(Intercept)	9.939	2.943				3.378	0.001
smoking(yes current)	1.512	0.678	4.536	1.201	17.131	2.230	0.026
smoking(yes past)	-0.335	0.881	0.715	0.127	4.022	-0.381	0.703
civil(married)	-2.048	1.370	0.129	0.009	1.891	-1.495	0.135
civil(divorced/separated)	-5.894	1.592	0.003	0.000	0.062	-3.702	<0.001
school year	-0.968	0.222	0.380	0.246	0.587	-4.359	<0.001
ethnicity	-1.242	0.603	0.289	0.089	0.942	-2.061	0.039
age	0.135	0.039	1.145	1.060	1.235	3.419	0.001
Cluster 2	-1.066	0.78	0.344	0.075	1.589	-1.366	0.172
Cluster 3	-1.599	0.783	0.202	0.044	0.938	-2.041	0.041

Table 4.16: Results of Logistic Regression without Smoking Status with CGR Algorithm Cluster Information

Variables	Estimate (β)	Std. Error	exp(β)	95% of Confidence Interval		z value	p-value
				Lower Limit	Upper Limit		
Intercept	10.416	2.610				3.990	<0.001
civil(married)	-2.014	1.391	0.133	0.009	2.039	-1.447	0.148
civil(divorced/separated)	-5.118	1.410	0.006	0.000	0.095	-3.631	<0.001
school year	-0.982	0.196	0.375	0.255	0.550	-5.009	<0.001
ethnicity	-1.123	0.576	0.325	0.105	1.006	-1.951	0.051
age	0.129	0.037	1.138	1.058	1.223	3.479	0.001
Cluster 2	-0.932	0.746	0.394	0.091	1.699	-1.249	0.212
Cluster 3	-1.304	0.736	0.271	0.064	1.149	-1.771	0.077

The results of this approach show similar results with logistic regression models. In Table 4.15, some levels of variables related with cluster are statistically significant. Additionally, some levels of variables related with cluster are barely significant in Table 4.16

4.3.5 Performance Measures

In Chapter 3.2, performance measures that are used in this thesis are given. Moreover, as we mentioned before, the data set is split in two parts as train data set and test data set. In this chapter, the results of performance measures are supplied in Tables 4.17 and 4.18 for train and test data sets, respectively.

Table 4.17: Results of Performance Measures for Train Data Set

Models	Accuracy	Sensitivity	Specificity	PPV	NPV	AIC
Logistic Regression Models						
With Smoking Status	0.803	0.900	0.763	0.610	0.945	103.480
Without Smoking Status	0.788	0.925	0.732	0.587	0.960	104.640
Marginal Models						
With Smoking Status	0.854	0.750	0.897	0.750	0.897	NA
Without Smoking Status	0.825	0.700	0.876	0.700	0.876	NA
Only BOLD Signals	0.292	1.000	0	0.292	NA	NA
Random Effects Models						
With Demographic Covariates	1.000	1.000	1.000	1.000	1.000	157.500
Only BOLD Signals	1.000	1.000	1.000	1.000	1.000	154.000
K-means Clustering+Logistic Regression Modelling						
With Smoking Status	0.832	0.925	0.794	0.649	0.963	93.392
Without Smoking Status	0.839	1.000	0.773	0.645	1.000	92.982
Hierarchical Clustering+Logistic Regression Modelling						
With Smoking Status	0.847	0.975	0.794	0.661	0.987	101.010
Without Smoking Status	0.810	0.950	0.753	0.613	0.973	101.260
CGR Clustering+Logistic Regression Modelling						
With Smoking Status	0.832	0.950	0.784	0.644	0.974	102.970
Without Smoking Status	0.796	0.90	0.753	0.600	0.948	105.330

Table 4.18: Results of Performance Measures for Test Data Set

Models	Accuracy	Sensitivity	Specificity	PPV	NPV
Logistic Regression Models					
With Smoking Status	0.794	0.800	0.792	0.615	0.905
Without Smoking Status	0.765	0.800	0.750	0.571	0.900
Marginal Models					
With Smoking Status	0.794	0.600	0.875	0.667	0.840
Without Smoking Status	0.765	0.600	0.833	0.600	0.833
Only BOLD Signals	0.294	1.000	0	0.294	NA
Random Effects Models					
With Demographic Covariates	0.706	0	1.000	NA	0.706
Only BOLD Signals	0.706	0	1.000	NA	0.706
K-means Clustering+Logistic Regression Modelling					
With Smoking Status	0.794	0.800	0.792	0.615	0.905
Without Smoking Status	0.794	0.900	0.750	0.600	0.945
Hierarchical Clustering+Logistic Regression Modelling					
With Smoking Status	0.765	0.700	0.792	0.583	0.864
Without Smoking Status	0.702	0.700	0.708	0.500	0.850
CGR Clustering+Logistic Regression Modelling					
With Smoking Status	0.794	0.800	0.792	0.615	0.905
Without Smoking Status	0.735	0.700	0.750	0.539	0.857

