MULTIMODAL MR PREDICTION MODELS FOR LATE-LIFE DEPRESSION AND TREATMENT RESPONSE

by

Meenal J. Patel

B.S. in Biomedical Engineering, Purdue University, 2009

Submitted to the Graduate Faculty of

Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Meenal J. Patel

It was defended on

March 25, 2014

and approved by

Carmen Andreescu, MD, Assistant Professor, Department of Psychiatry

Aaron Batista, PhD, Assistant Professor, Department of Bioengineering

Jiantao Pu, PhD, Assistant Professor, Department of Radiology and Bioengineering

George Stetten, MD, PhD, Professor, Department of Bioengineering

Dissertation Director: Howard Aizenstein, MD, PhD, Associate Professor, Department of Psychiatry and Bioengineering

MULTIMODAL MR PREDICTION MODELS FOR LATE-LIFE DEPRESSION AND TREATMENT RESPONSE

Meenal J. Patel, PhD

University of Pittsburgh, 2014

Currently, depression diagnosis relies primarily on behavioral symptoms and signs, instead of underlying brain characteristics, and treatment is guided by trial and error instead of individual suitability associated with underlying brain characteristics. Also, previous brain-imaging studies attempting to resolve this issue have traditionally focused on mid-life depression using a single imaging modality and region-based approach, which may not fully explain the complexity of the underlying brain characteristics; especially for late-life depression. We aimed to evaluate and compare underlying brain characteristics of late-life depression diagnosis and treatment response by estimating accurate prediction models using multi-modal magnetic resonance imaging and non-imaging measures. Based on our finding, late-life depression diagnosis and treatment response predictors involve measures from different imaging modalities, which are indicative of differences in underlying brain characteristics.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Many people have contributed to the successful completion of this dissertation. First and foremost, my advisor, Dr. Howard Aizenstein, has provided me with great support and guidance. I have learned a great deal from his mentorship and am grateful for the opportunities he provided me with to grow, learn, and experiment in his lab. I am also grateful for the support, advice, and feedback that my committee members (Dr. Carmen Andreescu, Dr. George Stetten, Dr. Aaron Batista, and Dr. Jiantao Pu) have given me; especially Dr. Carmen Andreescu for the time she took to guide me on my research projects.

Also, I would like to thank all the Geriatric Psychiatry Neuroimaging (GPN) Lab students, members, and collaborators that I have worked with. They have all provided great encouragement and made working at GPN a fun and unforgettable experience. I am grateful for all the time they have spent in helping me with my research studies and for everything I have learned from them.

Most importantly, I am indebted for the tremendous love, support and wisdom I have received from my family and friends. They have put up with me through this journey and have guided me in countless ways.

Last but not least, I would also like to thank the National Institutes of Health who have helped fund my research studies.

1.0 INTRODUCTION

In a given year, approximately 6% of the US population aged 65 and older (i.e. 2 million people) is diagnosable with depression not associated with normal aging [Mental Health America]. This elderly population above the age of 65 represented 13% (40.3 million people) of the U.S. population according to the 2010 census, and is predicted to increase to 19% (72.1 million people) in 2030 by the Administration of Aging [Vincent & Velkoff, 2010]. This portion of the population is growing fast due to the aging of the baby boomers and the increasing life expectancy. As a result, the number of elderly people with late life depression (LLD) is also increasing rapidly.

The current method for classifying mental disorders including LLD is to use the guidelines provided by Diagnostic and Statistical Manual of Mental Disorder 5 (DSM 5). However, the DSM categorizes mental disorders based on the symptoms experienced by the patient. The DSM 5 is criticized for being solely based on observable behavioral patterns of the disorders it classifies. It lacks the reliability and validity that could accrue via the use of biomarkers of the underlying brain changes. Thus, in order to advance the agenda of personalized medicine for persons with mental disorders, it is important to identify biomarkers reflecting the neural circuit abnormalities that characterize a given disorder and/or that provide the neurobiological basis of spectra of related disorders.

1.1 SPECIFIC AIMS

The goal of this dissertation is to successfully estimate predictive models using multimodal magnetic resonance imaging measure and machine learning methods to help better understand late-life depression and associated treatment response. We pursued this goal in three aims:

- **Aim 1:** To compare linear vs. non-linear model for studying: (1) brain functional connectivity in the elderly, (2) late-life depression diagnosis, and (3) late-life depression treatment response
  - Hypothesis: (1) Better model = non-linear for both cases
- **Aim 2:** To determine which features best predict (1) brain functional connectivity in the elderly, (2) late-life depression diagnosis, and (3) late-life depression treatment response
  - Hypothesis: Imaging features will have a greater significant influence then the demographic and cognitive ability features, which may exert their influence though changes in brain structure.
- **Aim 3:** To accurately predict (1) brain functional connectivity in the elderly, (2) late-life depression diagnosis, and (3) late-life depression treatment response
  - Hypothesis: There exists a model for each outcome variable that can accurately predicted it with the given features within a 10-20% error margin.

1.2 OUTLINE OF CHAPTERS

The experiments that address these aims are described in chapters 5-7. The next three chapters of this dissertation outline the background knowledge needed to understand the experiments. The second chapter describes late-life depression, its treatment, associated underlying neural circuit abnormalities, and other related factors. The third chapter discusses multimodal magnetic resonance imaging including image acquisition, image analysis methods, and image processing methods specific to neuroimaging. This chapter also discusses underlying brain structure and function alterations associated with late-life depression and its treatment response by past magnetic resonance imaging studies. The fourth chapter describes various machine learning methods that can be used to estimate accurate prediction models. The fifth chapter describes an experiment that shows how brain structure can affect the function magnetic resonance imaging signal acquired to study functional brain activation in the elderly. The sixth chapter describes an experiment that attempts to better understand the relationship between brain function and brain structure. The motivation behind studying this relationship is to identify if and how it varies between normal aging and late-life depression. The experiment uses multimodal imaging and machine learning to estimate a prediction model that would explain the relationship between brain function and structure. However, the study has several limitations and an accurate prediction model could not be estimated. Thus, a follow-up study in pursued by addressing the limitations, but again with little success (see “Appendix A”). So, in the seventh chapter, we attempt to directly estimate prediction models for late-life depression diagnosis and treatment response with greater success.

2.0 LATE-LIFE DEPRESSION

This chapter gives a background understanding of late-life depression and its treatment. It primarily describes the underlying brain changes and characteristics associated with late-life depression and its treatment. This chapter also describes potential demographic, clinical, and cognitive ability predictors of late-life depression and its treatment response based on past studies.

2.1 INTRODUCTION

When major depressive disorder (MDD) occurs in older adults it is often referred to as late-life depression (LLD). The age cut-off for LLD varies by research group, and has ranged from older than 55, to older than 70. Age-related neuropathology, along with other biological, psychological, and social factors have been identified as important contributors to the development and phenomenology of LLD.

2.1.1 Risk Factors of LLD

Biological risk factors for LLD include genetics, anatomical and physiological abnormalities (see "Anatomy and Physiology of LLD" section for more details), and medical (including myocardial

infarctions and stroke) and psychiatric co-morbidity (including dementia, alcohol abuse, and anxiety related disorders). Psychological risk factors include personality disorder, cognitive distortions (e.g. overreaction and exaggeration), and lack of emotional control and self-efficacy. Social risk factors include stressful life events, bereavement, socio-economic disadvantage, and impaired social support [Blazer & Hybels, 2005; Gottfries 2001].

2.1.2 Early-Life vs. Late-Life Depression

LLD is different from early-life depression in characteristics including late onset, executive dysfuntion, and/or vascular disease [Lebowitz et al., 1997, Alexopoulos et al., 2009]. Treatment of LLD is similar to early-life depression, but there is a greater risk of adverse events in the elderly [Gottfries 2001].

2.2 DIAGNOSIS OF LLD

Currently, the diagnosis of LLD is based on the Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM V). The DSM V determines diagnosis of mental disorders based on the symptoms experienced by the patient. Nine criteria proposed by the DSM V for diagnosis of depression are primarily depressed mood, reduced interest or pleasure in most activities, substantial unintended weight loss or gain, insomnia or excess sleeping, agitation or psychomotor retardation observed by other, fatigue or energy loss, feelings of worthlessness or excessive guilt, indecisiveness or reduced ability to think or concentrate, and reoccurring thoughts of death. The occurrence of at least five of these criteria approximately every day for a

two-week period, a score of at least 10 on the Beck Depression Inventory, or a score of at least 10 on the Geriatric Depression Scale supports the diagnosis of LLD [American Psychiatric Association].

DSM V is solely based on observable behavioral patterns of the disorders it classifies. It lacks the reliability and validity that would come from a model that considers underlying brain changes. Improvement of such deficiencies in the DSM could not only improve the diagnosis of mental health disorders like LLD, but also improve the remission rates for the disorders by helping make the treatment more personalized.

2.2.1 Relative Prevalence of LLD and Potential Predictors

Several recent late-life studies have shown relative prevalence of LLD to be associated with demographics (e.g. age, gender, and education) and cognitive ability. In regards to age, these late-life studies have shown that the prevalence of depression in the elderly population decreases with an increase in age [Forlani et al., 2013], increases with an increase in age [Luppa et al., 2012], and has a U-shaped relationship with age [Wild et al., 2012; Wu et al., 2012]. In regards to gender, these late-life studies have shown that women are more at risk of depression than men [Luppa et al., 2012], as well as a decrease in these differences with age [Forlani et al., 2013]. A study reviewing the literature for association between education and LLD has concluded that less education is related to a greater risk of LLD [Chang-Quan et al., 2010]. Another study has shown that the level of education does not affect the cognitive decrements related to LLD [Bhalla et al., 2005]. Other studies have also supported a relationship between reduced cognitive ability and LLD [Ganguli et al., 2006; Kohler et al., Apr 2010; Wilkins et al., 2009].

2.3 ANATOMY AND PHYSIOLOGY OF LLD

2.3.1 Neurotransmitter-Specific Systems' Decline

Several behavioral changes commonly found in LLD patients are most often associated with reductions in activation of certain neurotransmitter-specific systems (i.e. circuitry associated with a specific neurotransmitter) due to neurotransmitter loss [Meltzer et al., 1998]. Neurotransmitters are chemical substances that transfer information between neurons to allow different brain regions of the corresponding neurotransmitter-specific systems to communicate. This allows the nervous system to process sensory information and control behavior associated with the brain regions of the corresponding neurotransmitter-specific systems [Hyman, 2005].

Other important components of the neurotransmitter-specific systems that modulate the transfer of information between brain regions include synaptic vesicles, and neurotransmitter-specific receptors and transporters—both of which are proteins [Hyman, 2005; Meltzer et al., 1998]. The neurotransmitter-filled synaptic vesicles are responsible for releasing the neurotransmitter from a neuron into the synaptic cleft (space between two neighboring neurons) [Blakely & Edwards, 2012]. The released neurotransmitter then excites the neighboring neuron to continue relaying information via the help of neurotransmitter-specific receptors. The neurotransmitter-specific receptors are located at presynaptic—information transferring end of neuron that released the neurotransmitter, i.e. axon terminal—and/or postsynaptic—information receiving end of neighboring neuron, i.e. dendrites—sites of the systems. Both types of receptors bind with the appropriate neurotransmitters released into the synaptic cleft. Upon binding with the neurotransmitter, the presynaptic receptor inhibits further release of that neurotransmitter to modulate excitation of the neurotransmitter-specific system, while the postsynaptic receptor

excites the postsynaptic neuronal cell continuing the signal transfer and excitation of the neurotransmitter-specific system [Hyman, 2005; Raiteri, 2001]. In order to subdue the extent and duration of the signaling amongst the neurotransmitter-specific system, the neurotransmitter-specific transporter perform the reuptake of the neurotransmitter released in the synaptic cleft back and store it for future usage [Blakely & Edwards, 2012].

The “Pharmacotherapy” section further explains how neurotransmitters, neurotransmitter-specific receptors, and neurotransmitter-specific transporters of certain neurotransmitter-specific systems are targeted by common antidepressants to treat LLD. The mostly frequently studied neurotransmitter-specific systems in relation to LLD include the serotonin, norepinephrine, and dopamine systems.

Serotonin is often connected to mood, aggression, feeding and sleep. There are two known serotonin systems. The first system has serotonin neuronal cell bodies located primarily in the dorsal and median raphe nuclei of the caudal midbrain. These neurons project extensively through the thalamus, hypothalamus, basal ganglia, basal forebrain, and the neocortex. The organization of the projections, interaction with postsynaptic elements, and the distribution of terminals in cortical and limbic regions of the serotonin neurons suggests that this system is associated with regulation of behavioral state and modulation of more specific behaviors. The second system has serotonin neuronal cell bodies located primarily in the pontine and medullary raphe. This system projects through the brainstem, cerebellum, and spinal cord. As a result, it seems to be associated with modulation of sensory input and motor control. Overall, serotonin is associated with mediating various behaviors including mood, anxiety, sleep, temperature, appetite, sexual behavior, eating behavior, movements, gastrointestinal motility, and more [Meltzer et al., 1998; Stahl, 1998].

Norepinephrine is linked to alertness, energy, anxiety, attention, and interest in life [Nutt, 2008]. The norephinephrine system is formed of noradrenergic neurons. These neurons originate at the locus coeruleus (a region in the brain stem) and project to the neocortex, hippocampus, cerebellum, and thalamus. Inputs projections from the locus coeruleus are denser in brain regions associated with spatial attention. Overall, this system is associated with modulation of arousal, attention, and stress response [Benarroch, 2009].

Dopamine is primarily related to motivational control in regards to rewarding, aversive, and alerting events [Bromberg-Martin et al., 2010]. There are several dopamine systems and they are formed of the dopaminergic neurons. One of the major systems associated with depression, originates at the ventral tegmental region of the midbrain and projects to the ventral pallidum and limbic system regions including the hippocampus, amygdala, medial prefrontal cortex, and nucleus accumbens. Thus, the dopamine system plays an important role in modulating the information flow across the limbic circuitry, which is responsible for influencing motivational, emotional, contextual, and affective behavior [Pierce & Kumaresan, 2006].

2.3.2 Fronto-Striatal and Fronto-Limbic Circuitry Dysfunction

Several imaging studies have found associations of LLD with the fronto-striatal and fronto-limbic circuitry dysfunction. See chapter 3 for more details on these studies. Below is a description of the fronto-striatal and fronto-limbic circuitry. Abnormalities in these circuits are associated with several functions and/or behaviors that are altered in LLD patients.

The fronto-striatal brain circuitry gets inputs from different neurotransmitter systems (including those associated serotonergic, noradrenergic, and dopaminergic neurons), which modulate corticostriatal information processing. There are about five known fronto-striatal

circuits. Two of these circuits—motor and oculomotor circuits—are responsible for motor functions, while the other three circuits—dorsolateral prefrontal, orbital frontal, and anterior cingulate circuits—are responsible for executive functions (e.g. selecting and perceiving essential information, handling information in the working memory, planning and organizing, controlling behavior, adapting to changes, and making decisions), social behavior and motivational states. The overall anatomy for all five circuits includes a closed loop circuitry connecting regions of the following structures in the given order: frontal cortex; striatum (caudate, putamen, ventral striatum); globus pallidus and substantia nigra; and thalamas [Chudasama & Robbins, 2006; Tekin & Cummings, 2002]. Compromised structural integrity of fronto-striatal circuits are associated with executive dysfunction in LLD [Alexopoulos, 2002].

The fronto-limbic circuitry consists of neuronal pathways connecting the frontal lobe areas to the limbic regions. This circuitry is responsible for emotional and motivational processing. The anatomy of the circuit is majorly composed connections between the frontal cortex (e.g. prefrontal cortex and orbitofrontal cortex) and limbic lobe (e.g. hippocampus, amygdala, and anterior cingulate cortex) regions [Hart & Rubia, 2012]. Alterations of this circuitry are associated with induced sadness in non-depressed individuals [Alexopoulos, 2002].

2.3.3 Cerebrovascular Disease

LLD is also believed to be associated with vascular disease. This association is proposed by the vascular depression hypothesis [Taylor et al., 2013]. One potential cause of LLD is thought to be atherosclerosis, which is an underlying cause of vascular disease. The cerebral lesions formed by atherosclerosis may lead to depression by either disrupting pathways linked to mood regulation or amassing of lesions beyond an acceptable threshold. It may also lead to a stroke, which would

then result in ischemic lesions. On the contrary, LLD may also be the cause of vascular disease. Vascular disease may be a result of LLD related hypercortisolemia effects, immune activation, increased thrombosis due to platelet accumulation, loss of arterial endothelial functioning, or irregular metabolism of folate or homocysteine. Greater evidence of associations between LLD and vascular disease is given by Magnetic Resonance Imaging (MRI) studies (see chapter 3 for more details) [Kales et al., 2005].

2.4 TREATMENT OF LLD

Various treatment options exist for late-life depression. For the elderly, compared to the younger patients, the post-treatment risks of adversities are greater and side effects are not endured as well [Gottfries 2001]. However, when effective, the treatment can help improve the quality of life emotionally, socially, and physically. Treatment recommendations for patients should consider personal preferences, effectiveness of previous treatment(s), and comorbid conditions. Effectiveness of a given treatment is individual dependent and thus is typically guided by a trial and error process that requires monitoring especially during the first 8-10 weeks to reduce chances of premature discontinuation. Pharmacotherapy (i.e. antidepressants) is the most common form of treatment. Other treatment options proven to be effective include psychotherapy, electroconvulsive therapy, and physical exercise programs.

2.4.1 Pharmacotherapy

As a part of pharmacotherapy for the LLD patients, the Food and Drug Administration has approved twenty antidepressant medications. These antidepressants include the selective serotonin-reuptake inhibitors (SSRIs), serotonin-norepinephrine reuptake inhibitors (SNRIs), tricyclic antidepressants, monoamine oxidase inhibitors (MAOIs), norepinephrine-dopamine reuptake inhibitors (NDRIs), and more.

Most commonly, patients are asked to try SSRIs as a first attempt at treatment. The physiological mechanism behind SSRIs involves blocking the serotonin transporter to inhibit the reuptake of serotonin neurotransmitters after being released into the synaptic cleft. This causes greater concentration of serotonin to remain in the synaptic cleft. The increased serotonin level in the synaptic cleft is thought to first desensitize the presynaptic serotonin receptors to decrease inhibition of serotonin release. Consequently, function of the serotonin system in enhanced by an increase in serotonergic neurotransmission. The results include improved mood and other depressive symptoms like reduction of depression-related anxiety [Nutt et al., 1999; Stahl, 1998]. SSRIs include antidepressants like Fluoxetine (Prozac), Sertraline (Zoloft), Citalopram (Celexa), Escitalopram (Lexapro), and Paroxetine (Paxil).

The physiological mechanism behind SNRIs is similar to SSRIs, except it involves inhibition of the reuptake of both serotonin and norepinephrine neurotransmitters after being released into the synaptic cleft. This causes concentrations of both neurotransmitters to increase in the synaptic cleft and subsequently enhances the function of both serotonin and norepinephrine systems. The results include improvement of depressive symptoms including lower anxiety and improved general life functioning [Lambert & Bourin, 2002]. SNRIs include antidepressants like Venlafaxine (Effexor) and Duloxetine (Cymbalta).

Other antidepressants that work as reuptake inhibitors include tricyclic antidepressants like Nortriptyline (Pamelor) and Desipramine (Norpramin), and NDRIs like Bupropion (Wellbutrin). Tricyclic antidepressants behave similar to SNRIs by inhibiting uptake of serotonin and norepinephrine by blocking the respective transporters, except they also have an affinity to other receptors, thus increasing chances of adverse side effects [Lambert & Bourin, 2002]. They used to be one of the first antidepressants recommended to patients, but because of their side effects—including sedation, weight gain, dry mouth, urinary retention, constipation, blurry vision, orthostatic hypotension, and impairment of cardiac conduction— they no longer are. Because of their side effects, studies have associated tricyclic antidepressants with greater dropout rates than SSRIs [Unutzer et al., 2007]. Now, they are recommended for patients who have had previous successful responses with them or for patients who do not respond to other antidepressants. NDRIs are involved in inhibiting the reuptake of norepinephrine and dopamine neurotransmitters, thus increasing their availability for enhancing function of norepinephrine and dopamine systems respectively. Bupropion is the only Food and Drug Administration (FDA) approved NDRI. It is associated with anti-craving and anti-withdrawal effects as well as improved attention [Stahl et al., 2004].

Other types of antidepressants include MAOIs like Phenelzine (Nardil) and Tranylcypromine (Parnate), Mirtazapine (Remeron), and Nimodipine (Nimotop). MAOIs work by inhibiting monoamine oxidase activity, which prevents the breakdown of monoamine neurotransmitters thereby increasing their concentrations and chances of respective neurotransmitter system excitations. Monoamine neurotransmitters include serotonin, norepinephrine, and dopamine [Racagni & Popoli, 2008; Quitkin et al., 1979]. Mirtazapine helps increase both serotonin and norepinephrine related activity by acting as an antagonist to

respective presynaptic receptors (i.e. marginally blocks the presynaptic receptors to have weaker affects compared to SNRIs on inhibition of serotonin and norepinephrine reuptake) [Stimmel et al., 2012]. Its effects, unlike Bupropion, are related to sedation and help increase appetite as well as gain weight. Thus, it is beneficial to patients with insomnia and weight loss. Nimodipine is a calcium channel blocker, which is believed to improve blood flow and guard neurons from injury or degeneration, thus also benefiting cognition and brain function. As a result, it is hypothesized to be beneficial for vascular depression associated with ischemic lesions in the brain [Whyte et al., 2009].

Treatment response to a given antidepressant differs from person to person. It is essentially guided by a trial and error process. To help improve the chances of a positive response, some possible selection considerations to make other than side effects include past treatment response(s), potential interactions with other drugs, frequency of dosing, overdose effects, cost, and/or treatment response(s) of close relatives with depression. Other than that, the common procedure is to first test an SSRI and then continue with or change the medication based on the response, symptoms, and resulting side effects. To determine if a medication is working or not, it takes up to 4-6 weeks of treatment. Only 40-65% of the patients show signs of adequate treatment response to any given antidepressant. Thus, alternative treatment methods are often required, which is not ideal as it increases cost and chances of side effects especially since older adults need to take full doses for determination of adequate response. Also, on top of the low response rate, there is also a 70% chance of recurrence after remission. Thus, to avoid recurrences, a medication that has shown signs of success should be continued for 6-12 months after remission for a 60% decrease in the chances of recurrence according to a study in a sample patient population [Unutzer et al., 2007].

2.4.2 Other Forms of Therapy

As mentioned earlier, other treatment options for LLD are psychotherapy, electroconvulsive therapy, and physical exercise programs. Psychotherapy is recommended for patients who fail to respond to pharmacotherapy. Patients with chronic depression are often recommended a combination of pharmacotherapy (i.e. antidepressants) and psychotherapy. Psychotherapy often takes about 6-12 sessions by a trained therapist, although some patients benefit significantly from longer-term therapy. Effective types of psychotherapy include cognitive behavior therapy, interpersonal psychotherapy, and problem-solving therapy. Cognitive behavior therapy focuses on correcting negative depression related thoughts, interpersonal psychotherapy deals with interpersonal causes of depression, and problem-solving therapy teaches ways to solve daily depression related problems. Such methods of psychotherapy have been related to an increase in the density of serotonin receptors may be the underlying cause of improved social and occupational functioning associated with psychotherapy [Karlsson, 2012]. Statistically, these methods of treatment have been shown to reduce depressive symptoms by 50% with a success rate of 45-70% in a sample population [Unutzer et al., 2007].

Electroconvulsive therapy involves electrical induction of 6-8 seizures in patients to produce behavioral changes similar to antidepressants. The mechanisms of electroconvulsive therapy's efficacy in treating depression are not fully understood. It is believed that changes in the neurotransmitter systems play a role. These include the following effects of electroconvulsive therapy: (1) enhances dopamine receptor function (after seizures 1 to 2) leading to increased duration of interest in activities and restored appetite and drive; (2) increases synaptic norephinephrine (after seizures 3 to 5) leading to increase in energy and attention; and (3) increases serotonin function (after seizures 6 to 8) leading to positive change and loss of

negativity in cognition as well as resolved co-existing anxiety [Madsen et al., 2000; Nutt, 2008]. Electroconvulsive therapy has success rates of 60-80% and consists of 6-12 treatment sessions within 2-4 weeks. However, due to its high recurrence rate of 84%, it is usually followed by a pharmacology treatment, which has shown to reduce the recurrence rate to 39-60% in a sample population [Unutzer et al., 2007]. It is more beneficial for patients with psychotic depression, suicidal thoughts, or severe malnutrition. Most common side effects for this form of therapy include headache, temporary confusion, and or memory impairment.

For mild or moderate forms of depression, exercise programs have been shown to be effective. Exercising is known to release endorphins in the brain, which bind with neuronal receptors to reduce the perception of pain and elicit positive feelings resulting in a more positive and energized attitude towards life [Zetin et al., 2010]. However, sometimes LLD patients may find it difficult to participate in exercise programs, in which case they could also try other methods of treatment. This program requires about 12 weeks of participation under supervision in group-based aerobic exercises like walking. It has been shown to significantly decrease depressive symptoms in 45-65% of the patients in a sample population [Unutzer et al., 2007].

2.4.3 LLD Treatment Response and Potential Predictors

Not many recent studies have shown any results on the association of demographics and cognitive ability with the treatment response of LLD. One late-life study showed age and gender are not significantly related to treatment response [Katon et al., 2010], while another study showed that antidepressants are affective for ages 55 and up, but the effectiveness may reduces after 65+ years [Blazer et al., 2012]. Another study has also shown older age of onset in conjunction to early symptom improvement (based on change in Hamilton Rating Scale for

Depression) and lower baseline anxiety to be predictors of treatment response [Andreescu et al., 2008]. Similarly, a recent study has also shown higher baseline depressive symptom severity, smaller improvement of symptoms in the first two weeks after treatment, the male gender, a duration of current episode greater than two years, and sufficient past depression treatment to predict a lower probability of treatment remission [Joel et al., 2014]. In regards to cognitive ability, one study showed lower cognitive impairment (i.e. higher MMSE) may lead to a more positive treatment response [Ribeiz et al., 2013]. To the best of our knowledge, there are also no studies comparing years of education to the treatment response of LLD.

3.0 MAGNETIC RESONANCE IMAGING

This chapter gives a background understanding of magnetic resonance (MR) multi-modal imaging. It primarily describes how different MR modalities are acquired and methods used to extract essential brain structure and function measures from each modality. This chapter also describes potential MR imaging biomarkers of late-life depression and its treatment response based on past studies.

3.1 INTRODUCTION

Imaging is a non-invasive technique commonly used to study brain structure and function for normal biological process as well as pathology-related processes. It has been proven to be safe and is widely used for clinical purposes and research. For extracting brain structure and function measures, there are two commonly used imaging options: (1) a combination of Computer Tomography (CT) and Positron Emission Tomography, or (2) Magnetic Resonance Imaging (MRI). MRI is the focus of this chapter because its advantages outweigh the disadvantages for performing research studies attempting to gain a better understanding of the brain and model pathology. Disadvantages of MRI include its lesser availability and greater imaging time. However, unlike CT and PET, it does not involve the use of harmful radiation or a contrast agent, and offers a greater variety of modalities. This allows one to study the structure and

function of the brain more extensively without exposing participants to even marginally harmful chemicals [Butcher et al. 2010; Srinivasam et al., 2006].

3.2 NEUROIMAGING USING MRI MODALITIES

The most common MRI modalities used to study brain structure include T1-weighted imaging, T2-weighted imaging, and Diffusion Tensor Imaging (DTI). To study brain function, functional MRI (fMRI) is generally used. Each of the MRI modalities helps examine different aspects of the brain. T1-weighted images are used to study the differences and changes in cortical regions/structures, due to its high gray-white tissue contrast, which allows for more accurate labeling of gray matter regions and defining their boundaries. These images can be used to study the severity of atrophy in cortical regions by studying regional volume differences and changes. T2-weighted images are used to study white matter hyperintensities (WMHs), indicating the presence of ischemic or pre-ischemic white matter lesions. Both local and global volume measures of WMHs are used to study their affect on cognition. DTI images are used to gain an understanding of the brain from a microscopic level and study the diffusion of molecules in brain tissues. Two important measures acquired from DTI images include mean diffusivity (MD) and fractional anisotropy (FA), which signify the displacement and directionally of diffusion in tissue, respectively. These measures help evaluate the tissue integrity by helping determine cortical regions where diffusion is significantly decreased and dispersed due to lesions. Functional MRI images are used to study brain activity as well as functional connectivity between different cortical regions [Bihan et al., 2001; Blink, 2004; Vink, 2007].

3.3 MRI IMAGE COMPONENTS

An MRI image is usually 3-dimensional. To form the 3D MRI image, multiple 2D images—each representing a slice of the brain—are concatenated in the same order as their location in the brain. The plane in which the brain is sliced during image acquisition is pre-defined by the user. Since the images are 3D, they are composed of voxels instead of pixels. Voxels are 3D pixels. Each voxel of the image is given an intensity value. The intensity value is based on tissue characteristics and imaging modality (see the "MRI Image Acquisition for Neuroimaging" section for more details). For a visual representation of the image, the intensity value is translated into a gray scale color [McRobbie et al., 2007].

3.4 MRI IMAGE ACQUISITION FOR NEUROIMAGING

Different MRI modalities described earlier are acquired by using different image acquisition pulse sequences. Different pulse sequences can be acquired by varying certain image acquisition parameters. These details along with a generalized version of an image acquisition protocol are provided in this section.

3.4.1 Generalized Image Acquisition Protocol

MRI takes advantage of the large presence of protons in the body—from water, fat tissue, etc.—and their magnetic properties. To acquire an image, a MRI scanner (with a superconducting magnet and gradient coils), a transmitter radiofrequency (RF) coil, and a receiver RF coil—

which may be combined with or separate from the transmitter coil—are used. For neuroimaging, the head is the body part of interest to be scanned for an image of the brain. The procedure for image acquisition of the brain is as follows (Note: for the procedure, the directions are defined in terms of the Cartesian coordinate system as follows: X-direction is from left to right of scanner, Y-direction is from bottom to top of scanner, and Z-direction is along the center of the scanner from the foot to the head) [Blink, 2004; Hornak, 1996]:

1. The head is enclosed in a RF coil and placed inside the scanner.
2. The superconducting magnet of the MRI scanner applies a strong uniform magnetic field (known as the B_0 field) to align all the protons in the same direction—the Z-direction. The number of protons that fully align with the B_0 field depends on the corresponding tissue characteristics. The aligned protons, in addition to spinning, also start precessing (i.e. wobbling like a spinning top).
3. Slice encoding/selection is performed by:
    3a. Switching on the Gz gradient coil of the MRI scanner to apply a gradient magnetic field in the Z-direction. This gradient causes protons at different locations along the Z-direction to precess at different frequencies.
    3b. Using the transmitter RF coil to apply an RF pulse at a pre-defined frequency to generate a weak magnetic field (B_1 field), which is at a pre-defined angle (e.g. 90^0) from the B_0 field. The B_1 magnetic field only excites protons at the location along the Z-direction that are precessing at the same frequency as the transmitted pulse. The excited protons then align with the B_1 field.
    3c. Switching off the Gz gradient coil after a slice of the brain in the X-Y plane at a specific location along the Z-direction has been selected via RF pulse excitation.

4. Phase encoding is performed by:

    4a. Switching on the Gy gradient coil of the MRI scanner to apply a gradient magnetic field in the Y-direction. This gradient causes protons at different locations along the Y-direction to precess at different frequencies.

    4b. Switching off the Gy gradient coil to cause the protons to precess at same frequencies as before but be out-of-phase (i.e. out of sync) with each other. In other words, now the protons at different locations along the Y-direction are precessing with different phases.

5. Frequency encoding and signal readout are performed by:

    5a. Switching on Gx gradient coil of the MRI scanner to apply a gradient magnetic field in the X-direction. This gradient causes protons at different locations along the X-direction to precess at different frequencies.

    5b. Turning on the receiver RF coil to read the emitted RF waves of the excited protons in the process of relaxing back to align with B_0 field. The emitted RF waves are received by the RF coil in the form of a signal with a complex mixture of frequencies, phases, and amplitudes.

        - During the relaxation process, the protons experience 2 forms of independent yet simultaneous relaxations: T1 relaxation and T2 relaxation. T1 relaxation—i.e. spin-lattice relaxation—is the recovery of the net magnetization of all the protons in the given slice back to Z-direction of the B_0 field. This relaxation occurs because of the strong magnetic B_0 field forcing the protons to re-align along the Z-direction. T2 relaxation—i.e. spin-spin relaxation—is the decay (i.e. becoming

out-of-phase) of the protons away from the direction of the previously applied B_1 field in the X-Y plane. This relaxation occurs because of a sudden absence of a magnetic field in the X-Y plane causing the protons to scatter in all different directions or de-phase.

6. Steps 3-5 are repeated for varying amplitudes of phase encoding gradients to gather information about the rate of change of phases. This will help gather enough information to sufficiently distinguish locations of received signals along the Y-direction.
7. Frequency and Phase information acquired for each slice is organized in a k-space image. Each row of the k-space image represents the data collected from each repetition of step 6. Also, the k-space image is organized such that at the frequency value at the center of k-space image is zero. The phase is also zero at the center of k-space image because the image is real.

    - There are many different k-space filling techniques. The easiest to understand is the linear methods, in which the acquired data from each repetition of step 6 is filled from top to bottom. Another variation is filling the information from center out. The fastest technique is to use Echo Planar Imaging. For this technique, multiple phase encoding gradients are consecutively applied after the same excitation pulse and a read-out of the resulting change in signal is acquired each time. This technique has a poorer spatial resolution, but requires fewer repetition of step 6 and makes the whole process faster.

8. Steps 3-7 are repeated to select different slices along the Z-direction in step 3 until information from every location of the head—or a pre-defined specific portion of the head—is acquired.
9. Two-dimensional Fourier transforms are applied to each 2D k-space image formulated in step 7 to acquire 2D images of the brain slices selected in the corresponding step 3. All the 2D images are compiled to form a 3D image of the brain. The intensity value at each voxel in the image is determined by the signal's amplitude—which reflects the number of protons emitting energy—and the spatial distribution of the intensity values is determined by the signal's frequency and phase—which reflect the location on the body part of interest.

3.4.2 Important Image Acquisition Parameters

In the process of acquiring MR images, several acquisition parameters can be altered to control either image properties or representation.

Primary parameters that control image properties include: field of view, resolution, and slice thickness. All three of these parameters adjust appropriate settings of the gradient coils in the MR scanner to alter the overall resolution of the acquired image. The field of view parameter affects the overall size of the image. The resolution parameter affects the image's voxel size in the X-Y plane (determines the number of repetitions of step 6), while the slice thickness parameter affects the image's voxel size in the Z plane of the Cartesian coordinate system (determines the number of repetitions of step 8).

Primary parameters that control image representation include: repetition time, echo time, inversion time, and flip angle. Repetition time (TR) is the time between each repetition of RF

pulse transmission by the transmitter RF coil (step 3b). Thus, it determines the time allotted to the protons for T1 relaxation to flip back to the B_1 field. Echo time (TE) is the amount of time between the RF pulse transmission (step 3b) and data acquisition (step 5b). Thus, it determines the time allotted to the protons for T2 relaxation for them to become de-phased. Together the TR and TE will determine whether the image acquired is a T1-weighted, T2-weigthed, or Proton Density image (see "T1-weighted vs. T2-weighted vs. Proton Density Imaging" section for more details). The inversion time (TI) is the time allowed for T1 relaxation in the inversion recovery sequence before the spin echo sequence is applied. For the inversion recovery sequence, the echo time begins with the initiation of the incorporated spin echo sequence. The flip angle is the angle at which the protons are flipped (i.e. the magnetization vector is rotated) towards the X-Y plane by the transmission of the RF pulse in step 3b of the above-described procedure. It is mostly 90^0 for the spin echo and inversion recovery sequences, but for the gradient echo sequence it varies within a range of 1^0 to 90^0 [Blink, 2004; Hornak, 1996].

3.4.3 Image Acquisition Pulse Sequences

The above-described procedure is the general protocol used to acquire MRI images. However, there are several variations that allow for improved image acquisition. Three well-known acquisition pulse sequences are spin echo, gradient echo, and inversion recovery sequences.

The spin echo sequence requires the transmitter RF coil to apply an additional re-phasing RF pulse of 180^0 between the transmission of the 90^0 RF pulse (step 3b) and signal read out (step 5b). This additional RF pulse causes the de-phasing protons during T2 relaxation to re-phase and emit a stronger signal to the receiver RF coil, and in the process, compensating for local magnetic field inhomogeneities.

The gradient echo sequence, instead of applying an additional 180^0 RF pulse, applies a gradient polarity reversal using the frequency encoding gradient coil of the MRI scanner to re-phase the protons during T2 relaxation. The gradient polarity reversal requires the frequency encoding gradient coil to first apply a negative polarity, followed by a positive polarity. Thus, the protons precessing faster during the negative polarity begin to precess slower in the positive polarity and eventually all the protons re-phase to emit a stronger signal to the receiver RF coil. The gradient echo sequence is much faster than the spin echo and inversion recovery sequences, because it uses small flip angles ($< 90^0$) and very short recovery times. Nevertheless, it does not compensate for local magnetic field inhomogeneities leading to more artifacts in the image.

The inversion recovery sequence is the same as the spin echo sequence except for the addition of an extra 180^0 RF pulse that starts the sequence. This 180^0 RF pulse at the beginning of the sequences allow for a large T1 relaxation period with no T2 relaxation since the protons are not flipped to the X-Y plane, but are instead flipped to the negative Z-direction. Following a long period of T1 relaxation, the spin echo sequence is applied within a short time frame during which the emitted signal is read by the receiver RF coil. This sequence takes much longer than the spin echo sequence since the T1 relaxation takes twice as long, but it has a greater T1 contrast—i.e. allows for greater tissue distinction based on the T1 relaxation. The inversion recovery sequence is used to suppress a certain tissue type or cerebrospinal fluid (see "T1-weighted vs. T2-weighted vs. Proton Density Imaging" and "T2-weighted Fluid Attenuated Inversion Recovery (FLAIR) Imaging" sections for more details) [McRobbie et al., 2007].

3.4.4 Imaging Modalities

3.4.4.1 T1-weighted vs. T2-weighted vs. Proton Density Imaging

T1-weighted images are acquired to delineate anatomical structures and study pathology, where as T2-weighted and Proton Density images are acquired primarily to study pathology.

The differences between the T1-weighted, T2-weighted, and Proton Density imaging are based on the differences in T1 and T2 relaxation among various tissue types. Additionally, the T1 and T2 relaxation time allotted to the protons of all the tissue types are controlled by the TR and TE parameters respectively as mentioned earlier. Thus, to acquire the desired imaging modality—T1-weighted, T2-weighted, or Proton Density—the appropriate TR and TE parameters must be defined.

In regards to T1 relaxation, the differences between tissues are as follows: (1) fluids like the cerebrospinal fluid and urine have a slow T1 relaxation rate, (2) tissue like gray matter has a medium T1 relaxation rate that is slower than the T1 relaxation rate of white matter, and (3) the most fibrous tissues like white matter, tendons, and fat have fast T1 relaxation rate. A greater T1 relaxation leads to a greater energy emission from the protons. Consequently, greater energy emission leads to a larger intensity value—and brighter gray-scale color—observed on the acquired image. Thus, when the image acquisition sequence allows for a shorter period of T1 relaxation (i.e. short TR), the different tissue types are more distinguishable in the acquired image-based variations in intensity values. However, if the image acquisition sequence allows for a longer period of T1 relaxation (i.e. long TR), the protons from all tissue types reach full relaxation and thus they have similar intensity values making it harder to identify them on the image.

