EARLY PREDICTION OF LATE-LIFE DEPRESSION REMISSION: MULTI-FACTOR KERNEL-BASED MACHINE LEARNING UTILIZING SINGLE DOSE PHARMACOLOGICAL FUNCTIONAL MAGNETIC RESONANCE IMAGING

by

Helmet Talib Karim

B.Sc. in Theoretical Mathematics and Biology, University of Pittsburgh, 2012

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

2017

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Helmet T. Karim

It was defended on

May 26, 2017

and approved by

Carmen Andreescu, MD, Assistant Professor, Department of Psychiatry

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

John Galeotti, PhD, Adjunct Assistant Professor, Department of Bioengineering

Dissertation Director: Howard J. Aizenstein, MD, PhD, Professor, Departments of Psychiatry and Bioengineering

EARLY PREDICTION OF LATE-LIFE DEPRESSION REMISSION: MULTI-FACTOR KERNEL-BASED MACHINE LEARNING UTILIZING SINGLE DOSE PHARMACOLOGICAL FUNCTIONAL MAGNETIC RESONANCE IMAGING

Helmet T. Karim, Ph.D.

University of Pittsburgh, 2017

Treatment of major depressive disorder (MDD) currently relies on a prolonged trial and error process to identify the best pharmacological regimen. This process is further prolonged in older adults with major depressive disorder (Late-Life Depression or LLD), where it is associated with a host of negative outcomes, including suicide, worsening medical comorbidity, and poor quality of life. Functional magnetic resonance imaging (fMRI) brain changes have been associated with depression severity and treatment outcomes. Previous studies have shown that recovery from depression can be predicted using both pre-treatment neuroimaging as well as follow-up scans from the early treatment period. Pharmacological functional magnetic resonance imaging (phMRI) is an approach that utilizes multiple fMRI scans to investigate changes in functional neuroimaging following acute doses of pharmacotherapy. It has been demonstrated that antidepressants have a fast uptake period, effecting resting state networks as well as functional brain activation after only a single dose. We aimed to evaluate the efficacy of phMRI to identify these very early (single dose) functional changes, and use these to predict remission. Data was collected from an open-label pharmacologic treatment study of LLD (N=51). Multi-modal MRI, including phMRI, were acquired at 5 time-points. Results showed accurate prediction of depression remission from pre-treatment, as well as phMRI after only a single dose of pharmacotherapy. The trajectory of the neuroimaging changes across the treatment trial suggest an initial engagement of large scale resting networks, followed by engagement of implicit emotion control networks, and later changes in explicit emotion regulation. Utilizing kernel-

based (multi-factor principal components) machine learning, we found that leveraging both pharmacological neuroimaging and clinical data improved prediction efficacy of remission. In this body of work, we have integrated multiple imaging modalities to explain the long delay in clinical response to antidepressants, and to identify early markers of response.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Many people have contributed to the successful completion of this dissertation. My advisor, Dr. Howard Aizenstein, provided me with so much support, guidance, and wisdom. I am also grateful to my committee (Dr. Carmen Andreescu, Dr. George Stetten, and Dr. John Galeotti) for the support and advice they gave throughout this project, especially Dr. Carmen Andreescu for being so heavily involved in my projects and work. Howard and Carmen provided me with invaluable advice and wisdom, I am also grateful for their friendship – it is rare to have such great advisors and even more rare to have such great friends.

I would also like to thank all the lab members of the Geriatric Psychiatry Neuroimaging (GPN) lab for all of their support and help. They have made the time in the lab unforgettable.

Finally, I would like to express my gratitude to my friends and family for their love, unfailing encouragement, and support.