For all performance measures, the closest values to 1 indicate better model except for

AIC values. The lower AIC value indicate better model. Since marginal model estimation is based on generalized estimation equation, AIC values for marginal model cannot be calculated. According to the results of performance measures in Tables 4.17 and 4.18 we can conclude that:

- For train data set, it is interesting to observe that both random effects models classify the subjects as patient or healthy with 100% correctness. The model conducted with only BOLD signals also discriminate subjects with 100% correctness. This implies that the heterogeneity among subjects is high. However, for test data set, in both of the models, all subjects are classified as healthy. This implies that instead of using only BOLD signals or demographic information, using these information together contributes to classification more.

- For train data set, marginal models with smoking status and without smoking status classify subjects as patient or healthy with 85.4% and 82.5% correctness, respectively. The accuracy rate has increased approximately 5% compared to logistic regression approaches although the estimates of BOLD signals are very small in Tables 4.5 and 4.6. Moreover, in the model conducted with only BOLD signals, all subjects are classified as patient. This implies instead of using only BOLD signals or demographic information, using these information together contributes to classification more. For test data set, accuracy rates for logistic regression and marginal model with smoking status are same and accuracy rates for logistic regression and marginal models without smoking status are same. Although the accuracy rates are same, sensitivity is high in logistic regression models while specificity is high in marginal models. Moreover, in the model conducted with only BOLD signals, all subjects are classified as patient just as in the train data set.

- Logistic regressions with k-means, hierarchical and CGR clustering approaches give similar results in both train and test data sets. However, k-means clustering algorithm proceeded with logistic regression model without smoking variable is one of the best models according to AIC values and performance measures in both test and train data sets.

- To sum up, random effects models perform best in train data set but do not per-

form good in test data set according to performance measures. If we take into account both train and test data sets, with all model diagnostics, two models seem to give promising results. These are marginal model with smoking status and k-means clustering algorithm followed with logistic regression model excluding smoking status.

CHAPTER 5

CONCLUSION AND DISCUSSION

Discovering the brain function, structure and connectivity and their effects on attitudes is one of the fascinating scientific progress in this century (Components of the Human Connectome Project,2017). Furthermore, fMRI is non-invasive procedure to work and map on brain functions. R-fMRI might offer new understandings on brain connectivity.

Schizophrenia is a grave mental illness distinguished by multiple symptoms; positive symptoms such as hallucinations, delusions and racing thoughts; negative symptoms such as lack of enthusiasm and lack of interest, confusion and disorder; last but not least cognitive deficits like lack of perception and lack of apprehension. Cognitive deficiencies are often determined by neuropsychological tests yet the rest of the symptoms are quite challenging to review and examine(National Institute of Mental Health, 2017).

Neurological data can contribute in the scrutiny of a patient's brain and shows us where it would eventually drift off, therefore, it can contribute to perceiving and finding the origins of the problems in a sick brain; furthermore schizophrenia is known to disrupt proper DMN and the ability to focus on some tasks, this phenomenon is known as hyperactivation. Note that this was discovered upon consecutive analysis of neurological data amongst schizophrenia patients, together with the auditory oddball task (Garrity et al., 2007), working tasks(Meyer-Lindenberg et al., 2005, Pomarol-Clotet et al., 2008, Whitfield-Gabrieli et al., 2009) including language tasks also known as semantic priming (Jeong & Kubicki 2010).

Furthermore, when working memory requests are parametrically raised, healthy people display higher suppression of the DMN during working tasks. On the other hand patients fail to show this sequence(Meyer-Lindenberg et al., 2005, Pomarol-Clotet et al., 2008, Whitfield-Gabrieli et al., 2009). Whereas higher DMN activation is linked with poor cognitive performance in healthy people(Whitfield-Gabrieli et al., 2009), it can also be combined with cognitive deficits, working memory and language tasks in schizophrenia.

Current examinations on DMN liveliness on schizophrenia implied medial prefrontal cortical regions of the DMN grid disclosing abnormal integration or activeness, still evidence is not focalized on this section and angles alter in every examination (Zhou et al., 2007; Kim et al., 2009; Whitfield-Gabrieli et al., 2009; Ongur et al., 2010; Woodward et al., 2011). Nonetheless, these researches highlight the association of cognitive deficits and symptoms whichever connects them to the significant pathophysiology of schizophrenia (Rotarska-Jagiela et al., 2010). In addition to that, other research advocates corresponding detailed anomalies for schizophrenia (Calhoun et al., 2008). Zhang et al. (2010) stated that damaging of DMN might has been connected with depression, schizophrenia, Alzheimer's disease and autism.