In regards to T2 relaxation, the differences between tissues is as follows: (1) fluids have a slow T2 relaxation rate; (2) tissues like gray matter have a T2 relaxation rate more comparable to the T2 relaxation rate of white matter than fluids; and (3) the most fibrous tissue like muscle and fat have a fast T2 relaxation rate. A greater T2 relaxation leads to poorer strength of the signal emitted from the protons since they are more out-of-phase. Consequently, poorer signal strength leads to a smaller intensity value—and darker gray-scale color—observed on the acquired image. Thus, when the image acquisition sequence allows for a longer period of T2 relaxation (i.e. long TE), the different tissue types are more distinguishable in the acquired image based variations in intensity values. However, if the image acquisition sequence allows for a shorter period of T2 relaxation (i.e. short TE), the protons from different tissue types have not been given enough time to become sufficiently out-of-phase; thus the tissues have similar intensity values and cannot be easily distinguished.

Therefore, to acquire a T1-weighted image, a short TR and short TE must be used. This allows for the tissue contrast to be weighted by the T1 contrast (i.e. T1 relaxation differences amongst the tissues). Typically, these images have the highest contrast between the cerebrospinal fluid, gray matter, and white mater. The cerebrospinal fluid appears the darkest and the white matter appears the brightest on the acquired image. Due to the greater contrast between tissues observed in these images, they are acquired with a high resolution to outline anatomical structures and study tissue atrophy.

To acquire a T2-weighted image, a long TR and TE must be used. This allows for the tissue contrast to be weighted by the T2 contrast (i.e. T2 relaxation differences amongst the tissues). Typically, these images are better for distinguishing the cerebrospinal fluid from the gray and white matter. The gray and white matter cannot be separated as well as the gray matter

appears only slightly brighter than the white matter. The cerebrospinal fluid appears to be the brightest on the acquired image. Due to the greater distinction between fluids and other tissues observed in these images, they are acquired with a sufficiently high resolution to study pathology—which also appears bright like the cerebrospinal fluid—that may be associated with excess water in tissue (e.g. edema), loss of fibrous tissue, degeneration of myelin resulting in axons with greater intracellular and extracellular water content, gliosis, and/or infarction.

To acquire a Proton Density image, a long TR and a short TE must be used. These images are thus independent of T1 and T2 relaxation. They are instead dependent on the number (i.e. density) of protons in the tissue. Tissues with a greater number of protons appear brighter in the image. This image modality is also used to study pathology that may be better observed than on the T2-weighted images [Blink, 2004; McRobbie et al., 2007; Wahlund et al., 2001].

3.4.4.2 T2-weighted Fluid Attenuated Inversion Recovery (FLAIR) Imaging

T2-weighted FLAIR images are acquired to better isolate age- or pathology-related tissue lesions—similar to T2-weighted imaging—from the rest of the brain regions. They are acquired using an inversion recovery image acquisition sequence with a long TE. The inversion recovery sequence is used to suppress the cerebrospinal fluid signal, while the long TE is used to acquire T2-weighted tissue contrast for highlighting the tissue lesions. To suppress the cerebrospinal fluid, the spin echo sequence is initiated when the protons in the cerebrospinal fluid have relaxed 90^0 and are aligned with the X-Y plane during the inversion time. This is measured to occur at approximately 2000ms (i.e. TI = 2000ms). By initiating the spin echo sequence at this time, there are no protons in the cerebrospinal fluid along the Z-direction to be flipped on to the X-Y plane, and consequently there is no signal emission from the cerebrospinal fluid. Thus, the acquired image looks like a T2-weighted image, but the cerebrospinal fluid appears dark. This allows the

pathology-based ischemic white matter regions to appear more distinct and pronounced with hyperintensities (i.e. larger intensity values and brighter in color) compared to the rest of the image content [Blink, 2004; McRobbie et al., 2007].

3.4.4.3 Diffusion Tensor Imaging (DTI)

DTI measures are estimated using raw diffusion-weighted imaging. Diffusion-weighted imaging is used to characterize and track the 3D diffusion of water within the body. It can be used to determine tissue integrity by identifying development-, age- or pathology-related—e.g. ischemic stroke, demyelination, inflammation, edema, neoplasia, etc.—altered diffusion of water with the tissue. Anatomically, these alterations in water diffusion may be indirectly caused by changes in tissue microstructure and organization.

To acquire diffusion-weighted images, a pulsed gradient, spin eco sequence is used. This sequence is similar to the spin echo sequence except for the addition of a pair of large gradient pulses with a pre-defined direction placed on both sides of the 180^0 re-phase RF pulse. The first of the two gradient pulses de-phases the magnetization, while the second pulse re-phases it. This helps determine the amount of water diffusion that occurred in the direction of the gradient pulses. If the water molecule were stationary, the phases induced by the two gradient pulses would cancel each other out resulting in a stronger signal emission. Thus, greater amounts of diffusion would lead to greater signal attenuation. This protocol is repeated at least 6 different times. Each time the pair of gradient pulses is configured to a different direction and a new 3D diffusion-weighted image is acquired. To more accurately track the diffusion of water, it is beneficial to gather information from more directions by acquiring more diffusion-weighted images with directionally varying gradient pulses. To improve the signal-to-noise of DTI measures, one acquires and averages over multiple images for each direction.

Diffusion-weighted imaging acquisition also widely involves the use of echo planar imaging. The advantages of echo planar imaging are that it is fast, efficient and insensitive to small motion. This is beneficial since diffusion-weighted imaging is very sensitive to head motions.

A problem with diffusion-weighted imaging, though, is that diffusion-gradient eddy currents can lead to misalignment of diffusion-weighted images. This issue is addressed during image processing of diffusion-weighted images [Alexander et al., 2007].

3.4.4.4 Functional MRI (fMRI)

Functional MRI (fMRI) is thought to indirectly measure neuronal activity. Both spin echo and gradient echo sequences have been tested for acquiring fMRI images. Most studies have used gradient echo sequences, but some studies have shown that the spin echo sequence may be advantageous [Budde et al., 2014; Miyapuram et al., 2009]. Using the desired pulse sequence, multiple 3D brain images are acquired over a period of time to create 4D fMRI data where the 4^{th} dimension represents time. Thus, at every voxel of the 3D brain volume, a time series thought to indirectly measure neuron activity is captured.

Primarily, fMRI is acquired using blood-oxygen level dependent (BOLD) contrast. The process of measuring neuronal activity starts with performance of a task, followed by corresponding change in regional neural activity, and ends with a change in the acquired MR signal. If the performed task increases the regional neural activity, the MR signal is also increased in that region in the following order of BOLD-related events: (1) increase in neuronal activity, (2) increase in local cerebral blood flow, (3) decrease of deoxyhemoglobin compared to oxygenated hemoglobin (Note: deoxyhemoglobin, unlike oxygenated hemoglobin, is paramagnetic and creates magnetic field distortions by altering local magnetic susceptibility, thus

reducing strength of received MR signal), and (4) increase in strength (i.e. amplitude) of the received MR signal due to a greater alignment of the spins with the applied magnetic field. Thus, acquired MR BOLD signal reflects the change in blood flow, which is related to the change in regional neural activity. The blood flow and consequently measured BOLD response to a brief stimulus (i.e. excitation of a neuron) is known as a hemodynamic response. It consists of a 1-2 seconds delay, potentially a 1-2 seconds initial dip, a 4-6 seconds response in which it increases and then decreases, and a post-stimulus undershoot which may be 30 seconds or more. By evaluating BOLD-related presence of hemodynamic responses in the MR signal, studies are able to map patterns thought to indirectly represent neuronal activity during specific tasks and resting state [Buxton et al., 2004]. Tasks utilized for task-based fMRI and networks studied for resting state fMRI are described below. A variant of BOLD fMRI is arterial spin labeling perfusion MRI [Detre et al., 2009].

Task-based fMRI

One of the simplest and most common tasks used in fMRI studies is the finger-tapping task. It requires the participant to tap their fingers while being scanned and is used to study the human motor system [Witt et al., 2008]. Specific to late-life depression, the common tasks include: executive tasks, working memory tasks, and affective tasks.

Executive tasks involve executive control, such as integrating other cognitive activities. The prefrontal cortex plays a prominent role in the performance of executive tasks. Due to the involvement of the prefrontal cortex, executive tasks help analyze reflecting conscious, strategic, and goal-directed cognitive activity [Bryan & Luszcz, 1999]. Examples of executive tasks used in functional MR studies include: the Stroop task—which is known to test response inhibition, interference resolution, and behavioral conflict resolution [Adleman et al., 2001], and task-

switching tests—which help test cognitive efficiency, processing speed, performance, as well as controlling and coordinating execution of goal-directed behavior [Dove et al, 2000; DiGirolamo et al., 2001; Sylvester et al., 2003];

Working memory is a function of the brain that involves temporary storage and manipulating information for complex cognitive tasks like language comprehension, learning and reasoning. Working memory has been subdivided into three subcomponents: (1) central executive system responsible for attentional-control system, (2) visuospatial sketchpad responsible for manipulating visual information, and (3) phonological loop responsible for storing and rehearsing verbal information. The central executive subsystem also integrates and controls the other two subsystems. The regions associated with these subcomponents respectively include: (1) prefrontal cortex for updating or actively maintaining visual and/or verbal stimuli, (2) bilateral parietal and occipital cortices for perceptual processing of visual stimuli, and (3) left inferior frontal gyri and left inferior parietal cortex (specifically left supramarginal gyrus) for processing verbal stimuli. Functional MRI tasks used to study working memory involve remembering and later recalling verbal and/or visual stimuli. An example of such tasks includes item-recognition tasks, which help test accurate recollection of visual and/or verbal stimuli [Baddeley, 2008; Na et al., 2000; Salmon et al., 1996].

Affective tasks are used to evaluate neural systems associated with mood and emotion. Commonly associated regions with affective processes—e.g. emotion regulation—include limbic regions, such as the hippocampus, amygdala and anterior cingulate cortex. Also, because of its role in affect regulation, the prefrontal cortex (especially ventral prefrontal cortex) is also often studied with affective tasks. Functional affective tasks used to study these regions involve arousal and/or control of emotion and mood (e.g. emotional faces task) [Davidson et al., 2002].

Resting State fMRI

There are several resting state networks identified by fMRI studies. These are networks of brain regions which show correlated functional activity while a participant is resting (i.e. not performing a task). To identify these regions, the participants are asked to lie in the scanner, thinking of nothing in particular and resting with eyes focused on a fixation cross. Regions found to be significantly active with highly correlated BOLD signals during the resting state are defined to form a resting state network. Well-studied resting state networks include the default mode network and the salience network.

The default mode network mainly consists of the medial prefrontal cortex, posterior cingulate cortex, medial temporal cortex, inferior parietal lobule, and hippocampus. The involvement of this network in functional processes is still up for debate, however, there are two possibilities: 1) it plays a role in forming dynamic internal mental images of perspectives and scenarios not related to the present while the mind is detached from the external world (e.g. planning future actions based on past experiences), or 2) it plays a role in monitoring the external environment in an exploratory manner whenever focused external attention and sensory processes are relaxed [Buckner et al., 2008; Greicius et al., 2009].

The salience network primarily consists of the anterior insula and anterior cingulate cortex. The function of this network is to identify most relevant internal (e.g. self-regulated cognition) and extrapersonal (e.g. attention) stimuli for guiding behavior [Menon & Uddin, 2010]. Additionally, the salience network may also play role in modulating the activity of other large-scale networks, including the default mode network [Chiong et al., 2013].

3.5 MRI IMAGE ANALYSIS FOR NEUROIMAGING

Image analysis techniques are used to acquire various summary measures effectively and with reduced bias from MR images during image processing. Essential image analysis techniques include filtering, registration, and segmentation.

3.5.1 Filtering

Filtering is used to remove artifacts (i.e. reduce noise) from acquired images. It is generally used as a pre-processing step, because removing artifacts helps improve the accuracy of other image analysis and processing methods. Artifacts are present in medical images potentially due to imperfect hardware characteristics (e.g. gradient nonlinearities, concomitant gradients, timing errors, RF field non-uniformity, and limited dynamic range), resonant offsets (e.g. B_0 field inhomogeneity, magnetic susceptibility, and chemical shift), intrinsic tissue properties and biological behavior (e.g. respiration, cardiac cycle), and voluntary patient motion [Smith & Nayak, 2010]. To remove these artifacts, common MR neuroimaging preprocessing steps involve bias field correction, smoothing, and/or sharpening. For 4D medical images (e.g. fMRI, DTI), additional forms of filtering and de-noising are required. Though not a filtering method, realignment is also used in 4D medical imaging to de-noise spatial artifacts due to head motions and thus is also mentioned in this section. Also, for fMRI images used to study time series, additional 2D signal-processing filters (e.g. low-pass, high-pass, band-pass) may be used.

Bias field correction methods are used to reduce the low frequency intensity inhomogeneity artifacts (i.e. bias field) present in MRI images during scanning due to poor radio frequency coil uniformity, eddy currents driven by the gradients applied by the scanner, and the

participant's anatomy. A popular method used to remove these artifacts is called nonparametric nonuniform intensity normalization (N3). N3 requires the selection of only two parameters: (1) defining the smoothness of the bias field to be estimated and (2) controlling for the accuracy versus convergence rate tradeoff. Using these pre-defined parameters and an iterative approach, N3 estimates and the underlying bias field from the noisy image acquired during scanning. Then, it divides the original noisy image by the estimated bias field to produces a new image with the inhomogeneity artifacts filtered out [Sled et al., 1998].

Smoothing filters are used to reduce sharp changes in intensities (i.e. high frequencies in the signals), which make the image look grainy. This filter blurs the boundaries and conceals subtle details of the structures in the images. The most commonly used smoothing filter is the Gaussian blurring filter. It recalculates each voxel's intensity value by taking into account neighboring voxels' intensity values, giving the greatest consideration to most immediate neighbors. In other words, it performs a weighted average where the weightage depends on distance—shorter distance equals greater weight—from the voxel whose intensity is being recalculated. By taking the neighbors' intensity values into consideration, the filter makes sure all sharp changes are blurred out to produce a smoother signal. There are several smoothing filter variants, which differ in the weights' distribution (i.e. kernel function) and combination (e.g. averaging or computing the median) to recomputed intensity values at each voxel [Lee, 1983].

Sharpening filters are opposite in function to the smoothing filters. They are used to accentuate sharp changes in intensities, like those at edges (i.e. boundaries) in the image. This exaggerates the contrast at the boundaries of the structures in the images. The simplest sharpening filter method is to (1) high-pass filter the image (i.e. to extract the high frequencies in the signals that represent boundaries), (2) scale the high-pass filtered image to a desired amount

of sharpening, and (3) add the scaled version of the high-pass filtered image to the original image. However, this process is very sensitive to noise due to the high pass filtering [Polesel et al., 2000]. Thus, most commonly used sharpening filter is the unsharp filter. It does the following: (1) creates a new blurred version of the image using a smoothing filter (i.e. a new image containing the low frequencies of the original image), (2) subtracts it from the original image to create a new image with only high frequencies that appear as edges, (3) scales the new image by the desired percentage of sharpening of the edges, and (4) adds the new scaled image of only high frequencies to the original image, thereby enhancing the affects of the higher frequencies in the image. Variants of this filter include an adaptive unsharp filtering method [Singh, 2013].

Medical images that are 4D consist of multiple 3D volumes of the same brain with the 4^{th} dimension in functional images often representing time. Since there is a time component involved, there is also a greater risk of head movement in between resulting in anatomical artifacts—i.e. incorrect alignment of brain regions. To correct this, realignment is used. Realignment is simply a rigid-body linear registration (see "Registration" section for more details) used to align all 3D volumes across the dimension of time to the first 3D volume acquired [Friston et al., 1996]. Additionally, for fMRI images, the time series at every voxel is also shown to contain cardiac, respiratory, and other scanner-related low frequency noise. Thus, a high pass filter is usually used to filter out this noise. For resting state studies, the time series signal is further filtered with a low pass filter to extract the signal specific for the resting-state frequency range [Weissenbacher et al., 2009].

3.5.2 Registration

Registration is used to transform one image—known as the moving image—to match the spatial information of another image—known as the fixed image. There are two broad categories of registration: area-based and feature-based. Area-based registration uses the variations in intensities across images to perform registration, while feature-based registration uses features of structures within images (e.g. end points or centers of line features, centers of gravity of regions, etc.) to perform registration. Area-based registrations methods are more commonly used for neuroimaging related MR image analysis [Zitova & Flusser, 2003]. Thus the focus of this section will be on area-based registration.

Area-based registration is an iterative process that, given two input images (i.e. fixed and moving images), iterates through the following: (1) a metric is used to compare the two images and thus determine a cost function (i.e. objective function) to minimize the difference between the them, (2) an optimizer is used to optimize the cost function and thus determine transform parameters, (3) a transform is used to map points from the moving image to the fixed image using the transform parameter, and (4) an interpolator is used to interpolate voxel values of the transformed fixed image that are not exactly mapped to original grid positions. This iterative process stops when the change in transformation between consecutive iterations becomes minimal. Thus, the four important components of the area-based registration process include the metric, optimizer, transform, and interpolator [Johnson et al., 2013].

The selection of a metric depends on the image types of the fixed and moving images. If they are of the same image type and thus have a similar intensity distribution, common metrics include: least squares and normalized correlation. To compare the input images, the least squares metric computes sum of squared differences and the normalized correlation metric computes

correlation coefficients between their intensity values at every voxel respectively. The optimizer will need to minimize the cost function for both metrics. However, if the fixed and moving images are of different image types (i.e. multi-modal images), then the common metrics include: mutual information and correlation ratio. These metrics do not directly look for linear relationships between intensity values, but focus more on the dependency between intensity values of the two images since different image types, particularly among medical images, have different intensity distributions for the same structures. Thus, to compare the input images, the mutual information metric determines how much uncertainty about one image's intensity values is reduced by the knowledge of the other image's intensity values; and the correlation ratio metric measures the relationship between the statistical dispersion of various regions—in regards to the number of voxels—within each individual image and across both images. The optimizer will need to maximize the cost function for both metrics [Jin & Yang, 2013; Johnson et al., 2013; Roche et al., 1998].

There are primarily two types of optimizers: continuous and discrete. Continuous optimizers are limited to problems that involve real-valued transformation parameter values and a differentiable cost function. These methods estimate optimal transform parameters using an iterative process that starts with a best guess estimate of the transform parameter. Then, the transform parameters are updated every iteration based on a computed search direction and step size—which controls the amount of change of the optimal transform parameters in the computed direction. The search direction is re-computed every iteration using the cost function and an optional regularization term—i.e. which imposes constraints on the transformation based on prior knowledge. The step size can also be recomputed every iteration if preferred. The approach used to compute the step size and search direction every iteration is what distinguishes various

continuous optimization methods. Most common continuous optimization methods include gradient descent, conjugate gradient, Powell's conjugate directions, Quasi-Newton, Gauss-Newton, Levenberg-Marquardt, and stochastic gradient descent. On the contrary, discrete optimizers are limited to problems that involve discrete-valued transformation parameter values. These methods use graph representation to compute optimal transform parameters (i.e. Markov Random Field formulations). The nodes (i.e. vertices) of the graph represent the parameter values, while the edges connecting the nodes represent the similarity costs based on variations in labels of adjacent nodes. The goal of these methods is to determine the optimal label for each node from a predefined set of label options by optimizing the sum of the edge costs. There are three common types of discrete optimization methods: graph-based, message passing, and linear-programming [Sotiras et al., 2013; Zikic et al., 2010].

The selection of the transform depends on the degree of variations between the two input images. A linear transformation—which is computationally less expensive—may suffice for a lesser degree of variations (e.g. two different images of the same individual subject), while a nonlinear transformation may be required for a greater degree of variations (e.g. between a template and individual subject image). Linear transformations are restricted by the pre-selected degrees of freedom (i.e. the flexibility in deformation). The degrees of freedom may be increased as the degree of variations between the two input images increases. The number of degrees of freedom and their representations in relation to the three axes of the 3D Cartesian coordinate system are as follows: 3 degrees of freedom for translation along the three axes, 3 degrees of freedom for rotation about the three axes, 3 degrees of freedom for scaling along the three axes, and 3 degrees of freedom for skewing along the three axes. Well-known linear transformations include the rigid body (6 degrees of freedom in translation and rotation) and affine (12 degrees of

freedom in translation, rotation, scaling, and skewing) transformations [Hill et al., 2001]. A 3x3 matrix is often used to represent such transformations. On the other hand, nonlinear transformations have much greater degrees of freedom—up to millions of degrees of freedom—and have the greatest flexibility with fewer constraints; thus better for tackling larger deformations. Non-linear transformations include curved or elastic transformations, which can map lines to curves. Nonlinear transformations are utilized for deformable registration models including elastic body models, viscous fluid flow models, diffusion models (e.g. Demons), curvature registration, and flows of diffeomorphisms. Either local vector displacement fields or polynomial transformations in terms of original coordinates are often used to represent such transformations [Maintz & Viergever, 1998; Sotiras et al., 2013].

The selection of the interpolator depends on the degrees of freedom selected for the transformation and precision required for the recomputation of a transformed image's intensity values. Generally, with an increase in the degrees of freedom for the transform, there is also a need for more precision in the interpolation. There is also trade-off between precision and computation time. Common interpolators in order of increasing precision include nearest neighbor, trilinear, and B-spline. Nearest neighbor interpolation assigns each new voxel of the transformed image the intensity value of the spatially closest voxel from the original image before transformation. This is most useful when the original set of unique intensity values are better left unchanged—e.g. for the transformation of binary images (e.g. region masks) [Parker et al., 1983]. Trilinear interpolation is a linear interpolation method for 3D images. To determine the intensity values at each voxel of the transformed image, trilinear interpolation takes a weighted average of intensities from the nearest eight neighboring voxels of the original image. The weighting of the intensities is inversely proportional to the distance of the corresponding

voxel from the original image to the new voxel of interest in the transformed image [Hill et al., 2001]. A linear interpolation is a B-spline interpolation of the first order [Thevenaz et al., 2000]. Thus the B-spline interpolation is similar to the trilinear interpolation. The differences are that the B-spline interpolation (1) uses polynomial functions, instead of a linear function, to weight the intensities of neighboring voxels, and (2) incorporates intensities from a larger neighborhood of voxels from the original image to compute the intensity value at each new voxel of the transformed image [Mahmoudzadeh & Kashou, 2013].

3.5.3 Segmentation

Segmentation is used to isolate or classify regions of interest. It can be performed on the basis of image properties. Most commonly used image properties in MR neuroimaging include intensities, gradient, energy, region, shape, etc. Each image property is associated with different segmentation methods described below. Furthermore, these image properties in addition to others like texture are also used in combination with machine learning techniques for segmentation.

For medical images, intensity-based segmentations are the most common. The simplest intensity-based segmentation technique is thresholding, which zeros all voxels with an intensity value that does not meet the thresholding criteria. Another well-known technique is region-growing, which extends thresholding by also considering the connectedness of regions. Region-growing starts with a seed and iteratively grows by included neighbor voxels that meet a homogeneity criteria. It stops growing when there are no more neighboring voxels that meet the homogeneity criteria, thus segmenting a connected region. Another common variant of intensity-

based segmentation techniques is fuzzy connectedness [Balafar et al., 2010, Johnson et al., 2013].

A well-known gradient-based segmentation method is the watershed technique. A gradient is the rate of change of intensity values. For the watershed technique, the gradient value at each voxel is represented as a height measurement. Thus, the boundaries in an image—where the greatest change in intensity values occur (i.e. gradient value is the largest)—will represent local maximum heights. On the other hand, the homogeneous connected regions—i.e. regions with similar intensity values and low gradient values depicting the same structure within the image—will represent local minimum heights. These represent the regions to be segmented. Together, the boundaries form the watersheds and the regions to be segmented form catchments basins. Then each region separated by the watersheds is segmented as a separate structure and assigned the average intensity of the region [Balafar et al., 2010].

Energy-based segmentation methods include active contours. For implementing active contours, there are two principal techniques: snakes and level sets. Both start off with an initial estimate of a contour representing the structure of interest to be segmented and then grow the contours to more accurately segment the structure. Snakes grow active contours by trying to minimize the sum of the internal and external energy. The internal energy controls the rigidity of the deforming curve and increases when the curvature increases. The external energy guides the deforming curve to the target and increases when the gradient in the image decreases (i.e. when the contour edge is in a homogenous region as opposed to the structure boundary). In doing so, the contours are attracted toward the desired structure's boundaries. Level sets grow active contours with a more implicit approach that involves minimizing scalar function. Nevertheless, level sets also grow the active contours by taking into consideration the mean curvature of the

contour and the gradient information in the image. The mean curvature of the contours controls the speed of growth of the contour. The more curved the contour is, the faster it grows. The gradient information in the image determines when the contour stops growing. If the gradient is high enough (i.e. structure boundary is reached), then the contour stops growing [Maistrou, 2008].

Regions-based segmentations use pre-defined masks (i.e. binary images) of a region of interest to isolate or classify it. The technique simply multiplies the given region mask with the image, thereby zeroing out all voxels except those identified as the region of interest by the mask. The difficulty with this technique is to create the region masks. One tedious way to do it is to manually draw the images. Another way to do it is to use automated template to individual image registration and apply the transformation to the regions of interest in template space (see "Registration" section). This results in less bias and greater efficiency. Also, the regions will only need to be manually traced once on the template. The disadvantage to this technique is that its accuracy is depends on the accuracy of the registration [Rosano et al., 2005; Wu et al., 2006].

Shape-based segmentations use prior shape knowledge to perform segmentation. For MRI image analysis, this has generally been done using machine learning techniques (see chapter 4). In short, prior shape knowledge has been provided via a large ranging of similar segmentation examples from past images to predict the segmentation of the region of interest from a new image. Variants of shape-based segmentation methods have been proposed by Abd El Munim et al. (2005), Rousson et al. (2004), and Tsai et al. (2004).

Machine learning can also be used in different ways to perform segmentation. Information used to help achieve a good segmentation may include any of the above-defined image properties as well as texture information. Additionally, examples of benchmark

segmentations (e.g. manual segmentations) from similar images can also be inputted to help segment new images using supervised machine learning techniques. If no benchmark segmentation examples exist, unsupervised machine learning techniques (e.g. clustering methods) can be used instead. These techniques require only the input of image properties from the new image to be segmented. See chapter 4 for more details on these techniques. Note that for these techniques, image properties denote the input features and the image properties of each voxel in the image denote each individual data instance. Also for the supervised machine learning techniques, the benchmark segmentation classification of each voxel in the image denotes the corresponding label value [Kruggel et al., 2008; Punia & Singh, 2013].

3.6 MRI IMAGE PROCESSING FOR NEUROIMAGING

Image processing is performed on an image to obtain valuable information about the brain. The image processing performed varies for different imaging modalities since different modalities provide information about the brain in different ways. More specific examples of image processing pipelines are described in the chapters 5-7.

3.6.1 T1-weighted, T2-weighted, Proton Density, & T2-weighted FLAIR Imaging

Image processing is performed on T1-weighted, T2-weighted, Proton Density and T2-weighted FLAIR imaging to acquire volume-based information. For T1-weighted imaging, the information acquired is the volume of anatomical structures, while for the other images it is the volume of pathology-related abnormalities (e.g. ischemic lesions). These volume measures are acquired via

segmentation of the regions of interest. Sometimes, for segmentation, registration methods can also be utilized as described earlier [Wu et al., 2006]. To improve segmentation results, it may also be beneficial to first filter the image(s) using N3 correction and/or any other filtering method [Garg & Kaur, 2013].

3.6.2 Diffusion Tensor Imaging (DTI)

Image processing of diffusion-weighted images first requires a filtering method to remove the misalignment created by the presence of eddy currents during scanning. This can be resolved by using image registration methods [Alexander et al., 2007]. Then, multiple linear regression is used to obtain diffusion tensor components from the set of diffusion weighted images with varying directional/orientation information regarding diffusion of water. A tensor is a 3x3 matrix that represents molecular mobility along each direction and correlation between these directions at each voxel of the image. Next, “diagonalization” of the tensors is performed to obtain eigenvectors and eigenvalues that represent the main diffusion directions and related diffusivities respectively. These main directions and eigen diffusivity are used to depict tensors in the form of diffusion ellipsoids [Bihan et al., 2001].

The tensors and their ellipsoid representations define various DTI measures that relate to the tissue microstructure and architecture at each voxel. These measures including mean diffusivity, fractional anisotropy, and main direction of diffusivity. Mean diffusivity describes the overall mean-squared displacement of water molecule. It may be affected by the presence of obstacles that could impede diffusion. It is represented by size of the ellipsoid, where a larger size indicates greater displacement. Fractional anisotropy describes the degree to which molecular displacements vary in space (i.e. degree of anisotropy in water diffusion). It may be

affected by the presence of oriented structures (e.g. bundles of myelinated axonal fibers running in parallel to form white matter tracks) that would increase anisotropy of diffusion. It is represented by the eccentricity of the ellipsoid, where a greater degree of eccentricity indicates greater anisotropy. The main direction of diffusivity is represented by the main axis of the ellipsoid (i.e. primary eigenvector of the tensor). It is associated with the orientation of the tissue microstructures at each voxel and useful for performing brain fiber tracking of white matter tracks to infer brain connectivity [Bihan et al., 2001].

Thus, once the tensors are constructed, the next processing steps include: (1) computing mean diffusivity and fractional anisotropy maps for statistical analysis, and (2) performing tractography (i.e. fiber tracking). For the statistical analysis (e.g. tract-based spatial statistics), the mean fractional anisotropy maps are registered to a common template space (i.e. fractional anisotropy skeleton) for appropriate group comparisons. For tractography, the tensors are used to locate axonal tracts via a deterministic or probabilistic approach. The deterministic approach attempts to determine the exact path of the axonal tracts in a 3D continuous manner by following the main direction of diffusion from voxel to voxel. It requires more detailed information from a greater number of directions to accurately perform tracking. Thus, the number of images acquired to perform deterministic tractrography is significantly larger. When the necessary amount if information is not available, the probabilistic approach is recommended. The probabilistic approach determines the probabilistic path of axonal tracts by computing a probability density function of the neuronal fiber orientation. In the process, it results in more dispersed trajectories of probable axonal tracts [Mukherjee, Chung et al., Apr 2008; Mukherjee, Berman et al., May 2008].

One limitation of DTI is its lack of ability to directly image multiple fiber orientations within a single voxel. To solve this limitation, alternative approaches including diffusion spectrum MRI and Q-ball methods have been studied [Wedeen et al., 2008].

3.6.3 Functional MRI (fMRI)

Image processing for fMRI images begins with pre-processing. The first pre-processing step is to correct for head motion-related artifacts by realigning all images to the first 3D fMRI image as discussed in the "Filtering" section. Then, the all fMRI images are linearly co-registered to the same subject's high-resolution image. Next, the high-resolution image is normalized using a combination of linear and nonlinear registration to a high-resolution template brain. The corresponding transformation from the normalization is applied to the co-registered fMRI image. Thus, the high-resolution images act as a mediatory image that helps register the fMRI to template space because of its greater resolution. Lastly, the fMRI images are smoothed using a Gaussian filter to increase the signal-to-noise ratio, increase inter-subject overlap, and increase validity of analysis. Pre-processing prepares the fMRI images for appropriate individual and/or group level analyses. There are two forms of analyses that can be performed including activation-based and connectivity-based analysis [Vink, 2007].

For activation-based individual level analysis, the time series (i.e. 4^{th} dimension of the fMRI images) from the images are evaluated for probability of task-specific activation using general linear model analysis. The general linear model attempts to perform multiple linear regression to compute the scalar values in the following equation: (actual time series at a voxel) = (expected time series from a region with task-based neuronal activation)*(scalar_0) + (covariate_1)*(scalar_1) + …+ (covariate_N)*(scalar_N) + (a noise term). The expected time

series is formulated by convolving the design matrix (which defines the time points at which the task was expected to have occurred based on the experimental design) with the hemodynamic response (which is the expected MR signal from an activated neuron). The covariates include factors that may be affecting the actual time series but are not of interest like head motion artifacts, age, etc. Regions that show strong correlations between the actual and expected time series resulting in large scalar values are considered to be active during and play an important role for the task performed during image acquisition. For activation-based group level analysis, the regions of activation found in the individual level analysis are compared across or between group(s) of subjects to find regions of strong overlap. For all analyses, t-tests are commonly used to determine significance of regional activation [Vink, 2007]. Though the general linear model technique is most widely used, the field is progressing towards finding more accurate techniques for determining regions of activation. Such techniques include information theory and/or machine learning approaches [Ostwald & Bagshaw, 2011].

For connectivity-based analysis, the above-described pre-processed time series is further processed before analysis. First, the effects of non-interest like head motion artifacts, average white matter BOLD signal, and average cerebrospinal fluid BOLD signal are co-varied out of the time series signal using regression. Then, the new time series signal is band pass filtered as mentioned in the "Filtering" section. Using the temporally processed signal, individual level connectivity analysis is performed. Various pre-defined combinations of region pairs are compared using correlation or regression methods. Regions that show high correlations between their respective time series are considered as connected and part of the same neural network for the task—e.g. resting state—performed during image acquisition. Then, if required the analysis is extended to group level comparisons for determining significant overlap in functionally

connected regions across or between group(s) of subjects [Whitfield-Gabrieli & Nieto-Castanono, 2012].

3.7 LLD BIOMARKERS

Biomarkers are frequently studied in clinical research studies. They are evaluated as indicators of normal biological processes, pathogenic processes, or responses to pathology-related treatment. They are quantifiable and reproducible characteristics or medical signs that are objectively measured. Overall, they play an important role in helping better understand the normal physiology and pathophysiology, as well as improve processes for the treatment of pathologies [Strimbu & Tavel, 2010].

3.7.1 LLD Diagnosis

A number of recent late-life studies have shown an association between imaging measures and LLD diagnosis. In regards to structural measures, these studies indicate that LLD is associated with the following: (1) gray matter volume reductions mostly within the frontal-subcortical and limbic networks [Chang et al., 2011; Ribeiz et al., 2013; Sexton et al., 2013], (2) greater WMH burden supporting the vascular depression hypothesis [Aizenstein et al., 2011; Crocco et al., 2010; Disabato et al., 2012; Firbank et al., 2012; Gunning-Dixon et al., 2010; Kohler et al., Feb 2010; Teodorczuk et al., 2010], and (3) abnormalities in DTI measures [Colloby et al., 2011; Mettenburg et al., 2012; Sexton et al., 2013; Shimonv et al., 2009]. In regards to functional task-based activation, past studies have related LLD with both increase and decrease in task-related

activity within different regions of the fronto-striatal and fronto-limbic circuitry. For the fronto-striatal circuitry, executive tasks have been used to show hypoactivation in the dorsolateral prefrontal cortex and hyperactivity in the striatum (caudate and putamen). The hypoactivation of the dorsolateral prefrontal cortex may be due to the executive function deficits associated with LLD. The hyperactivity of the striatum may indicate a greater response to negative rewards and altered emotional processes in LLD patients [Aizenstein et al., 2005; Aizenstein et al., 2009; Bobb et al., 2011; Wang et al., 2008]. For the fronto-limbic circuitry, affective tasks have been used to show attenuated activation in the ventromedial prefrontal cortex and increased limbic activity in LLD patients. The attenuated activation in the ventromedial prefrontal cortex may be due to its role in evaluating and regulating emotional occurrences and contextual reward processing [Brassen et al., 2008]. The increased limbic activity may be due to its role in responding to emotional stimuli [Aizenstein et al., 2011]. In regards to functional resting state connectivity measures, past studies have shown greater, lower, and non-significant resting state functional connectivity difference between LLD and controls in varying regions—including some regions from the dDMN and aSN [Alalade et al., 2011; Alexopoulos et al., 2012; Andreescu et al., 2013; Bohr et al., 2012; Crocco et al., 2010; Steffens et al., 2011; Wu et al., 2011].

3.7.2 LLD Treatment Response

Several recent late-life studies have shown an association between imaging measures and LLD treatment response. In regards to High-Resolution structural measures, a couple studies indicated that LLD remission is associated with higher baseline gray matter volumes [Marano et al., 2013; Ribeiz et al., 2013]. In regards to WHM burden and DTI measures, there are varying findings in

the literature. Some studies have shown LLD remission to be associated with low baseline WMH severity [Disabato et al., 2012; Gunning-Dixon et al., 2010], while others have shown no significant association [Disabato et al., 2012]. Similarly, LLD remission is shown to be associated with greater baseline WM integrity by some studies [Alexopoulos et al., 2008], and with lower baseline WM integrity by another study in similar ROIs [Taylor et al., 2008]. To the best of our knowledge there are no studies evaluating the association of baseline functional task-based activation with LLD treatment response. However, there are studies that have focused on associating changes in functional task-based activation from pre- to post-treatment with LLD treatment response. These studies have shown increased activation post-treatment compared to pre-treatment in the dorsolateral prefrontal cortex (fronto-striatal circuitry) and ventromedial prefrontal cortex (fronto-limbic circuitry) [Aizenstein et al., 2011; Brassen et al., 2008]. The functional resting state connectivity studies have shown decreased connectivity between the dorsal anterior cingulated cortex and dorsolateral prefrontal cortex as well as increased connectivity between the posterior cingulated cortex and striatum to be associated with poorer LLD treatment response [Andreescu et al., 2013; Alexopoulos et al., 2012].

4.0 MACHINE LEARNING

This chapter gives a background understanding of machine learning methods. It primarily describes learning methods and potential methods for improving these methods to estimate accurate prediction models for a given framework or problem. This chapter also describes potential predictors and biomarkers for prediction models of depression and its treatment response based on past studies.

4.1 INTRODUCTION

Machine learning consists of a group of methods used to find relationships or patterns from empirical data for a given framework (i.e. problem). The input data used for the learning is made up of instances, i.e. samples. The number of instances in the input data defines the sample size of the data. Every instance of the input data is defined by a feature vector that describes the instance by the values assigned for each feature in the feature vector. The length of the feature vector (i.e. number of scalar values contained in it) defines the dimensionality of the data (i.e. number of features used to describe the data set). When the data is labeled, a label (i.e. output/outcome variable) is also assigned to every instance of the data. When developing a generalized model, a training data set is used as the input data. Once the model is created, it is used to predict the label or category of any new unseen data samples, also known as the test data set. The training and test

data are similar in the number and type of features used to define each instance and represent the same type of empirical data. The test data set is used to determine how well the generalized model represents this type of empirical data [Kapitanova & Son, 2012; Taskar et al., 2003].

4.2 TYPES OF LEARNING

Depending on the data, three possible types of learning include supervised learning, semi-supervised learning, and unsupervised learning. Supervised learning is performed if all of the data is labeled, semi-supervised learning is performed when there is unlabeled data along with labeled data, and unsupervised learning is performed when all of the data is unlabeled. For each type of learning, there are linear and nonlinear methods. Linear methods are simpler, while nonlinear methods are more flexible in nature. For supervised and semi-supervised learning, the methods can be further categorized as classification- or regression-based methods. Classification-based methods attempt to classify the data by discrete and categorical labels, while regression-based methods attempt to fit the data to a continuous function and thus work with continuous labels for the data. For unsupervised learning, the methods can be primarily categorized as clustering methods—which attempt to group the data into clusters based on underlying similarities [Ghahramani et al., 2004; Kapitanova & Son, 2012; Muller et al., 2003]. A list of common methods for each type of learning is summarized in table 1.