1.0 INTRODUCTION AND SPECIFIC AIMS

Major depressive disorder (MDD) is a complex neuropsychological disorder that has a second peak of incidence in late-life, known as late-life depression (LLD). LLD carries additional risk of suicide, worsening comorbidity, and care-giving burden (Katon *et al*, 2010; Mulsant *et al*, 2006; Nelson *et al*, 2013). While research has made significant strides in our understanding of depression and its treatment, its translation has lagged severely as currently there are not any accepted neural or genetic biomarkers to aid in the diagnosis, treatment, or management. This may be the result of the complex nature of the underlying etiology and pathophysiology (high heterogeneity) as well as the complexity of the available treatments. Currently, clinicians work to find an effective regimen (of antidepressants) or other treatment option using a prolonged trial and error process that delays overall improvement, increases risk of suicide, and may lead to patient dropping from care (Andreescu and Reynolds, 2011; Reynolds *et al*, 2006). In LLD, approximately 6-8 weeks are needed to identify whether the current regimen is effective (Patel *et al*, 2017), and if it is not then they will be tapered off and started on a new regimen. This period worsens risk of suicide especially in LLD (Katon *et al*, 2010; Mulsant *et al*, 2006; Nelson *et al*, 2013), thus it is critical to find early treatment biomarkers. Previous work suggests that functional magnetic resonance imaging (fMRI) may be a potentially useful tool in finding such markers and by utilizing machine learning methods we may further improve this search.

While markers pre-treatment are often an important predictor of overall improvement (remission), they may not be sufficient. Recent research has found that following single doses of antidepressants there are significant changes in brain activation and connectivity (Bruhl *et al*, 2010; Loubinoux *et al*, 2002; Miskowiak *et al*, 2007; Murphy *et al*, 2009; Rawlings *et al*, 2010; Schaefer *et al*, 2014). These changes may be an important clinical predictor of remission. This may reflect an early engagement of functional networks but may have a significant latency period to translate to behavioral changes. This is further supported in studies using positron emission tomography (PET) that have found that antidepressant uptake occurs acutely (Meyer *et al*, 2001; Parsey *et al*, 2006). We collected fMRI data in an LLD sample during and after a full trial of pharmacotherapy. The fMRI data was collected pre-treatment and post-treatment, but also following acute periods (after a single dose and following a week). We investigated changes in functional brain connectivity and brain activation (during an emotion reactivity task) and whether any acute changes occurred and critically whether machine-learning approaches could be applied to predict remission (using only acute data).

Thus, the goal of this dissertation is to:

- **Aim 1:** Investigate changes in brain activation and connectivity in LLD. Determine acute and chronic changes in the brain.
    - **Hypothesis 1a:** We hypothesize that there will be acute and chronic changes in resting state connectivity, specifically increased executive control network connectivity paired with decreased default mode network connectivity.
    - **Hypothesis 1b:** Functional activation during an emotional reactivity task will show differentially acute and chronic changes in activation.

- **Aim 2:** Develop and test machine-learning models that utilize the multi-modal nature of the neuroimaging data that leverages clinical measures.
    - **Hypothesis 2a:** Certain features will be more predictive of remission, mainly that structural markers may not be great markers of state (since this is expected to change rapidly). Further, some pre-treatment markers will work well to predict remission, however acute changes may act as better markers since they reflect the acute pharmacological change.
    - **Hypothesis 2b:** Leveraging the multi-modal nature of neuroimaging data and clinical features will provide the best predictive capability.

2.0 MAGNETIC RESONANCE IMAGING (MRI)

This chapter is intended to introduce the basic concepts of magnetic resonance imaging (MRI). It describes the hardware components, basic MR physics, and generation of the MR signal as well as details the different imaging modalities: structural imaging, diffusion weighted sequences, arterial spin labeling, and functional imaging. These sequences are utilized throughout to extract essential structural and functional measures. We also describe functional sequences used as well as functional tasks used throughout the study this dissertation utilized.

2.1 MRI: NON-INVASIVE IMAGING OF THE BRAIN

MRI is a non-invasive technique that has become one of the most commonly used methods to study the human body and brain. It is safe, non-invasive, and non-ionizing (no radiation or harmful contrast agents needed) and thus can be utilized heavily to study the structure and function of the human brain with low risk (Kanal *et al*, 2002). The technique utilizes several components to generate a signal based on the spin of protons in the body and has a wide range of imaging modalities that can be generated, including: structural (images that enhance gray/white matter contrast), diffusion weighted imaging (study of white matter tracts), arterial spin labeling (measuring perfusion), and functional activation (blood oxygen-level dependent response, BOLD).