In the light of these information, in this study, we aim statistical disease detection of patients diagnosed as schizophrenia and healthy subjects is investigated via demographic covariates and BOLD signals of R-fMRI in DMN regions. To classify subjects, first, we conducted logistic regression on demographic information of subjects. Next, we obtained BOLD signals in DMN regions via connectivity of the brain. Furthermore, by using these BOLD signals, we conducted marginal models, random effects models and logistic regression models together with k-means, hierarchical and CGR clustering methods. The reason of classifying with and without BOLD signals is to see how BOLD signals contribute to classification.

To summarize all results, one should consider BOLD signals together with demographic information for diagnosing patients since better results are obtained from those models. As we discussed in Chapter 4.4, random effects models give best results in train data set whereas they cannot give good results in test data set. For overall evaluation of the all models, marginal model with smoking status and k-means clus-

tering algorithm followed with logistic regression model excluding smoking status gives best results.

Li and her friends (2007) suggested generalized sufficiency score and generalized conditional score approaches for joint models. Both proposed scores need to satisfy neither a distributional nor a covariance structural assumption. Authors of this paper run algorithms of these scores in the SAS program. In future, we intend to work on the datat set which we study on this thesis by transferring to RStudio.

REFERENCES

Ashby, F. G. (2011). *Statistical analysis of fMRI data*. Cambridge, Massachusetts : The MIT Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4.*Journal Of Statistical Software, 67*(1), 1-48.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345-370. doi:10.1007/BF02294361

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease.*Annals Of The New York Academy Of Sciences, 11241*-38.

Calhoun, V. D., Maciejewski, P. K., Pearlson, G. D., & Kiehl, K. A. (2008). Temporal lobe and 'default' hemodynamic brain modes discriminate between schizophrenia and bipolar disorder. *Human Brain Mapping, 29*(11), 1265-1275.

Carey V. J, Lumley T, & Ripley B. gee: Generalized Estimation Equation Solver. (2015). gee: Generalized Estimation Equation Solver. R package version 4.13-19. https://CRAN.R-project.org/package=gee.

Cinar, O., Ilk, O., & Iyigun, C. (2015). cgr: Clustering Genes with Replication in Short Time Series. R package version 0.1.

Cinar, O., Ilk, O., & Iyigun, C. (2017). Clustering of short time-course gene expression data with dissimilar replicates. *Annals of Operations Research*.

Components of the Human Connectome Project. Retrieved May 5, 2017, from http://www.humanconnectome.org/about/project/resting-fmri.html.

Diggle, P. J., Heagerty, P. J., Liang, K., & Zeger, S. L. (2013). *Analysis of longitudi-*

nal data. Oxford University Press, Oxford.

Garrity, A. G., Pearlson, G. D., McKiernan, K., Lloyd, D., Kiehl, K. A., & Calhoun, V. D. (2007). Aberrant "default mode" functional connectivity in schizophrenia. *The American Journal Of Psychiatry, 164*(3), 450-457.

Han, J., & Kamber, M. (2006). *Data Mining, Southeast Asia Edition*. Amsterdam: Morgan Kaufmann.

Jeong, B., & Kubicki, M. (2010). Reduced task-related suppression during semantic repetition priming in schizophrenia. *Psychiatry Research, 181*(2), 114-120.

Kim, D, Manoach, D, Mathalon, D, Turner, J, Mannell, M, Brown, G, Ford, J, Gollub, R, White, T, Wible, C, Belger, A, Bockholt, H, Clark, V, Lauriello, J, O'Leary, D, Mueller, B, Lim, K, Andreasen, N, Potkin, S, & Calhoun, V (2009). Dysregulation of working memory and default-mode networks in schizophrenia during a Sternberg item recognition paradigm. *Human Brain Mapping, 30*(11), 3795–3811.

Lee, M. H., Smyser, C. D., & Shimony, J. S. (2013). Resting state fMRI: A review of methods and clinical applications. AJNR. *American Journal of Neuroradiology, 34*(10), 1866–1872.

Li, E., Wang, N., & Wang, N.-Y. (2007). Joint Models for a Primary Endpoint and Multiple Longitudinal Covariate Processes. *Biometrics, 63*(4), 1068–1078.

Lindquist, M. A. (2008). The Statistical Analysis of fMRI Data. *Statistical Science, 23*(4), 439-464.

Malika, C., Nadia, G., Véronique, B., & Azam, N. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal Of Statistical Software, Vol 61, Iss 1, Pp 1-36 (2014), (1),*1. doi:10.18637/jss.v061.i06

MATLAB and Statistics Toolbox Release 2012b, The MathWorks, Inc., Natick, Massachusetts, United States.