Table 1. Common Machine Learning Methods

Supervised Learning Methods		
	Linear	***Nonlinear***
Classification	• Logistic Regression • Support Vector Machines (Linear Kernel) • Bayesian Networks	• Artificial Neural Networks (with discrete output(s)) • Support Vector Machines (Radial Basis Function (RBF) or Polynomial Kernel) • Bayesian Networks • K-Nearest Neighbor • Decision Trees
Regression	• Linear Regression • Support Vector Regression (Linear Kernel) • Bayesian Networks	• Artificial Neural Networks (with one continuous output) • Support Vector Regression (RBF or Polynomial Kernel) • Bayesian Networks • K-Nearest Neighbor • Decision Trees
Semi-Supervised Learning Methods		
	Linear	***Nonlinear***
Classification	• Expectation Maximization + Generative Model • Transductive Support Vector Machines	• Expectation Maximization + Generative Model • Graph Mincut
Regression	• Transductive Regression	• COREG
Unsupervised Learning Methods		
	Linear	***Nonlinear***
Clustering	• K-Means Clustering	• Self-Organizing Maps

4.2.1 Supervised Learning Methods

Supervised learning methods consist of discriminative models, generative models, and more. Discriminative models (e.g. logistic regression, linear regression, artificial neural networks, support vector machines, support vector regression, etc.) are either (1) used to directly model the conditional distribution probability (i.e. p(y|x) where y = output variable, and x = input data) without knowing anything about the distribution of input features (i.e. p(x)); or (2) used to identify a representation of a function that maps input features to output variable(s). On the other hand, generative models (e.g. Bayesian Networks) attempt to estimate the underlying unknown probability distribution from the data. They first require the computation of the joint distribution between the output variable and input data (i.e. p(x,y)), from which they determine the model for the dependence of the output variable on the input data (i.e. conditional probability p(y|x) = p(x,y)/p(x)). More details on generative models and how they compute the joint distribution for classification-based frameworks are given in the "Semi-Supervised Learning Methods" section [Ng & Jordan, 2002; Peharz et al., 2013; Xue & Titterington; 2008].

Among the supervised learning methods listed in table 1, the first method in each section (i.e. logistic regression, linear regression, and artificial neural networks) is an example of a discriminative model and is similar in approach for creating a prediction model. Each of these methods attempts to fit a pre-defined function(s) (e.g. sigmoid for logistic regression, linear for linear regression, etc.) to the data. In the process, optimum weights are assigned to each input feature such that the combination of the weighted inputs results in predictions of the output variable. The artificial neural networks are multiple layered versions of the linear models. These additional layer(s) are known as hidden layer(s), which allow the weighted inputs to be combined in a linear and/or non-linear fashion—depending on the pre-defined function(s) for

each layer—to predict the output variable [Dreiseitl & Ohno-Machoda, 2002; Sarle, 1994; Zhao & Yu, 2006]. Variations to logistic and linear regression include an addition of a regularization or penalty term explained more extensively in “Appendix B” [Liu & Zhang, 2008]. Several variations of the artificial neural networks also exist as a part of the deep learning family of algorithms [Le et al., 2011].

The second method in each section of table 1 (i.e. support vector machines and support vector regression) is also an example of a discriminative model and is similar in the make-up of the prediction model it develops, but differs in model objectives. The components of the prediction models they create consist of hyperplane(s) and corresponding equidistance support vectors that define error margins based on the training data. For support vector machines the model components form a decision boundary in attempts to divide the data into label-based categories, while for support vector regression the model components attempt to fit the data. Both methods are part of a group of algorithms called kernel methods because they depend on the data only through dot products computed using kernel functions. Kernel functions help reduce computation time—especially for high-dimensional data—by allowing for non-linear model components when required without explicitly mapping the data to high-dimensional feature space. The utilized kernel function determines whether the model components will be linear or non-linear in the original input data space [Ben-Hur & Weston, 2010; Smola & Scholkopf, 2004]. Other kernel methods include least squares support vector machines and least squares support vector regression. More recent popular kernel methods include relevance vector machines and relevance vector regression [Tipping, 2001].

Other supervised learning methods listed in table 1 include Bayesian networks, which are an example of generative models and can be used for both classification- and regression-based

frameworks. Bayesian networks graphically depict probabilistic dependencies between random variables (e.g. features and outcome labels). Bayesian network methods first compute the joint probability distribution between features and labels in the training data. Then, using Bayes theorem, these methods compute the conditional distribution of labels given features to predict the outcome variable from the input test data. Naïve Bayes is a static variation of Bayesian network classifiers that assumes that the input features are independent of one another. Further variations of this method involve assumptions of various input features' distribution (e.g. Gaussian distribution) [Friedman et al., 1997; Pavlovic et al., 2002]. Other variations of the Bayesian networks classifiers include the Hidden Markov model that can be considered as a simple dynamic Bayesian network [Jing et al., 2008]. A form of Bayesian network method that can also be used for regression-based frameworks is the Tree-Augmented Naïve Bayes [Fernandez et al., 2007].

Similar to Bayesian networks, k-nearest neighbors can also be used for classification- or regression-based frameworks. This method uses a distance metric (most commonly the Euclidean distance) to find the k training samples most similar to the test sample. The most frequent occurring label from the k-nearest neighbor training samples is then assigned to the test sample as the predicted label [Dreiseitl & Ohno-Machado, 2002; Weinberger et al., 2006]. Variations of this method include distance-weighted k-nearest neighbor and large margin nearest neighbor classifiers [Domeniconi et al., 2005; Gou et al., 2012]. There are also k-nearest neighbor regression methods [Maltamo & Kangas, 1998].

Last, but not least, decision trees can also be used for classification- or regression-based frameworks. This method creates a tree with nodes and edges (i.e. branches) to predict the output variable. The nodes of the tree represent an input feature and the edges, which branches off the

nodes, split the input feature values by a threshold. Each edge is then connected to another new node, which represents another input feature. The node from which the edge branches is known as the parent node, while the new nodes added to the branched edges are known as the children nodes. This pattern continues until adding more nodes to the edges will no longer produce optimal prediction results. The input feature at every node and corresponding threshold values at the edges are selected based on a metric. A commonly used metric is information gain, which represents the reduction in uncertainty of the outcome predictions after adding a particular input feature as a child node. The aim would be to select input features for the children nodes such that the information gain is maximized. After the decision tree algorithm terminates, the ending edges that have no nodes attached to them are assigned a leaf node with a prediction value for the dependent variable (i.e. outcome). For classification trees, this value is the most likely label value in the form of a finite number of unordered values, while for regression trees it is in the form of a continuous or ordered discrete values [Friedman et al., 1996; Loh, 2011]. There are several variations of decision tree methods that incorporate approaches used by other learning algorithms. For example, Naïve Bayes is combined with decision trees to form the Naïve Bayes decision tree method. Other variations of decision trees include methods that use different strategies to combine decision trees and produce a more accurate prediction model. Examples of such methods are random forests—which randomly creates multiple decision trees and selects a label for the output variable based on the most frequent label predicted by these trees—and alternating decision trees—which use boosting to combine multiple weak classifiers to form a stronger classifier in the form of a generalized decision tree [Kingsford & Salzberg, 2008; Kohavi, 1996].

4.2.2 Semi-supervised Learning Methods

Several different techniques have been used to solve semi-supervised learning problems. These include generative models, self-training, co-training, avoiding dense regions changes (e.g. transductive support vector machines), and graph-based methods. Generative models, self-training, co-training, transductive support vector machines tend to be more inductive in nature (i.e. learner can predict unseen test data after being trained on the labeled and unlabeled training data), while some graph-based methods are more transductive in nature (i.e. learner can only work with labeled and unlabeled training data, thus it cannot predict future unseen data) [Zhu, 2006]. Table 1 lists some commonly used methods related to these techniques.

Generative models assume a mixture distribution [$p(x|y)$ where x is the input data and y the output variable/label], for instance mixture of Gaussian distribution (Gaussian mixture models). Based on this assumption, the joint probability distribution [$p(x,y) = p(y)p(x|y)$] is determined. Then, the test data is classified by assigning each instance in the test data with the label that produces the greatest joint probability given its input features (i.e. by computing $p(y|x) = p(x,y)/p(x)$). However, in the presence of unlabeled data and insufficient labeled data, it may be beneficial to incorporate the unlabeled data for determining the mixture distribution. In order to do so, the Expectation Maximization (EM) algorithm is used. This algorithm is used to identify the mixture components (i.e. individual distributions of the input features; when these distributions are combined, they form the mixture distribution) of the assumed mixture model using both the labeled and unlabeled data. The EM algorithm first initiates with an estimate of a classifier approximating the mixture distribution, possibly using the labeled data. Then, until the classifier is optimized by finding the maximum likelihood estimates of its parameters, the method iterates with primarily 2 steps: (1) Expectation step: uses current classifier parameters to

compute the expected values of the unlabeled data, and (2) Maximization step: re-computes the classifier parameters using the expected values found in the expectation step. An alternative method for the combined approach of generative models and EM algorithm is the combined approach of Hidden Markov Models and Baum-Welch algorithm [Nigam et al., 2006; Zhu, 2006].

Self-training methods use an iterative process to classify the unlabeled data and thereby increase the sample size of the labeled data set. This process involves first training a classifier on the small amount of labeled data set. Second, the trained classifier is used to predict the classification of the unlabeled data and the most confident of the classified unlabeled data are added to the labeled data set. Then, this two-step process is reiterated by re-training the classifier on the newly increased labeled data set until a pre-defined heuristic convergence criterion is met [Didaci & Roli, 2006]. The combined approach of generative models and EM algorithm is considered a special case of ‘soft’ self-training [Pise & Kulkarni, 2008].

Co-training, like self-training, methods also uses an iterative process to classify the unlabeled data and thereby increase the sample size of the labeled data set. However, instead of using one classifier to train the entire labeled data set, for this method two classifiers are first trained on respectively two sub-feature sets (i.e. features of the data set are split into two subsets) of the labeled data set. Secondly, after training on the small amount of labeled data set, both classifiers are individually used to predict the classification of the unlabeled data. The most confident of the classified unlabeled data are added to the labeled data set. Similar to the self-training, for co-training also this two-step process is reiterated by re-training the classifier on the newly increased labeled data set until a pre-defined convergence criterion is met. For this method to work well, two assumptions about the sub-features sets should be made: (1) both classifiers

are compatible such that they produce the same classification labels for all test patterns, and (2) the two sub-features sets are conditionally independent such that they can both individually help train an optimal classifier if there is sufficient labeled data [Didaci & Roli, 2006; Zhou & Li, 2005]. Also, for this method, regressors can also be used in place of classifiers. COREG, described by Zhou & Li (2005), is an example of a co-training method that uses regressors, specifically k-nearest neighbor regressors with different distance metrics.

Another semi-supervised learning technique involves using discriminative models and classifying the data according to regional data density (i.e. amount of data points in a given region). Data density is involved because without it a discriminative method directly predicts $p(y|x)$ to classify the data and does not consider if it shares parameters with $p(x)$ or the input data distribution. When $p(x)$ is not related to the classification made by a trained discriminative model classifier, semi-supervised learning doesn't perform as well. Thus, the incorporation of data density as an important factor in the classifier training process of discriminative methods is essential. Transductive support vector machines (TSVMs) is an example of such a method. TSVMs is a modified version of the discriminative model called support vector machines that determines classification boundaries—defined by a hyperplane and support vectors (as described earlier)—by avoiding high density regions. In order to do so, TSVMs attempts to classify the data by finding a linear boundary that has maximum margin (i.e. largest error margin between linear support vectors and hyperplane) for both the labeled and unlabeled data. The inclusion of unlabeled data instinctively guides the linear boundary towards low-density regions. Other similar methods include Gaussian Processes, Information Regularization, and Entropy Minimization [Pise & Kulkarni, 2008; Zhu, 2006].

Graph-based methods also have a discriminative and transductive nature like TSVMs. Graphs used for these methods consist of nodes—that represent labeled and unlabeled instances—connected by weighted edges—that represent how similar the nodes are that it connects. Similarity weights of the edges are computed using features (e.g. Euclidean distance, etc.). A graph-based method used to classify data using graphs is mincut. Mincut first determines the labels for each node of the labeled data set. It then weights the edges connecting nodes with the same labels infinitely high. The edges connected to nodes of unlabeled data instances are weighted according to some relationship with other nodes. For example, if distance is used to represent the relationship between nodes, then edges connecting two nearby nodes will be weighted higher than edges connecting two far apart nodes. This is based on the notion that nearby nodes will generally have the same label. After weighting all the edges, a minimum cut is applied to the graph by removing minimum total weight set of edges such that nodes of different labels are disconnected. Then, the unlabeled nodes in each new graph set are labeled according to the labels of the labeled nodes in the graph set [Blum & Chawla, 2001]. Variations of the mincut methods include Gaussian random fields and harmonic function methods, and discrete Markov random fields. In addition to classification-based frameworks, graph-based methods can also be used for regression-based frameworks since they estimate a function for the graph [Zhu, 2006].

Regression-based frameworks can also be attempted using other types of methods that are discriminative and transductive in nature. An example of such a method is transductive regression. This method first locally estimates the labels for the unlabeled data by using its position information. Essentially, the unlabeled data is estimated to have labels corresponding to the weighted average of label values from neighboring labeled data. Then, the method uses a

discriminative method called ridge regression (see “Appendix B” for more details) with an extra regularization term that globally optimizes the label estimates [Cortes & Mohri, 2006].

4.2.3 Unsupervised Learning Methods

Unsupervised learning methods primarily consist of clustering-based methods. Clustering-based methods are used for categorizing data into groups or labels since there is no labeled data and thus there are no labels to begin with.

One well-known unsupervised clustering method is k-means clustering. The most common algorithm used for this method is called Lloyd’s algorithm. This algorithm uses an iterative process to partition the input data into k clusters. It begins by initializing k centers for a preliminary estimate of k clusters. The centers are computed such that the mean squared Euclidean distance between it and each data instance in the cluster is minimized. Then, the data instances from the training data are redistributed. Each instance is assigned to the cluster with which the shortest Euclidean distance is computed using the cluster’s center. The algorithm re-iterates between redistributing the data instances and re-computing the k centers until a predefined convergence criterion is reached. There are many other centroid models for clustering that cluster based on information at the center of each cluster including k-medians clustering, fuzzy-c-means clustering, etc [Bezdek et al., 1984; Kanungo et al., 2002]. Also, there are several alternative clustering methods including connectivity models that cluster based on distance connectivity, distribution models that cluster based on statistical distributions of the data, and density models that cluster based on regional data density [Kapitanova & Son, 2012].

Another type of unsupervised clustering method is self-organizing maps. Self-organizing maps is an unsupervised version of Artificial Neural Networks. They are used to find hidden

patterns in unlabeled data. Self-organizing maps consist of nodes (i.e. neurons) and each node is assigned a weight vector the size of an input data instance. These weight vectors are either initialized with random small values or sampled uniformly from the subspace defined by the input data's two largest principal component eigenvectors. To begin with these maps are either rectangular or hexagonal in shape. The method goes through an iterative process until the nodes of the self-organizing maps are aligned with the input training data. The process consists of three phases: competitive phase, cooperative phase, and adaptive phase. To start the process, for the competitive phase, an instance from the training data is selected and the Euclidean distance between it and every node's weight vector is computed. In the next phase, the cooperative phase, the center of the topological neighborhood formed by the winning node (i.e. the nodes whose weight vector is the shortest distance away from the selected instance) and its neighboring nodes (i.e. cooperating nodes) is determined. Note that over time the size of the topological neighbor decreases. Then in the adaptive phase, the weights of the winning and cooperating nodes are updated to move closer to the selected instance of the training data using the information from the previous phase. This process is repeated for every instance in the training data until the weight vectors of all the nodes follow the distribution of the input training data [Anvar et al., 2013; Sathya & Abraham, 2013]. Variations of this method include generative topographic maps, adaptive self-organizing maps, and growing hierarchical self-organizing maps [Bishop & al., 1998; Rauber et al., 2002; Wang et al., 2005].

4.3 VALIDATION MEASURES

Validation measures are used to assess how well the learning methods developed a generalized model for any given data set. To compute these measures, the trained prediction model is first applied to a test data set and predictions of labels/categories for each instance is acquired. Then, the validation measures are computed by comparing these predictions with actual labels if they are available. The validation measures differ based on the type of framework used for the learning method.

For classification-based frameworks, some common validation measures include accuracy, specificity, sensitivity, and receiver operating characteristic curve—which consists of true positive rates (i.e. sensitivity) as a function of false positive rates (i.e. 1 - specificity). The accuracy measure helps evaluate how accurately the prediction model classifies the test data overall, the specificity and sensitivity measures respectively help evaluate how accurately the prediction model classifies each label of the test data, and the receiver operating characteristic curve illustrates the overall performance of the classifier. Confusion matrices can also be used when labeled data is available, especially for models with more than two labels, to display how well each label was classified and the distribution of misclassification. The confusion matrix is a KxK matrix for K labels, where one side of the matrix represents actual labels and the other side represents predicted labels [Baldi et al., 2000].

For regression-based frameworks, some common validation measures include correlation coefficients and mean squared error. The correlation coefficients and their corresponding significance values help evaluate how well the model predictions are correlated with the actual label values, and the mean squared error helps evaluate the level of error in the model predictions [Baldi et al., 2000; Meyer, 2014].

For clustering-based frameworks, some common validations measures can be sorted into three types: external, internal, and relative. External measures help evaluate how well the clusters were predicted based on pre-specified external information. An example of an external measure is entropy. Entropy helps evaluate how varied the categorization is of the data instances in each cluster. Other similar external measures include mutual information and purity. Internal measures help quantitatively evaluate how well the model formed clusters from the data. Examples of internal measures include the Davies-Bouldin index that measures average similarity between clusters, the Dunn index that measures the ratio of inter-cluster to intra-cluster distance between data instances, and the Bayesian information criterion that evaluates how well the model fits the data including its complexity. Relative measures (e.g. can be an external or internal measure) help compare two different sets of clusters formed by using the same learning method, but different parameter values of the learning method. These measures are usually defined based on compactness—how close data instances in each cluster are to each other—and separability—how distinct two clusters are [Halkidi et al., 2001, Rendon et al., 2011].

4.4 PRACTICAL PROBLEMS

Due to the nature of real-world data, several problems are encountered when trying to use learning methods to estimate a generalized prediction model. More often than not, real-world data is high dimensional (i.e. has too many features) and limited in sample size: both of which can cause problems with estimating an optimal learner. Problems can also be encountered if the learning methods are not correctly selected and/or setup (e.g. regularized, combined with a filter reduction technique, etc.) based on the nature of the data.

4.4.1 Bias vs. Variance

With empirical data, learning methods often face a trade-off between high bias and high variance. High bias indicates that the learning method is learning an incorrect model, while high variance indicates that the learning method is learning a random model. When the prediction model created by a learning method is too simple and results in a high error when predicting labels/categories for the training data to begin with, the learning method is considered to have high bias in its ability to make predictions. In such a case, the prediction model is underfitting the data and most likely the prediction error for the test data will also be high. Linear learning methods tend to suffer more from a high bias. On the contrary, when the prediction model created by a learning method is too complex and fits the training data very accurately but the test data poorly (i.e. the model does not generalize well), the learning method is considered to have high variance in its ability to make predictions. In such a case, the prediction model is overfitting the data. Nonlinear learning methods tend to suffer more from a high variance [Domingos, 2012; Yu et al., 2006]. Solutions to this problem can be found in the "Parameter(s) Selection for Learning Methods", "Boosting", and "Feature Reduction" sections under "Practical Solutions".

4.4.2 High Dimensionality

High dimensionality (i.e. large number of features) is another real-world problem often faced by learning methods. When the data has a very high dimensionality, it becomes harder to develop a generalized model. This is because it is harder to learn and understand what is happening when the data has high dimensions. For example, with the involvement of too many features, there is a chance that a large number of the features are noise (i.e. irrelevant), thus making it harder to

accurately model the data. On the other hand, there is also a possibility of a larger number of relevant features in high dimensionality data. If there are too many relevant features, there is also a greater change of redundant information, which can lead to random learning. In both cases, there is a risk of overfitting [Domingos, 2012; Mwangi et al., 2013]. Solutions to this problem can be found in the “Feature Reduction” section under “Practical Solutions”.

4.4.3 Sample Size

A small sample sized data can also pose a problem for learning methods. The number of instances required for a learning method is considerably more than the number of features included. Generally, the greater the ratio of sample size to feature size, the better the results. Reasonably for optimal results, this ratio should be greater than three. However, when the sample size is small, it is harder for the learning method to build an accurate prediction model due to lack of information and sufficient representation of the framework it is trying to learn. Therefore, on a training data set with a small sample, a learning method will most likely build a high variance prediction model that represents the training data very well since there is less information to account for. In other words, the prediction model will show signs of overfitting by performing very accurately when predicting on the training data, but poorly when predicting on the test data since there is a greater chance that the information contained in the test data was missing in the training data. On the contrary, extremely large sized data can also pose a problem of scalability. Due to the large amount of information present in large sized data, it can cost both computation time and memory, which may be limited [Domingos, 2012; Foley, 1972]. Solutions to this problem can be found in the “Feature Reduction” and “Cross Validation” sections under “Practical Solutions”.

4.5 PRACTICAL SOLUTIONS

One solution to most real world learning problems is to add more training data and increase the sample size. However, in practice, this is not always possible. Thus, in this section, we explore other options for solving learning problems with empirical data.

4.5.1 Boosting

Boosting methods are used in conjunction with learning methods to reduce the risk of high bias. These methods work by iteratively combining weak prediction rules to form a strong (i.e. very accurate) prediction rule for estimating a model that represents the given data. To generate these weak prediction rules, a weak learning method is used. The weak learning method iterates through different subset of the training data or differently weighted training data and trains on the modified training set to develop a new weak prediction rule each time. There are several boosting methods and the main differences between these methods include the process of modifying the training set every iteration and the technique used to combine all the weak prediction rules into one strong prediction rule. The well-known boosting method is adaptive boosting (AdaBoost), which uses weights to modify the training data at every iteration and combines the weak predictions rules to form a more accurate prediction rule. This method assigns the largest weights to the most misidentified data instances and thereby focuses the attention of the learning method on the most difficult examples. AdaBoost has also been used to form a modified version of the decision trees learning method called alternating decision trees [Freund & Schapire, 1999; Opitz & Maclin, 2011; Schapire et al., 2003].

4.5.2 Feature Reduction

Feature reduction methods are used to reduce the number of features in high-dimensional data to a limited number of most relevant features for estimating a more accurate prediction model. These methods can be primarily categorized into supervised and unsupervised methods. Supervised methods require labeled data as they perform feature reduction with the help of the labels. Unsupervised methods, on the other hand, perform feature reduction based solely on information available in the features included in the data. Additionally, there is an alternative option of forced feature selection, which can be computationally expensive [Mwangi et al., 2013; Reif & Shafait, 2014].

4.5.2.1 Supervised Feature Reduction Methods

Supervised feature reduction methods are primarily used to perform feature selection (i.e. select the most relevant features from a larger set of input features) and thus reduce the noise in the input data. These methods include filter techniques, wrapper techniques, and embedded techniques. Each technique also includes different types of methods described below [Mwangi et al., 2013].

Filter techniques select features independent of the learning method and based on relationships determined by the labels. These techniques differ slightly between classification- and regression-based frameworks. For classification-based frameworks, filter techniques involve categorizing the training data instances into groups by labels. Then for each feature, a statistical test (e.g. t-test, ANOVA, correlation, etc.) is performed to determine the significant difference or degree of correlation between the groups. The features for which the groups demonstrate the greatest significant difference or correlation based on a predefined threshold are selected for

developing the prediction model. For regression-based frameworks, filter techniques involve comparing each feature with the corresponding labels of the training data using correlational statistical analysis (e.g. Pearson's correlation, Kendall tau rank correlation, etc.) and selecting the features that demonstrate the greatest significant correlation based on a predefined threshold for developing the prediction model [Ladha et al., 2011; Mwangi et al., 2013; Saeys et al., 2007; Zeng et al., 2012].

Wrapper techniques depend on the selection of the learning method since features are selected based on how well they help a particular learning method develop accurate prediction models. These techniques do not vary much based on the type of framework and include methods that perform either backward elimination or forward selection of features. Backward elimination methods (e.g. Recursive Feature Elimination) start out with the inclusion of all the features in the data set. Then, using an iterative process the following is implemented: (1) a small subset of features is removed at every iteration, (2) a prediction model is developed with the remaining features using an appropriate learning method, and (3) a cross validation is used to test the accuracy of this model. This process stops when a predefined termination criterion is reached or all features are eliminated. In the end, the features selected are based on the features used in the iteration that produced the most accurate prediction model. Forward selecting methods (e.g. Searchlight, a method for Neuroimaging studies), on the other hand, start out with an empty data set with no features. Then, they use an iterative process similar to the backward elimination methods, but instead of removing features, one feature is added at every iteration. To determine which feature to add at the first iteration, every feature is individually used to develop a prediction model. Then, cross validation is used to determine the feature that produces the most accurate prediction model and that feature is added as the first selected feature. For every other

iteration, each remaining feature is individually incorporated with the already selected feature(s) to develop prediction models and cross validation is used to determine the most optimal prediction model. The feature whose incorporation results in the most accurate prediction model is added to the data. Similar to the backward elimination method, the iterative process stops when a predefined termination criterion is reached or all features have been added. In the end, the feature set that resulted in the most accurate prediction model is selected. Individual and combined variations of both above described backward elimination and forward selection methods also exist. An example of a method that implements a combination of the two methods is the Plus-L-Minus-R Selection [Kohavi & John, 1997; Ladha et al., 2011; Mwangi et al., 2013].

Embedded techniques include learning methods—for both the classification- and regression-based frameworks—that inherently perform feature selection in the process of developing an optimal prediction model. Examples of such methods include decision trees. Other examples include regularized or penalized discriminative methods. These types of methods include L1-regularized logistic regression and least absolute shrinkage and selection operator (LASSO) regression (i.e. L1-regularized linear regression). In general, a regularization term penalizes the learning method when its complexity increases. The regularization term can sometimes help further reduce the complexity of the resulting prediction model by allowing the weights associated with some feature(s) to go to zero (i.e. by eliminating the least important features). Thus, it helps reduce the variance in the learning method's ability to predict and risk of overfitting. Variations of these methods include the L1/2-regularized logistic regression and Elastic Nets (i.e. linear regression with a L1- and L2-regularization term) [Chen et al., 2013; Grabczewski & Jankowski, 2005; Ladha et al., 2011; Mwangi et al., 2013]. For more detailed information on regularization, see "Appendix B".

4.5.2.2 Unsupervised Feature Reduction Methods

Unsupervised feature reduction methods are primarily used for feature extraction (i.e. extracting features that are different from the original input features, yet formed by patterns found among the input features) in order to reduce the dimensionality of the input data. Nevertheless, variations of unsupervised feature reduction methods are also used to reduce data dimensionality by performing feature selection. Most widely used unsupervised feature reduction methods include Principal Components Analysis and Independent Components Analysis [Mwangi et al., 2013].

Principal component analysis (PCA) extracts uncorrelated features that are a weighted linear combination of the input features. It does so by first finding orthogonal (i.e. uncorrelated) directions that represent the input data. The top directions that account for the most variance in input data are then selected such that they span a lower dimensional space than the input features (i.e. there are a fewer number of directions selected than the number of features in the input data). Next, the input features are linearly projected onto this lower dimension subspace spanned by the selected directions to form the extracted features, also known as the principal components. There are several variations of the principal component analysis method. One example is the kernel principal component analysis method, which can extract features that are a nonlinear combination of the input features. Additionally, modifications can also be made to the principal components analysis method to perform feature selection. One such modification involves simply selecting one feature from the input data for each selected eigenvector such that the selected feature is the most similar in directionality to the eigenvector (i.e. the selected feature's axis is the most dominating in the direction of the eigenvector). Put another way, the feature vectors that require the least amount of transformation to be projected onto the first eigenvector

are selected. This is performed in place of projecting all the input features on the selected eigenvectors [Lu et al., 2007]. Additionally, there is a variant of principal component analysis method in the form of a supervised feature extraction method called linear discriminant analysis (LDA) [Khan & Farooq, 2011].

Independent component analysis (ICA) extracts independent features, which form the source of the input features. In other words, these new features are extracted such that the input features are a weighted linear combination of the extracted features. The method first assumes that the source of the input features consists of independent features [Mwangi et al., 2013]. Then, the method attempts to extract components (i.e. features) such that the components have maximum statistical independence. Different independent component analysis methods vary in the ways by which the method achieves maximum statistical independence. These include but are not limited to minimizing mutual information and maximizing non-Gaussianity among the components [Langlois et al., 2010]. There also exist other variations including non-linear independent component analysis methods, which extracts new features such that the input features are decomposed into nonlinear components [Hyvarinen & Pajunen, 1999].

4.5.2.3 Forced Feature Reduction

Another way to reduce the dimensionality of the data is to perform forced feature reduction (i.e. manually remove features thought to be irrelevant based on prior information). A computationally intensive version of forced feature reduction is known as brute-force feature selection. For this method, all possible subsets of the input features are tested to determine the optimal subset, which achieves the highest prediction accuracy [Reif & Shafait, 2014]. Less computationally intensive variations of this method involve reducing the number of input features' subsets tested. This can be done in several ways including using the knowledge of the

relationships among the features or optimal features found in past studies. However, it is important that the reduction method is unsupervised and independent of the labels, if any, to avoid biasing the results. A similar feature reduction technique has been utilized by Muller et al. (2005).

4.5.3 Selection of Learning Method(s)

To accurately learn a framework or problem, it is not only important to select the right features, but it is also important to select the right learning method. The first step is to simply determine whether the given data to be learned consist of labeled instances only, a mixture of labeled and unlabeled instances, or unlabeled instances only. Consequently, this will determine whether to use a supervised, semi-supervised, or unsupervised learning method respectively. If data consist of a mixture of labeled and/or unlabeled instances, it would be beneficial to determine whether or not the unlabeled would help the learner. If the unlabeled data does not sufficiently increase the overall sample size, it may be better to exclude it. The second step is to determine the goal (e.g. classification, regression, or clustering) of the learner. The third step is to determine whether the nature of the data is linear or non-linear. Generally, when the data size is small, it is better to use a linear method to avoid overfitting. However, if the data size is sufficiently large, it may be beneficial to test non-linear methods to allow for more flexibility in the learning. The fourth step is then to decide the learning method from the narrowed down options.

Since no one learning method is the best for all application, it may be useful to test multiple methods. When selecting a learning method(s) for a given framework or problem, one should consider evaluating several different aspects of the method(s) including: computation time, underlying assumptions, interpretability, complexity, flexibility, optimization ability, and

tested applications by past studies. If there are still too many options of methods to choose from, it may be helpful to use machine learning libraries (e.g. LIBSVM, LIBLINEAR, etc.) or software (e.g. MATLAB, Python scikit-learn, WEKA, etc.) for testing the performance of different methods on the data. On the other hand, if there are too few options, it may be beneficial to modify—e.g. add constraints, regularize, combine methods (including learning, feature reduction, and/or boosting methods), etc.—existing learning methods to make them more suitable for learning the given data. Similar techniques for selecting learning methods have been utilized by Bibi & Stamelos (2006), Frank et al. (2004) and Kotthoff et al. (2012).

4.5.4 Cross-Validation

Cross-validation is used to estimate the accuracy of a prediction model created by the learning method(s) of choice. There are several techniques for performing cross-validation including holdout, k-fold cross-validation, and leave-one-out cross-validation. These techniques can be essentially considered as variants of the k-fold cross-validation technique. For all techniques, the respective validation measures described earlier are used to assess performance of the learning method(s).

When the available data set has a considerably large sample size, one way to perform cross-validation is to take the holdout approach by first splitting the data set into training and test sets without repetition of instances. Then, the learning method is used to estimate a model that describes the data by training the learner on the training set. Lastly, the estimated model is tested on the test set and the performance of the learning method is evaluated using appropriate validation measure(s). This is essentially a k-fold cross-validation technique (described below) where k equals one.

When the available data set is not sufficiently large enough to be split in two, a k-fold cross-validation technique is used. This technique first divides the data into k equal sized sets. Then, it does the following: (1) classifies one of the k sets as the test set, while combining the others to form the training set; (2) uses the learning method to estimate a model that describes the data by training the learner on the training set; (3) tests the estimated model on the test set; and (4) computes appropriate validation measure(s) to determine the precision of the model. This 4-step process is reiterated for k-iterations, each time classifying a different set as the test set without repetition. Lastly, the validation measure(s) values from all the iterations are averaged to evaluate the overall performance of the learning method.

When the available data set has a small sample size, a leave-one-out cross-validation method is used. This method is essentially a k-fold cross-validation method with k equal to the sample size of the data. In other words, for each iteration of the above-described 4-step process, one instance of the data is classified as the test set, while the rest of the instances are used for the training set.

A variation of these cross-validations methods includes bootstrapping. Bootstrapping methods are similar to cross-validation methods, except they increase the number of iterations for every fold of cross-validation by resampling with replacement from the given data, instead of using it as given. For example, a bootstrapping method using k-fold cross-validation replaces the above-described 4-step process with the following: (1) classifies one of the k sets as the test set, while combining the others to form the training set; (2) generates a pre-defined number of training and corresponding test set instances of the same size from the existing test and training set respectively with replacement; and (3) performs for each pair of newly generated training and test sets an iterative process that (a) uses the learning method to estimate a model that describes

the data by training the learner on the training set, (b) tests the estimated model on the test set, and (c) computes appropriate validation measure(s) to determining the precision of the model [Kohavi, 1995]. A variation of bootstrapping is bagging [Maclin & Opitz, 2011].

4.5.5 Parameter(s) Selection for Learning Methods

When slight changes to a certain parameter's values of a given learning method cause considerable variability in the resulting prediction model, it may be useful to perform a parameter selection process. Selection of a parameter(s) that somehow regulate the complexity (e.g. regularization parameters, which penalize complexity and target the overfitting problem) of the prediction model developed by the learning method is especially important. This is because, as discussed earlier, the complexity of a prediction model determines whether it adequately generalizes, overfits, or underfits the data. The most common approach used for parameter selection is cross-validation to determine optimal parameter values [Lim & Yu, 2013].

However, most likely, a cross-validation technique is already being used to evaluate the overall generalization-based performance of a learning method. Thus to perform parameter selection, a nested inner cross-validation loop would need to be implemented. For this inner cross-validation loop, the training set at every iteration of the outer cross-validation loop is used as the full data set on which parameter selection is performed. This inner cross-validation loop would be implemented between steps one and two of the above-described 4-step process of the k-fold cross-validation technique.

Any of the cross-validation techniques described in the "Cross-Validation" section or any variant of these techniques (e.g. estimation stability with cross validation) can be used for the inner cross-validation looped to perform parameter selection. The only difference is that instead

of iterating through different test sets, the parameter selection method iterates through each of the pre-defined set of possible parameter values. At every iteration, it uses the same training and test set to estimate a model that describes the data and assess the precision of the model by computing appropriate validation measure(s) respectively. The optimal parameter value—i.e. the parameter value that results in the most precise model—is then selected. The selected parameter value is then used to train the full data set (i.e. the training set of the outer cross-validation loop) for step 2 from the 4-step process of the k-fold cross-validation technique [Kohavi & John, 1995; Lim & Yu, 2013].

The process of selecting multiple parameters' optimal values is similar to the process used to select one parameter's optimal value. The only difference is that the cross-validation method iterates through each possible set of values from each parameter to find the optimal set. All possible combinations of a set of parameters' values are identified from pre-defined options of values for each parameter using the grid search technique [Bergstra & Bengio, 2012].

4.6 DEPRESSION PREDICTION MODELS

To the best of our knowledge, there are no past studies that have attempted at establishing a predictive model using MR neuroimaging for the elderly population. However, there have been several past studies that have successfully explored predictive models for diagnosis and treatment response of depression in the younger populations. Below is a survey of the studies that have used magnetic resonance imaging measures for estimating the prediction models in depression diagnosis and treatment response.

4.6.1 Depression Diagnosis

Studies of depression in younger populations involving prediction models have used both functional [Fu et al., 2008; Hahn et al., 2011; Marquand et al., 2008; Nouretdinov et al., 2011; Zeng et al., 2012] and structural [Costafreda et al., 2009; Mwangi et al., May 2012] imaging measures to obtain accurate classifications. Most of these studies have focused on utilizing support vector machines as their classifier, with the exception of one that successfully used Gaussian process classifiers [Hahn et al., 2011]. The highest classification accuracy (94.3%) among all these studies was achieved by using support vector machines with a filter feature reduction method (Kendall-tau). In this study, the biomarkers of depression diagnosis were found to be functional connections in the default mode network, affective network, visual cortical areas and cerebellum [Zeng et al., 2012].

4.6.2 Depression Treatment Response

Studies of depression remission after treatment in younger populations that successfully obtained accurate classification models have majorly utilized T1-weighted Hi-Res structural imaging measures [Costafreda et al., 2009; Liu et al., 2012; Nouretdinov et al., 2011]. One study that attempted to use a task-based functional measure did not achieve very high accuracy [Marquand et al., 2008]. All of these studies have focused on utilizing support vector machines as their classifier. The highest classification accuracy (88.9%) among all these studies was achieved by combining support vector machines with a filter feature reduction method (ANOVA). In this study, the biomarkers of depression treatment response were found to be whole brain structural neural correlates—especially greater grey matter density in the right rostral anterior cingulate

cortex, left posterior cingulated cortex, left middle frontal gyrus, and right occipital cortex [Costafreda et al., 2009].

5.0 ASSOCIATION OF SMALL VESSEL ISCHEMIC WHITE MATTER CHANGES WITH BOLD FUNCTIONAL MR IMAGING IN THE ELDERLY

This chapter describes an experiment that shows how alterations in brain structure observed in structural MR images can affect the acquisition of functional MR images in late-life depression. More specifically, it shows how lesions in the white matter are associated with the acquired functional task-based activation signal in late-life depression.

5.1 INTRODUCTION

In the elderly, magnetic resonance imaging (MRI)—particularly T2-weighted images—often reveal white matter hyperintensities (WMHs), which indicate the presence of ischemic or pre-ischemic white matter lesions. The lesions are generally associated with myelin pallor, tissue rarefraction, and mild gliosis [Gunning-Dixon et al., 2009; Madden et al., 2009; Debette and Markus, 2010]. Neuroimaging studies have shown that WMH burden is associated with cognitive changes of aging, as well as neuropsychiatric disability in the elderly [Wen and Sachdev, 2004]. Past studies have indicated an association between greater WMH burden and poorer global cognitive performance, executive function, and processing speed, as well as an increased risk of stroke, dementia, and death [de Groot et al., 2000; Gunning-Dixon et al., 2009; Debette and Markus, 2010]. Similarly, diffusion tensor imaging (DTI) studies have shown a

direct correlation between white matter integrity and cognitive performance, executive function, and information processing speed [Gunning-Dixon et al., 2009; Madden et al., 2009; Vernooij et al., 2009]. A DTI study by Taylor et al. (2001) also showed that WMHs are associated with damage to tissue structure, thus suggesting disruption of white matter tracts. These studies suggest that the white matter lesions underlying the WMHs affect neuronal activity.

Other studies have shown how cerebrovascular disease influences the coupling between neural activity and corresponding hemodynamics (i.e. cerebral blood flow, cerebral blood volume, and cerebral metabolic rate of oxygen consumption) [Carusone et al., 2002; Rossini et al., 2004]. Thus, considering WMHs as a marker for cerebrovascular disease, one would predict that WMHs might contribute to altered hemodynamic coupling, and the neural activity interpreted by blood oxygen level dependent (BOLD) functional magnetic resonance imaging (fMRI) might also be affected in the presence of WMHs. Additionally, the white matter lesions associated with the WMHs affect the T2* BOLD signal itself. On the T2* functional images, the areas with WMHs have increased intensity, similar to T2-weighted fluid attenuated inversion recovery (FLAIR) images (see Figure 1). The presence of WMHs on the T2*-weighted images may alter the sensitivity of the regional T2* BOLD signal.