2.2 MR SCANNER, PHYSICS, AND IMAGE ACQUISITION

2.2.1 MR Components

There are several critical components to the MR scanner: the magnet, gradient coils, and the (transmitter/receiver) radiofrequency (RF) coil. The magnet is used to generate a uniform magnetic field and is typically in the range of 0.5 Tesla (T or 5,000 Gauss) to 7T (for reference, earth's magnetic field is ~0.5 Gauss). This aligns protons in the same direction as the magnetic field (i.e. this is the lowest energy state). The RF coil is used to force the protons into a high-energy state and as they return to the lower energy state (equilibrium), they release RF energy that can be picked up the RF receiver. The gradient coils are used to localize the signal in three dimensions.

2.2.1 MR Physics and Signal

MRI takes advantage of precessing water in the body and their magnetic properties (mainly differences in relaxation times). Without the presence of a uniform magnetic field, water precesses randomly (direction and frequency) in the body with no net magnetic moment. However, a superconducting magnet can be used to generate a homogenous magnetic field that aligns the protons' precession creating a net magnetic moment. The number of protons that align in that direction is proportional to the strength of the scanner, thus a stronger magnet is able to generate a stronger signal.

The aligned protons are in a low-energy state that is then disrupted using a RF pulse that energizes them into a high-energy state (reverse direction of the net magnetic field and in-phase).

Protons then seek to return to an equilibrium (low-energy) state and relax. This relaxation is the basis of the MR signal, as they return they release RF energy that is then measured by an RF receiver coil. There are two types of relaxation: T1 and T2 relaxation. T1 (spin-lattice) relaxation is the recovery of the net magnetization in the direction of the net magnetic field. T2 (spin-spin) relaxation is due to the de-phasing of the protons as they precess out of phase. Critically, different tissues have differences in the T1 and T2 relaxation, thus allowing for a natural contrast between different tissue types.

2.2.1 Image Acquisition and Parameters

An MR image is usually a 3-dimensional (3D) image that is generated by collecting multiple 2D images that contain a value representing intensity at each volumetric element (voxel or 3D pixels). In neuroimaging, the individual's head is placed in the scanner and RF coil. Typically, the Z-direction is along the direction of the scanner (head to toe) and the X-/Y- is left/right and bottom/top of the scanner, respectively. The following sections describe the general linear procedure that generates an image.

After, the magnet applies a net magnetic field (B_0) that aligns protons along the Z-direction (where they continue to precess); slice encoding is performed to choose which "slice" of the brain is going to be imaged. The Gz gradient coil applies a gradient magnetic field that causes protons to precess at different frequencies along the Z-direction. The transmitter RF coil applies an RF pulse at a pre-defined frequency that generates a magnetic field (B_1) that excites only protons with the same frequency (thus choosing the slice to excite). These protons are now aligned along the B_1 field. The Gz gradient coil is then turned off.

Phase encoding is done to encode the location in the Y-direction by turning on the Gy gradient coil that applies a gradient magnetic field in the Y-direction. This causes protons to along the Y-direction to precess at differing frequencies and once the Gy gradient coil is turned off then the protons precess at the same frequencies however they are now out of phase. Similarly, this allows protons along the Y-direction to be precessing with different phases, which encodes the information along the Y-direction.

The Gx gradient is then turned on to apply a gradient in the X-direction causing protons along that direction to precess at different frequencies. At this point, the RF receiver coil is used to read the emitted RF energy of the relaxing protons. This signal however is complicated and is composed of differing frequencies, phases, and amplitudes. This process of slice, phase, and frequency encoding are conducted at different amplitudes of phase encoding gradients.

This data is best represented in a k-space image that represents the frequency and phase information for each slice (where the center represents zero frequency and phase). The most common way to fill this k-space image is known as echo-planar imaging (EPI) due to its speed where multiple phase encoding gradients are applied consecutively with the RF receiver coil collecting data after each consecutive excitation. Each 2D complex k-space image can be reconstructed into a corresponding image (from frequency domain into original space domain) using a 2D Fourier transform (which encodes the relationship between the original signal and frequency domain information) (Bracewell, 1989). The multiple 2D images are then stacked to create a single 3D image with intensities that reflect the emitted energy at each voxel.