Max, K. (2008). Building Predictive Models in R Using the caret Package. *Journal Of Statistical Software, 28*(5).

Meyer-Lindenberg, A., Olsen, R., Kohn, P., Brown, T., Berman, K., Egan, M., & Weinberger, D. (2005). Regionally specific disturbance of dorsolateral prefrontal-hippocampal functional connectivity in schizophrenia. *Archives Of General Psychiatry, 62*(4), 379-386.

Milligan, G., & Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159-179. doi:10.1007/BF02294245

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis*. Hoboken, NJ : Wiley, 2012.

Nash, J., & Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: Optimx for R. *Journal Of Statistical Software, 43*(9), 1-14.

National Institute of Mental Health. (n.d.). Retrieved 3 September 2016, from https://www.nimh.nih.gov/health/statistics/prevalence/schizophrenia.shtml.

Ongur, D., Shinn, A., Cohen, B., Lundy, M., Greenhouse, I, Menon, V, & Renshaw, P. (2010). Default mode network abnormalities in bipolar disorder and schizophrenia. *Psychiatry Research - Neuroimaging, 183*(1), 59-68.

Poldrack, R. A., Congdon, E., Triplett, W., Gorgolewski, K. J., Karlsgodt, K. H., Mumford, J. A., Sabb, F. W., Freimer, N. B., London, E. D. , Cannon, T. D., & Bilder, R. M. (2016). A phenome-wide examination of neural and cognitive function. *Scientific Data, 3*, 160110.

Pomarol-Clotet, E., Salvador, R., Sarró, S., Gomar, J., Vila, F., Martínez, Á., & McKenna, P. J. (2008). Failure to deactivate in the prefrontal cortex in schizophrenia: Dysfunction of the default mode network? *Psychological Medicine, 38*(8), 1185-1193.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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

Raichle, M. E., & Snyder, A. Z. (2007). A default mode of brain function: a brief history of an evolving idea. *Neuroimage, 37*(4), 1083-1090.

Rotarska-Jagiela, A., van de Ven, V., Oertel-Knöchel, V., Uhlhaas, P. J., Vogeley, K., & Linden, D. E. (2010). Resting-state functional network correlates of psychotic symptoms in schizophrenia. *Schizophrenia Research, 117*, 21-30.

RStudio Team (2017). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. http://www.rstudio.com.

Whitfield-Gabrieli, S., and Nieto-Castanon, A. (2012). Conn: A functional connectivity toolbox for correlated and anticorrelated brain networks. Brain Connectivity.

Whitfield-Gabrieli, S., Thermenos, H. W., Milanovic, S., Tsuang, M. T., Faraone, S. V., McCarley, R. W., & Seidman, L. J. (2009). Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. *PNAS Proceedings Of The National Academy Of Sciences Of The United States Of America, 106*(4), 1279-1284.

Wickham, H. (2009). ggplot2. *Elegant Graphics for Data Analysis*. New York, NY : Springer-Verlag New York, 2009.

Woodward, N. D., Rogers, B., & Heckers, S. (2011). *Functional resting-state networks are differentially affected in schizophrenia. Schizophrenia Research, 130*(1-3), 86-93.

Zhang, Z, Lu, G, Zhong, Y, Tan, Q, Liao, W, Wang, Z, Wang, Z, Li, K, Chen, H, & Liu, Y (2010). Altered spontaneous neuronal activity of the default-mode network in mesial temporal lobe epilepsy.*Brain Research, 1323*, 152-160.

Zhou, Y., Liang, M., Tian, L., Wang, K., Hao, Y., Liu, H., & Jiang, T. (2007). Functional disintegration in paranoid schizophrenia using resting-state fMRI. *Schizophrenia Research, 97*(1-3), 194-205.

A Merged Categories of Demographic Information

Table A.1 Merged Categories of Demographic Information

Original Values		**Merged Values**	
Race Main	Frequency	Race Main	Frequency
1: American Indian or Alaskan Native	36	*1: White*	128
2: Asian	3	*2: Other*	43
3: Native Hawaiian/Pacific Islander	0		
4: Black/African American	3		
5: White	128		
6: More than one race	1		
Civil Status	Frequency	Civil Status	Frequency
1:Married	18	*1:Married*	18
2:Separated	2	*2:Separated/Divorced*	20
3:Divorced	18	*5:Never married*	133
5:Never married	133		
Religion	Frequency	Religion	Frequency
1:Catholic	86	*1:Christian*	128
2:Protestant	42	*2:Not Affiliated*	24
3:Jewish	11	*3:Other*	19
4: Muslim	1		
5:Not Affiliated	24		
6: Other	7		