As a summary, Figure 2 demonstrates the three stages where WMHs may influence the study of brain function (neuronal activity) using fMRI BOLD signals. Some past studies have studied the association between WMHs and functional activity based on specific tasks using BOLD fMRI [Nordahl et al., 2006; Aizenstein et al., 2011; Hedden et al., 2011; Linortner et al., 2012], however the relationship between WMHs and the BOLD fMRI signal is underexplored. Thus, this study evaluates how WMH burden in the elderly is associated with the BOLD signal change determined using a sensory-motor task, which is known to not be significantly associated

with WMH burden in task related regions [Linortner et al., 2012]. The simple finger-tapping fMRI task was chosen for this study because of its known reliability and reproducibility. Also, we used total WMH burden to represent WMH burden for each subject to reduce the number of independent variables and based on evidence indicating global WMH burden is associated with local WMH burden [DeCarli et al., 2005].

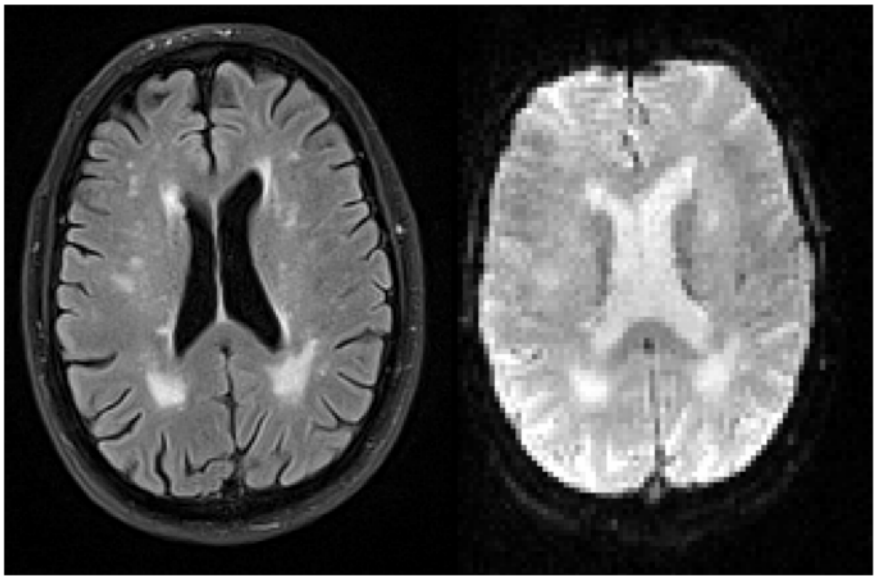


Figure 1. Presence of WMHs on T2-weighted FLAIR (left) and T2*-weighted images (right) of the same subject

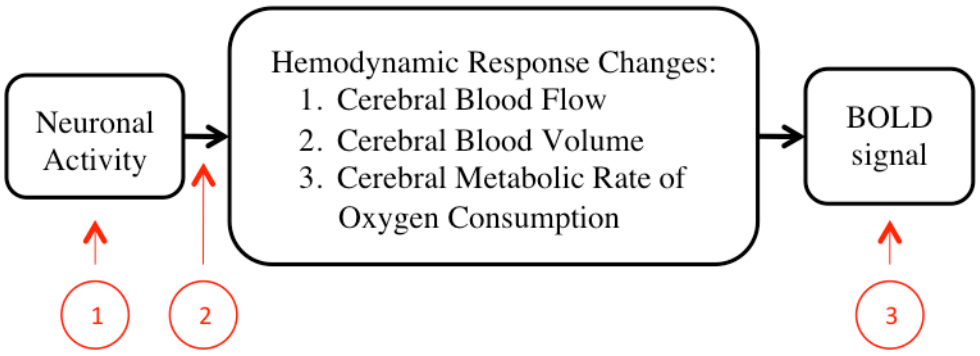


Figure 2. Flow chart of the physiology behind the BOLD fMRI signal in black. Three ways in which WMHs can affect the BOLD signal in red: 1) by affecting neuronal activity; 2) by affecting the coupling between neural activity and corresponding hemodynamics; and 3) by altering the sensitivity of the regional T2* BOLD signal

5.2 METHODS

5.2.1 Subject Recruitment

Elderly non-psychotic, unipolar major depressive disorder (MDD) patients and non-depressed individuals were recruited from the community for a late-life depression study and the same data was used for this structural and functional MRI study. All participants were required to undergo a SCID-IV evaluation. The exclusion criteria included: history of Axis I disorders (except MDD and anxiety disorders for the depressed patients only), stroke, significant head injury, Alzheimer's, Parkinson's, and/or Huntington's disease. Individuals were also excluded if they had taken psychotropic medications within 2 weeks prior the scans. Ten subjects (2 non-depressed and 8 depressed) were excluded from this analysis due to motion artifacts or poor image registration results (evaluated visually using Statistical Parametric Mapping 5 software (SPM5) [Friston et al., 1994] running on MATLAB (Math Works, Natick, Massachusetts, USA))

during the normalization step. Forty-one non-depressed and 33 depressed elderly individuals were included in this analysis. Demographics of the included subjects are shown in Table 2.

Table 2. Demographics of the included depressed and non-depressed subjects

	Depressed	**Non-Depressed**
Age in years, mean (SD)	68.3 (6.6)	71.7 (7.9)
Sample Size	33	41
Gender	13 Males & 20 Females	12 Males & 29 Females
Mini Mental Score[(a)], mean (SD)	27.2 (4.0)	29.9 (1.7)
Hamilton D[(b)], mean (SD)	20.0 (5.1)	——— [(c)]
Normalized WMH Volume, mean (SD)	0.00197 (0.0034)	0.00379 (0.0067)

(a) Missing data from 4 depressed & 1 non-depressed subjects is not included in average & standard deviation calculations
(b) Missing data from 1 depressed subject is not included in average & standard deviation calculations
(c) Depression screening through psychiatric interviews was performed, but Hamliton Depression Rating Scale was not performed on these subjects

5.2.2 Image Acquisition and Data Collection

Subjects were scanned on a 3T Siemens TIM TRIO scanner. T1-weighted images were acquired with a 1 mm slice thickness, 256x224mm resolution, 256x224mm field of view (FOV), 2300ms repetition time (TR), 900ms inversion time (TI), 3.43ms echo time (TE), and 9 degrees flip angle (FA) in the axial plane. T2-weighted images were acquired with a 3 mm slice thickness, 256x224mm resolution, 256x224mm FOV, 3000ms TR, 100ms TI, 101ms TE, and 150 degrees FA in the axial plane. T2-weighted Fluid Attenuated Inversion Recovery (FLAIR) images were acquired with a 3 mm slice thickness, 256x240mm resolution, 256x212mm FOV, 9160ms TR, 2500ms TI, 88ms TE, and 150 degrees FA in the axial plane. Functional images were acquired

using a gradient-echo echo planar imaging sequence with a 3 mm slice thickness, 128x128mm resolution, 256x256mm FOV, 2000ms TR, 34ms TE, integrated parallel acquisition technique (IPAT) = 2, and 90 degrees FA in the axial plane. The paradigm performed by the subjects during the functional image acquisition was a 5-minute block-design. There were five 30 seconds long experimental blocks (during which the tapping cue was presented 40 times), and it was alternated with five 30 seconds long control blocks. During the experimental blocks subjects tapped the right hand index finger for 30 seconds while looking at a cue (the word tap). During the control blocks, subjects rested while looking at a fixation-cross in the center of the screen [Howseman et al., 1997]. During the performance of this task, behavioral data pertaining to task performance—including accuracy and reaction times—were collected for each subject. For our statistical analyses, we computed the median of reaction times corresponding to accurate tapping responses by each participant (except for 3 depressed subjects for whom we are missing behavioral data).

5.2.3 Image Processing and Analysis

The following image processing steps were performed for each subject using SPM5. All functional images were realigned to the first image in the sequence and then co-registered with the subject's structural grey matter. The co-registered structural MR image and all realigned functional images were normalized to Montreal Neurological Institute (MNI) space using the a priori grey matter template in SPM5. The normalized functional images were smoothed with a Gaussian kernel (full width at half maximum (FWHM) = 10mm) to account for the greater morphologic variability in elderly subjects [Reuter-Lorenz PA and Lustig, 2005]. The effect of task on BOLD signal intensity was examined by general linear modeling (GLM) in SPM5. The

tapping block was modeled by a box function and convolved with a hemodynamic response function then submitted to the GLM. Prior to the GLM, a high pass filter of 128 seconds was applied to the image to correct low frequency drift. The serial autocorrelation was corrected using an autoregressive model.

Additionally, an automated WMH segmentation method was used to obtain whole-brain WMH volumes from each subject's T2-weighted FLAIR images [Wu et al., 2006]. Acquired WMH volume measurements were then normalized by total brain volume [Wu et al., 2006]. In addition to WMHs, ventricles were manually segmented from each subject's T1-weighted image and the computed volume was also normalized by total brain volume. A final analysis included the computation of a WMH map indicating the number of subjects with WMH in various regions of the white matter above the cerebellum. To obtain this map, each subject's T2-weighted FLAIR along with the corresponding WMH segmentation was co-registered to the structural image, and all co-registered images were registered to the MNI template.

5.2.4 Statistical Analysis

First, the general linear model was used to estimate task-related significant BOLD signal changes for each subject. Then, group level analyses were performed using a two-sample t-test, Wilcoxon's rank-sum tests, and a regression analysis. The two-sample t-test compared the differences in the BOLD signal change between the non-depressed and depressed groups. Due to the skewed distribution of the reaction time and WMH, we performed Wilcoxon rank-sum tests comparing these variables: (1) median reaction times to analyze differences in task-related behavior; and (2) normalized WMH volume measures in the whole brain and in each of the 20 regions from the Johns Hopkins University White Matter Atlas [Wakana et al., 2004] to analyze

differences in global and local WMH burden distribution among subjects respectively. Since no significant differences were detected in task-related BOLD signal change, median reaction times, and normalized WMH volumes between the two groups, they were combined. Voxel-wise regression analysis was performed with BOLD signal change on finger-tapping versus fixation as the response variable, and the normalized WMH volume as a predictor. Normalized ventricle volume and group (a dichotomous variable categorizing depressed and non-depressed participants) were also included in the regression analysis as covariates of non-interest. The normalized ventricle volume was included as a covariate to ensure that the results are not biased by ventricle size since large ventricles are associated with high WMH burden and are prone to cause poor registration, which may in turn significantly affect the statistical analysis. Similarly, group was also included as a covariate to control for any group differences. Lastly, we also performed two post-hoc analyses: (1) we performed a Wilcoxon rank-sum test comparing BOLD signal change at the peak coordinate found to be significant in our regression analysis between participants with and without co-localized WMHs within a one-voxel neighborhood of the peak coordinate; and (2) we performed a Spearman's rank correlation analysis between the medians of reaction times and WMH burden to ensure that the performance of the task was not associated with WMH severity. All statistical parametric analyses were performed in SPM5 and non-parametric tests were performed using MATLAB.

5.3 RESULTS

No significant task-related difference in BOLD signal change was found between the depressed and non-depressed groups from the two-sample t-test ($t(1,72) = 3.0$, SPM Family wise error

(FWE) corrected $p = 0.92$). Similarly, no significant group difference was found from the Wilcoxon rank-sum tests comparing median reaction times ($z = 0.70$, $p = 0.48$), global normalized WMH volumes ($z = -1.78$, $p = 0.07$), and local WMH volume in all 20 regions ($z(\text{min,max,std}) = (-2.17,-0.03,0.53)$, $p(\text{min,max,std}) = (0.03,0.97,0.26)$; none of them survive Bonferroni correction) between the two groups. For both groups, most subjects (> 80% of the subjects) had a normalized WMH volume within a range of 0 to 0.005 (see Figure 6). Therefore, the non-depressed and depressed groups' data were pooled.

As expected, the group random-effects analysis between all subjects indicate presence of significant positive BOLD signal change in regions of the motor cortex (including primary motor, premotor, and supplementary motor regions) during motor activity of tapping ($t(1,73) > 4.48$, $k = 100$, FWE corrected $p < 0.05$). Additionally, regions known to be a part of the default mode network (including the mid-temporal, prefrontal, and posterior cingulate regions) were significantly deactivated during the tapping task when compared to the baseline fixation period. ($t(1,73) > 4.48$, $k = 100$, FWE corrected $p < 0.05$) predominantly in the left hemisphere (see Figure 3).

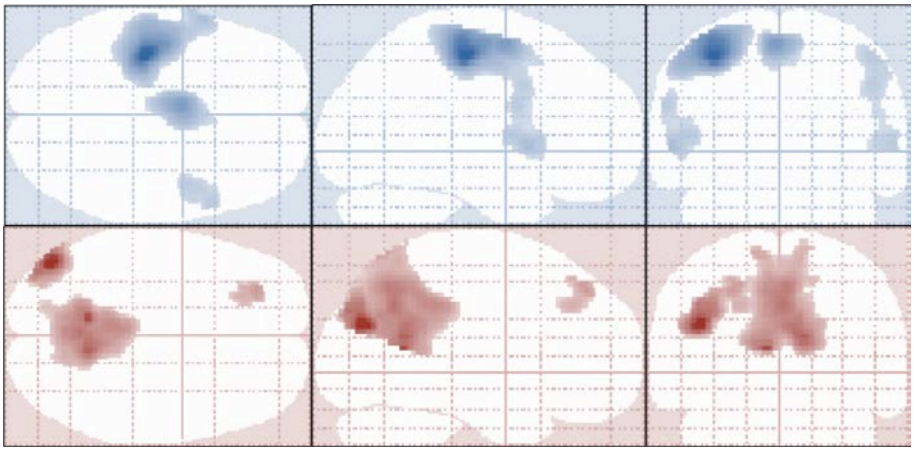


Figure 3. Projected activation maps in neurologic orientation showing significant regions ($t(1,73) > 4.48$, $k = 100$, FWE corrected $p < 0.05$) for main effect of tap (top 3 blue images) & main effect of fixation (bottom 3 red images)

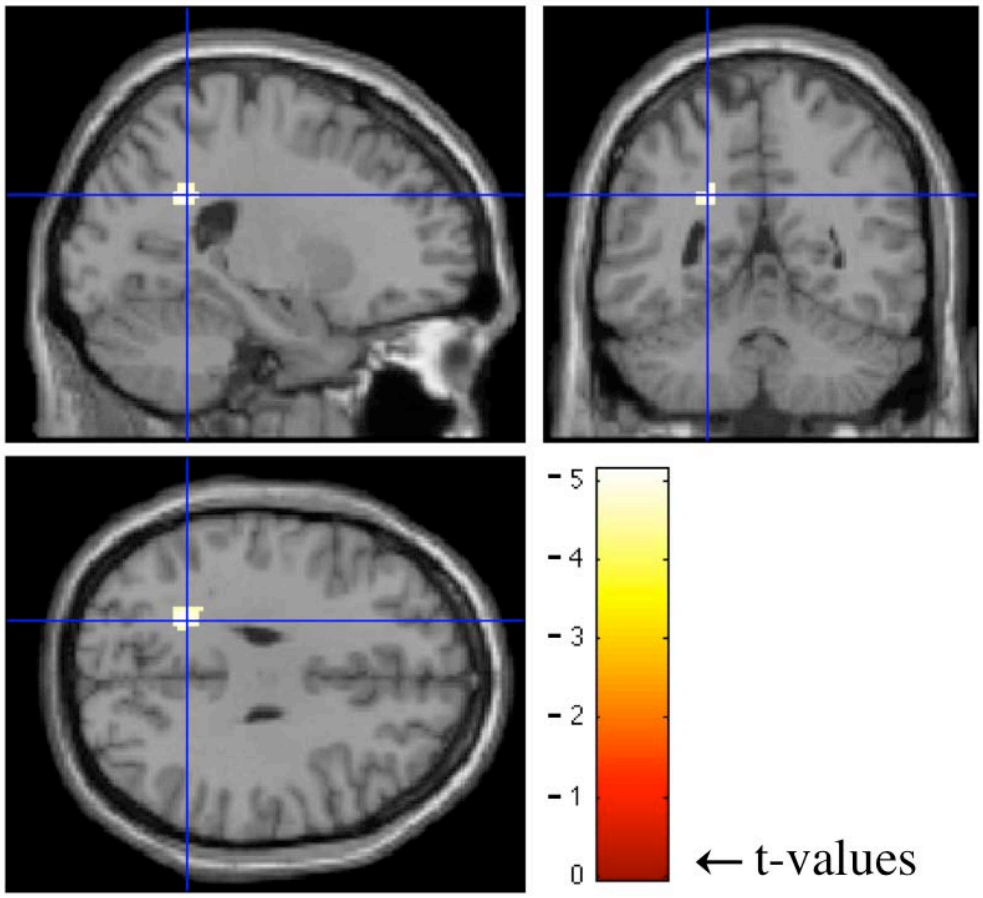


Figure 4. Results of regression analysis performed usign SPM5. Crosshairs indicate the region of significance ($t(1,70) = -5.13$, $k = 60$, FWE corrected $p < 0.05$, peak coordinate MNI -22 -50 30)

The regression analysis results indicate a significant negative correlation ($t(1,70) = -5.13$, $k = 60$, FWE corrected $p < 0.05$, peak coordinate MNI -22 -50 30) between the whole-brain normalized WMH burden and BOLD signal change during finger-tapping (see Figure 4), and no group based significant differences ($t(1,70) = 2.92$, FWE corrected $p = 0.95$) were found. The region of significance is located in the parietal white matter and largely overlaps with regions where at least two subjects presented with WMHs as shown in Figure 5. As shown by the histogram in figure 6, essentially all of the subjects with the highest WMH burden have WMH lesions near (within 1 voxel of) the peak of the ROI identified in the fMRI analysis. Although, the overall number of subjects with co-localized WMHs is limited (13 out of 71 subjects), these subjects account for much of the overall WMH burden in the sample (69%). Thus, the distribution of WMH across subjects in this sample does mostly co-localize with fMRI activity. In support of our assertion, the Wilcoxon rank-sum test results also showed significantly greater ($z = -3.07$, $p < 0.005$) BOLD signal change in participants with co-localized WMHs within a one-voxel neighborhood of the peak coordinate in the ROI. Also, based on the Spearman's rank correlation analysis results, there is no significant correlation between median reaction times and WMHs ($r = -0.08$, $r^2 = 0.006$, $t = 0.65$, $p = 0.52$), indicating that the participant's performance on the task was not significantly associated with WMH severity and thus had no significant affect on the regression analysis results.

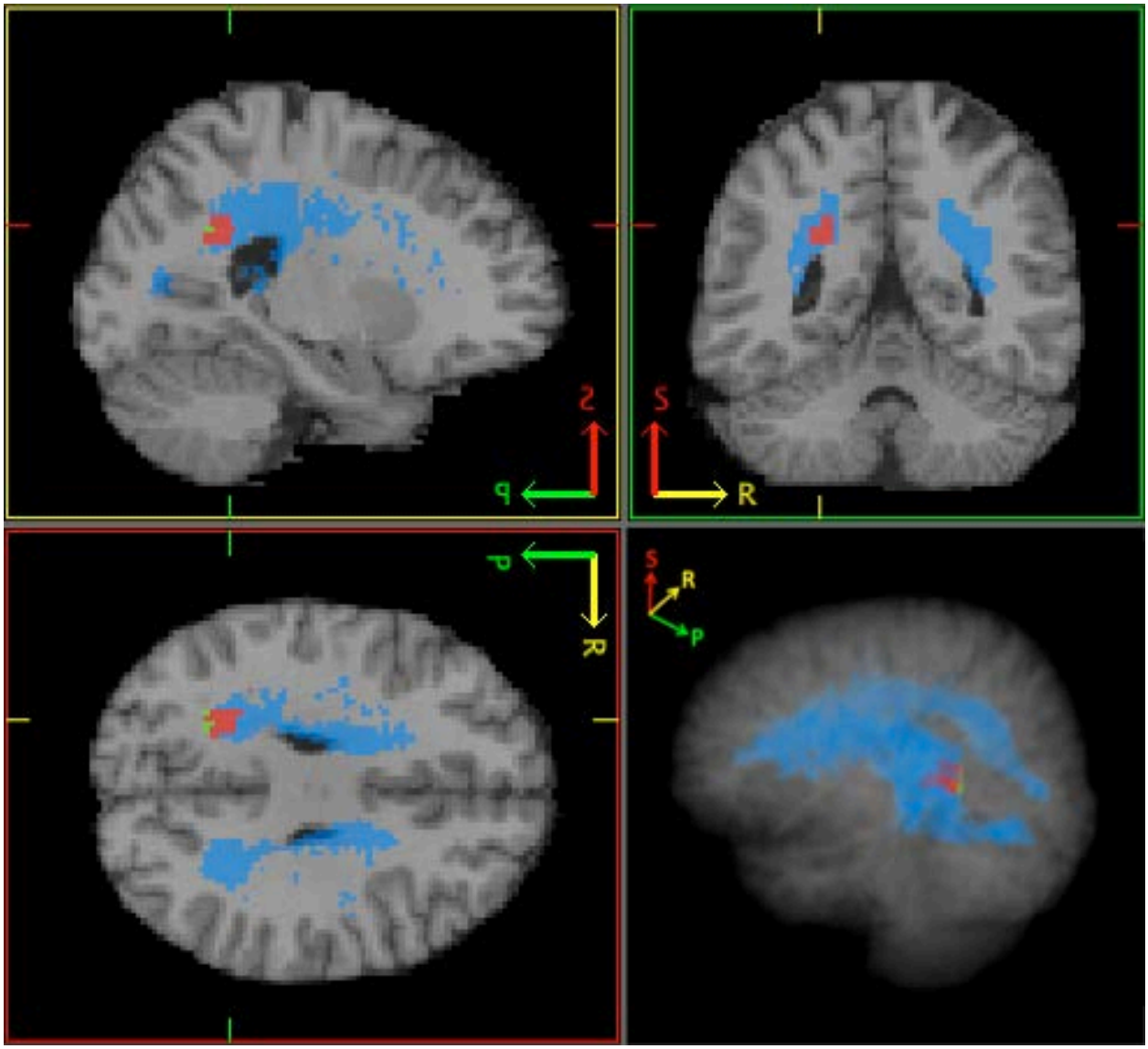


Figure 5. This figure shows the region of significance from the regression analysis (in green), areas where at least 2 subjects had WMHs (in blue), and region of overlap between the two (in red)

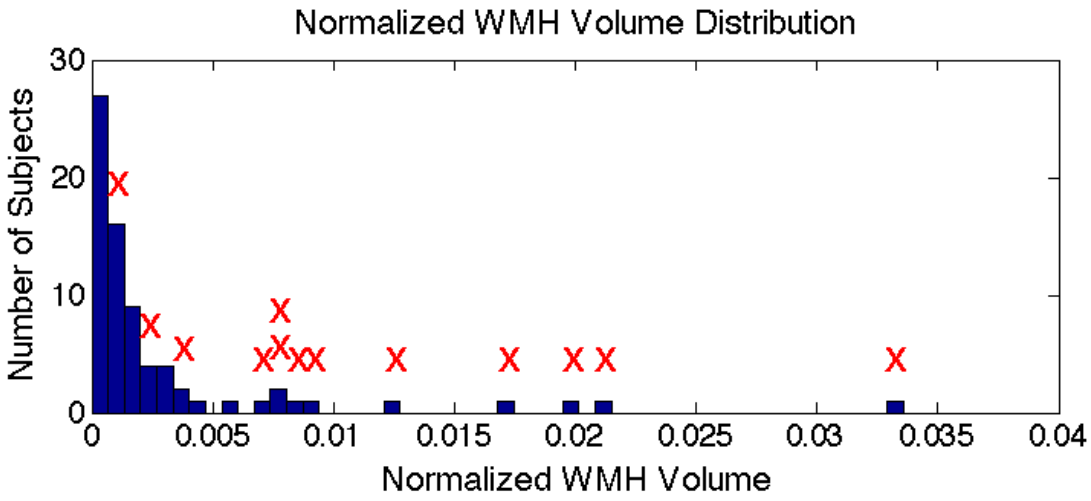


Figure 6. Histogram displaying normalized WMH volume distribution of subjects included in this study. The red "X"s indicate subjects with co-localized WMHs in region of significance from the regression analysis.

5.4 DISCUSSION

In this study, we found WMH burden in the elderly was inversely associated with BOLD signal change on a simple finger-tapping task in a small region of the white matter. Individuals with higher WMH burden showed a decreased BOLD signal change on tapping (relative to fixation condition). The area of significance was located in the parietal white matter: an area that was not strongly associated with the task based on the BOLD signal analysis, but is nevertheless a region where WMHs are found in several individuals. Thus, it is not clear whether WMHs are associated with a global decrease of BOLD signal change, or perhaps have local effects where the WMHs are most prominent. The significant results in the white matter suggest one of two ideas: fMRI is able to detect BOLD signal change reflective of individual differences in neural activation in the white matter, or the presence of WMHs significantly affects the BOLD MR contrast leading to individual differences not necessarily related to neural activation.

Mazerolle et al., 2008 have shown support for detecting a significant BOLD signal change indicating activation in the white matter. Also, Brickman et al., 2009 studied the cerebral blood flow (CBF) using arterial spin labeling (ASL) and concluded that CBF is significantly lower in WMH laden areas compared to normal appearing white matter and grey matter. If indeed it is possible to generate a detectable BOLD fMRI response in the white matter, our results that show a decrease in BOLD signal change (also indicative of decreased CBF) with an increase in WMH burden, would then be consistent with the conclusions of Mazerolle and Brickman. Our results also would agree with conclusions of other perfusion-weighted MRI studies [Marstrand et al., 2002; Sachdev et al., 2004] with similar pathologies (i.e. hypoperfusion) associated with WMHs.

Nevertheless, detection of the BOLD signal in the white matter is controversial. Two main reasons for this are: 1. Cerebral blood flow and volume are lower in the white matter compared to the grey matter, and 2. Post-synaptic potentials, instead of action potentials, are thought to be associated with the BOLD signal [Gawryluk et al., 2009]. Therefore, we believe the stronger argument explaining the significant inverse correlation results is that the T2* BOLD MR contrast is significantly affected by the presence of WMHs. The WMHs are visible on the T2* weighted images and could be significantly affecting preprocessing steps and/or distorting the BOLD signal. Thus, the WMHs and their relation to the functional images need to be further studied. Limitations of this study to fully understand the relation also necessitate future studies. Longitudinal studies and additional BOLD independent studies involving event-related potentials (ERPs) measured using electroencephalogram (EEG), event-related fields (ERFs) measured using magnetoencephalography (MEG), and/or metabolically based ^{18}F-fluorodeoxyglucose positron emission tomography (PET) may help understand the affects of WMHs better. Some

other limitations of this study include the limited sample size and the skewed distribution of WMH burden across subjects. An additional limitation is a possible selection bias, which is suggested by the lack of difference in WMH burden between depressed and control subjects. However, since depression is not a variable of interest for the current report, we do not think the possible subject selection influences these results.

6.0 RELATING STRUCTURAL MRI, DEMOGRAPHIC AND COGNITIVE ABILITY MEASURES TO FUNCTIONAL MRI MEASURES

This chapter describes an experiment that attempts to directly study the relationship between brain structure and function using magnetic resonance imaging. For the study, resting state functional connectivity was used to study brain function because it is simple, allows for a whole brain analysis, and is widely studied for late-life depression in the literature.

6.1 INTRODUCTION

Magnetic resonance imaging (MRI) is used by researchers to study various facets of the human body, especially the brain. Different aspects of the brain are studied using different MR pulse sequences. T1-weighted images are used to study the anatomy, specifically differences and changes in brain regions/structures, due to its high-resolution, which allows for more accurate labeling of regions and defining their boundaries. These images can be used to study the severity of atrophy in brain regions by studying regional volume difference and changes. Diffusion Tensor Imaging (DTI) images are used to gain an understanding of the brain from a microscopic level and study the diffusion of molecules in brain tissues. Two important measures acquired from DTI images include mean diffusivity (MD) and fractional anisotropy (FA), which signify the amount and directionality of

diffusion in tissue respectively. These measures help evaluate the tissue integrity by helping determine brain regions where diffusion is significantly decreased and dispersed due to lesions. T2-weighted images are used to study white matter hyperintensities (WMHs), indicating the presence of ischemic or pre-ischemic white matter lesions. Both local and global volume measures of WMHs are used to study their affect on cognition. Functional MRI (fMRI) images are used to study brain activity as well as functional connectivity between different brain regions. Using each of these MR modalities, the goal of this study is to analyze the relationship of the structure measures acquired from the structural images (i.e. T1-weighted, DTI, and T2-weighted images) in addition to demographic and cognitive ability measures (e.g. Age, Education, Gender, and Mini-Mental Score Examination (MMSE) score) with the functional connectivity measure acquired from the functional images (specifically resting state function images).

Past studies have shown a relationship between functional measures and normal aging differing from neuropsychiatric disorders of aging [Fox & Greicius, 2010; Greicius, 2008]. Several past studies have shown a relation of only a select few of structural [Greicius, et al., 2009; Honey et al., 2009; Steffens et al., 2011; Teipel et al., 2010; Wu et al., 2011], demographic [Weissman-Goel et al., 2010], and/or cognitive ability measures with functional measures. However, to the best of our knowledge, this is the first study to relate all of these non-functional measures with functional measures. This study aims to determine potential biomarkers and achieve a more comprehensive understanding of which non-functional measures are most associated with functional circuit abnormalities in the elderly. To study this association, we used linear regression and artificial neural networks (ANN) as a non-linear alternative for comparison. Both methods were used to attempt to

predict function from structure, demographics, and cognitive ability as well as determine which features (i.e. structural, demographic, and/or cognitive ability) best help predict function. For the functional measure, we chose to study resting state functional connectivity—due to its well-known usage for studying interactions between brain regions—and focus on connectivity between regions involved in one of the major resting state networks: the default mode network. These regions include the amygdala, Brodmann's area 23 (which includes the posterior cingulate cortex (PCC)), hippocampus, inferior parietal, and rectus.

In the long-term, we believe an accurate predictive model will help us better understand the relationship between the structural and functional brain changes associated with normal aging, and also to better identify predictive biomarkers to improve prevention and treatment strategies for the neuropsychiatric disorders of aging. Such a model can help better classify neuropsychiatric disorders and thus lead to the helping provide personalized treatment.

6.2 METHODS

6.2.1 Subject Recruitment

Elderly individuals were recruited for this MRI study from the community and from the healthy controls registry of the Pittsburgh Alzheimer's Disease Research Center. All participants underwent a SCID-IV evaluation. The exclusion criteria included: history of Axis I disorders, stroke, significant head injury, Alzheimer's, Parkinson's, and/or

Huntington's disease. Thirty elderly individuals were included in this analysis. Their demographics and cognitive ability data was also acquired during recruitment.

6.2.2 Image Acquisition

Subjects were scanned on a 3T Siemens TIM TRIO scanner using a 12-channel Siemens head coil. T1-weighted images were acquired with a 1 mm slice thickness, 256x224mm resolution, 256x224mm field of view (FOV), 2300ms repetition time (TR), 900ms inversion time (TI), 3.43ms echo time (TE), and 9° flip angle in the axial plane. T2-weighted Fluid Attenuated Inversion Recovery (FLAIR) images were acquired with a 3 mm slice thickness, 256x240mm resolution, 256x212mm FOV, 9160ms TR, 2500ms TI, 88ms TE, and 150° flip angle in the axial plane. DTI images were acquired with a 3 mm slice thickness, 128x128mm resolution, 256x256mm FOV, 5300ms TR, 2500ms TI, 88ms TE, and 90° flip angle in the axial plane. Functional images were acquired using a gradient-echo-planar imaging sequence with a 3 mm slice thickness, 128x128mm resolution, 256x256mm FOV, 2000ms TR, 34ms TE, integrated parallel acquisition technique = 2, and 90° flip angle in the axial plane.

6.2.3 Image Processing: Feature & Expected Output Values

6.2.3.1 T1-weighted Images

Features extracted from the T1-weighted images include the ratio and sum total of volumes for each pair of region of interest (ROI)s. Together, these measures were included to represent the regional atrophy that may affect functional connectivity. To obtain these

volume measures, the T1-weighted images were first preprocessed: skull stripped using Medical Image Processing, Analysis, and Visualization (MIPAV) software [McAuliffe et al., 2001] and ACPC aligned using Analysis of Functional NeuroImages (AFNI) software [Cox, 1996]. Then, the similarly preprocessed Montreal Neurological Institute (MNI) colin template and its corresponding labeled regions were registered to the preprocessed T1-weighted image. The volume measure for each ROI was calculated by dividing the number of voxels in the registered and thresholded ROI by the number of voxels in the whole brain.

6.2.3.2 DTI Images

Features extracted from the DTI images include, the average MD in each pair of ROI, the weighted average FA in the tracks connecting each pair of ROI, and the approximate number of tracks connecting each pair of ROI. Together, these measures were included to represent the diffusivity and integrity of the gray and white matter that may affect functional connectivity. To obtain these measures, FMRIB Software Library (FSL) [Jenkinson et al., 2012] was used to perform eddy current correction, dti (diffusion tensor imaging) reconstruction, and tractography. In addition, the subject's T1-weighted image and previously registered ROIs were registered to the subject's DTI space using FSL. For each pair of registered ROIs, the resulting MD map was used to calculate the average MD, FA map was used to compute the weighted average FA among the tracks found during tractography, and the number of tracks found was also recorded.

6.2.3.3 T2-weighted FLAIR Images

Features extracted from the T2-weighted FLAIR images include the amount of global and local track-based WMHs. Together, these measures were included to represent the amount of

white matter lesions that may affect functional connectivity. To obtain these measures, the insight toolkit (ITK) [Yoo et al., 2002] was used to segment the WMHs. Then, the subject's DTI and corresponding ROI masks were registered to the subject's T2-weighted FLAIR image. The global and local WMH volume measure was computed by dividing the number of voxels part of the WMH segmentation and the number of these voxels intersecting the tracks connecting each pair of ROI respectively by the number of voxels in the whole brain.

6.2.3.4 Resting State Functional Images

Expected output values were extracted from the resting state function images and comprised the Fisher transformed correlation coefficient representing the functional connectivity between each pair of ROI. To obtain this measure, all functional images were realigned to the first image in the sequence and then co-registered with the subject's structural grey matter. The co-registered structural MR image and all realigned functional images were normalized to Montreal Neurological Institute (MNI) space using the a priori grey matter template in Statistical Parametric Mapping 5 software (SPM5) [Friston et al., 1994]. The normalized functional images were smoothed with a Gaussian kernel (full width at half maximum (FWHM) = 10mm) to account for the greater morphologic variability in elderly subjects [Reuter-Lorenz PA and Lustig, 2005]. The affects of the head motion parameter were then regressed out of the resulting data and it was high pass filtered at a cut off of 100 Hz to obtain the resting state brain activity at each voxel. Then, Region-of-interest extraction toolbox (REX) was used to get the 1st eigenvariate time series from the processed data for each ROI and Fisher transformed correlation coefficients were computed between the resulting time series for each pair of ROIs. These coefficients represent how well two

regions are connected functionally (greater coefficient value = greater functional connectivity), and serve as functional connectivity indices (FCI).

6.2.4 Statistical Learning

For this study, we compared results from two methods: linear regression and artificial neural networks. For each method, the features included—which were all scaled to a range of [0,1] before input—and corresponding actual/expected output variable are listed in table 3. A training dataset of 150 samples, validation dataset of 50 samples, and test dataset of 100 samples was randomly selected from the larger data set. First to obtain optimal values of each parameter involved in the respective methods tested, the training dataset was trained using each method. Then, the trained weights were used to obtain prediction outputs for the validation dataset and obtain predicted outputs. The predictions were used to compute the mean squared error (MSE) for the chosen parameters. The combination of parameter values that resulted in the smallest MSE were selected to obtain prediction outputs for the training and test datasets. Lastly, to analyze the performance of each method, we did the following between the predicted and expected outputs of both the train and test datasets: (1) performed t-tests to study the significance difference of means, and (2) computed the train and test MSEs of the predicted outputs to study the overall difference taking into account the variance and bias.

Table 3. Feature inputs and expected output variable

	Dataset:
Feature Inputs	**Feature Type**
Constant	
Age	Demographic
Gender (F=1/M=0)	Demographic
Education	Demographic
MMSE score	Cognitive Ability
ROIs volume ratio	Structural (T1)
ROIs volume total	Structural (T1)
Global WMH volume	Structural (T2 FLAIR)
Local WMH volume	Structural (T2 FLAIR)
MD (average)	Structural (DTI)
FA (weighted)	Structural (DTI)
# of Tracks (average)	Structural (DTI)
Expected Output	**Output Type**
FCI	Functional

6.2.4.1 Linear Regression

A gradient descent algorithm with a prior was performed to calculate the weights θs in equation 1 that would help predict the output y (where y is the FCI) given features $x_{k|k = 0,\ldots,12}$ (listed in table 3) such that the MSE between the prediction and actual values of y is minimized. The parameters varied to obtain the optimal set of $\theta_{k|k = 0,\ldots,12}$ included the prior factor (λ) and the step size α. For each value of α and the optimal prior, the objective function value was plotted as function of number of iteration (max number of iterations was set to 500) [Kivinen, 1997].

$$y = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_6 + \theta_7 x_7 + \theta_8 x_8 + \ldots$$

$$\ldots + \theta_9 x_9 + \theta_{10} x_{10} + \theta_{11} x_{11} + \theta_{12} x_{12} \qquad [1]$$

6.2.4.2 Artificial Neural Networks

A backpropagation algorithm was used to compute the optimal weights for a 3-layered artificial neural network with one hidden layer. All weights were trained using sigmoid functions. The algorithm iteratively trains the dataset and terminates when the MSE stabilizes and is less than a set error margin = 0.0001. The parameters varied to obtain the optimal set of weights between each layer included the learning rate, the starting value range of every weight, and the number of hidden layer nodes. Based on the starting value range, each weight was initialized to a random value within this range. The algorithm was run five times and the iteration where the predicted outputs of the validation set produced the smallest MSE was used to select optimal weights for the network [Gershenson, 2003].

6.3 RESULTS

6.3.1 Linear Regression

For the linear regression model, the following parameter values were found to be most optimal and resulted in the least MSE: (1) prior value of $\lambda = 0.1$, and (2) step size of $\alpha = 0.01$. The optimal weights computed by training the model on these parameter values are shown in table 4 along with the mean, standard deviation, t-test results, and MSE of the predicted outputs. The corresponding objective function is shown in figure 7.

6.3.2 Artificial Neural Networks

For the ANN model, the following parameter values were found to be most optimal and resulted in the least MSE: (1) learning rate = 86, (2) starting value range = [0.07-0.9], and (3) number of hidden layer nodes = 6. The optimal weights computed by training the model on these parameter values are show in table 5 along with the mean, standard deviation, t-test results, and MSE of the predicted outputs.