The procedure involves several key parameters that control the size of the image, resolution, contrast, and modality (e.g., T1- vs. T2-weighted). The number of slices, field of view, and matrix size affect the resolution of the image. The greater the number of slices, the

greater the resolution in the Z-direction, while a greater matrix size and/or field of view has a similar effect in the X-/Y-direction.

Several other parameters affect the contrast of the image and can be used to create different images that weight different properties of the tissues. The repetition time (TR) is the time between each RF pulse. The longer the RF pulse, the greater the time allotted for the T1 signal to relax. Conversely, the echo time (TE) is the time between the RF pulse and data acquisition, which affects the amount of time for T2 relaxation. Thus, different combinations of TR and TE can generate T1-weighted or T2-weighted images. When the RF pulse is applied, the B_1 field is generated at an angle (which affects signal to noise, SNR) from the B_0 field and is called the flip angle (FA).

There are several variations on this image acquisition process that are important to define: spin echo, gradient echo, and inversion recovery sequences. Spin echo involves applying an additional re-phasing RF pulse between the initial pulse and data acquisition. The FA of this excitation depends on the FA of the first (180^o if initially 90^o) and causes the de-phasing protons to re-phase that improves the signal and compensates for local field inhomogeneities (from several sources). The gradient echo sequence uses the frequency-encoding gradient to rephase the protons by applying a negative polarity and then a positive polarity (fast precessing protons take longer to rephase than slow precessing protons), which causes a rephasing of the protons and emits a stronger signal. While this sequence is faster than spin echo, it does not compensate for local inhomogeneities leading to greater artifacts in the image. The inversion recovery sequences are exactly the same as the spin echo, however there is an initial RF pulse (180^0) before the sequence starts and allows for a long T1 relaxation period (no T2 relaxation). This allows for greater T1 contrast, however is much longer as the T1 relaxation is doubled.

2.3 STRUCTURAL NEUROIMAGING

Utilizing differences in T1- and T2-relaxation between different tissues, we can generate T1-weighted vs. T2-weighted images. Short TR sequences allow for greater tissue contrast in T1-weighted images, while longer TR (longer T1 relaxation) measures the number of protons (proton density). Consequently, short TE (short T2 relaxation) does not allow for sufficient dephasing of protons thus tissues have similar intensities compared to longer TE that allow for greater differences between tissue intensities. T1-weighted images have short TR and TE and typically have the greatest tissue contrast. T2-weighted images have long TR and TE and allow for greater contrast between CSF and other brain tissue, but lower gray/white matter contrast. Consequently, T1-weighted images are used to study structure of the brain, while T2-weighted images are utilized for studying pathology. The magnetization prepared rapid gradient echo (MPRAGE) is a structural T1-weighted sequence that is heavily used due to the rapid acquisition (Mugler and Brookeman, 1990). These images are typically used to segment the gray matter and determine gray matter density within a voxel.

The fluid attenuated inversion recovery (FLAIR) is a T2-weighted image that improves the visualization of age-related tissue lesions, specifically white matter hyperintensities (WMH) (Hajnal *et al*, 1992). The sequence is used specifically as it suppresses signal from the CSF (appearing dark). WMH appear white (hyperintense) and can be used to segment these lesions more accurately. This can be used to measure global WMH burden.

2.4 DIFFUSION WEIGHTED IMAGING

Diffusion weighted imaging is used to track the diffusion of water within the body specifically the white matter (Bammer, 2003). This is possible as free water diffuses randomly in most spaces, however within the white matter tracts it is constrained to diffuse along the tract, which can be detected and measured. Structural damage and pathological changes may alter the diffusion of water within the white matter but also the diffusion in the gray matter.

A pulsed gradient spin echo sequence is used to acquire diffusion-weighted images. This is very similar to the gradient spin echo sequence, however two gradient pulses with a defined direction are placed on both sides of the rephasing pulse. If water molecules do not move, then the phases induced by the two gradient pulses would cancel out resulting in greater signal. Due to this, greater diffusion would attenuate the signal. Usually multiple images of multiple directions are acquired, where a greater number of directions can resolve greater diffusion information. Using this acquired data, it is possible to identify the mean diffusivity and primary direction of tracts and can be used to measure tract integrity.

2.5 ARTERIAL SPIN LABELING (ASL)

Arterial spin labeling involves the labeling (or tagging) of proximally flowing blood, which after a transit time flows into a slice or area of interest and is paramagnetically labeled (Detre *et al*, 1992). This inflowing tagged blood alters tissue magnetization, reducing the overall signal. An unlabeled (control) image is collected as well, where the difference in the unlabeled and labeled

image is the cerebral blood flow/perfusion of blood in that region. This requires no contrast as the labeled blood acts like a contrast agent altering local tissue magnetization.

Pulsed continuous ASL (pCASL) is the currently most preferred method to acquire perfusion imaging (Dai *et al*, 2008). A train of short RF pulses is used to first invert the blood proximally, followed by a full brain acquisition. The refocusing portion of the slice selection gradient (Gz) is not balanced, which causes an accumulation of additional phase after each RF pulse. The control sequence, however, uses a balanced gradient that removes this effect. Typically, tagged and untagged images are collected one after another. This process is repeated over an extended period of time (at rest) and generates a 3D perfusion image across time (4D) that is typically averaged to measure mean perfusion.

2.6 FUNCTIONAL MRI (FMRI)

This section introduces the blood oxygen level dependent (BOLD) response and how this is measured using fMRI. Similar to ASL, a 4D image is generated that measures activation across time – however it is divided into task-based and resting state fMRI.

2.6.1 Blood Oxygen Level Dependent (BOLD) Response

The BOLD response is measured using fMRI, which utilizes T2*-weighted imaging (Ogawa and Lee, 1990a; Ogawa *et al*, 1990b). The BOLD response is thought to be an indirect measure of neuronal activity. When an individual performs a task (e.g., tapping their right finger), neuronal activity within the left motor cortex increases and due to increased activation there is a change in

metabolism as well as an increase in deoxygenated blood in the venous capillaries (Attwell and Iadecola, 2002). This is typically followed by an increase in local cerebral blood flow (and possibly volume) that results in an overall increase in oxygenated hemoglobin (venous) and decrease in deoxygenated hemoglobin (known as 'washout') (Attwell *et al*, 2002). While oxygenated hemoglobin is diamagnetic, deoxygenated hemoglobin is paramagnetic, which creates magnetic field distortions (altering local magnetic susceptibility) thus reducing the MR signal. Specifically, the T2*-weighted signal is a combination of the T2-signal and inhomogeneities (hence T2*). As the deoxygenated hemoglobin decreases, the MR signal increases in amplitude. Thus the BOLD signal is an *indirect* and delayed measure of neuronal activity. Usually this delay is approximately 6-8 seconds, thus we can construct a canonical hemodynamic response function (HRF). Multiple of these images are acquired while participants perform tasks. Typically a task of interest and a control are performed (e.g. viewing faces vs. viewing houses).

By designing certain tasks (as well as sufficient control tasks), we can understand and map functional activation of the brain. This process has been used heavily in the past decade to map the human brain and more importantly understand how these are altered in certain disease states (e.g. depression). However, even without performing a task – the brain has intrinsic changes in brain activation. It is now well-known that this intrinsic activation can be used to better understand the connectivity of the brain (at rest) and is essential for understanding depression as it is thought that resting state activation represents introspective functioning. The next two sections describe in greater detail resting state and task-based designs (specifically the tasks used in this study).