Table 4. Optimal weights for the linear regression model with corresponding predicted output analysis

Feature Inputs	**θs**	
Constant	2.499	
Age	-0.683	
Gender (F=1/M=0)	-0.011	
Education	-0.130	
MMSE score	-1.728	
ROIs volume ratio	0.178	
ROIs volume total	0.303	
Global WMH volume	-0.123	
Local WMH volume	-0.379	
MD (average)	-0.096	
FA (weighted)	0.134	
# of Tracks (average)	1.236	
Predicted Output	**Train Dataset**	**Test Dataset**
FCI Mean (STD)	0.503 (0.240)	0.481 (0.235)
t-value (p-value)	-0.052 (0.959)	0.527 (0.599)
Mean Squared Error	0.110	0.105
Correlation Coefficient (p-value)	0.55 (<0.0001)	0.39 (<0.0001)

Table 5. Optimal weights (top 12x6 matrix: weights connecting input nodes to hidden layer nodes; middle 1x6 matrix: weights connecting hidden layer nodes and output node) for the ANN model with corresponding predicted output analysis

WEIGHTS	**Hidden Layer Nodes:**					
Feature Inputs:	**1**	**2**	**3**	**4**	**5**	**6**
Constant	-2.067	-0.462	-1.829	-1.756	-2.710	-1.928
Age	-1.776	-0.528	-0.792	-1.012	-2.069	-1.480
Gender (F=1/M=0)	-1.646	-0.405	-1.518	-0.717	-1.210	-1.336
Education	-0.836	-0.314	-1.079	-1.180	-1.896	-0.762
MMSE score	-1.752	0.119	-1.745	-1.535	-2.360	-1.589
ROIs volume ratio	0.263	0.949	0.133	0.085	0.086	-0.494
ROIs volume total	-0.329	0.006	-0.935	-0.881	-1.076	-0.514
Global WMH volume	0.450	1.167	0.306	0.341	-1.088	-0.061
Local WMH volume	0.600	0.649	0.212	0.213	-0.119	-0.012
MD (average)	-1.272	-0.147	-1.026	-0.609	-1.276	-1.232
FA (weighted)	-0.420	-0.300	-0.666	-1.196	-1.658	-0.723
# of Tracks (average)	0.105	2.736	0.374	0.506	1.313	0.915
	Hidden Layer Nodes:					
Output:	**1**	**2**	**3**	**4**	**5**	**6**
FCI	0.927	1.930	0.607	0.450	1.167	1.004
Predicted Output:	**Train Dataset**	**Test Dataset**				
FCI Mean (STD)	0.642 (0.057)	0.635 (0.056)				
t-value (p-value)	4.17 (4.03e-05)	5.20 (4.86e-07)				
Mean Squared Error	0.165	0.132				
Correlation Coefficient (p-value)	0.31(<0.0001)	0.34 (0.0005)				

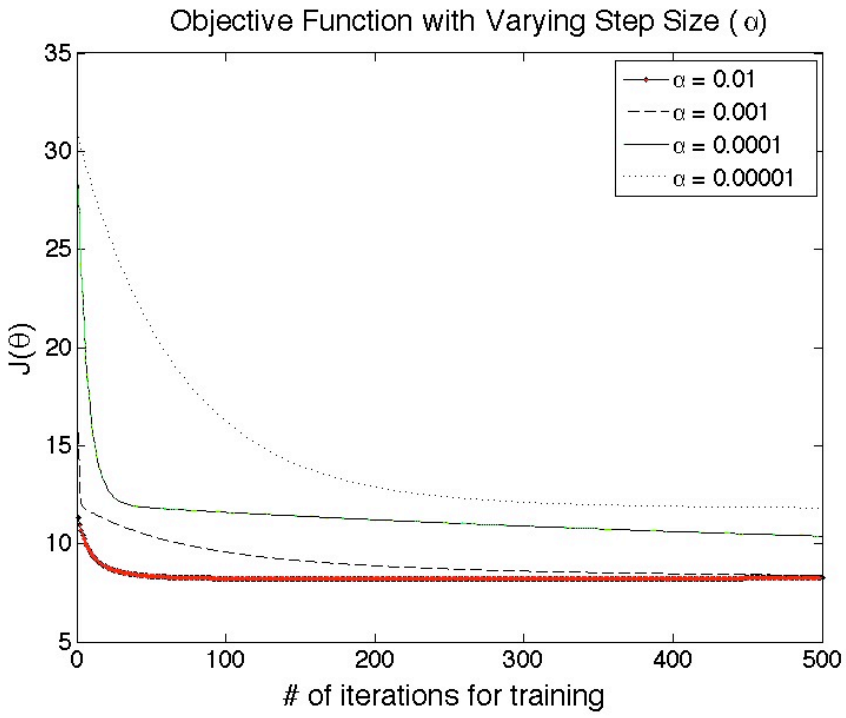


Figure 7. Objective Function plots for the various step sizes (alpha) tested at a prior value of lambda = 0.1 (Note: for alpha = 1 and 0.1 the values are significantly larger and are thus left out; the optimal alpha value is 0.01).

6.4 DISCUSSION

In terms of mean squared error, the results indicate that linear regression performed better than the ANN model. This may suggest that this non-linear model is overfitting the model studying the relationship of brain structure, demographics, and cognitive ability with brain function. Most notably, the ANN predicted outputs have significantly different means from the expected outputs based on the t-test results ($p < 0.0001$) unlike the linear regression model (p (train) = 0.959, p (test) = 0.599).

However, when evaluating the correlation between the predicted and actual labels for both training and test datasets, both methods seem to be underperforming. Even though the

correlation coefficients are found to be highly significant ($p < 0.001$), the coefficients themselves are low ($r < 0.6$); this suggests discrepancies between the predicted and actual labels. Thus, we speculate that either (1) it cannot be assumed that the association between brain structure and regional connectivity is independent of the selection of regions, (2) the high sparsity of input feature data does not allow the learning methods to generalize a good prediction model, (3) there is an added bias due to using multiple regions from the same subject as individual training samples, (4) there is a lack of relevant features in the data set since most of them have a logarithmic non-Gaussian distribution, or (5) even with an ideal model structure is not fully predictive of function as measured by MRI markers used in this study. For future work, we believe either subject-wise analysis or an improved set of features with less sparsity may help estimate more accurate prediction models.

7.0 PREDICTING LATE-LIFE DEPRESSION AND TREATMENT RESPONSE

This chapter describes an experiment that attempts to estimate prediction models for late-life depression and treatment response using both structural and functional imaging measures, as well as non-imaging measures. Also, another portion of this study is explained in “Appendix A”. This portion of the study is performed as follow-up to the experiment defined in the previous chapter.

7.1 INTRODUCTION

In a given year, approximately 2 million people aged 65+ suffer from late-life depression (LLD) not associated with normal aging [Mental Health America]. The current diagnosis and treatment of LLD is based on behavioral symptoms and signs. It lacks the reliability and validity that could accrue from biomarkers of underlying brain characteristics. To advance towards personalizing medicine, it is important to identify biomarkers reflecting the neural circuit abnormalities that characterize LLD.

For this study, we focused on the association of brain structure and function to late-life depression (LLD) and its treatment response in an elderly population. The following measures were used: demographic characteristics, cognitive ability, brain structure, and brain function. Demographic and cognition measures included: age, education, gender, and Mini-Mental Score

Examination (MMSE). Magnetic resonance imaging (MRI) were used to extract brain structure and function measures.

Past studies have evaluated the association of the diagnosis and treatment response of LLD with select few of the demographic [Blazer, 2012; Chang-Quan et al., 2010; Forlani et al., 2013; Katon et al., 2010; Luppa et al., 2012; Wild et al., 2012; Wu et al., 2012], clinical [Andreescu et al., 2008], cognition ability [Bhalla et al., 2005; Ganguli et al., 2006; Kohler et al., Apr 2010; Ribeiz et al., 2013; Wilkins et al., 2009], MR structural [Alexopoulos et al., 2008; Aizenstein et al., 2011; Change et al., 2011; Colloby et al., 2011; Crocco et al., 2010; Disabato et al., 2012; Firbank et al., 2012; Gunning et al., 2009; Gunning-Dixon et al., 2010; Kohler et al., Feb 2010; Mettenburg et al., 2012; Sexton et al., 2013; Shimony et al., 2009; Taylor et al., 2008; Taylor et al., 2011; Teodorczuk et al., 2010], and/or MR functional measures [Alalade et al., 2011; Alexopoulos et al., 2012; Andreescu et al., 2011; Andresscu et al., 2013; Bohr et al., 2012; Colloby et al., 2012; Liu et al., 2012; Steffens et al., 2011; Wang et al., 2008; Wu et al., 2011]. However, to the best of our knowledge this is the first study to explore a wide range and combinations of demographic, cognitive ability, MR structural, and MR functional measures in association to LLD diagnosis and treatment response. These measures include: (1) normalized total gray plus white matter volume, and average normalized regional volume to measure brain tissue atrophy using T1-weighted images; (2) average mean diffusivity (MD), average weighted fractional anisotropy (FA), and average number of tracts connecting regions of interest to measure white matter integrity using DTI images; (3) normalized global and average local track-based white matter hyperintensity volume to measure white matter lesions using T2-weighted images; (4) age, gender and education; (5) Mini-Mental State Examination score; and (6) average Fisher transformed correlation coefficients to measure resting state functional connectivity using

fMRI images. By using a broader spectrum of features, we hope to get a more complete and accurate understanding of the underlying mechanisms of the brain associated with LLD. An international study in progress by Grieve et al., 2013 has presented some preliminary results showing the importance of multimodal MRI measures as potential biomarker for depression in a younger population. Compared with mid-life depression, LLD has a different neural signature including gray matter (GM) and white matter (WM) structural changes [Aizenstein et al., 2014] along with a more difficult treatment response [Andreescu et al., 2011]. Thus, we believe that the underlying circuitry associated with the diagnosis and treatment response of LLD may involve a combination of structural and functional biomarkers. Also, we aimed to better understand how the features relate to and affect one another.

Using the unique set of measures described above, we aimed to develop a model that can accurately predict the diagnosis and treatment response of LLD. By creating a predictive model, we hope to increase our understanding of LLD and aid in the progress towards personalized treatment. There have been several past studies that have successfully explored predictive models for diagnosis [Costafreda et al., 2009; Fu et al., 2008; Hahn et al., 2011; Marquand et al., 2008; Mwangi et al., Jan 2012; Mwangi et al., May 2012; Nouretdinov et al., 2011; Zeng et al., 2012] and treatment response [Costafreda et al., 2009; Liu et al., 2012; Marquand et al., 2008; Nouretdinov et al., 2011] of depression in the younger populations. Also, most of these studies have focused on utilizing support vector machines as their classifier. However, to the best of our knowledge, there are no past studies that have attempted at establishing a predictive model for the elderly population by comparing multiple classification methods. Given that past studies have been successfully able to form predictive model for diagnosis (94.3% classification accuracy using fMRI by Zeng et al., 2012) and treatment response (88.9% classification accuracy

using structural MRI by Costafreda et al., 2009) of depression in younger populations, we also believe it is possible to do so for LLD. Considering the increased complexity of brain structure and function in the elderly population (resulting from age and disease), we studied multiple classification methods for predictive models including: L1-regularized Logistic Regression, Support Vector Machines, and Alternating Decision Trees. Additionally, we focus on resting state networks including the default mode network and salience network, which have been studied in LLD [Alalade et al., 2011; Alexopoulos et al., 2012; Andreescu et al., 2011; Andreescu et al., 2013; Bohr et al., 2012; Gunning et al., 2009; Steffens et al., 2011; Wu et al., 2011]. Unlike past studies that focused on region-based approaches (e.g. regions resulting from voxel-wise analysis [Costafreda et al., 2009; Fu et al., 2008; Hahn et al., 2011; Marquand et al., 2008; Liu et al., 2012; Mwangi et al., Jan 2012; Mwangi et al., May 2012; Nouretdinov et al., 2011] or anatomical regions of interest (ROIs) [Zeng et al., 2012]), we perform a whole brain and network analyses using functional ROIs in order to reduce data complexity and more precisely represent the brain areas activated during a functional activity (e.g. resting state) of interest [Nieto-Castanon et al., 2003].

To our knowledge, this is the first study to estimate prediction models for the diagnosis and treatment response in late-life depression (LLD), by evaluating: (1) the potential of multimodal magnetic resonance imaging (MRI) measures as biomarkers; (2) combinations and interactions between potential imaging and non-imaging predictors; (3) multiple learning methods; and (4) whole brain and network analyses using functional ROIs.

7.2 METHODS

7.2.1 Subject Recruitment

Non-psychotic, unipolar LLD patients (n = 33) and elderly non-depressed (ND) (n = 35) individuals were recruited from the Pittsburgh's Advanced Center for Intervention and Services Research for Late-Life Mood Disorders, and Alzheimer Disease Research Center's healthy controls registry respectively. Each participant provided written informed consent after receiving a full description of the study. All participants were paid $50. Participants were evaluated using the Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders 4th edition (SCID-IV) [First et al., 1995]. Based on their SCID-IV evaluation, participants were excluded if they had a history of Axis I disorders other than major depressive disorder and anxiety disorders, stroke, significant head injury, Alzheimer's, Parkinson's, and/or Huntington's disease. The recruited patients were treated with duloxetine, venlafaxine, nimodipine, or escitalopram. Upon recruitment and after 12 weeks of treatment during a follow-up, these participants were assessed for LLD severity using the Hamilton Depression Rating Scale (HAM-D). If a participant showed an improvement in their HAM-D score of less than ten [Roose et al., 1994], she or he was classified as a responder to treatment. If the necessary HAM-D scores were not available (n = 3), other depression scales (i.e. Montgomery–Åsberg Depression Rating Scale or Patient Health Questionnaire-9) were used in conjunction to clinician notes of patient progress. Any subject without sufficient response data was excluded from the treatment response based analyses.

From all recruited individuals, one LLD individual was excluded from the analysis due to excessive functional MRI (fMRI) head motion artifacts. Two more LLD and four ND individuals

were excluded due to bad fMRI registration to template. Additionally, individuals were excluded due to incorrect ROI(s) registration to structural images (this includes three LLD and three ND individuals for the dorsal default mode network (dDMN) analysis; and five LLD individuals for the anterior salience network (aSN) analysis). For the treatment response analysis, three additional LLD individuals were excluded due to missing or partial treatment response. The demographics and cognitive ability data for each individual also were acquired during recruitment and are summarized in table 6.

Table 6. Summary of Participants-Related Information

	Depression Analysis			**Treatment Response Analysis**		
Network	dDMN	aSN	Both	dDMN	aSN	Both
# of ND	28	31	28	n/a	n/a	n/a
# of LLD	27	25	22	24	22	19
# of Responders	n/a	n/a	n/a	11	11	9
# of Non-Responders	n/a	n/a	n/a	13	11	10
Age [avg (stdev)] years	70.20 (7.98)	70.00 (7.85)	69.78 (7.81)	68.83 (7.52)	68.36 (6.54)	67.37 (6.45)
Gender (% Female)	76.36%	78.57%	78.00 %	75.00%	81.82%	78.95%
Education [avg (stdev)] years	14.36 (2.45)	14.43 (2.51)	14.18 (2.40)	14.75 (2.75)	14.45 (2.69)	14.37 (2.75)
MMSE [avg (stdev)] score	28.40 (2.01)	28.32 (2.08)	28.36 (2.10)	27.58 (2.60)	27.41 (2.65)	27.26 (2.83)
HAM-D at baseline[(a)] [avg (stdev)] score	20 (4)	20 (4)	21 (4)	20 (4)	20 (4)	21 (4)
HAM-D at 12 week follow-up[(a)] [avg (stdev)] score	10 (5)[(b)]	10 (6)[(b)]	10 (6)[(b)]	10 (5)	10 (6)	10 (6)

(Note: In the above table, “avg” is short for average and “stdev” is short for standard deviation)

[(a)] Information regarding HAM-D scores is presented for LLD participants only.

[(b)] HAM-D scores are missing for 3 participants and thus they were not included in the calculations.

7.2.2 Image Acquisition

A 3T Siemens TIM TRIO scanner with a 12-channel Siemens head coil was used to scan the subjects. T1-weighted, T2-weighted, T2-weighted fluid attenuated inversion recovery (FLAIR), diffusion tensor imaging (DTI) and resting state fMRI (rs-fMRI) images were acquired for each subject. The parameters for T1-weighted images were: 1mm slice thickness, 256x224mm resolution, 256x224mm field of view (FOV), 2300ms repetition time (TR), 900ms inversion time (TI), 3.43ms echo time (TE), and 9° flip angle in the axial plane. The parameters for T2-weighted images were: 3mm slice thickness, 256x224mm resolution, 256x224mm field of view (FOV), 3000ms repetition time (TR), 100ms inversion time (TI), 11/101ms echo time (TE), and 150° flip angle in the axial plane. The parameters for T2-weighted FLAIR images were: 3mm slice thickness, 256x240mm resolution, 256x212mm FOV, 9160ms TR, 2500ms TI, 88ms TE, and 150° flip angle in the axial plane. The parameters for DTI images were: 3mm slice thickness, 128x128mm resolution, 256x256mm FOV, 5300ms TR, 2500ms TI, 88ms TE, and 90° flip angle in the axial plane. The parameters for the rs-fMRI images acquired using a gradient-echo-planar imaging sequence were: 3mm slice thickness, 128x128mm resolution, 256x256mm FOV, 2000ms TR, 34ms TE, integrated parallel acquisition technique = 2, and 90° flip angle in the axial plane. For the rs-fMRI scans, the subjects were asked to stay awake, think of nothing in particular and rest with eyes focused on a fixation cross.

7.2.3 Regions of Interest (ROIs) Selection

For this study, functional ROIs from the dorsal default mode network (dDMN) and anterior Salience Network (aSN) were used. All methods described below were repeated for the dDMN,

aSN, and both networks combined. ROIs for both networks were obtained from the FIND Lab at Standford University [Shirer et al., 2012].

7.2.4 Image Processing: T1-weighted High Resolution (Hi-Res) Image Features

From the T1-weighted images the following 2 features were extracted: (1) normalized whole brain gray and white matter volume, and (2) average of normalized ROIs' gray matter volumes. These measures represent the whole brain and regional atrophy respectively. To obtain these volume measures, the T1-weighted images were first preprocessed: aligned along the anterior and posterior commissure line using 3DSlicer [Pieper et al., 2004] software and skull stripped using ITK-SNAP software [Yushkevich et al., 2006]. Then, the skull-stripped Montreal Neurological Institute (MNI) colin template and the ROIs in colin space were registered to the preprocessed T1-weighted image. The T1-weighted images were also segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) using FMRIB Software Library (FSL) [Jenkinson et al., 2012]. Each registered ROI was then thresholded by the respective gray matter segmentation. Normalized volume measures were computed by dividing the number of voxels in each region by the number of voxels in the intracranial volume of the brain.

7.2.5 Image Processing: DTI Image Features

From the DTI images the following 3 features were extracted: (1) the average of all ROIs' mean MD, (2) the average of all ROIs' weighted average FA in the tracks connecting each pair of ROIs, and (3) the average of the approximate number of tracks connecting each pair of ROIs.

These measures represent the diffusivity, integrity, and structural connectivity respectively of the gray and white matter. These measures were acquired using FSL to perform eddy current correction, dti reconstruction, and tractography. FSL was also used to register each subject's T1-weighted image and ROIs in T1-weighted subject space to the subject's DTI space. For each pair of ROIs in DTI space, average MD was calculated from the resulting MD map, the weighted average FA among the tracks found during tractography were calculated from the resulting FA map, and the number of tracks found during tractography was also recorded.

7.2.6 Image Processing: T2-weighted FLAIR Image Features

From the T2-weighted FLAIR images the following 2 features were extracted: (1) the amount of global WMHs, and (2) the average of the amount of local track-based WMHs between each pair of ROIs. These measures represent the amount of white matter lesions. For these measures, we first segmented the WMHs using the insight toolkit (ITK) [Yoo et al., 2002]. Next, the subject's DTI image was registered to the subject's T2-weighted FLAIR image space and the transformation was applied to the corresponding ROI masks using FSL. Then, we divided the number of voxels in the WMH segmentation and the number of these voxels intersecting the tracks connecting each pair of ROIs by the number of voxels in the whole brain to obtain the global and local WMH volume measures respectively.

7.2.7 Image Processing: Resting State Functional Images

From the rs-fMRI images, we extracted the average of Fisher transformed correlation coefficients between each pair of ROIs. For the purposes of this study, these coefficients are

referred to as functional connectivity indices (FCI). This measure represents the average overall network functional connectivity. For this measure, all functional images were processed using CONN [Whitfield-Gabrieli et al., 2012]. First, they were preprocessed: slice timing corrected, realigned to the first image in the sequence, co-registered with the subject's T2-weighted structural gray matter, normalized to template space, and smoothed with a Gaussian kernel. Then, the head motion artifacts were regressed out and the data were band pass filtered at cut offs of 10 and 100 Hz to acquire the resting state brain activity at each voxel. Then, the 1st eigenvariate time series from the processed data for each ROI and Fisher transformed correlation coefficients between each pair of ROIs were computed.

7.2.8 Feature Selection

After acquiring all the required features (see table 7 for a summary of all features), 13 different sets of features selected by force (i.e. a feature(s) was explicitly chosen to be removed from the full set of features) were analyzed using the statistical learning methods described below. Table 8 describes the features removed for each set. Force feature selection using these specific 13 feature sets was performed to study the influence of (1) different MRI modalities, (2) imaging vs. non-imaging measures, and (3) each individual feature on the prediction of LLD diagnosis and treatment response. Furthermore, different feature reduction methods including principal component analysis (unsupervised method) and a filter technique using Kendall's tau correlation coefficient (supervised method) were tested using these feature sets. However, except for support vector machines, these feature reduction methods did not improve results—possibly due to the embedded feature selection properties of alternating decision trees and L1 regularized Logistic Regression—and thus are not presented in this article. With the SVM methods, only the affects

of the filter technique are presented since it produced the best results. Lastly, we also tested a region-wise analysis instead of whole network analysis by including features for each region pair instead of the overall averages, but this also did not produce high accuracy results possibly due to high data sparsity, dimensionality and overfitting. So, for this study, we decided to limit the analysis to as few features as possible by performing whole network analysis.

7.2.9 Statistical Learning

For this study, we compared results between both generalized linear (L1 Regularized Logistic Regression (L1-LR) and Support Vector Machines with Linear Kernel (SVM-L)) and nonlinear (Alternating Decision Tree (ADTree) and Support Vector Machines with Radial Basis Function Kernel (SVM-RBF)) classification-based learning methods. A variety of both generalized linear and nonlinear methods were included to more precisely determine the nature of the data. Furthermore, SVM methods were chosen due to their popularity in the current literature [Costafreda et al., 2009; Fu et al., 2008; Liu et al., 2012; Marquand et al., 2008; Mwangi et al., May 2012; Nouretdinov et al., 2011; Zeng et al., 2012], versatility for kernel functions plus features' dimensionality, and convergence speed [Cortes & Vapnik, 1995]. L1-LR and ADTree were chosen due to their embedded feature reduction abilities, simplicity in interpretation of results, and fast convergence speed [Pfahringer et al., 2001; Yuan et al., 2010].

For all learning methods except ADTree—which is not affected by variations in distribution of values between the features—all the features were standardized before input. Each method was used to predict two expected output variables separately: depression and treatment response. Due to the small sample size, a leave-one-out cross validation (LOOCV) method was used to determine classification accuracy of each expected output variable. Additionally, optimal

values of each method's varied parameter were chosen at every LOOCV iteration based on greatest classification accuracy from a nested-LOOCV on the train set. Both the average train and test set accuracy from the LOOCV were recorded for all tests performed; this includes combinations of 13 feature sets, 3 networks, 4 learning methods, and 2 expected output variables. In addition, the respective specificity, sensitivity, and ROC curve measures were also recorded.

7.2.9.1 L1-Regularized Logistic Regression (L1-LR)

For L1-LR, a coordinate descent method using one-dimensional Newton directions as described by Yuan et al., 2010 was coded and implemented in-house in Python. This learning method attempts to fit the data to a regularized logistic loss function f(w) by finding minimum values of w that best fit equation 1—where n is the total number of instances (i.e. examples of the input data described by feature vectors) in the data, y_i represents the label of the expected output variable for the i^{th} instance, x_i is the i^{th} instance of the input data, w are the weights associated with the input features and estimated by the learning method to obtain a best fit logistic model, and the constant variable C is used to balance the regularization term ($\|w\|_1$) and loss term ($\Sigma\log(1+e^{-y((w\hat{}T)x)})$). The regularization term prevents the model from becoming too complex and thus reduces the risk of overfitting. The loss term helps optimize the prediction model—i.e. compute the optimal weights—to achieve the highest classification accuracy. For this method, the parameter varied was the variable C because it is essential in controlling the sparsity of the final model weights and consequently the features selected by the algorithm. Values tested for this variable include: $\{2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{4}\}$. One variation made to the algorithm described by Yuan et al., 2010 was an addition of an input feature of constant value equal to one for each i^{th} instance. This additional feature acts as a bias or intercept term that also affects the regularization term. This variation improved the overall classification accuracy.

$$f(w) \equiv \|w\|_1 + C\sum_{i=1}^{n} \log(1 + e^{-y_i(w^T x_i)}) \quad [1]$$

7.2.9.2 Support Vector Machines (SVM)

The Sci-Kit Learn Python [Pedregosa et al., 2011] library was used to implement SVM with both a linear and nonlinear radius basis function (RBF) kernel. This learning method attempts to find a hyperplane and corresponding support vectors (i.e. train set samples closest to the hyperplane) that best divide the data by their label values. The varied parameter for this method was the penalty parameter of the error term because it controls for the amount of noise in the data by affecting the margin size between support vectors and orientation of the hyperplane. Values tested for this variable include: $\{2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{4}\}$. To improve the performance of the SVM algorithm, we also implemented a supervised feature selection, as presented by Zeng et al., 2012, during every iteration of the LOOCV in Python. This feature selection method is a filter technique that first divides the data instances into two groups by corresponding labels. Then, it computes the Kendall tau correlation coefficient between the groups for every feature and selects features with the highest correlation coefficients based on a predefined threshold. For our study, we set the threshold to equal half of the maximum correlation coefficient among the features for every LOOCV iteration.

7.2.9.3 Alternating Decision Tree (ADTree)

The optimized version of ADTree presented by Pfahringer et al., 2001 and provided in WEKA [Hall et al., 2009] was coded and implemented in-house in Python. This learning method combines ADA boost and decision tree methods in attempts to create a tree with multiple paths

and optimal splitting criterion (i.e. select features and corresponding threshold values that determine the path taken by a given data instance) that nonlinearly classifies the data. The varied parameter for this method was the number of boosting iterations. Values tested for this variable include: {3,4,5,6,7,8,9,10}. The smallest value that gave the greatest accuracy was selected for every iteration of the LOOCV. This variable was varied because it determines the number of branches of the final tree and behaves like a post-pruning method to reduce the risk of overfitting (i.e. the chances of developing too complex of a prediction model such that it represents more noise than the underlying relationship). Since ADTree is a nonlinear method and this study uses a small sample size, it is important to take measures that reduce the risk of overfitting. To further reduce the risk of overfitting as well as increase the convergence speed, we modified the ADTree algorithm presented by Pfahringer et al., 2001 to produce a less complex prediction model in the following ways: (1) a pre-pruning restriction was applied to prevent the tree from growing more than 3 branches in depth, and (2) the number of splitting criterion thresholding options for each feature was minimized to x/2—where x is the number of unique values of a given feature in the training set—by using a modified version of the method described by Quinlan, 1996; the method described by Quinlan 1996 sets the averages of adjacent value pairs from an array of sorted unique values x as the splitting criterion options; we further modified this method by continuing to take averages of the resulting averages until only x/2 splitting criterion thresholding options remain.

Table 7. Summary of Features

Feature Type	Feature	Feature Short Forms	Representation
Demographics	Age	Age	Whether younger or older old adult
Demographics	Gender	Gender	Whether female or male
Demographics	Education	Level of Education	Number of years formal education was received
Cognitive Ability	MMSE	Mini-Mental State Examination score	Whether strong or poor cognitive ability
Functional Imaging (rs-fMRI)	Functional Connectivity Index (FCI)	Average of Fisher transformed correlation coefficients between each pair of ROIs	Degree of functional connectivity—allowing for communication between ROIs—within network
Structural Imaging (T1-weighted Hi-Res)	Normalized whole brain tissue volume (NWBTV)	Normalized whole brain gray and white matter volume	Degree of whole brain atrophy
Structural Imaging (T1-weighted Hi-Res)	Normalized ROIs' tissue volume (NRTV)	Average of normalized ROIs' gray matter volumes	Degree of network-based regional brain atrophy
Structural Imaging (DTI)	Mean Diffusivity (MD)	Average of all ROIs' mean gray matter MD	Amount of diffusivity within gray matter of network
Structural Imaging (DTI)	Number of tracks	Average of the approximate number of tracks connecting each pair of ROIs	Amount of structural connectivity—allowing for communication between ROIs—within network
Structural Imaging (DTI)	Fractional Anisotropy (FA)	Average of all weighted mean FA computed along tracks connecting each pair of ROIs	Amount of white matter integrity within network
Structural Imaging (T2-weighted FLAIR)	Global White Matter Hyperintensities (WMHs)	WMH burden in cerebral cortex	Amount of global (i.e. whole brain) WMH burden or white matter lesions
Structural Imaging (T2-weighted FLAIR)	Local WMHs	Average WMH burden among tracks connecting each pair of ROIs	Amount of WMH burden or white matter lesions within network

Table 8. Description of Feature Sets (Note: Blocks in gray indicate the features removed for each set)

	Demographics			Cognitive Ability Measure	fMRI Feature	Structural MRI Features		
Feature Sets	Age	Gender	Education	MMSE	FCI	Hi-Res Features	DTI Features	FLAIR Features
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								

7.3 RESULTS

Figure 8-10 show a summary of all the results produced using the 4 methods: L1-LR, SVM with a linear kernel, SVM with a nonlinear RBF kernel, and ADTree. Since their fundamentals vary in approaches of classification, these methods were not fully comparable in their results for each outcome variable, network, and feature set combination. Nevertheless, there are some striking patterns among the feature sets for classifying the diagnosis and treatment response of LLD. The 4 methods are most in sync when both networks are analyzed together for both outcome variables. Thus, for each outcome variable, we present the results in following ways: (1) compare the performance of each method utilized, (2) focus our comparisons across different feature sets on the results obtained using features from both networks, and (3) take a closer look at the classification model that resulted in the best accuracy.

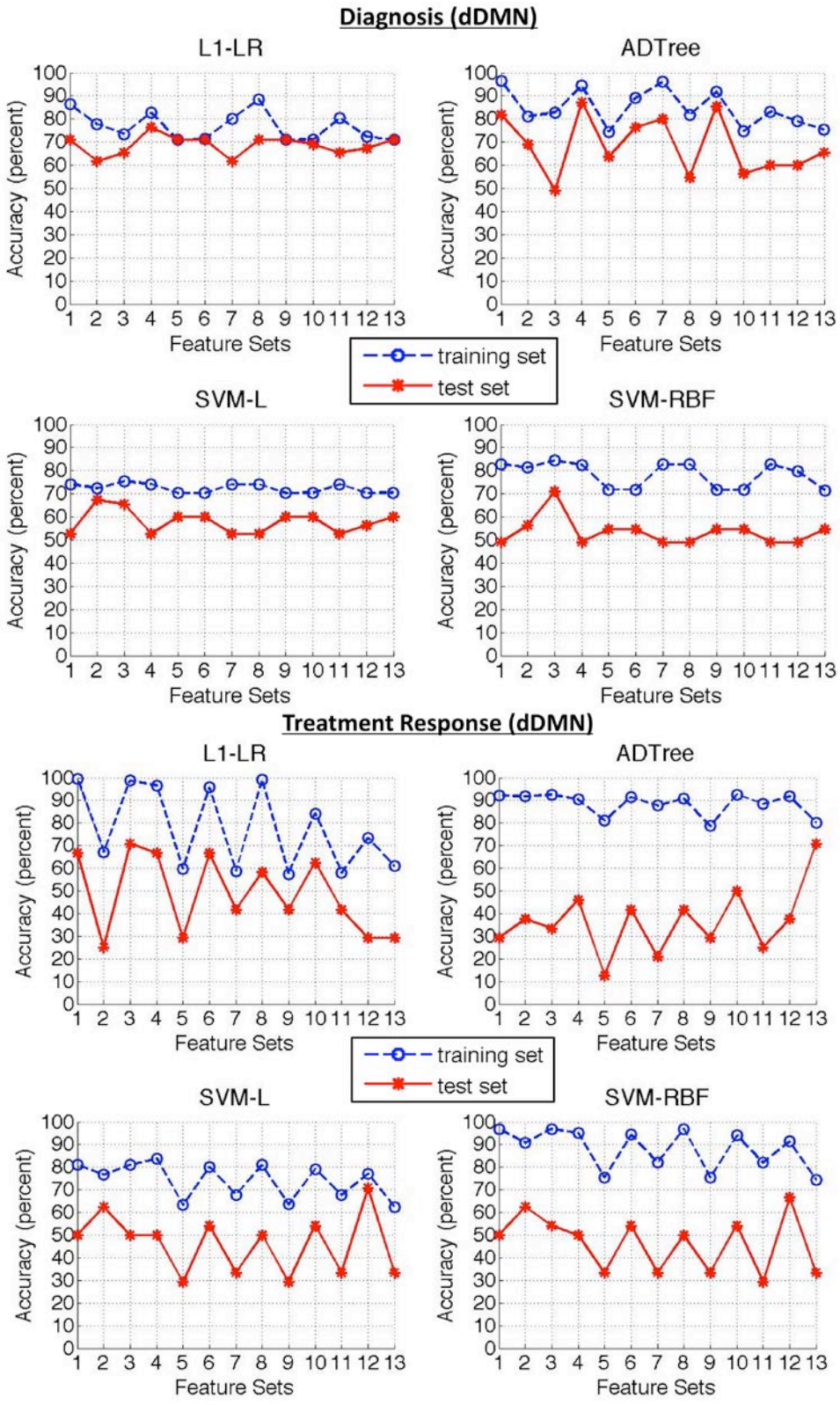


Figure 8. Feature sets' classification accuracies for dDMN analysis

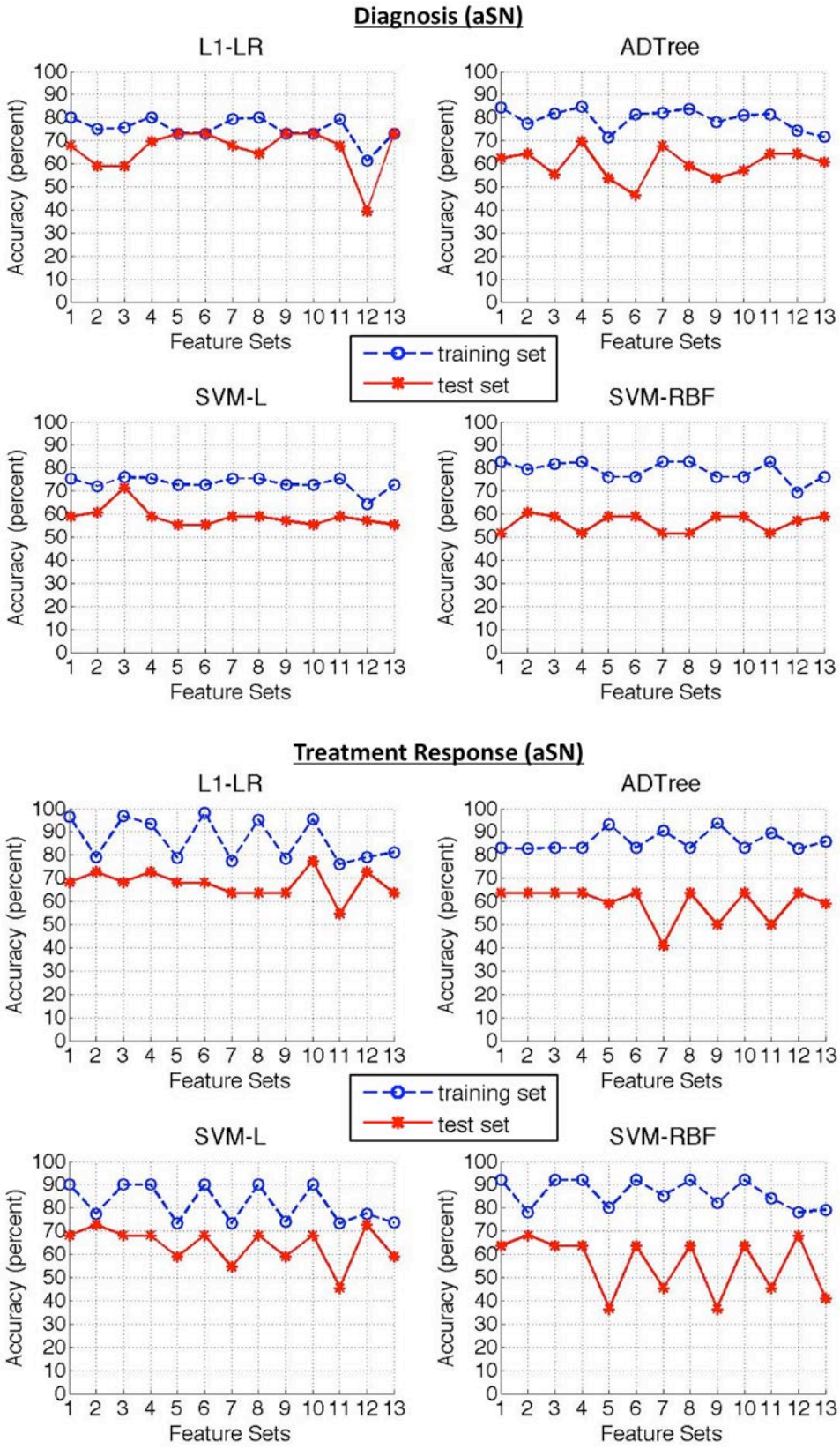


Figure 9. Feature sets' classification accuracies for aSN analysis

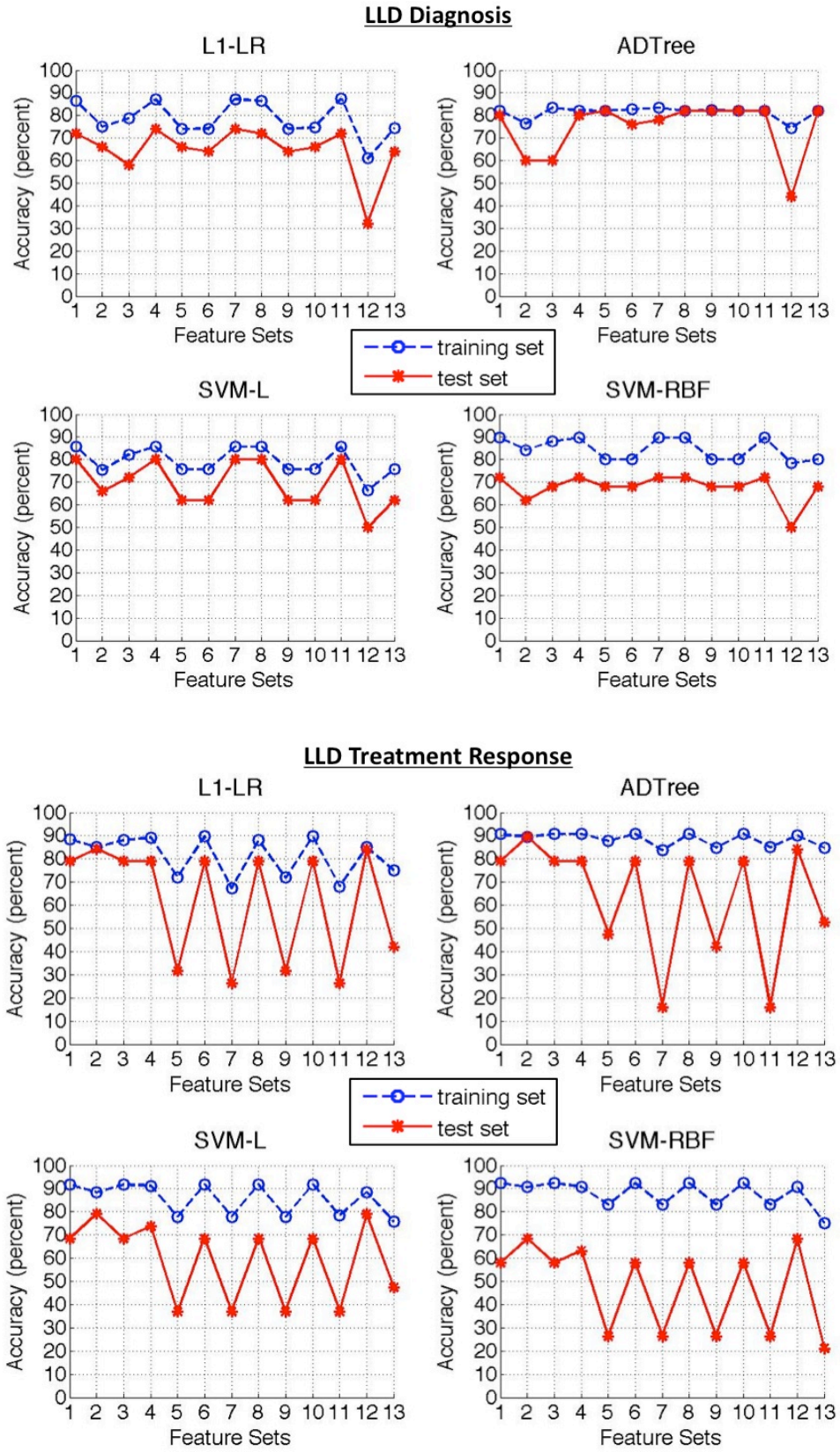


Figure 10. Feature sets' classification accuracies for both networks analysis

7.3.1 Comparing Methods

In comparing the learning methods, we evaluate how well the prediction models developed by the method classify the data. This involves comparing the test set classification accuracies as well as assessing whether the prediction models overfit or underfit the data. Overfitting occurs when a learner develops an excessively complex prediction model that over-represents the training data and does not generalize well, thereby poorly predicting future test data. Overfitting can occur when the sample size is too small or the input training data has too many noisy and/or not enough relevant features. On the other hand, underfitting occurs when a learner develops an overly simple prediction model that under-represents the training data and thus poorly predicts both the training and future test data. Underfitting can happen when the learning method's parameters are too relaxed or the learner is too simple (e.g. linear learners). When comparing learning methods by assessing overfitting and underfitting, we focus on the training and test set classification accuracy in feature set 1, when the input data comprises of all features. This is because the results from the rest of the feature sets are used to study the relevance of each type of feature (see "Both Networks Analysis" section).