2.6.2 Intrinsic Resting State Activation

Even without an explicit task, the brain is intrinsically active and this spontaneously fluctuating signal can be measured using fMRI. Usually participants are asked to lie awake in the scanner viewing a fixed object (e.g., a cross-hair to prevent sleep). This has been used to measure the intrinsic connectivity of brain regions as well as define stable networks (Fox *et al*, 2005). This is especially important in MDD, as it is thought that a large change in the intrinsic activation in the brain is altered as the majority of intrinsic activation is thought to reflect an introspective aspect of human thought. MDD is associated with an alteration in this introspective thought process as it becomes ruminative in nature leading to changes in resting state activation (Hamilton *et al*, 2011; Sheline *et al*, 2010b; Zhu *et al*, 2012). The data utilized acquired a resting state scan where participants were instructed to lie awake in the scanner viewing a white cross hair (black background) for five minutes.

2.6.3 Task-Based Activation

Task-based fMRI designs typically involve a task of interest and a control task (to control for unrelated effects). In this section we describe several of the tasks used throughout this study: emotion reactivity task (face/shapes), explicit emotion regulation task, and a memory-encoding task.

Emotion Reactivity Task (Face/Shapes): During functional scanning participants performed the face/shapes emotion reactivity task to study the effect of emotional reactivity to faces (Hariri *et al*, 2003). Participants were instructed to match either a face cue or a shapes cue. A cue was shown on the center of the screen and they were instructed to respond with an MR-

compatible glove (left or right index finger) by matching to one of two simultaneously presented faces. The facial expressions shown were either angry or fearful. During the shapes, they matched a shape to one of two simultaneously presented shapes. The shapes task (5 blocks) was interleaved with the faces task (4 blocks) and each block lasted 24 seconds containing 6 trials (4 seconds each). Before the beginning of each block participants are instructed visually to "match emotion" or "match form" (2 seconds). The faces images are presented from a set 12 different images (six per block, three of each gender) and are all derived from a standard set of pictures of facial affect. Stimulus presentation and responses were controlled using E-prime software (Psychology Software Tools, Inc., Pittsburgh). This task has been shown to robustly activate the human amygdala (generates emotional responses as part of the limbic system) even in healthy participants (Hariri *et al*, 2003). Critically, this is thought to have an implicit regulation aspect, where participants regulate their emotional responses automatically.

Explicit Emotion Regulation Task: Participants were shown emotionally neutral or negative images from the standardized International Affective Picture System (IAPS) and were instructed to either "Look" or "Decrease" (Lang *et al*, 2008). During the look instruction, participants were to view content naturally. During the decrease instruction, participants were instructed to reappraise the image to actively alter the elicited emotion. Reappraisal involves actively altering emotional responses to a viewed image (e.g., a picture of someone in the hospital may be reappraised as someone who is getting life-saving treatment). Participants were sufficiently trained in this process prior to scanning. After each image they were asked to rate how negatively they felt from 1 to 5. The neutral (11 events), negative (15 events), negative regulate (15 events) conditions were interleaved and each event lasted 6 seconds. The stimulus presentation and responses were controlled using E-prime software (Psychology Software Tools,

Inc., Pittsburgh). This task measures their ability to explicitly regulate their emotions and significantly activates explicit areas implicated in regulation (dorsolateral prefrontal cortex) (Gyurak *et al*, 2011; Ochsner *et al*, 2002).

Memory Encoding Task: This task has been used widely to measure memory encoding and significantly activates the hippocampus (Sperling *et al*, 2001). Participants first encoded two face-name pairs outside of the scanner (i.e., two faces were presented each with a name). Participants are shown face-name pairs in a block design. During the control blocks, the two already encoded face-name pairs are shown. During the novel blocks, new pairs of face-names are presented. They are instructed during both blocks to state whether the name "fits" and are told that this helps with encoding the face-name pairs. They are told that they will have to recall the name that was presented with each face at the end of the scan – thus they are instructed to encode these faces into memory. Each run contains four blocks alternating between the control and novel blocks (48-seconds each). Each block presents 8 faces for five seconds each (1 second fixation in between) that results in 32 novel and 2 familiar faces. Between each block is a 25 second fixation. Stimulus presentation and responses were controlled using E-prime software (Psychology Software Tools, Inc., Pittsburgh).