To assess these measures, we look at (1) the difference between the training and test set classification accuracy in feature set 1, and (2) the difference in classification accuracies (training and test) between linear and nonlinear methods, especially among the SVM methods since the underlying method is the same. A greater difference between the training and test set classification accuracies indicates a greater probability of overfitting. Also, an improvement of classification accuracies with nonlinear methods in comparison to the linear methods indicates a greater probability of underfitting by the linear methods. On the other hand, an improvement of

classification accuracies with linear methods in comparison to the nonlinear methods indicates a greater probability of overfitting by the nonlinear methods.

7.3.1.1 LLD Diagnosis

When comparing test set classification accuracies for the diagnosis of LLD, the L1-LR and ADTree method mostly outperforms the SVM methods. The ADTree produced the optimal prediction model with an accuracy of 87.27% (sensitivity = 88.89%, specificity = 85.71%) using feature set 4 in dDMN analysis (see figure 8) for predicting LLD diagnosis (see corresponding ROC curve in figure 11).

Overall, the linear classification methods showed signs of less overfitting than the nonlinear classification methods, among which ADTree overfits less. Overfitting is observed most in the aSN analysis. In the dDMN analysis ADTree outperforms the linear methods, suggesting a possibility of underfitting among the linear models. This may also be an indicator of ADTree being a better learning method for predicting LLD diagnosis.

7.3.1.2 LLD Treatment Response

When comparing test set classification accuracies for the treatment response of LLD, all methods perform poorly for the dDMN analysis, the linear methods perform better for the aSN analysis with ADTree being a close second best, and non-SVM methods (L1-LR and ADTree) perform better overall for both networks analysis. Again, the ADTree produced the optimal prediction model with an accuracy of 89.47% (sensitivity = 88.89%, specificity = 90.00%) using feature set 2 in both networks analysis (see figure 10) for predicting LLD treatment response (see corresponding ROC curve in figure 12).

Overall, the linear classification methods showed signs of less overfitting than the nonlinear classification methods, among which ADTree overfits less. Overfitting is observed most in the dDMN analysis. Generally, all methods show signs of greater overfitting compared to LLD diagnosis—possibly due to the smaller sample size. In the dDMN and aSN analysis L1-LR outperforms the nonlinear methods, suggesting a possibility of overfitting among the nonlinear models. However, in the dDMN analysis all methods show signs of greater overfitting, and in the aSN analysis L1-LR shows signs of greater overfitting than the ADTree. Thus, based on this assessment, it is difficult to determine if the nonlinear methods are overfitting because of increased complexity of their prediction models, L1-LR is overall a better learning method for predicting LLD treatment response, or the problem is with the features used. Nevertheless, considering that the ADTree produced the best classification model when both networks were included, the answer maybe a lack of relevant features.

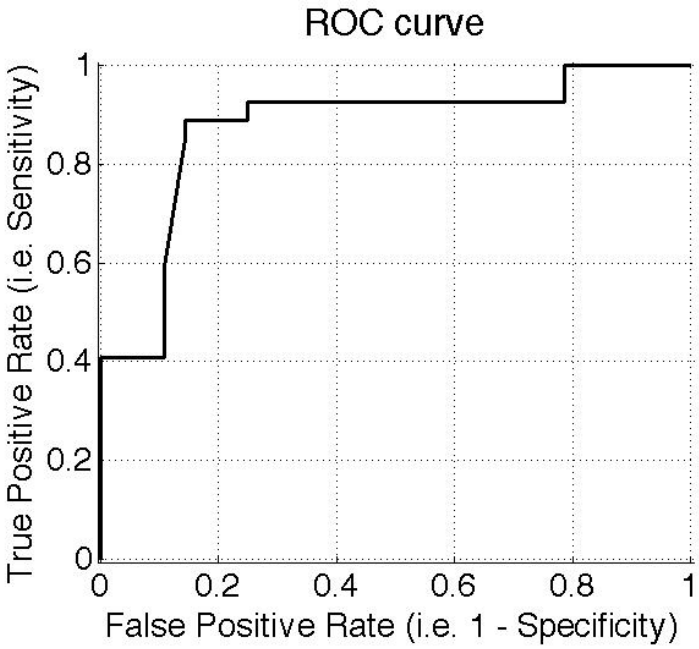


Figure 11. ROC curves for optimal ADTree models predicting LLD diagnosis

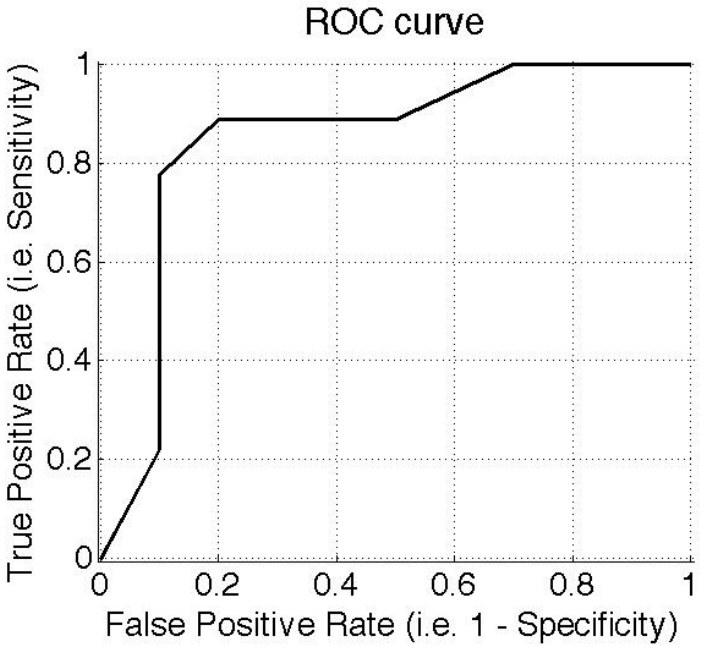


Figure 12. ROC curves for optimal ADTree models predicting LLD treatment response

7.3.2 Both Networks Analysis

As mentioned earlier, all methods performed better and more consistently with the inclusion of features from both networks (dDMN and aSN). Below, we present common patterns across the 4 methods in feature sets' classification accuracy results for diagnosis and treatment response of LLD. When comparing the importance vs. insignificance of the feature(s) removed in a given feature set, we compare its classification accuracy to that of feature set 1 or the norm for this analysis. Feature set 1 is the norm for this study because it includes all features in the analysis, thus giving each feature an equal chance of affecting the classification results. In comparing with feature set 1, one can make four possible observations: (1) if a feature set's training and test set classification accuracies are greater than or equal to feature set 1, the feature(s) removed in the

feature set is considered to be an unnecessary predictor/biomarker for predicting the outcome variable (Note: improved accuracies from the norm may suggest that the feature(s) removed was causing underfitting); (2) if the feature set's training and test set classification accuracy is less than that of the norm, the feature(s) removed in the feature set is considered to be a potentially important predictor/biomarker for predicting the outcome variable (Note: worsened accuracies from the norm may suggest that the removal of the feature(s) is resulting in underfitting); (3) if the feature set's training set classification accuracy is less than the norm while the test set classification accuracy is greater than or equal to the norm, the feature(s) removed in the feature set is considered to be an unnecessary predictor/biomarker for predicting the outcome variable (Note: an decrease in training accuracy and increase in test accuracy from the norm may suggest that the feature(s) removed was causing overfitting); and (4) if the feature set's training set classification accuracy is greater than or equal to the norm while the test set classification accuracy is less than the norm, the feature(s) removed in the feature set is considered to be a potentially important predictor/biomarker for predicting the outcome variable (Note: an increase in training accuracy and decrease in test accuracy from the norm may suggest that the removal of the feature(s) is resulting in overfitting).

In short, if the test set classification accuracy of a given feature set is less then the norm, then the feature(s) removed in that feature set is a potentially important predictor/biomarker for predicting the outcome variable; otherwise the feature(s) removed is unnecessary for predicting the outcome variable. Based on this method of assessing feature sets, the common patterns among the two outcome variables are strikingly of exactly opposite natures in terms of the importance of imaging vs. non-imaging features as potential biomarkers.

7.3.2.1 LLD Diagnosis

One commonality among the results of 4 methods for diagnosis of LLD is the increase in the training and test set classification accuracies from feature set 12 to 13. This increase suggests that overall non-imaging features (i.e. demographics and cognitive ability measures) are more important biomarkers for predicting diagnosis of LLD than imaging features.

These observations are also reflected in the results of feature sets 2-5. These feature sets indicate that when either of the non-imaging features is removed, the test set classification accuracy decreases from the norm indicating importance of these features. Also, if at least one of the imaging features is removed the test set classification accuracy increases from or remains approximately similar to that of the norm indicating insignificance of these feature(s). However, this observation is made only for the removal of functional imaging feature except for the ADTree algorithm results, which show it for both the removal of functional and structural imaging features. For the other methods, the removal of structural imaging feature in fact causes the test set classification accuracy to decrease from that of the norm. Looking, more specifically into the removal of the individual and combinations of the different structural imaging features, feature sets 6, 9 and 10 also cause the test set classification accuracy to decrease from that of the norm. These feature sets include the removal of at least one of the structural imaging features (mostly the Hi-Res T1-weighted image features and/or the FLAIR T2-weighted image features), indicating the importance of each of the individual structural imaging features as a potential biomarker.

In summary, the potential biomarkers among the features tested in this study for the diagnosis of LLD include: demographics, cognitive ability, and Hi-Res T1-weighted plus FLAIR T2-weighted structural imaging measures. One of the least important features, as indicated by a

consensus among all the 4 methods, appears to be the functional imaging measure. As mentioned earlier, the highest test set classification accuracy for the diagnosis of LLD was also achieved by removing only the functional imaging measure.

7.3.2.2 LLD Treatment Response

Contrary to the diagnosis prediction models, the commonality among the results of 4 methods for treatment response of LLD is the large decrease in the training and test set classification accuracies from feature set 12 to 13. This decrease suggests that overall imaging features are more important biomarkers for predicting treatment response of LLD than non-imaging features (i.e. demographics and cognitive ability measures).

These observations are also reflected in the results of feature sets 2-5. These feature sets show that when the non-imaging demographics features are removed, the test set classification accuracy increases from that of the norm. Along similar lines, when the non-imaging cognitive ability feature is removed, the test set classification accuracy remains the same to that of the norm. Both these observations indicate the irrelevance of non-imaging features. On the other hand, feature set 5 indicates the importance of imaging features, specifically structural imaging features, by showing how the test set classification accuracy decreases from that of the norm when structural imaging features are removed. However, according to feature set 4, the functional imaging features does not seem to be as important since by removing it, the test set classification accuracy either increases from or remains the same to that of the norm. Looking, more specifically into the removal of the individual and combinations of the different structural imaging features, feature sets 7, 9 and 11 also cause the test set classification accuracy to decrease from that of the norm. These features all have in common the removal of the DTI image features, indicating its potential importance as a biomarker.

In summary, the potential biomarkers among the features tested in this study for the treatment response of LLD include imaging features, specifically the DTI structural imaging features. The least important features, as indicated by a consensus among all the 4 methods, appear to be the demographic measures. As mentioned earlier, the highest test set classification accuracy for the treatment response of LLD was also achieved by removing only the demographics measures.

7.3.3 Optimal Prediction Models

Now we look closer at the precise features selected by the model that produced the optimal classification accuracy for each outcome variable. Since the ADTree method produced the optimal model for both outcome variables, the models studied will be in the form of ADTrees. These optimal models presented below were obtained by retaining only the most frequently occurring branches amongst each of the ADTree models created during the LOOCV iterations. For interpreting these models, one must sum up the rule values associated with each attribute and if the total is positive, the individual is more likely to be an LLD patient or a positive responder to treatment for LLD depending on the outcome variable predicted by the ADTree. Also see table 7 to better understand what each feature represents in the optimal ADTrees and the interpretations provided below.

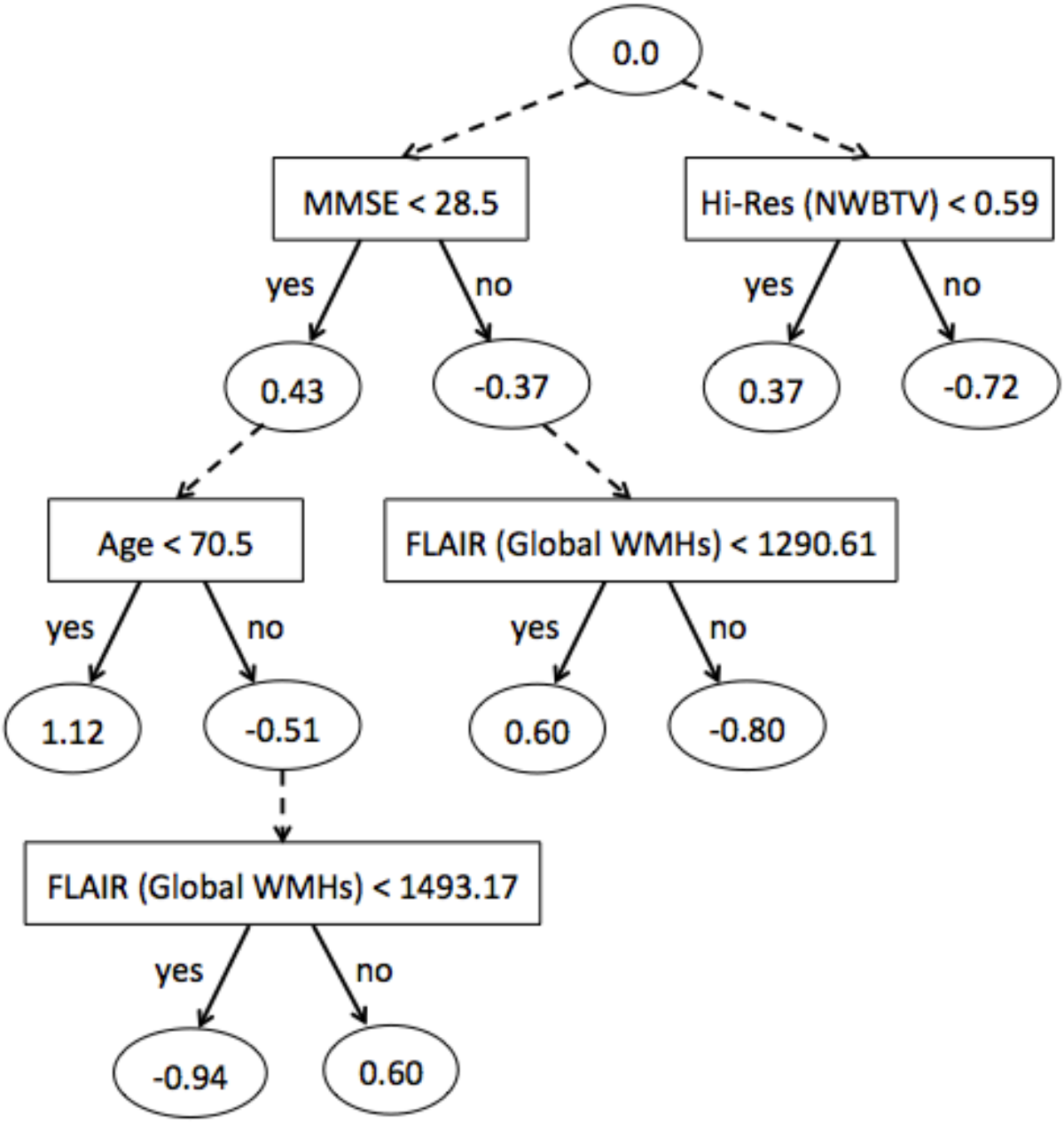


Figure 13. Optimal prediction models in the form of alternating decision trees for predicting late-life depression diagnosis [Legend: Square = Splitting Criterion; Oval = Rules]

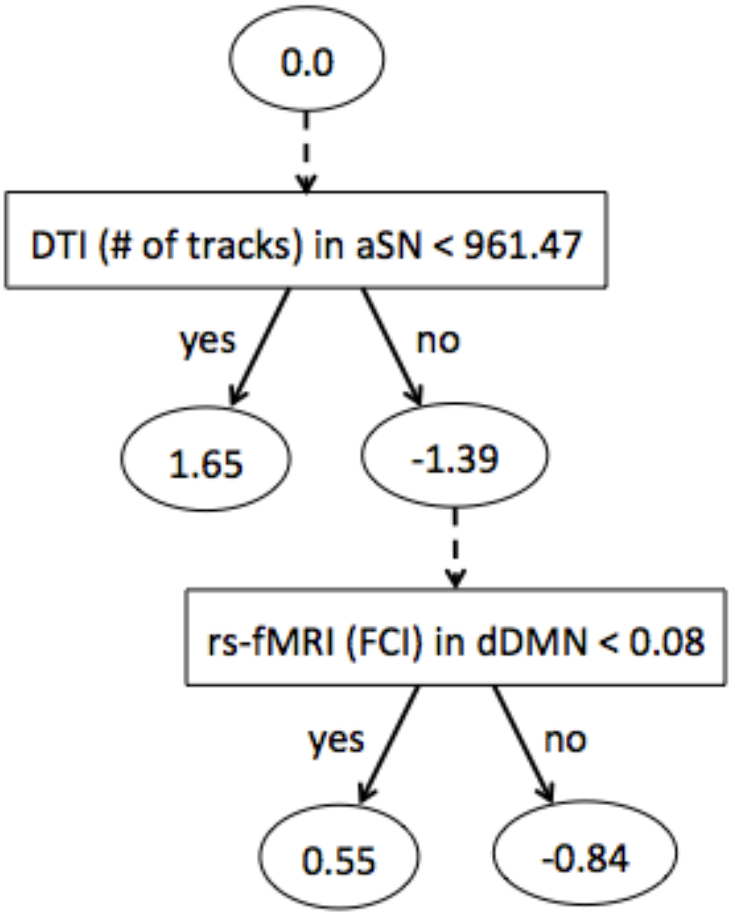


Figure 14. Optimal prediction models in the form of alternating decision trees for predicting late-life depression treatment response [Legend: Square = Splitting Criterion; Oval = Rules]

7.3.3.1 LLD Diagnosis

With the ADTree method, the highest classification accuracy for the diagnosis of LLD is obtained by inputting the non-imaging and structural imaging features while removing the functional imaging feature (i.e. FCI) as mentioned earlier. The optimal ADTree model that produced this classification accuracy is shown in figure 13.

The ADTree model in figure 13 indicates that a high accuracy model to predict the diagnosis of LLD can be created using the following features: MMSE (cognitive ability measure), age (demographic measure), Hi-Res normalized whole brain gray and white matter

volume (structural imaging measure), and Flair global WMH count (structural imaging measure). These results are in agreement in terms of features selected to those observed when analyzing the different feature sets in the "Both Networks Analysis" section. However, one important observation to be made here is that the features selected by the optimal model are not dependent on the network even though the optimal model was acquired during the dDMN analysis. The possible explanation for the results achieving a higher accuracy for the dDMN may be the even distribution of subjects among the depressed and non-depressed group in comparison to the other two analyses, and also a larger sample size plus fewer features in comparison to the both networks analysis.

Based on the optimal ADTree model in figure 13, an individual who is more likely to be diagnosed with LLD will have one of the following attributes: (1) low cognitive ability + younger old adult, (2) low cognitive ability + older old adult + high global WMH burden + high whole brain atrophy, or (3) high cognitive ability + low global WMH burden + high whole brain atrophy. On the other hand, an individual who is not likely to diagnosed with LLD will have one of the following attributes: (1) low cognitive ability + older old adult + low global WMH burden, (2) low cognitive ability + older old adult + high global WMH burden + low whole brain atrophy, (3) high cognitive ability + low global WMH burden + low whole brain atrophy, or (4) high cognitive ability + high global WMH burden.

7.3.3.2 LLD Treatment Response

With the ADTree method, the highest classification accuracy for the treatment response of LLD is obtained by inputting features from both networks and removing the demographic features (i.e. age, gender, education) as mentioned earlier. The optimal ADTree model that produced this classification accuracy is shown in figure 14.

The ADTree model in figure 14 indicates that a high accuracy model to predict the treatment response of LLD can be created using the following features: average # of tracks from DTI images in aSN (structural imaging measure), and average FCI from rs-fMRI images in dDMN (structural imaging measure). These results also agree in terms of both the features and networks selected with the results observed when analyzing the different feature sets in the "Both Networks Analysis" section.

Based on the optimal ADTree model in figure 14, an individual who is more likely to be a positive responder to treatment for LLD will have fewer structural connections—indicative of a lower WM integrity—in the aSN before treatment is administer. Additionally, an individual who had a lower functional connectivity in the dDMN before the administration of treatment is less likely to be a negative responder to treatment for LLD.

7.4 DISCUSSION

In this study, we showed how nonlinear combinations of imaging and/or non-imaging measures can be used to develop classification models that can successfully predict the diagnosis and treatment response of LLD outcome variables with high accuracies of 87.27% and 89.47% respectively. When studying the specific features selected to optimally classify each outcome variable, no overlap in features was found. In fact, the prediction models for each outcome variable was strikingly opposite in nature. While demographics—primarily age—were found to be one of the more important features for predicting diagnosis, they were also found to be the least important for predicting treatment response. On the other hand, the functional imaging feature was found to be the least important feature for predicting diagnosis, while it was an

important feature for predicting treatment response. Additionally, the diagnosis prediction model was network independent, while the treatment response prediction model depended on information from both the dDMN and aSN. Below we evaluate our findings further using past studies for comparison.

7.4.1 Optimal Predictors/Biomarkers

7.4.1.1 LLD Diagnosis vs. LLD Treatment Response

Non-imaging (i.e. MMSE and age) and global volume-based imaging (i.e. whole brain atrophy and global WMH burden) measures combined were found to the optimal predictors/biomarkers of LLD diagnosis. Agreeing with past studies, poor cognitive ability [Ganguli et al., 2006; Kohler et al., Apr 2010; Wilkins et al., 2009] and greater whole brain atrophy [Chang et al., 2011; Ribeiz et al., 2013; Sexton et al., 2013] indicated LLD. Our findings also suggest that high MMSE could still indicate LLD if accompanied with low global WMH burden and high whole brain atrophy. Possibly explaining the discrepancies between past studies, age [Forlani et al., 2013; Luppa et al., 2012; Wild et al., 2012; Wu et al., 2012] and global WMH burden [Aizenstein et al., 2011; Greenwald et al., 1998; Gunning-Dixon et al., 2010; Firbank et al., 2012; Teodorczuk et al., 2010] were fully dependent on the other measures in regards to their association with LLD diagnosis. We speculate that the primary role of non-imaging measures in predicting diagnosis suggests that the current neuroimaging methods cannot – yet – capture the neural complexity associated with the etiopathogenesis of LLD. The involvement of structure-related neural biomarkers (global atrophy and WM burden) in diagnosing LLD supports past studies that suggest vascular and atrophic changes trigger mood disorder in late-life [Aizenstein et al., 2014].

Optimal biomarkers of LLD treatment response included connectivity-based imaging measures. Specifically, lower structural connectivity—supporting the more recent of the two [Taylor et al., 2008] contradicting past findings [Alexopoulos et al., 2008; Taylor et al., 2008]—and lower functional connectivity—supporting compensation theories [Stern et al., 2003]—indicate a greater probability of treatment remission. This dependency of LLD treatment response on global network health (i.e. communication strength between network regions) may serve as a biomarker for future personalized care studies.

Overall, the mix of features predictive of diagnosis likely reflects that LLD is heterogeneous. Our observation that these particular features were not predictive of treatment response suggests that there may be a more proximal mediator of depression recovery, and perhaps the features reflecting LLD heterogeneity lead to a set of global network changes (indexed by rs-fMRI and DTI). It is intriguing that it is these global network biomarkers that were identified as most predictive of treatment response.

7.4.1.2 Mid-Life vs. Late-Life Depression Prediction Models

Unlike past studies of depression in younger populations involving prediction models, this is the first study to accurately model both diagnosis and treatment response using the same approach. While past studies have used a single imaging modality and region-based approach [Costafreda et al., 2009; Fu et al., 2008; Hahn et al., 2011; Marquand et al., 2008; Liu et al., 2012; Mwangi et al., Jan 2012; Mwangi et al., May 2012; Nouretdinov et al., 2011; Zeng et al., 2012], we used a multi-modal imaging with whole brain and network-based approach that also included non-imaging measures. Our results may suggest that biomarkers of disease diagnosis and remission possibly differ on the basis of brain structure and function—i.e. the different representations of MRI modalities—as opposed to brain regions. It is possible that regional changes do not fully

reflect the underlying neural vulnerabilities associated with LLD. This is supported by recent studies [Ajilore et al., 2014; Tadayonnejad et al., 2013] that describe associations of global brain networks alterations with LLD.

Past prediction model studies of mid-life depression diagnosis have shown accurate classifications can be obtained using functional [Fu et al., 2008; Hahn et al., 2011; Marquand et al., 2008; Nouretdinov et al., 2011; Zeng et al., 2012] or structural [Costafreda et al., 2009; Mwangi et al., Jan 2012; Mwangi et al., May 2012] imaging. Our study in LLD found structural volume-based measures in conjunction to non-imaging measures to be better predictors. We speculate that these differences in prediction factors may suggest that LLD diagnosis is primarily related to impaired structure (GM and WM), while midlife depression may stem from aberrant communication/activation of various brain regions. This hypothesis will require further testing.

Past prediction model studies of mid-life depression treatment response have primarily utilized T1-weighted Hi-Res structural imaging measures [Costafreda et al., 2009; Liu et al., 2012; Nouretdinov et al., 2011]. One study [Marquand et al., 2008] that attempted to use a task-based functional imaging measure did not achieve very high accuracy. Our study in LLD found structural and functional connectivity measures to be better predictors. Since connectivity-related imaging measures have not been tested for prediction models of mid-life depression treatment response, it is difficult to draw any conclusions.

7.4.2 Learning Methods

Most of the past studies described above involving prediction models for both diagnosis and treatment response of depression in the younger populations have mostly used SVM as the learning method. Based on our findings, modified versions of decision tree and logistic

regression are potential alternative learning methods—to the traditionally used SVMs—that can accurately predict diagnosis and treatment response of depression, at least in late-life. Modified decision tree methods with embedded feature selection capabilities, especially, may be a useful tool for studying real-world nonlinear relationships in high-dimensional data.

7.4.3 Limitations and Future Work

Limitations to this study include: small sub-sample size for treatment response prediction (nevertheless the results were cross checked using four different learning methods) and higher percentage of women (reflecting the naturalistic gender distribution in LLD [Luppa et al., 2012]). Another limitation is the heterogeneous treatment. However, this may not have affected our results since all administered antidepressants except nimodipine (used only for one subject) are either selective serotonin reuptake inhibitors or serotonin-norepinephrine reuptake inhibitors, and the efficacy difference between the two is still a matter of debate [Papakostas et al., 2007; Taylor et al., 2004; Taylor et al., 2006; Thase et al., 2011]. Future work includes extensive studies verifying, improving as necessary, and testing the real-world applicability of the optimal prediction models found in our study.

8.0 SUMMARY AND CONCLUSIONS

Currently, late-life depression (LLD) is diagnosed based on behavioral symptoms and signs. Treatment of LLD is guided by trial and error. Both the diagnosis and treatment procedures lack reliability and could be improved with additional knowledge of associated underlying brain characteristics and changes. The goal of this dissertation is to identify biomarkers reflecting the neural circuit abnormalities that characterize LLD and its treatment response.

In regards to underlying brain characteristics and changes, LLD has been associated with neurotransmitter-specific system decline, fronto-striatal and fronto-limbic circuitry dysfunction, and cerebrovascular disease. Neurotransmitters involved for the LLD related neurotransmitter-specific system decline primarily include serotonin, norepinephrine, and dopamine. Loss of these neurotransmitters is associated with alterations in mood, stress response, motivational control, etc. Fronto-striatal and fronto-limbic dysfunction respectively alter executive control and emotional processing. Cerebrovascular disease is thought to disrupt pathways in the brain associated with mood regulation. Treatment of LLD predominantly consists of antidepressants, which focus on controlling for the loss of neurotransmitters associated with LLD.

In this dissertation, magnetic resonance imaging (MRI) is used to study the underlying brain characteristics and changes associated with LLD and its treatment. The different underlying brain characteristics associated with LLD are studied using different MRI modalities, which vary based on the MR scanning parameters and sequence used. MRI modalities include

both structural and functional imaging. Using different image analysis and processing methods, information including degree of regional atrophy in neural circuits, lesions due to cerebrovascular disease, integrity of the connections within the neural circuits, amount of dysfunction with the neural circuits, etc. is extracted from the various MRI modalities.

The information extracted from the different MRI modalities is used to study how the brain structure affects brain function, as well as how both brain structure and function can help determine LLD diagnosis and its treatment response. Statistical analysis is used to study the affect of brain structure on function, and machine learning methods are used to estimate predictions models that can better estimate and explain all above-described relationships.

The results of this dissertation suggest that (1) brain structure may not be directly related to functional connectivity in the elderly, and (2) LLD diagnosis and treatment response may be better predicted using a combination of multi-modal MRI measures. The results also suggest that the incorporation of non-imaging predictors could also help improve prediction, at least for LLD diagnosis. Additionally, we speculate that whole brain and network related multi-modal MRI measures—as opposed to region-based single modality measures—may be more appropriate for comparing LLD diagnosis and treatment response in terms of associated underlying brain changes. The high accuracy of the prediction models estimated in this dissertation may be useful for better diagnosing and taking preliminary steps towards establishing personalized treatment for late-life depression patients in the future.

8.1 FUTURE WORK

Future work primarily includes studying prediction models of LLD treatment response more in depth with a larger sample size. Including participants for whom we do not have treatment response information can be one way to increase the sample size of the data. This would mean incorporating unlabeled data with the label data. With this new, larger data set, either semi-supervised or unsupervised learning methods can be tested to estimate accurate prediction models. Different imaging modalities (e.g. task-based functional MRI, proton density, etc.) and imaging features (e.g. shape of regions, texture of lesions, etc.) can also be tested to improve the estimation and generalization of prediction models. Additionally, longitudinal studies can also be performed for future work. These studies would include testing the accuracy of prediction models over a period of time to determine how well the models can predict future treatment response.

8.2 ACKNOWLEDGMENTS

This research was supported by NIH grants MH076079/MH076079-04S1, MH 086686, KL2 RR024154, R21 NS060184, R37 AG025516, P30-MH52247/P30 MH71944, MH K23 086606, P30 MH90333, and Brain and Behavior Research (NARSAD) Young Investigator Award (Dr. Andreescu). It was also supported by the John A. Hartford Center of Excellence in Geriatric Psychiatry at the University of Pittsburgh.

APPENDIX A

STUDY CONTINUED FROM CHAPTER 7

A.1 INTRODUCTION

This chapter describes a secondary set of machine learning analyses that were done to follow-up on the analyses of chapters 6. As in chapter 6 the experiments here, address the question of how well structural imaging features predict resting-state fMRI. But, unlike the repeated within-subject imaging approach of chapter 6, the analyses in this chapter use a subject-wise analysis using features with primarily non-zero values. This study is a continuation of the study described in chapter 7, except now we evaluate whole networks'—networks studied include the dorsal default mode network and anterior salience network—functional connectivity in the elderly (depressed and non-depressed) as an outcome variable instead of late-life depression diagnosis or treatment response.

A.2 METHODS

The same methods used in chapter 7 for diagnosis and treatment response were used to estimate prediction models for functional connectivity. The difference is that a median split criterion was used to form the two groups representing high versus low functional connectivity. Additionally, for the functional connectivity, the both network analysis was repeated twice; each time using functional connectivity measures from different networks (dDMN vs. aSN) to represent the outcome variable.

A.3 RESULTS

Figure 15-16 show a summary of all the results produced for functional connectivity as the outcome variables using the 4 methods: L1-LR, SVM with a linear kernel, SVM with a nonlinear RBF kernel, and ADTree. These results were analyzed the same way as those for LLD diagnosis and treatment response in chapter 7.

A.3.1 Comparing Methods for Functional Connectivity

See chapter 7 for details on how the results from feature set 1 were analyzed to compare the different learning methods.

When comparing test set classification accuracies for the functional connectivity in the elderly, the SVM-L seems to perform the best, but in general all methods perform poorly (accuracy < 70%) for all analysis. The only one time a prediction model achieves accuracy

greater than 70%, is using the ADTree method in the dDMN network analysis. The ADTree method produces the greatest classification accuracy of 74.55% (sensitivity = 77.78%, specificity = 71.43%) for feature set 10 using features from only the dDMN network (see figure 17 for the corresponding ROC curve).

Primarily, only SVM-L consistently shows less signs of overfitting and underfitting. When comparing between the linear and nonlinear methods, especially between the two SVM methods, primarily the nonlinear models show more signs of overfitting. Overall, SVM-L seems to be the best performing method.

A.3.2 Both Networks Analysis for Functional Connectivity

See chapter 7 for details on how the both networks analysis was used to access the results across the feature sets.

There are no meaningful patterns that can be observed among the feature sets across the learning methods for accessing functional connectivity in the elderly. In fact, most of the results indicate less than 60% accuracy of the predictions models. In summary, it is very difficult to determine potential biomarkers from these results.

A.3.3 Optimal Prediction Models for Functional Connectivity

The ADTree produced the optimal model for function connectivity. Thus, the optimal model studied for function connectivity will be in the form of an ADTree. Detailed description of how the optimal ADTree was formed and how to interpret is provided in chapter 7.

With the ADTree method, the highest classification accuracy for functional connectivity is obtained by inputting the non-imaging, structural DTI, and functional imaging features from only dDMN. This involves removing the structural Hi-Res and FLAIR imaging features from the full feature set described in table 7. The optimal ADTree model that produced this classification accuracy is shown in figure 18.

The ADTree model in figure 18 indicates that a high accuracy model to predict the functional connectivity of the dDMN in the elderly can be created using the following features: education, average # of tracks from DTI images in dDMN (structural imaging measure), and average weighted FA from DTI images in dDMN (structural imaging measure). However, these results are not reciprocated by the observations made when evaluating patterns across features sets. For example, feature set 2—in which demographics were removed—and feature set 7—in which DTI measures were removed—should be shown to perform poorly compared to the feature set 1 since the removed features are found to be essential by the optimal prediction model. However, these feature sets are not consistently shown to perform poorly across learning methods in any of the analyses.

Based on the optimal ADTree model in figure 18, an elderly individual who is more likely to have high dDMN functional connectivity will have either a high level of education or a low level of education with greater structural integrity in the dDMN. However, an elderly individual with a low level of education with lesser structural integrity in the dDMN is more likely to have low dDMN functional connectivity.

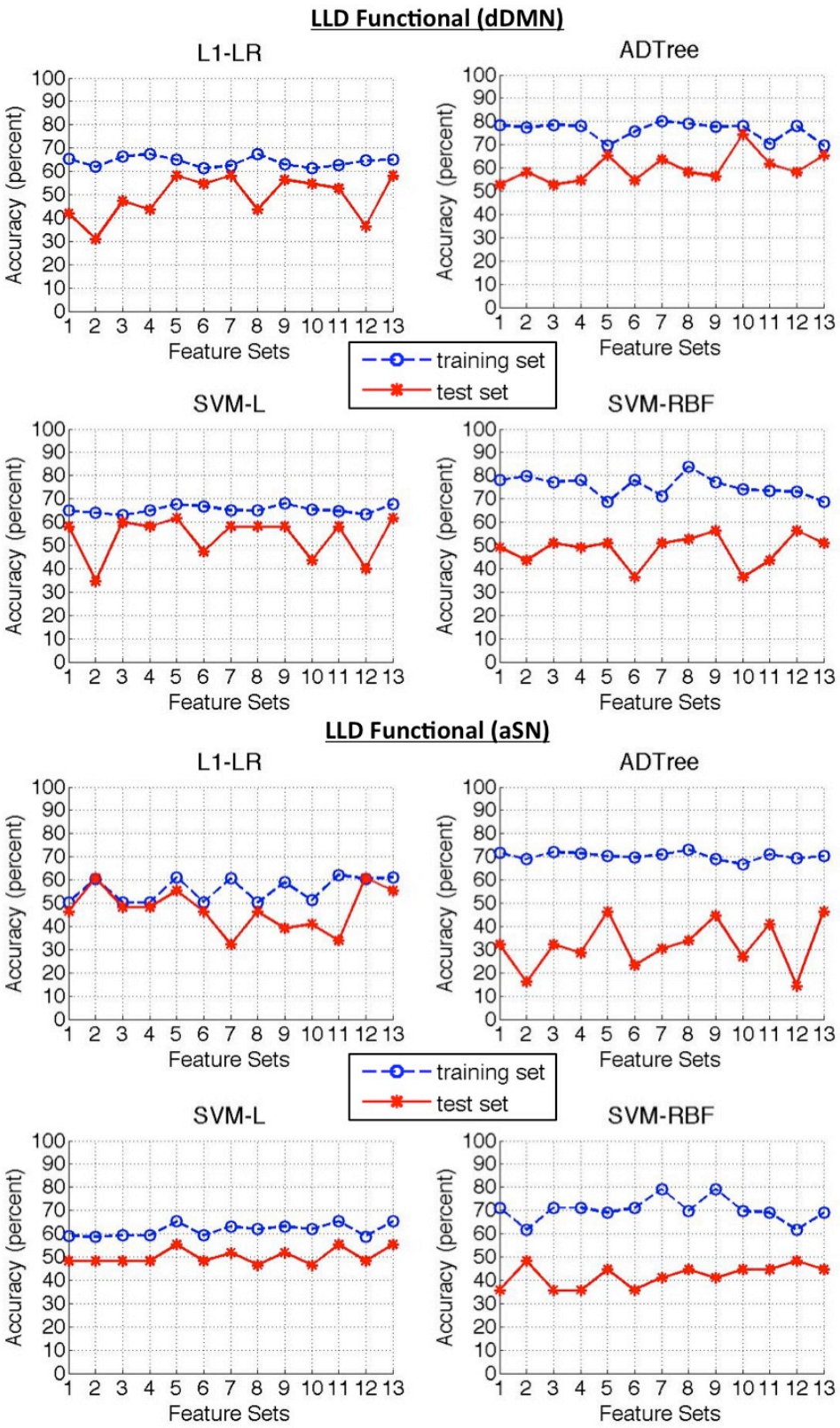


Figure 15. Feature sets' classification accuracies for individual dDMN and aSN network analyses

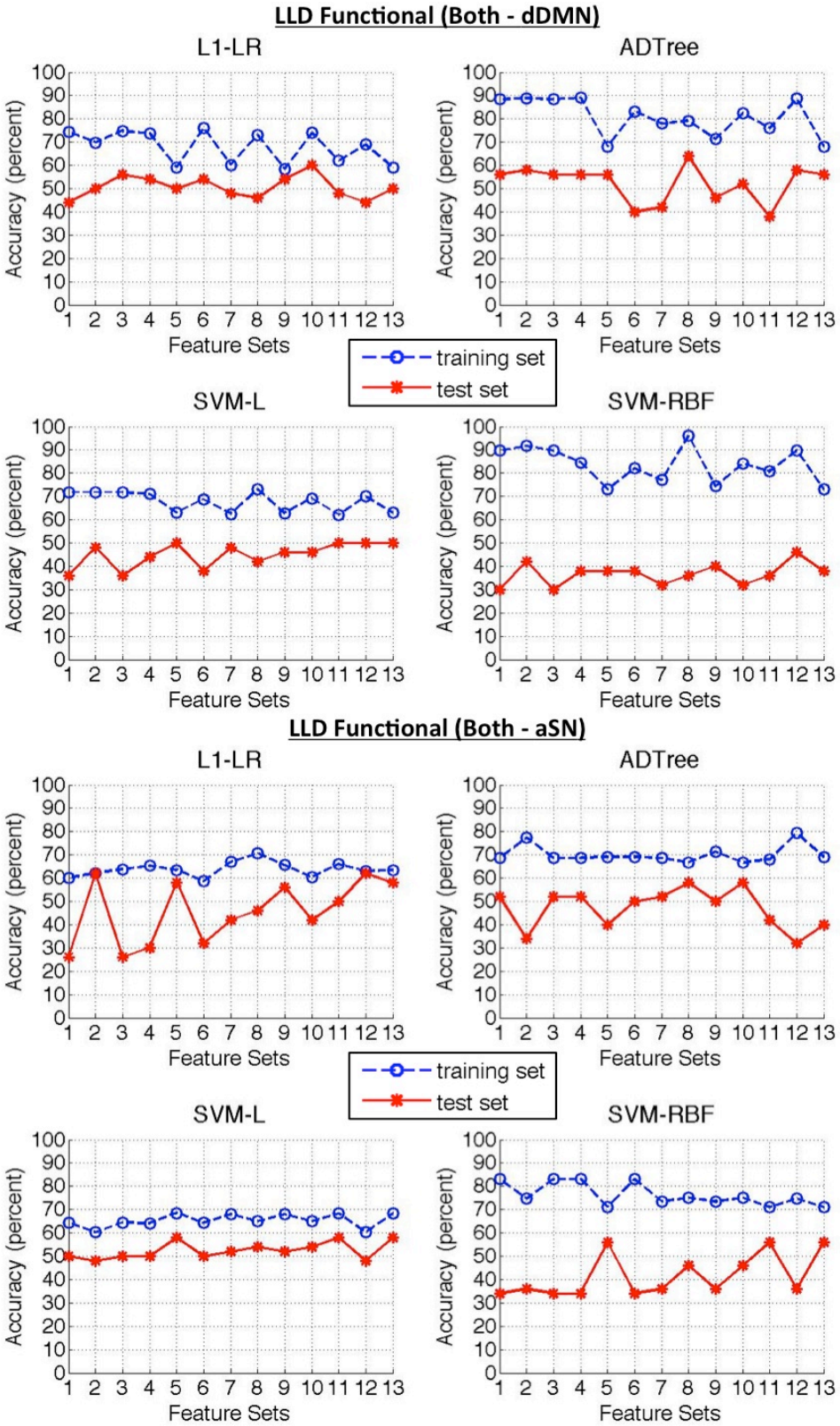


Figure 16. Feature sets' classification accuracies for both networks analyses

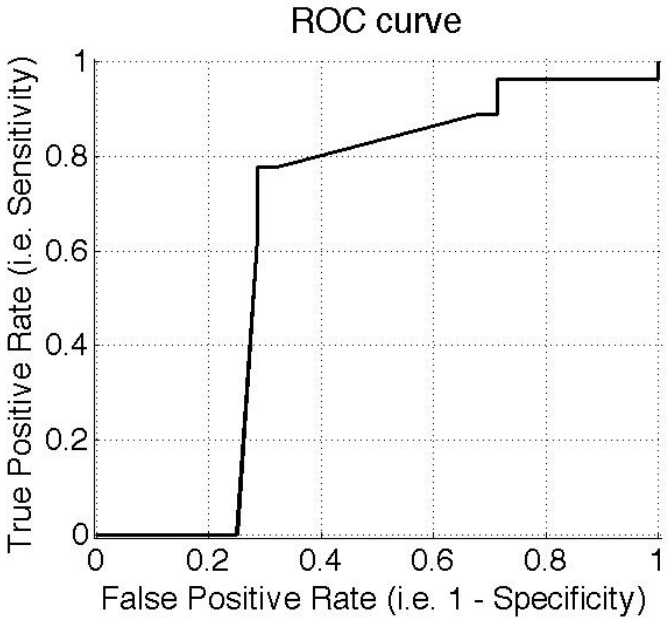


Figure 17. ROC curves for optimal ADTree models predicting functional connectivity in the elderly

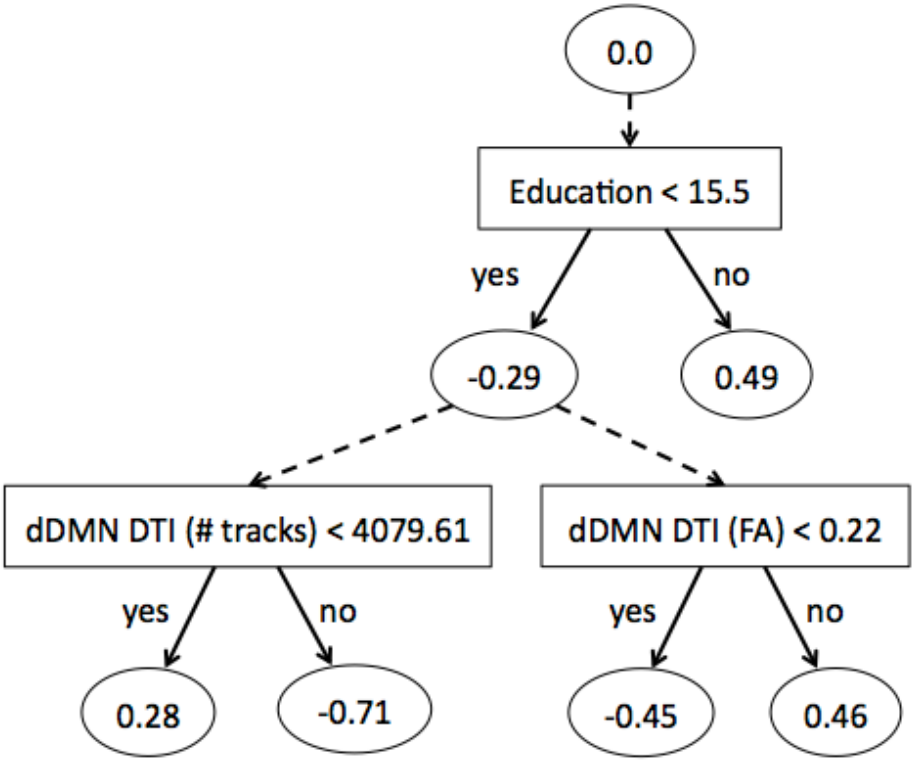


Figure 18. Optimal prediction models in the form of alternating decision trees for predicting functional connectivity in the elderly [Legend: Square = Splitting Criterion; Oval = Rules]

A.4 DISCUSSION

The results of this study suggest that brain functional connectivity may not be directly related to brain structure in the elderly. The results also suggest that functional connectivity is more related to structural connectivity than structural lesions or atrophy. Specifically, greater functional connectivity is associated with greater structural integrity in the default mode network. This possibility of an indirect relationship between resting state functional and structural connectivity has also been shown in younger populations by past studies [Deligianni et al., 2011; Honey et al., 2009]. Based on this study, we speculate that other non-brain related external factors (e.g. education) might help better learn this indirect relationship in addition to imaging measures.

Nevertheless, these results and corresponding suggestions are very preliminary, as a considerably high accuracy prediction model was not achieved. This may have been due to lack of understanding of the negative functional connectivity values, which affect how the separation of individuals into groups of high versus low functional connectivity. It could also be due to a limitation on the number and type of imaging and/or non-imaging measures used. Additionally, a stronger relationship may be found between structural imaging measure and task-based functional activation. Thus, future work could include studying a wider range of features, incorporating task-based functional imaging features, and/or testing other learning methods.

APPENDIX B

REGULARIZATION OF REGRESSION

The goal of linear and logistic regression is to estimate optimal weights for each input features such that and accurate label for the output variable can be predicted. Optimal weights are estimated by minimizing an objective function—which represents the sum of squared difference between the predicted labels and actual observed labels—for linear regression and logistic loss function—which is the same as maximizing the likelihood of the data given the prediction model—for logistic regression. Thus, both the objective and logistic loss function are convex in nature (see figure 19 for an example of a 3D convex shape) [Czepiel, 2002; Liu & Zhang, 2008; Yuan et al., 2010].

When, the input data has a high dimensionality, a regularization or penalty term is added to the objective or logistic loss function to perform embedded feature reduction (see "Feature Reduction" section in chapter 4). The most common regularization terms are L1- and L2-norm. L1-norm ($||w||_1$) is the sum of the weights, while L2-norm ($||w||_2$) is the square root of the sum of weights squared. Variations of linear and logistic regression that use L1-norm include L1-regularized Least Squares (LASSO) and L1-regularized logistic regression respectively. Variations of linear and logistic regression that use L2-norm include ridge regression and L2-

regularized logistic regression respectively [Czepiel, 2002; Liu & Zhang, 2008; Yuan et al., 2010].

Compared to L2-norm, L1-norm is more effective in performing feature reduction. This is because there is a greater chance for weights to attain a zero value based on the graphical nature of L1-norm (see figure 20 and 21). Thus, a greater number of weights are estimated to have a zero value. To further attain a greater reduction in features, L1/2-norms have also been used by past studies. Figure 21, illustrates how the chances of attaining zero values for weights increases further with a L1/2-norm regularization term; thus further increasing the number of weights estimated to have a zero value [Chen et al., 2013; Ng, 2004].

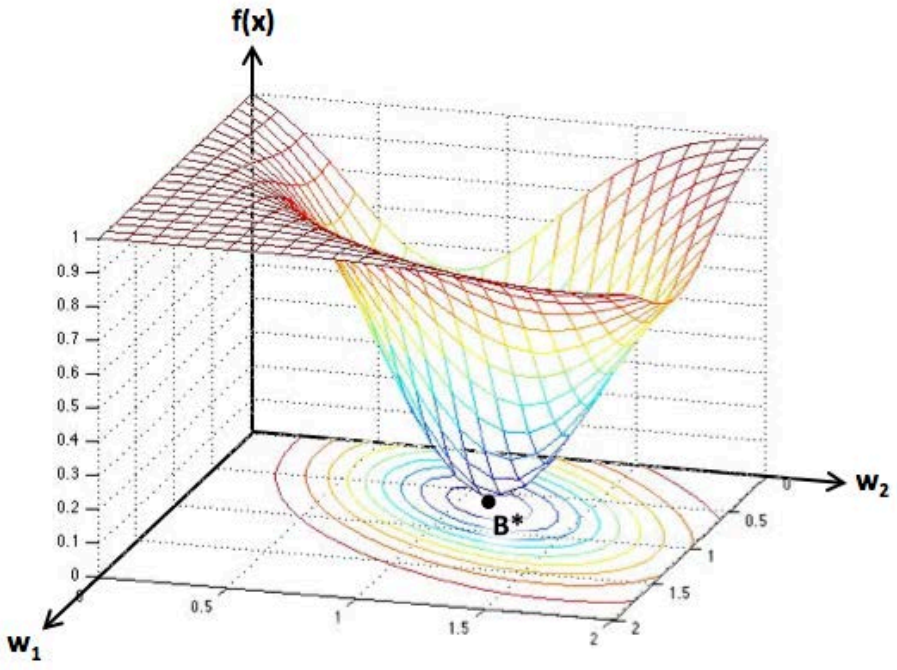


Figure 19. Example of a convex 3D function

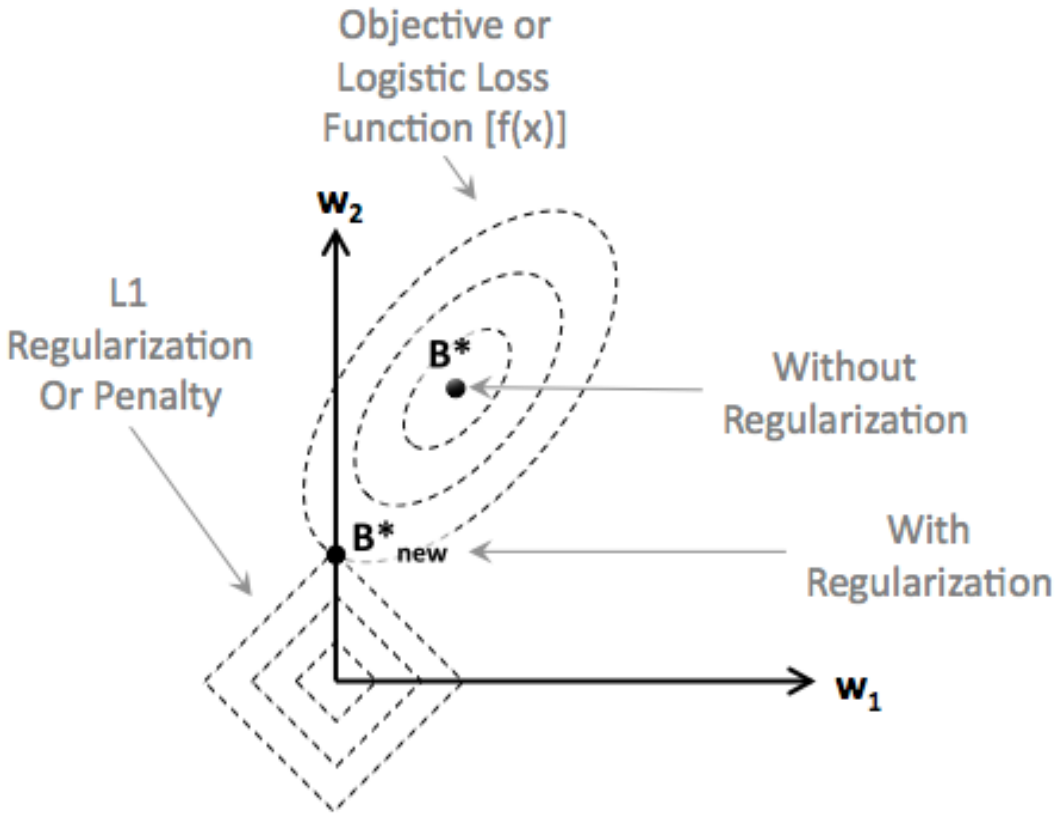
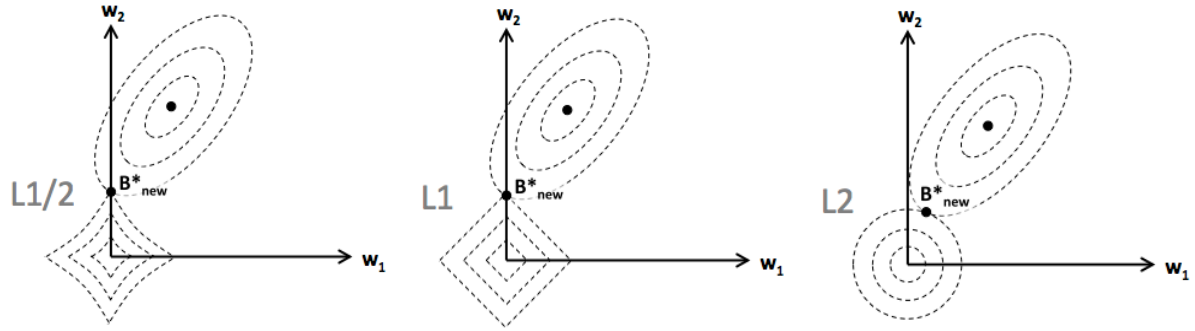


Figure 20. Regularization with L1-norm



a) b) c)

Figure 21. Comparison of regularization with (a) L1/2-norm, (b) L1-norm, and (c) L2-norm

BIBLIOGRAPHY

Abd El Munim, HE, & Farag, Aly A. (2005). *A shape-based segmentation approach: An improved technique using level sets.* Paper presented at the Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on.

Adleman, N. E., Menon, V., Blasey, C. M., White, C. D., Warsofsky, I. S., Glover, G. H., & Reiss, A. L. (2002). A developmental fMRI study of the Stroop color-word task. *Neuroimage, 16*(1), 61-75. doi: 10.1006/nimg.2001.1046

Aizenstein, H. J., Andreescu, C., Edelman, K. L., Cochran, J. L., Price, J., Butters, M. A., . . . Reynolds, C. F., 3rd. (2011). fMRI correlates of white matter hyperintensities in late-life depression. *Am J Psychiatry, 168*(10), 1075-1082. doi: 10.1176/appi.ajp.2011.10060853

Aizenstein, H. J., Butters, M. A., Figurski, J. L., Stenger, V. A., Reynolds, C. F., 3rd, & Carter, C. S. (2005). Prefrontal and striatal activation during sequence learning in geriatric depression. *Biol Psychiatry, 58*(4), 290-296. doi: 10.1016/j.biopsych.2005.04.023

Aizenstein, H. J., Butters, M. A., Wu, M., Mazurkewicz, L. M., Stenger, V. A., Gianaros, P. J., . . . Carter, C. S. (2009). Altered functioning of the executive control circuit in late-life depression: episodic and persistent phenomena. *Am J Geriatr Psychiatry, 17*(1), 30-42. doi: 10.1097/JGP.0b013e31817b60af

Aizenstein, H. J., Khalaf, A., Walker, S. E., & Andreescu, C. (2014). Magnetic resonance imaging predictors of treatment response in late-life depression. *J Geriatr Psychiatry Neurol, 27*(1), 24-32. doi: 10.1177/0891988713516541

Ajilore, O., Lamar, M., Leow, A., Zhang, A., Yang, S., & Kumar, A. (2014). Graph theory analysis of cortical-subcortical networks in late-life depression. *Am J Geriatr Psychiatry, 22*(2), 195-206. doi: 10.1016/j.jagp.2013.03.005

Alalade, E., Denny, K., Potter, G., Steffens, D., & Wang, L. (2011). Altered cerebellar-cerebral functional connectivity in geriatric depression. *PLoS One, 6*(5), e20035. doi: 10.1371/journal.pone.0020035

Alexander, A. L., Lee, J. E., Lazar, M., & Field, A. S. (2007). Diffusion tensor imaging of the brain. *Neurotherapeutics, 4*(3), 316-329. doi: 10.1016/j.nurt.2007.05.011

Alexopoulos, G. S. (2002). Frontostriatal and limbic dysfunction in late-life depression. *Am J Geriatr Psychiatry, 10*(6), 687-695.

Alexopoulos, G. S., Hoptman, M. J., Kanellopoulos, D., Murphy, C. F., Lim, K. O., & Gunning, F. M. (2012). Functional connectivity in the cognitive control network and the default mode network in late-life depression. *J Affect Disord, 139*(1), 56-65. doi: 10.1016/j.jad.2011.12.002

Alexopoulos, G. S., & Kelly, R. E., Jr. (2009). Research advances in geriatric depression. *World Psychiatry, 8*(3), 140-149.

Alexopoulos, G. S., Murphy, C. F., Gunning-Dixon, F. M., Latoussakis, V., Kanellopoulos, D., Klimstra, S., . . . Hoptman, M. J. (2008). Microstructural white matter abnormalities and remission of geriatric depression. *Am J Psychiatry, 165*(2), 238-244. doi: 10.1176/appi.ajp.2007.07050744

Andreescu, C., Mulsant, B. H., Houck, P. R., Whyte, E. M., Mazumdar, S., Dombrovski, A. Y., . . . Reynolds, C. F., 3rd. (2008). Empirically derived decision trees for the treatment of late-life depression. *Am J Psychiatry, 165*(7), 855-862. doi: 10.1176/appi.ajp.2008.07081340

Andreescu, C., & Reynolds, C. F., 3rd. (2011). Late-life depression: evidence-based treatment and promising new directions for research and clinical practice. *Psychiatr Clin North Am, 34*(2), 335-355, vii-iii. doi: 10.1016/j.psc.2011.02.005

Andreescu, C., Tudorascu, D. L., Butters, M. A., Tamburo, E., Patel, M., Price, J., . . . Aizenstein, H. (2013). Resting state functional connectivity and treatment response in late-life depression. *Psychiatry Res, 214*(3), 313-321. doi: 10.1016/j.pscychresns.2013.08.007

Andreescu, C., Wu, M., Butters, M. A., Figurski, J., Reynolds, C. F., 3rd, & Aizenstein, H. J. (2011). The default mode network in late-life anxious depression. *Am J Geriatr Psychiatry, 19*(11), 980-983. doi: 10.1097/JGP.0b013e318227f4f9

Anvar, Amir Maleki, Mohammadi, Alireza, & Pilevar, Abdolhamid. (2013). SOM Neural Network as a Method in Image Color Reduction. *International Journal of Computer Science & Network Security, 13*(2).

Association, American Psychiatric. (2013). *Diagnostic and Statistical Manual of Mental Disorders (Fifth ed.)*: American Psychiatric Publishing.

Baddeley, Alan. (1992). Working memory. *Science, 255*(5044), 556-559.

Baird, Abigail A, Colvin, Mary K, VanHorn, John D, Inati, Souheil, & Gazzaniga, Michael S. (2005). Functional connectivity: integrating behavioral, diffusion tensor imaging, and functional magnetic resonance imaging data sets. *J Cogn Neurosci, 17*(4), 687-693.

Balafar, Mohd Ali, Ramli, Abdul Rahman, Saripan, M Iqbal, & Mashohor, Syamsiah. (2010). Review of brain MRI image segmentation methods. *Artificial Intelligence Review, 33*(3), 261-274.

Baldi, Pierre, Brunak, Søren, Chauvin, Yves, Andersen, Claus AF, & Nielsen, Henrik. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics, 16*(5), 412-424.

Ben-Hur, Asa, & Weston, Jason. (2010). A user's guide to support vector machines *Data mining techniques for the life sciences* (pp. 223-239): Springer.

Benarroch, E. E. (2009). The locus ceruleus norepinephrine system: functional organization and potential clinical significance. *Neurology, 73*(20), 1699-1704. doi: 10.1212/WNL.0b013e3181c2937c

Bergstra, James, & Bengio, Yoshua. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research, 13*, 281-305.

Bezdek, James C, Ehrlich, Robert, & Full, William. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences, 10*(2), 191-203.

Bhalla, R. K., Butters, M. A., Zmuda, M. D., Seligman, K., Mulsant, B. H., Pollock, B. G., & Reynolds, C. F., 3rd. (2005). Does education moderate neuropsychological impairment in late-life depression? *Int J Geriatr Psychiatry, 20*(5), 413-417. doi: 10.1002/gps.1296

Bibi, Stamatia, & Stamelos, Ioannis. (2006). Selecting the appropriate machine learning techniques for the prediction of software development costs *Artificial Intelligence Applications and Innovations* (pp. 533-540): Springer.

Bishop, Christopher M, Svensén, Markus, & Williams, Christopher KI. (1998). GTM: The generative topographic mapping. *Neural computation, 10*(1), 215-234.

Blakely, R. D., & Edwards, R. H. (2012). Vesicular and plasma membrane transporters for neurotransmitters. *Cold Spring Harb Perspect Biol, 4*(2). doi: 10.1101/cshperspect.a005595

Blazer, D. G. (2012). Review: antidepressants are effective for the treatment of major depressive disorder in individuals aged 55 years or older. *Evid Based Ment Health, 15*(3), 72. doi: 10.1136/ebmental-2012-100629

Blazer, D. G., 2nd, & Hybels, C. F. (2005). Origins of depression in later life. *Psychol Med, 35*(9), 1241-1252. doi: 10.1017/S0033291705004411

Blink, Evert J. (2004). mri: Physics. *MRI-physics. net,(Nov 2004), second edition pages4-8*.

Blum, Avrim, & Chawla, Shuchi. (2001). Learning from labeled and unlabeled data using graph mincuts.

Bobb, D. S., Jr., Adinoff, B., Laken, S. J., McClintock, S. M., Rubia, K., Huang, H. W., . . . Kozel, F. A. (2012). Neural correlates of successful response inhibition in unmedicated patients with late-life depression. *Am J Geriatr Psychiatry, 20*(12), 1057-1069. doi: 10.1097/JGP.0b013e318235b728

Bohr, I. J., Kenny, E., Blamire, A., O'Brien, J. T., Thomas, A. J., Richardson, J., & Kaiser, M. (2012). Resting-state functional connectivity in late-life depression: higher global connectivity and more long distance connections. *Front Psychiatry, 3*, 116. doi: 10.3389/fpsyt.2012.00116

Brassen, S., Kalisch, R., Weber-Fahr, W., Braus, D. F., & Buchel, C. (2008). Ventromedial prefrontal cortex processing during emotional evaluation in late-life depression: a longitudinal functional magnetic resonance imaging study. *Biol Psychiatry, 64*(4), 349-355. doi: 10.1016/j.biopsych.2008.03.022

Brickman, A. M., Zahra, A., Muraskin, J., Steffener, J., Holland, C. M., Habeck, C., . . . Stern, Y. (2009). Reduction in cerebral blood flow in areas appearing as white matter hyperintensities on magnetic resonance imaging. *Psychiatry Res, 172*(2), 117-120. doi: 10.1016/j.pscychresns.2008.11.006

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron, 68*(5), 815-834. doi: 10.1016/j.neuron.2010.11.022

Bryan, Janet, Luszcz, Mary A, & Pointer, Sophie. (1999). Executive function and processing resources as predictors of adult age differences in the implementation of encoding strategies. *Aging, Neuropsychology, and Cognition, 6*(4), 273-287.

Buckner, R. L., Snyder, A. Z., Shannon, B. J., LaRossa, G., Sachs, R., Fotenos, A. F., . . . Mintun, M. A. (2005). Molecular, structural, and functional characterization of Alzheimer's disease: evidence for a relationship between default activity, amyloid, and memory. *J Neurosci, 25*(34), 7709-7717. doi: 10.1523/JNEUROSCI.2177-05.2005

Budde, J., Shajan, G., Zaitsev, M., Scheffler, K., & Pohmann, R. (2014). Functional MRI in human subjects with gradient-echo and spin-echo EPI at 9.4 T. *Magn Reson Med, 71*(1), 209-218. doi: 10.1002/mrm.24656

Butcher, K., & Emery, D. (2010). Acute stroke imaging. Part II: The ischemic penumbra. *Can J Neurol Sci, 37*(1), 17-27.

Buxton, R. B., Uludag, K., Dubowitz, D. J., & Liu, T. T. (2004). Modeling the hemodynamic response to brain activation. *Neuroimage, 23 Suppl 1*, S220-233. doi: 10.1016/j.neuroimage.2004.07.013

Carusone, L. M., Srinivasan, J., Gitelman, D. R., Mesulam, M. M., & Parrish, T. B. (2002). Hemodynamic response changes in cerebrovascular disease: implications for functional MR imaging. *AJNR Am J Neuroradiol, 23*(7), 1222-1228.

Chang, C. C., Yu, S. C., McQuoid, D. R., Messer, D. F., Taylor, W. D., Singh, K., . . . Payne, M. E. (2011). Reduction of dorsolateral prefrontal cortex gray matter in late-life depression. *Psychiatry Res, 193*(1), 1-6. doi: 10.1016/j.pscychresns.2011.01.003

Chang-Quan, H., Zheng-Rong, W., Yong-Hong, L., Yi-Zhou, X., & Qing-Xiu, L. (2010). Education and risk for late life depression: a meta-analysis of published literature. *Int J Psychiatry Med, 40*(1), 109-124.

Chiong, W., Wilson, S. M., D'Esposito, M., Kayser, A. S., Grossman, S. N., Poorzand, P., . . . Rankin, K. P. (2013). The salience network causally influences default mode network activity during moral reasoning. *Brain, 136*(Pt 6), 1929-1941. doi: 10.1093/brain/awt066

Chudasama, Y., & Robbins, T. W. (2006). Functions of frontostriatal systems in cognition: comparative neuropsychopharmacological studies in rats, monkeys and humans. *Biol Psychol, 73*(1), 19-38. doi: 10.1016/j.biopsycho.2006.01.005

Colloby, S. J., Firbank, M. J., He, J., Thomas, A. J., Vasudev, A., Parry, S. W., & O'Brien, J. T. (2012). Regional cerebral blood flow in late-life depression: arterial spin labelling magnetic resonance study. *Br J Psychiatry, 200*(2), 150-155. doi: 10.1192/bjp.bp.111.092387

Colloby, S. J., Firbank, M. J., Thomas, A. J., Vasudev, A., Parry, S. W., & O'Brien, J. T. (2011). White matter changes in late-life depression: a diffusion tensor imaging study. *J Affect Disord, 135*(1-3), 216-220. doi: 10.1016/j.jad.2011.07.025

Cortes, Corinna, Mohri, Mehryar, & Mohri, M. (2006). *On transductive regression.* Paper presented at the NIPS.

Cortes, Corinna, & Vapnik, Vladimir. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297. doi: 10.1007/BF00994018

Costafreda, S. G., Chu, C., Ashburner, J., & Fu, C. H. (2009). Prognostic and diagnostic potential of the structural neuroanatomy of depression. *PLoS One, 4*(7), e6353. doi: 10.1371/journal.pone.0006353

Cox, Robert W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research, 29*(3), 162-173.

Crocco, E. A., Castro, K., & Loewenstein, D. A. (2010). How late-life depression affects cognition: neural mechanisms. *Curr Psychiatry Rep, 12*(1), 34-38. doi: 10.1007/s11920-009-0081-2

Czepiel, Scott A. (2002). Maximum likelihood estimation of logistic regression models: theory and implementation. *Available at czep. net/stat/mlelr. pdf*.

Davidson, R. J., Pizzagalli, D., Nitschke, J. B., & Putnam, K. (2002). Depression: perspectives from affective neuroscience. *Annu Rev Psychol, 53*, 545-574. doi: 10.1146/annurev.psych.53.100901.135148

de Groot, J. C., de Leeuw, F. E., Oudkerk, M., van Gijn, J., Hofman, A., Jolles, J., & Breteler, M. M. (2000). Cerebral white matter lesions and cognitive function: the Rotterdam Scan Study. *Ann Neurol, 47*(2), 145-151.

Debette, S., & Markus, H. S. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ, 341*, c3666. doi: 10.1136/bmj.c3666

DeCarli, C., Fletcher, E., Ramey, V., Harvey, D., & Jagust, W. J. (2005). Anatomical mapping of white matter hyperintensities (WMH): exploring the relationships between periventricular WMH, deep WMH, and total WMH burden. *Stroke, 36*(1), 50-55. doi: 10.1161/01.STR.0000150668.58689.f2

Deligianni, Fani, Robinson, Emma, Beckmann, Christian F, Sharp, David, Edwards, A David, & Rueckert, Daniel. (2011). *Inference of functional connectivity from direct and indirect structural brain connections.* Paper presented at the Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on.

Detre, J. A., Wang, J., Wang, Z., & Rao, H. (2009). Arterial spin-labeled perfusion MRI in basic and clinical neuroscience. *Curr Opin Neurol, 22*(4), 348-355. doi: 10.1097/WCO.0b013e32832d9505

Didaci, Luca, & Roli, Fabio. (2006). Using co-training and self-training in semi-supervised multiple classifier systems *Structural, Syntactic, and Statistical Pattern Recognition* (pp. 522-530): Springer.

DiGirolamo, G. J., Kramer, A. F., Barad, V., Cepeda, N. J., Weissman, D. H., Milham, M. P., . . . McAuley, E. (2001). General and task-specific frontal lobe recruitment in older adults during executive processes: a fMRI investigation of task-switching. *Neuroreport, 12*(9), 2065-2071.

Disabato, B. M., & Sheline, Y. I. (2012). Biological basis of late life depression. *Curr Psychiatry Rep, 14*(4), 273-279. doi: 10.1007/s11920-012-0279-6

Domeniconi, Carlotta, Gunopulos, Dimitrios, & Peng, Jing. (2005). Large margin nearest neighbor classifiers. *Neural Networks, IEEE Transactions on, 16*(4), 899-909.

Domingos, Pedro. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87.

Dove, A., Pollmann, S., Schubert, T., Wiggins, C. J., & von Cramon, D. Y. (2000). Prefrontal cortex activation in task switching: an event-related fMRI study. *Brain Res Cogn Brain Res, 9*(1), 103-109.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform, 35*(5-6), 352-359.

Dreiseitl, Stephan, & Ohno-Machado, Lucila. (2002). Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform, 35*(5), 352-359.

Fernández, Antonio, Morales, María, & Salmerón, Antonio. (2007). Tree augmented naive Bayes for regression using mixtures of truncated exponentials: application to higher education management *Advances in Intelligent Data Analysis VII* (pp. 59-69): Springer.

Firbank, M. J., Teodorczuk, A., van der Flier, W. M., Gouw, A. A., Wallin, A., Erkinjuntti, T., . . . group, Ladis. (2012). Relationship between progression of brain white matter changes and late-life depression: 3-year results from the LADIS study. *Br J Psychiatry, 201*(1), 40-45. doi: 10.1192/bjp.bp.111.098897

First, MRJM, Spitzer, RL, Williams, J, & Gibbons, M. (1995). Structured clinical interview for DSM-IV-Patient version. *New York: Biometrics Research Department, New York State Psychiatric Institute*.

Foley, Donald H. (1972). Considerations of sample and feature size. *Information Theory, IEEE Transactions on, 18*(5), 618-626.

Forlani, C., Morri, M., Ferrari, B., Dalmonte, E., Menchetti, M., De Ronchi, D., & Atti, A. R. (2013). Prevalence and Gender Differences in Late-Life Depression: A Population-Based Study. *Am J Geriatr Psychiatry*. doi: 10.1016/j.jagp.2012.08.015

Fox, Michael D, Greicius, Michael, Fox, MD, & Greicius, M. (2010). Clinical applications of resting state functional connectivity. *Frontiers in systems neuroscience, 4*, 19.

Frank, Eibe, Hall, Mark, Trigg, Len, Holmes, Geoffrey, & Witten, Ian H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics, 20*(15), 2479-2481.

Freund, Yoav, Schapire, Robert, & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence, 14*(771-780), 1612.

Friedman, Jerome H, Kohavi, Ron, & Yun, Yeogirl. (1996). *Lazy decision trees.* Paper presented at the AAAI/IAAI, Vol. 1.

Friedman, Nir, Geiger, Dan, & Goldszmidt, Moises. (1997). Bayesian network classifiers. *Machine Learning, 29*(2-3), 131-163.

Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., & Turner, R. (1996). Movement-related effects in fMRI time-series. *Magn Reson Med, 35*(3), 346-355.

Friston, Karl J, Holmes, Andrew P, Worsley, Keith J, Poline, J-P, Frith, Chris D, & Frackowiak, Richard SJ. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Hum Brain Mapp, 2*(4), 189-210.

Fu, C. H., Mourao-Miranda, J., Costafreda, S. G., Khanna, A., Marquand, A. F., Williams, S. C., & Brammer, M. J. (2008). Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. *Biol Psychiatry, 63*(7), 656-662. doi: 10.1016/j.biopsych.2007.08.020

Ganguli, M., Du, Y., Dodge, H. H., Ratcliff, G. G., & Chang, C. C. (2006). Depressive symptoms and cognitive decline in late life: a prospective epidemiological study. *Arch Gen Psychiatry, 63*(2), 153-160. doi: 10.1001/archpsyc.63.2.153

Garg, Skimpy, & Kaur, Jagpreet. (2013). Improving Segmentation by Denoising Brain MRI Images through Interpolation Median Filter in ADTVFCM.

Gawryluk, J. R., Brewer, K. D., Beyea, S. D., & D'Arcy, R. C. (2009). Optimizing the detection of white matter fMRI using asymmetric spin echo spiral. *Neuroimage, 45*(1), 83-88. doi: 10.1016/j.neuroimage.2008.11.005

Gershenson, Carlos. (2003). Artificial neural networks for beginners. *arXiv preprint cs/0308031*.

Ghahramani, Zoubin. (2004). Unsupervised learning *Advanced Lectures on Machine Learning* (pp. 72-112): Springer.

Gottfries, C. G. (2001). Late life depression. *Eur Arch Psychiatry Clin Neurosci, 251 Suppl 2*, II57-61.

Gou, J, Du, L, Zhang, Y, & Xiong, T. (2012). New Distance-weighted knearest Neighbor Classifier. *Journal of Information and Computational Science, 9*, 1429-1436.

Grabczewski, Krzysztof. (2005). Feature selection with decision tree criterion.

Greenwald, B. S., Kramer-Ginsberg, E., Krishnan, K. R., Ashtari, M., Auerbach, C., & Patel, M. (1998). Neuroanatomic localization of magnetic resonance imaging signal hyperintensities in geriatric depression. *Stroke, 29*(3), 613-617.

Greicius, M. D., Supekar, K., Menon, V., & Dougherty, R. F. (2009). Resting-state functional connectivity reflects structural connectivity in the default mode network. *Cereb Cortex, 19*(1), 72-78. doi: 10.1093/cercor/bhn059

Greicius, Michael. (2008). Resting-state functional connectivity in neuropsychiatric disorders. *Curr Opin Neurol, 21*(4), 424-430.