3.0 PROCESSING NEUROIMAGING DATA

This chapter introduces the basic concepts of processing neuroimaging data as well as details regarding spatial pre-processing, coregistration and motion correction, as well as segmentation and normalization (coregistration to a standard anatomical space). It also describes how to compute outputs that are relevant to this dissertation including: gray matter density (structural), white matter hyperintensity (WMH) burden, fractional anisotropy and mean diffusivity (diffusion weighted imaging), perfusion (arterial spin labeling, ASL), eigenvector centrality (resting state functional MRI), and task-based fMRI activation.

3.1 PRE-PROCESSING NEUROIMAGING DATA

There are several key pre-processing methods that are essential for processing neuroimaging data. Slice time correction of functional data involves accounting for differences in the timing of slice acquisition during data collection, because during the functional sequence the first slice collected is an entire TR away from the last slice collected. Coregistration is the process of aligning two images together and can refer to registering similar images within a sequence (typically called realignment or motion correction in fMRI), or images within a subject of different modalities (e.g., MPRAGE and FLAIR), or even images between several subjects or a template space (usually referred to as normalization). Segmentation involves the labeling of

neuroimaging data into separate tissue classes (e.g., gray vs. white matter). Smoothing is a process that blurs the image but reduces spurious noise. We describe these processes in greater detail in the following sections.

3.1.1 Slice-Time Correction

Functional MRI scans usually have multiple 2D images (slices) collected over a short period (a single TR) to generate a single 3D image that is collected over a certain period of time. If the TR is long, then the collection of the first slice is approximately one TR away from the last slice. Slice-time correction is a method to resolve this issue following data acquisition (Sladky *et al*, 2011). If the TR is sufficiently short, or if the design of the task is presented in blocks rather than as single events, or if the data is not resting state – then this processing is not necessary. A short TR means that there are small differences in the timing of the acquisition and thus is not necessary in the described scenarios. The precise timing within a long block does not affect the determination of activation.

As we know the order of the collected slices, we can use this information to shift the signal in a certain direction. Critically, this process is performed using Fourier transforms. The Fourier transform represents *any* signal as a linear weighted combination of sinusoids. For each signal, depending on when they were collected – we can shift the sinusoid by adding some constant (dependent on slice order) to the phase of each frequency. This will have an effect of shifting the data in time (Sladky *et al*, 2011).

3.1.2 Coregistration and Motion Correction

Registration is a process where a transformation that spatially aligns two images (usually based on intensity): the reference (image that does not move) and the source (image that is transformed). The intensities of the images are used to help align the two images, usually involving several key components: similarity metric, optimizer of the alignment, transformation matrix, and the interpolation method.

The similarity metric is used to minimize the difference between two images' intensities. Thus, the similarity metric is the cost function that is to be minimized. The selection of the cost function is dependent on the image types of the reference and source image. If the reference and source are of a similar image type, then they have similar intensity distributions. Thus, we can compare their intensities directly using measures like least squares (sum of squared differences) or normalized correlation (correlation between each voxel intensity). However, if the images are of different types – then their intensities do *not* directly match. Normalized mutual information is a metric commonly used in neuroimaging methods. Mutual information determines how much uncertainty about the reference image's intensity values is reduced by the knowledge of the source image's intensity values (thus the image intensities do not need to match). This can assume a linear or non-linear relationship and is a robust measure of similarity. In this way, if two images are aligned well – then their intensities match (or are predictive) and its minimization will allow for a local minima that represents the coregistered images.

The optimizer is used to minimize the cost function (similarity metric). This process usually involves several key steps: initialization, iterative parameter optimization search involving assessment and modification, and finally convergence. In the first step, the similarity metric is calculated. The initial parameters are updated (both value and direction of change) and

are done using several possible methods (e.g., gradient descent, Quasi-Newton, Gauss-Newton, Levenberg-Marquardt). The similarity metric is thus calculated again and the process is iterated over several steps. Some convergence criteria determine when the algorithm should stop (e.g., difference of similarity metric between current and previous metric is sufficiently low).