Grieve, S. M., Korgaonkar, M. S., Etkin, A., Harris, A., Koslow, S. H., Wisniewski, S., . . . Williams, L. M. (2013). Brain imaging predictors and the international study to predict optimized treatment for depression: study protocol for a randomized controlled trial. *Trials, 14*(1), 224. doi: 10.1186/1745-6215-14-224

Gunning, F. M., Cheng, J., Murphy, C. F., Kanellopoulos, D., Acuna, J., Hoptman, M. J., . . . Alexopoulos, G. S. (2009). Anterior cingulate cortical volumes and treatment remission of geriatric depression. *Int J Geriatr Psychiatry, 24*(8), 829-836. doi: 10.1002/gps.2290

Gunning-Dixon, F. M., Brickman, A. M., Cheng, J. C., & Alexopoulos, G. S. (2009). Aging of cerebral white matter: a review of MRI findings. *Int J Geriatr Psychiatry, 24*(2), 109-117. doi: 10.1002/gps.2087

Gunning-Dixon, F. M., Walton, M., Cheng, J., Acuna, J., Klimstra, S., Zimmerman, M. E., . . . Alexopoulos, G. S. (2010). MRI signal hyperintensities and treatment remission of geriatric depression. *J Affect Disord, 126*(3), 395-401. doi: 10.1016/j.jad.2010.04.004

Hahn, T., Marquand, A. F., Ehlis, A. C., Dresler, T., Kittel-Schneider, S., Jarczok, T. A., . . . Fallgatter, A. J. (2011). Integrating neurobiological markers of depression. *Arch Gen Psychiatry, 68*(4), 361-368. doi: 10.1001/archgenpsychiatry.2010.178

Halkidi, Maria, Batistakis, Yannis, & Vazirgiannis, Michalis. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems, 17*(2-3), 107-145.

Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, & Witten, Ian H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*(1), 10-18.

Hart, H., & Rubia, K. (2012). Neuroimaging of child abuse: a critical review. *Front Hum Neurosci, 6*, 52. doi: 10.3389/fnhum.2012.00052

He, Jing, Carmichael, Owen, Fletcher, Evan, Singh, Baljeet, Iosif, Ana-Maria, Martinez, Oliver, . . . DeCarli, Charles. (2012). Influence of functional connectivity and structural MRI measures on episodic memory. *Neurobiol Aging, 33*(11), 2612-2620.

Hedden, T., Van Dijk, K. R., Shire, E. H., Sperling, R. A., Johnson, K. A., & Buckner, R. L. (2012). Failure to modulate attentional control in advanced aging linked to white matter pathology. *Cereb Cortex, 22*(5), 1038-1051. doi: 10.1093/cercor/bhr172

Hill, D. L., Batchelor, P. G., Holden, M., & Hawkes, D. J. (2001). Medical image registration. *Phys Med Biol, 46*(3), R1-45.

Honey, CJ, Sporns, O, Cammoun, Leila, Gigandet, Xavier, Thiran, Jean-Philippe, Meuli, Reto, & Hagmann, Patric. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences, 106*(6), 2035-2040.

Hornak, Joseph P. (1996). The Basics of MRI. *Web book available at http://www. cis. rit. edu/htbooks/mri/. Accessed July, 17*.

Howseman, Alistair M, Josephs, Oliver, Rees, Geraint, & Friston, Karl J. (1997). Special issues in functional magnetic resonance imaging. *SPM course, short course notes*.

Hyman, S. E. (2005). Neurotransmitters. *Curr Biol, 15*(5), R154-158. doi: 10.1016/j.cub.2005.02.037

Hyvärinen, Aapo, & Pajunen, Petteri. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks, 12*(3), 429-439.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage, 62*(2), 782-790. doi: 10.1016/j.neuroimage.2011.09.015

Jin, Jing, & Yang, Yunle. (2013). *Similarity metric in medical image registration.* Paper presented at the Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering.

Jing, Yushi, Pavlović, Vladimir, & Rehg, James M. (2008). Boosted Bayesian network classifiers. *Machine Learning, 73*(2), 155-184.

Joel, I., Begley, A. E., Mulsant, B. H., Lenze, E. J., Mazumdar, S., Dew, M. A., . . . Team, Irl Grey Investigative. (2014). Dynamic prediction of treatment response in late-life depression. *Am J Geriatr Psychiatry, 22*(2), 167-176. doi: 10.1016/j.jagp.2012.07.002

Johnson, H J, McCormick, M, & L, Ibanex. (2013). The ITK Software Guide.

Kales, H. C., Maixner, D. F., & Mellow, A. M. (2005). Cerebrovascular disease and late-life depression. *Am J Geriatr Psychiatry, 13*(2), 88-98. doi: 10.1176/appi.ajgp.13.2.88

Kanungo, Tapas, Mount, David M, Netanyahu, Nathan S, Piatko, Christine D, Silverman, Ruth, & Wu, Angela Y. (2002). *A local search approximation algorithm for k-means clustering.* Paper presented at the Proceedings of the eighteenth annual symposium on Computational geometry.

Kapitanova, Krasimira, & Son, Sang H. (2012). Machine Learning Basics. *Intelligent Sensor Networks: The Integration of Sensor Networks, Signal Processing and Machine Learning*, 13.

Karlsson, Hasse. (2012). Psychotherapy increases the amount of serotonin receptors in the brains of patients with major depressive disorder *Psychodynamic Psychotherapy Research* (pp. 233-238): Springer.

Katon, W., Unutzer, J., & Russo, J. (2010). Major depression: the importance of clinical characteristics and treatment response to prognosis. *Depress Anxiety, 27*(1), 19-26. doi: 10.1002/da.20613

Khan, Aamir, & Farooq, Hasan. (2011). Principal Component Analysis-Linear Discriminant Analysis Feature Extractor for Pattern Recognition. *International Journal of Computer Science Issues (IJCSI), 8*(6).

Kingsford, Carl, & Salzberg, Steven L. (2008). What are decision trees? *Nature biotechnology, 26*(9), 1011-1013.

Kivinen, Jyrki, & Warmuth, Manfred K. (1997). Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation, 132*(1), 1-63.

Kohavi, Ron. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the IJCAI.

Kohavi, Ron. (1996). *Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid.* Paper presented at the KDD.

Kohavi, Ron, & John, George H. (1995). *Automatic parameter selection by minimizing estimated error.* Paper presented at the ICML.

Kohavi, Ron, & John, George H. (1997). Wrappers for feature subset selection. *Artificial intelligence, 97*(1), 273-324.

Kohler, S., Thomas, A. J., Barnett, N. A., & O'Brien, J. T. (2010). The pattern and course of cognitive impairment in late-life depression. *Psychol Med, 40*(4), 591-602. doi: 10.1017/S0033291709990833

Kohler, S., Thomas, A. J., Lloyd, A., Barber, R., Almeida, O. P., & O'Brien, J. T. (2010). White matter hyperintensities, cortisol levels, brain atrophy and continuing cognitive deficits in late-life depression. *Br J Psychiatry, 196*(2), 143-149. doi: 10.1192/bjp.bp.109.071399

Kotthoff, Lars, Gent, Ian P, & Miguel, Ian. (2012). An evaluation of machine learning in algorithm selection for search problems. *AI Communications, 25*(3), 257-270.

Kruggel, F., Paul, J. S., & Gertz, H. J. (2008). Texture-based segmentation of diffuse lesions of the brain's white matter. *Neuroimage, 39*(3), 987-996. doi: 10.1016/j.neuroimage.2007.09.058

Ladha, L, & Deepa, T. (2011). FEATURE SELECTION METHODS AND ALGORITHMS. *International Journal on Computer Science & Engineering, 3*(5).

Lambert, O., & Bourin, M. (2002). SNRIs: mechanism of action and clinical features. *Expert Rev Neurother, 2*(6), 849-858. doi: 10.1586/14737175.2.6.849

Langlois, Dominic, Chartier, Sylvain, & Gosselin, Dominique. (2010). An introduction to independent component analysis: InfoMax and FastICA algorithms. *Tutorials in Quantitative Methods for Psychology, 6*(1), 31-38.

Le Bihan, D., Mangin, J. F., Poupon, C., Clark, C. A., Pappata, S., Molko, N., & Chabriat, H. (2001). Diffusion tensor imaging: concepts and applications. *J Magn Reson Imaging, 13*(4), 534-546.

Le, Quoc V, Ngiam, Jiquan, Coates, Adam, Lahiri, Ahbik, Prochnow, Bobby, & Ng, Andrew Y. (2011). *On optimization methods for deep learning.* Paper presented at the Proceedings of the 28th International Conference on Machine Learning (ICML-11).

Lebowitz, B. D., Pearson, J. L., Schneider, L. S., Reynolds, C. F., 3rd, Alexopoulos, G. S., Bruce, M. L., . . . Parmelee, P. (1997). Diagnosis and treatment of depression in late life. Consensus statement update. *JAMA, 278*(14), 1186-1190.

Lee, Jong-Sen. (1983). Digital image smoothing and the sigma filter. *Computer Vision, Graphics, and Image Processing, 24*(2), 255-269.

Li, Wenjun, Antuono, Piero G, Xie, Chunming, Chen, Gang, Jones, Jennifer L, Ward, B Douglas, . . . Li, Shi-Jiang. (2012). Changes in regional cerebral blood flow and functional connectivity in the cholinergic pathway associated with cognitive performance in subjects with mild Alzheimer's disease after 12-week donepezil treatment. *Neuroimage, 60*(2), 1083-1091.

Liang, Peipeng, Wang, Zhiqun, Yang, Yanhui, Jia, Xiuqin, & Li, Kuncheng. (2011). Functional disconnection and compensation in mild cognitive impairment: evidence from DLPFC connectivity using resting-state fMRI. *PLoS One, 6*(7).

Lim, Chinghway, & Yu, Bin. (2013). Estimation Stability with Cross Validation (ESCV). *arXiv preprint arXiv:1303.3128*.

Linortner, P., Fazekas, F., Schmidt, R., Ropele, S., Pendl, B., Petrovic, K., . . . Enzinger, C. (2012). White matter hyperintensities alter functional organization of the motor system. *Neurobiol Aging, 33*(1), 197 e191-199. doi: 10.1016/j.neurobiolaging.2010.06.005

Liu, F., Guo, W., Yu, D., Gao, Q., Gao, K., Xue, Z., . . . Chen, H. (2012). Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. *PLoS One, 7*(7), e40968. doi: 10.1371/journal.pone.0040968

Liu, F., Hu, M., Wang, S., Guo, W., Zhao, J., Li, J., . . . Chen, H. (2012). Abnormal regional spontaneous neural activity in first-episode, treatment-naive patients with late-life depression: a resting-state fMRI study. *Prog Neuropsychopharmacol Biol Psychiatry, 39*(2), 326-331. doi: 10.1016/j.pnpbp.2012.07.004

Liu, Han, & Zhang, Jian. (2008). On the l1− lq regularized regression: Technical Report.

Loh, Wei-Yin. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 14-23.

Lu, Yijuan, Cohen, Ira, Zhou, Xiang Sean, & Tian, Qi. (2007). *Feature selection using principal feature analysis.* Paper presented at the Proceedings of the 15th international conference on Multimedia.

Luppa, M., Sikorski, C., Luck, T., Ehreke, L., Konnopka, A., Wiese, B., . . . Riedel-Heller, S. G. (2012). Age- and gender-specific prevalence of depression in latest-life--systematic review and meta-analysis. *J Affect Disord, 136*(3), 212-221. doi: 10.1016/j.jad.2010.11.033

Maclin, Richard, & Opitz, David. (2011). Popular ensemble methods: An empirical study. *arXiv preprint arXiv:1106.0257*.

Madden, D. J., Bennett, I. J., & Song, A. W. (2009). Cerebral white matter integrity and cognitive aging: contributions from diffusion tensor imaging. *Neuropsychol Rev, 19*(4), 415-435. doi: 10.1007/s11065-009-9113-2

Madsen, T. M., Treschow, A., Bengzon, J., Bolwig, T. G., Lindvall, O., & Tingstrom, A. (2000). Increased neurogenesis in a model of electroconvulsive therapy. *Biol Psychiatry, 47*(12), 1043-1049.

Mahmoudzadeh, A. P., & Kashou, N. H. (2013). Evaluation of interpolation effects on upsampling and accuracy of cost functions-based optimized automatic image registration. *Int J Biomed Imaging, 2013*, 395915. doi: 10.1155/2013/395915

Maintz, J. B., & Viergever, M. A. (1998). A survey of medical image registration. *Med Image Anal, 2*(1), 1-36.

Maistrou, Aliaksei. (2008). Level set methods—overview: Technical Report, Computer Aided Medical Procedures, TUM.

Maltamo, Matti, & Kangas, Annika. (1998). Methods based on k-nearest neighbor regression in the prediction of basal area diameter distribution. *Canadian Journal of Forest Research, 28*(8), 1107-1115.

Marquand, A. F., Mourao-Miranda, J., Brammer, M. J., Cleare, A. J., & Fu, C. H. (2008). Neuroanatomy of verbal working memory as a diagnostic biomarker for depression. *Neuroreport, 19*(15), 1507-1511. doi: 10.1097/WNR.0b013e328310425e

Marstrand, J. R., Garde, E., Rostrup, E., Ring, P., Rosenbaum, S., Mortensen, E. L., & Larsson, H. B. (2002). Cerebral perfusion and cerebrovascular reactivity are reduced in white matter hyperintensities. *Stroke, 33*(4), 972-976.

Mazerolle, E. L., D'Arcy, R. C., & Beyea, S. D. (2008). Detecting functional magnetic resonance imaging activation in white matter: interhemispheric transfer across the corpus callosum. *BMC Neurosci, 9*, 84. doi: 10.1186/1471-2202-9-84

McAuliffe, Matthew J, Lalonde, Francois M, McGarry, Delia, Gandler, William, Csaky, Karl, & Trus, Benes L. (2001). *Medical image processing, analysis and visualization in clinical research.* Paper presented at the Computer-Based Medical Systems, 2001. CBMS 2001. Proceedings. 14th IEEE Symposium on.

McRobbie, Donald W. (2007). *MRI from Picture to Proton*: Cambridge University Press.

Meltzer, C. C., Smith, G., DeKosky, S. T., Pollock, B. G., Mathis, C. A., Moore, R. Y., . . . Reynolds, C. F., 3rd. (1998). Serotonin in aging, late-life depression, and Alzheimer's disease: the emerging role of functional imaging. *Neuropsychopharmacology, 18*(6), 407-430. doi: 10.1016/S0893-133X(97)00194-2

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Struct Funct, 214*(5-6), 655-667. doi: 10.1007/s00429-010-0262-0

MentalHealthAmerica.). Depression in Older Adults. from http://www.mentalhealthamerica.net/conditions/depression-older-adults#2

Mettenburg, J. M., Benzinger, T. L., Shimony, J. S., Snyder, A. Z., & Sheline, Y. I. (2012). Diminished performance on neuropsychological testing in late life depression is correlated with microstructural white matter abnormalities. *Neuroimage, 60*(4), 2182-2190. doi: 10.1016/j.neuroimage.2012.02.044

Meyer, David. (2012). Support Vector Machines. *The Interface to libsvm in package e1071. e1071 Vignette*.

Miyapuram, Krishna P, Tobler, Philippe N, Schultz, Wolfram, Osterbauer, Robert, & Schwarzbauer, Christian. (2009). Imaging Brain Regions with Susceptibility-induced Signal Losses using Gradient and Spin Echo Techniques.

Mukherjee, P., Berman, J. I., Chung, S. W., Hess, C. P., & Henry, R. G. (2008). Diffusion tensor MR imaging and fiber tractography: theoretic underpinnings. *AJNR Am J Neuroradiol, 29*(4), 632-641. doi: 10.3174/ajnr.A1051

Mukherjee, P., Chung, S. W., Berman, J. I., Hess, C. P., & Henry, R. G. (2008). Diffusion tensor MR imaging and fiber tractography: technical considerations. *AJNR Am J Neuroradiol, 29*(5), 843-852. doi: 10.3174/ajnr.A1052

Muller, K. R., Anderson, C. W., & Birch, G. E. (2003). Linear and nonlinear methods for brain-computer interfaces. *IEEE Trans Neural Syst Rehabil Eng, 11*(2), 165-169. doi: 10.1109/TNSRE.2003.814484

Müller, Meinard, Röder, Tido, & Clausen, Michael. (2005). Efficient content-based retrieval of motion capture data. *ACM Transactions on Graphics (TOG), 24*(3), 677-685.

Mwangi, B., Ebmeier, K. P., Matthews, K., & Steele, J. D. (2012). Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. *Brain, 135*(Pt 5), 1508-1521. doi: 10.1093/brain/aws084

Mwangi, B., Matthews, K., & Steele, J. D. (2012). Prediction of illness severity in patients with major depression using structural MR brain scans. *J Magn Reson Imaging, 35*(1), 64-71. doi: 10.1002/jmri.22806

Mwangi, Benson, Tian, Tian Siva, & Soares, Jair C. (2013). A Review of Feature Reduction Techniques in Neuroimaging. *Neuroinformatics*, 1-16.

Na, D. G., Ryu, J. W., Byun, H. S., Choi, D. S., Lee, E. J., Chung, W. I., . . . Han, B. K. (2000). Functional MR imaging of working memory in the human brain. *Korean J Radiol, 1*(1), 19-24.

Ng, Andrew Y. (2004). *Feature selection, L 1 vs. L 2 regularization, and rotational invariance.* Paper presented at the Proceedings of the twenty-first international conference on Machine learning.

Ng, Andrew Y, & Jordan, Michael I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems, 2*, 841-848.

Nieto-Castanon, A., Ghosh, S. S., Tourville, J. A., & Guenther, F. H. (2003). Region of interest based analysis of functional imaging data. *Neuroimage, 19*(4), 1303-1316.

Nigam, Kamal, McCallum, Andrew, & Mitchell, Tom. (2006). Semi-supervised text classification using EM. *Semi-Supervised Learning*, 33-56.

Nordahl, C. W., Ranganath, C., Yonelinas, A. P., Decarli, C., Fletcher, E., & Jagust, W. J. (2006). White matter changes compromise prefrontal cortex function in healthy elderly individuals. *J Cogn Neurosci, 18*(3), 418-429. doi: 10.1162/089892906775990552

Nouretdinov, I., Costafreda, S. G., Gammerman, A., Chervonenkis, A., Vovk, V., Vapnik, V., & Fu, C. H. (2011). Machine learning classification with confidence: application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. *Neuroimage, 56*(2), 809-813. doi: 10.1016/j.neuroimage.2010.05.023

Nutt, D. J. (2008). Relationship of neurotransmitters to the symptoms of major depressive disorder. *J Clin Psychiatry, 69 Suppl E1*, 4-7.

Ostwald, D., & Bagshaw, A. P. (2011). Information theoretic approaches to functional neuroimaging. *Magn Reson Imaging, 29*(10), 1417-1428. doi: 10.1016/j.mri.2011.07.013

Papakostas, G. I., Thase, M. E., Fava, M., Nelson, J. C., & Shelton, R. C. (2007). Are antidepressant drugs that combine serotonergic and noradrenergic mechanisms of

action more effective than the selective serotonin reuptake inhibitors in treating major depressive disorder? A meta-analysis of studies of newer agents. *Biol Psychiatry, 62*(11), 1217-1227. doi: 10.1016/j.biopsych.2007.03.027

Parker, J., Kenyon, R. V., & Troxel, D. E. (1983). Comparison of interpolating methods for image resampling. *IEEE Trans Med Imaging, 2*(1), 31-39. doi: 10.1109/TMI.1983.4307610

Pavlović, Vladimir, Garg, Ashutosh, & Kasif, Simon. (2002). A Bayesian framework for combining gene predictions. *Bioinformatics, 18*(1), 19-27.

Pedregosa, Fabian, Ga, #235, Varoquaux, l, Gramfort, Alexandre, Michel, Vincent, . . . Duchesnay, douard. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res., 12*, 2825-2830.

Peharz, Robert, Tschiatschek, Sebastian, & Pernkopf, Franz. (2013). *The most generative maximum margin Bayesian networks.* Paper presented at the Proceedings of The 30th International Conference on Machine Learning.

Pfahringer, Bernhard, Holmes, Geoffrey, & Kirkby, Richard. (2001). Optimizing the Induction of Alternating Decision Trees. In D. Cheung, G. Williams & Q. Li (Eds.), *Advances in Knowledge Discovery and Data Mining* (Vol. 2035, pp. 477-487): Springer Berlin Heidelberg.

Pieper, Steve, Halle, Michael, & Kikinis, Ron. (2004). *3D Slicer.* Paper presented at the Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on.

Pierce, R. C., & Kumaresan, V. (2006). The mesolimbic dopamine system: the final common pathway for the reinforcing effect of drugs of abuse? *Neurosci Biobehav Rev, 30*(2), 215-238. doi: 10.1016/j.neubiorev.2005.04.016

Pise, Nitin Namdeo, & Kulkarni, Parag. (2008). *A survey of semi-supervised learning methods.* Paper presented at the Computational Intelligence and Security, 2008. CIS'08. International Conference on.

Polesel, A., Ramponi, G., & Mathews, V. J. (2000). Image enhancement via adaptive unsharp masking. *IEEE Trans Image Process, 9*(3), 505-510. doi: 10.1109/83.826787

Punia, Ritu, & Singh, Shailendra. (2013). Review on Machine Learning Techniques for Automatic Segmentation of Liver Images. *International Journal, 3*(4).

Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *J. Artif. Int. Res., 4*(1), 77-90.

Quitkin, F., Rifkin, A., & Klein, D. F. (1979). Monoamine oxidase inhibitors. A review of antidepressant effectiveness. *Arch Gen Psychiatry, 36*(7), 749-760.

Racagni, G., & Popoli, M. (2008). Cellular and molecular mechanisms in the long-term action of antidepressants. *Dialogues Clin Neurosci, 10*(4), 385-400.

Raiteri, M. (2001). Presynaptic autoreceptors. *J Neurochem, 78*(4), 673-675.

Rauber, Andreas, Merkl, Dieter, & Dittenbach, Michael. (2002). The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. *Neural Networks, IEEE Transactions on, 13*(6), 1331-1341.

Reif, Matthias, & Shafait, Faisal. (2014). Efficient feature size reduction via predictive forward selection. *Pattern Recognition, 47*(4), 1664-1673.

Rendón, Eréndira, Abundez, Itzel, Arizmendi, Alejandra, & Quiroz, Elvia M. (2011). Internal versus External cluster validation indexes. *International Journal of computers and communications, 5*(1), 27-34.

Reuter-Lorenz, P. A., & Lustig, C. (2005). Brain aging: reorganizing discoveries about the aging mind. *Curr Opin Neurobiol, 15*(2), 245-251. doi: 10.1016/j.conb.2005.03.016

Ribeiz, S. R., Duran, F., Oliveira, M. C., Bezerra, D., Castro, C. C., Steffens, D. C., . . . Bottino, C. M. (2013). Structural brain changes as biomarkers and outcome predictors in patients with late-life depression: a cross-sectional and prospective study. *PLoS One, 8*(11), e80049. doi: 10.1371/journal.pone.0080049

Roche, Alexis, Malandain, Grégoire, Pennec, Xavier, & Ayache, Nicholas. (1998). The correlation ratio as a new similarity measure for multimodal image registration *Medical Image Computing and Computer-Assisted Interventation—MICCAI'98* (pp. 1115-1124): Springer.

Roose, S. P., Glassman, A. H., Attia, E., & Woodring, S. (1994). Comparative efficacy of selective serotonin reuptake inhibitors and tricyclics in the treatment of melancholia. *Am J Psychiatry, 151*(12), 1735-1739.

Rosano, C., Becker, J., Lopez, O., Lopez-Garcia, P., Carter, C. S., Newman, A., . . . Aizenstein, H. (2005). Morphometric analysis of gray matter volume in demented older adults: exploratory analysis of the cardiovascular health study brain MRI database. *Neuroepidemiology, 24*(4), 221-229. doi: 10.1159/000085140

Rossini, P. M., Altamura, C., Ferretti, A., Vernieri, F., Zappasodi, F., Caulo, M., . . . Tecchio, F. (2004). Does cerebrovascular disease affect the coupling between neuronal activity and local haemodynamics? *Brain, 127*(Pt 1), 99-110. doi: 10.1093/brain/awh012

Rousson, Mikaël, Paragios, Nikos, & Deriche, Rachid. (2004). Implicit active shape models for 3D segmentation in MR imaging *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004* (pp. 209-216): Springer.

Sachdev, P., Wen, W., Shnier, R., & Brodaty, H. (2004). Cerebral blood volume in T2-weighted white matter hyperintensities using exogenous contrast based perfusion MRI. *J Neuropsychiatry Clin Neurosci, 16*(1), 83-92.

Saeys, Yvan, Inza, Iñaki, & Larrañaga, Pedro. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507-2517.

Salmon, E., Van der Linden, M., Collette, F., Delfiore, G., Maquet, P., Degueldre, C., . . . Franck, G. (1996). Regional brain activity during working memory tasks. *Brain, 119 (Pt 5)*, 1617-1625.

Sarle, Warren S. (1994). Neural networks and statistical models.

Sathya, R, & Abraham, Annamma. (2013). Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *International Journal of Advanced Research in Artificial Intelligence*, 34-38.

Schapire, Robert E. (2003). The boosting approach to machine learning: An overview *Nonlinear estimation and classification* (pp. 149-171): Springer.

Sexton, C. E., Mackay, C. E., & Ebmeier, K. P. (2013). A systematic review and meta-analysis of magnetic resonance imaging studies in late-life depression. *Am J Geriatr Psychiatry, 21*(2), 184-195. doi: 10.1016/j.jagp.2012.10.019

Shimony, J. S., Sheline, Y. I., D'Angelo, G., Epstein, A. A., Benzinger, T. L., Mintun, M. A., . . . Snyder, A. Z. (2009). Diffuse microstructural abnormalities of normal-appearing white matter in late life depression: a diffusion tensor imaging study. *Biol Psychiatry, 66*(3), 245-252. doi: 10.1016/j.biopsych.2009.02.032

Shirer, W. R., Ryali, S., Rykhlevskaia, E., Menon, V., & Greicius, M. D. (2012). Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cereb Cortex, 22*(1), 158-165. doi: 10.1093/cercor/bhr099

Singh, Mr Harvinder. (2013). Image Enhancement using Sharpen Filters.

Sled, J. G., Zijdenbos, A. P., & Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans Med Imaging, 17*(1), 87-97. doi: 10.1109/42.668698

Smith, Travis B, & Nayak, Krishna S. (2010). MRI artifacts and correction strategies. *Imaging in Medicine, 2*(4), 445-457.

Smola, Alex J, & Schölkopf, Bernhard. (2004). A tutorial on support vector regression. *Statistics and computing, 14*(3), 199-222.

Sotiras, A., Davatzikos, C., & Paragios, N. (2013). Deformable medical image registration: a survey. *IEEE Trans Med Imaging, 32*(7), 1153-1190. doi: 10.1109/TMI.2013.2265603

Srinivasan, A., Goyal, M., Al Azri, F., & Lum, C. (2006). State-of-the-art imaging of acute stroke. *Radiographics, 26 Suppl 1*, S75-95. doi: 10.1148/rg.26si065501

Stahl, S. M. (1998). Mechanism of action of serotonin selective reuptake inhibitors. Serotonin receptors and pathways mediate therapeutic effects and side effects. *J Affect Disord, 51*(3), 215-235.

Stahl, S. M., Pradko, J. F., Haight, B. R., Modell, J. G., Rockett, C. B., & Learned-Coughlin, S. (2004). A Review of the Neuropharmacology of Bupropion, a Dual Norepinephrine and Dopamine Reuptake Inhibitor. *Prim Care Companion J Clin Psychiatry, 6*(4), 159-166.

Steffens, D. C., Taylor, W. D., Denny, K. L., Bergman, S. R., & Wang, L. (2011). Structural integrity of the uncinate fasciculus and resting state functional connectivity of the ventral prefrontal cortex in late life depression. *PLoS One, 6*(7), e22697. doi: 10.1371/journal.pone.0022697

Stern, Y. (2003). The concept of cognitive reserve: a catalyst for research. *J Clin Exp Neuropsychol, 25*(5), 589-593. doi: 10.1076/jcen.25.5.589.14571

Stimmel, G. L., Dopheide, J. A., & Stahl, S. M. (1997). Mirtazapine: an antidepressant with noradrenergic and specific serotonergic effects. *Pharmacotherapy, 17*(1), 10-21.

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Curr Opin HIV AIDS, 5*(6), 463-466. doi: 10.1097/COH.0b013e32833ed177

Sylvester, C. Y., Wager, T. D., Lacey, S. C., Hernandez, L., Nichols, T. E., Smith, E. E., & Jonides, J. (2003). Switching attention and resolving interference: fMRI measures of executive functions. *Neuropsychologia, 41*(3), 357-370.

Tadayonnejad, R., & Ajilore, O. (2013). Brain Network Dysfunction in Late-Life Depression: A Literature Review. *J Geriatr Psychiatry Neurol*. doi: 10.1177/0891988713516539

Taskar, Benjamin, Wong, Ming Fai, & Koller, Daphne. (2003). *Learning on the test data: Leveraging unseen features.* Paper presented at the ICML.

Taylor, M. J., Freemantle, N., Geddes, J. R., & Bhagwagar, Z. (2006). Early onset of selective serotonin reuptake inhibitor antidepressant action: systematic review and meta-analysis. *Arch Gen Psychiatry, 63*(11), 1217-1223. doi: 10.1001/archpsyc.63.11.1217

Taylor, W. D., Aizenstein, H. J., & Alexopoulos, G. S. (2013). The vascular depression hypothesis: mechanisms linking vascular disease with depression. *Mol Psychiatry, 18*(9), 963-974. doi: 10.1038/mp.2013.20

Taylor, W. D., & Doraiswamy, P. M. (2004). A systematic review of antidepressant placebo-controlled trials for geriatric depression: limitations of current data and directions for the future. *Neuropsychopharmacology, 29*(12), 2285-2299. doi: 10.1038/sj.npp.1300550

Taylor, W. D., Kuchibhatla, M., Payne, M. E., Macfall, J. R., Sheline, Y. I., Krishnan, K. R., & Doraiswamy, P. M. (2008). Frontal white matter anisotropy and antidepressant remission in late-life depression. *PLoS One, 3*(9), e3267. doi: 10.1371/journal.pone.0003267

Taylor, W. D., Macfall, J. R., Boyd, B., Payne, M. E., Sheline, Y. I., Krishnan, R. R., & Murali Doraiswamy, P. (2011). One-year change in anterior cingulate cortex white matter microstructure: relationship with late-life depression outcomes. *Am J Geriatr Psychiatry, 19*(1), 43-52. doi: 10.1097/JGP.0b013e3181e70cec

Taylor, W. D., Payne, M. E., Krishnan, K. R., Wagner, H. R., Provenzale, J. M., Steffens, D. C., & MacFall, J. R. (2001). Evidence of white matter tract disruption in MRI hyperintensities. *Biol Psychiatry, 50*(3), 179-183.

Teipel, Stefan J, Bokde, Arun LW, Meindl, Thomas, Amaro Jr, Edson, Soldner, Jasmin, Reiser, Maximilian F, . . . Hampel, Harald. (2010). White matter microstructure underlying default mode network connectivity in the human brain. *Neuroimage, 49*(3), 2021-2032.

Tekin, S., & Cummings, J. L. (2002). Frontal-subcortical neuronal circuits and clinical neuropsychiatry: an update. *J Psychosom Res, 53*(2), 647-654.

Teodorczuk, A., Firbank, M. J., Pantoni, L., Poggesi, A., Erkinjuntti, T., Wallin, A., . . . Group, Ladis. (2010). Relationship between baseline white-matter changes and development of late-life depressive symptoms: 3-year results from the LADIS study. *Psychol Med, 40*(4), 603-610. doi: 10.1017/S0033291709990857

Thase, M. E., Ninan, P. T., Musgnung, J. J., & Trivedi, M. H. (2011). Remission with venlafaxine extended release or selective serotonin reuptake inhibitors in depressed patients: a randomized, open-label study. *Prim Care Companion CNS Disord, 13*(1). doi: 10.4088/PCC.10m00979blu

Thevenaz, P., Blu, T., & Unser, M. (2000). Interpolation revisited. *IEEE Trans Med Imaging, 19*(7), 739-758. doi: 10.1109/42.875199

Tipping, Michael E. (2001). Sparse Bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research, 1*, 211-244.

Tsai, A., Wells, W., Tempany, C., Grimson, E., & Willsky, A. (2004). Mutual information in coupled multi-shape model for medical image segmentation. *Med Image Anal, 8*(4), 429-445. doi: 10.1016/j.media.2004.01.003

Unutzer, J. (2007). Clinical practice. Late-life depression. *N Engl J Med, 357*(22), 2269-2276. doi: 10.1056/NEJMcp073754

Vernooij, M. W., Ikram, M. A., Vrooman, H. A., Wielopolski, P. A., Krestin, G. P., Hofman, A., . . . Breteler, M. M. (2009). White matter microstructural integrity and cognitive function in a general elderly population. *Arch Gen Psychiatry, 66*(5), 545-553. doi: 10.1001/archgenpsychiatry.2009.5

Vincent, Grayson K, & Velkoff, Victoria Averil. (2010). *The next four decades: The older population in the United States: 2010 to 2050*: US Department of Commerce, Economics and Statistics Administration, US Census Bureau.

Vink, Matthijs, Raemaekers, M, van der Schaaf, A, Mandl, R, & Ramsey, N. (2007). Pre-processing and Analysis.

Wahlund, L. O., Barkhof, F., Fazekas, F., Bronge, L., Augustin, M., Sjogren, M., . . . European Task Force on Age-Related White Matter, Changes. (2001). A new rating scale for age-related white matter changes applicable to MRI and CT. *Stroke, 32*(6), 1318-1322.

Wakana, S., Jiang, H., Nagae-Poetscher, L. M., van Zijl, P. C., & Mori, S. (2004). Fiber tract-based atlas of human white matter anatomy. *Radiology, 230*(1), 77-87. doi: 10.1148/radiol.2301021640

Wang, L., Krishnan, K. R., Steffens, D. C., Potter, G. G., Dolcos, F., & McCarthy, G. (2008). Depressive state- and disease-related alterations in neural responses to affective and executive challenges in geriatric depression. *Am J Psychiatry, 165*(7), 863-871. doi: 10.1176/appi.ajp.2008.07101590

Wang, Yong, Yang, Chengyong, Mathee, Kalai, & Narasimhan, Giri. (2005). Clustering using adaptive self-organizing maps (asom) and applications *Computational Science–ICCS 2005* (pp. 944-951): Springer.

Wedeen, V. J., Wang, R. P., Schmahmann, J. D., Benner, T., Tseng, W. Y., Dai, G., . . . de Crespigny, A. J. (2008). Diffusion spectrum magnetic resonance imaging (DSI) tractography of crossing fibers. *Neuroimage, 41*(4), 1267-1277. doi: 10.1016/j.neuroimage.2008.03.036

Weinberger, Kilian, Blitzer, John, & Saul, Lawrence. (2006). Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems, 18*, 1473.

Weissenbacher, A., Kasess, C., Gerstl, F., Lanzenberger, R., Moser, E., & Windischberger, C. (2009). Correlations and anticorrelations in resting-state functional connectivity MRI: a quantitative comparison of preprocessing strategies. *Neuroimage, 47*(4), 1408-1416. doi: 10.1016/j.neuroimage.2009.05.005

Weissman-Fogel, Irit, Moayedi, Massieh, Taylor, Keri S, Pope, Geoff, & Davis, Karen D. (2010). Cognitive and default-mode resting state networks: Do male and female brains "rest" differently? *Hum Brain Mapp, 31*(11), 1713-1726.

Wen, W., & Sachdev, P. (2004). The topography of white matter hyperintensities on brain MRI in healthy 60- to 64-year-old individuals. *Neuroimage, 22*(1), 144-154. doi: 10.1016/j.neuroimage.2003.12.027

Whitfield-Gabrieli, S., & Nieto-Castanon, A. (2012). Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect, 2*(3), 125-141. doi: 10.1089/brain.2012.0073

Whyte, E. (2009). Effectiveness of Nimodipine Plus Antidepressant Medication in Treating Vascular Depression.

Wild, B., Herzog, W., Schellberg, D., Lechner, S., Niehoff, D., Brenner, H., . . . Raum, E. (2012). Association between the prevalence of depression and age in a large representative German sample of people aged 53 to 80 years. *Int J Geriatr Psychiatry, 27*(4), 375-381. doi: 10.1002/gps.2728

Wilkins, C. H., Mathews, J., & Sheline, Y. I. (2009). Late life depression with cognitive impairment: evaluation and treatment. *Clin Interv Aging, 4*, 51-57.

Witt, S. T., Laird, A. R., & Meyerand, M. E. (2008). Functional neuroimaging correlates of finger-tapping task variations: an ALE meta-analysis. *Neuroimage, 42*(1), 343-356. doi: 10.1016/j.neuroimage.2008.04.025

Wu, M., Andreescu, C., Butters, M. A., Tamburo, R., Reynolds, C. F., 3rd, & Aizenstein, H. (2011). Default-mode network connectivity and white matter burden in late-life depression. *Psychiatry Res, 194*(1), 39-46. doi: 10.1016/j.pscychresns.2011.04.003

Wu, M., Rosano, C., Butters, M., Whyte, E., Nable, M., Crooks, R., . . . Aizenstein, H. J. (2006). A fully automated method for quantifying and localizing white matter hyperintensities on MR images. *Psychiatry Res, 148*(2-3), 133-142. doi: 10.1016/j.pscychresns.2006.09.003

Wu, Z., Schimmele, C. M., & Chappell, N. L. (2012). Aging and late-life depression. *J Aging Health, 24*(1), 3-28. doi: 10.1177/0898264311422599

Xu, Chen, Peng, ZhiMing, & Jing, WenFeng. (2013). Sparse kernel logistic regression based on L 1/2 regularization. *Science China Information Sciences, 56*(4), 1-16.

Xue, Jing-Hao, & Titterington, D Michael. (2008). Comment on "on discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes". *Neural processing letters, 28*(3), 169-187.

Yoo, T. S., Ackerman, M. J., Lorensen, W. E., Schroeder, W., Chalana, V., Aylward, S., . . . Whitaker, R. (2002). Engineering and algorithm design for an image processing Api: a technical report on ITK--the Insight Toolkit. *Stud Health Technol Inform, 85*, 586-592.

Yu, Lean, Lai, Kin Keung, Wang, Shouyang, & Huang, Wei. (2006). A bias-variance-complexity trade-off framework for complex system modeling *Computational Science and Its Applications-ICCSA 2006* (pp. 518-527): Springer.

Yuan, Guo-Xun, Chang, Kai-Wei, Hsieh, Cho-Jui, & Lin, Chih-Jen. (2010). A Comparison of Optimization Methods and Software for Large-scale L1-regularized Linear Classification. *J. Mach. Learn. Res., 11*, 3183-3234.

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage, 31*(3), 1116-1128. doi: Doi 10.1016/J.Neuroimage.2006.01.015

Zeng, L. L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., . . . Hu, D. (2012). Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. *Brain, 135*(Pt 5), 1498-1507. doi: 10.1093/brain/aws059

Zetin, Mark, Hoepner, Cara T, & Kurth, Jennifer. (2010). *Challenging Depression: The Go-To Guide for Clinicians and Patients (Go-To Guides for Mental Health)*: WW Norton & Company.

Zhao, Peng, & Yu, Bin. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research, 7*, 2541-2563.

Zhou, Zhi-Hua, & Li, Ming. (2005). *Semi-Supervised Regression with Co-Training.* Paper presented at the IJCAI.

Zhu, Xiaojin. (2006). Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison, 2*, 3.

Zikic, D., Glocker, B., Kutter, O., Groher, M., Komodakis, N., Kamen, A., . . . Navab, N. (2010). Linear intensity-based image registration by Markov random fields and discrete optimization. *Med Image Anal, 14*(4), 550-562. doi: 10.1016/j.media.2010.04.003

Zitova, Barbara, & Flusser, Jan. (2003). Image registration methods: a survey. *Image and vision computing, 21*(11), 977-1000.