The transformation is the matrix that is applied to the source image to coregister it to the reference image. However, depending on the type of transformation – a different number of degrees of freedom may be allotted. Linear transformations refer to those with lower degrees of freedom allowing for only small alterations. Transformations may have the following linear changes: 3 translations, rotations, and scaling (total 9 degrees of freedom) as well as 3 degrees of freedom for skewing along all three axes (3D space). Some well known combinations include: rigid body (6 translations/rotations but no scaling) and affine (12 degrees of freedom). A simple 4x4 matrix can be used to represent such transformations. Non-linear transformations can have a much larger number of degrees of freedom – thus allowing for even small local changes. Linear transformations are well suited for coregistering images within an individual (as they have similar structural properties, e.g., gray matter folds), however non-linear transformations are better suited for coregistering across individuals with highly varying structural properties.

There are several types of interpolation methods that can be utilized during this process. Nearest-neighbor interpolation involves assigning each new voxel of the coregistered image the intensity value of the spatially closest voxel from the source image prior to the transformation. This is most commonly used when interpolating masks, which are binary images that represent some prior segmentation. B-spline interpolation is another common interpolation method that uses polynomial functions to weight the intensities of neighboring voxels from a large neighborhood of voxels. The degree of the B-spline is a reference to the size of the polynomial

function used. This is one of the most commonly used methods as it has high accuracy with low computation cost relative to other methods. The trade-off between accuracy and computational cost is an important consideration when determining the interpolation method.

Generally, in the processing of neuroimaging data we have several key types of coregistration that are performed: within session motion correction, within subject functional-structural or structural-structural coregistration, or across subject coregistration. Across subject coregistration is a process called *normalization* that we cover in the next section along with segmentation (due to the unified nature of some algorithms).

Within session motion correction refers to coregistration between images collected within a single 4D image (e.g. BOLD or ASL data) (Ashburner *et al*, 1999). The 4D image is a set of 3D images across time – and during this period the participant was likely to move their head. To correct for this we perform multiple coregistrations between each image and one of the images is used as a reference. Typically, the first image is chosen as the reference, the other images are coregistered to the reference, a mean image is computed and is now treated as the reference, then the true coregistration is performed between all images and the mean. This process corrects for the motion within a session and is usually done as a rigid body coregistration process (as the images are similar in type and size). This also outputs a six parameter matrix that represent the motion in each of the six directions that can be used to further remove residual motion.

Within subject coregistration (functional-structural or structural-structural) involves coregistering images that are different types (e.g., MPRAGE and FLAIR), however they are from the same participant thus they do not differ structurally (e.g., gray matter folds) (Ashburner *et al*, 1999). Typically, one of these images is chosen to be the reference and the other is coregistered using an affine transformation.

3.1.3 Unified Segmentation and Normalization

Segmentation is a process in which a set of tissues are identified in an image and classified. In neuroimaging this process usually involves segmenting images into six tissue types: gray matter, white matter, cerebrospinal fluid, soft-tissue, skull, and air. There are many different processes that are used to segment neuroimaging data, however we will focus solely on the unified segmentation and normalization algorithm as it is utilized most throughout the dissertation (Ashburner and Friston, 2005).

The segmentation algorithm utilizes a Gaussian mixture model based approach with tissue priors. The basic concept is that if we consider segmenting gray vs. white matter, then their intensities (depending on the imaging type) could be represented using two Gaussian distributions (bimodal) with one representing gray matter tissue and another representing white matter tissue. Thus, we can fit a Gaussian mixture model and then segment the entire brain into several tissue classes. This considers all the brain voxels as a mixture of multiple Gaussian distributions and attempts to identify them. To further improve that, we can give an initial guess as to the location of these tissues using an average tissue prior (a probability map indicating the likelihood that some tissue is represented at any one voxel) generated from healthy neuroimaging data of a large cohort of participants (e.g., we can give probability maps of where we expect the gray and white matter *should* be) (Penny *et al*, 2011). This involves coregistering the neuroimaging data to a standard anatomic space (with tissue probability maps) and using the probability maps to weight the classification according to Bayes theorem (describing the probability of some occurrence based on knowledge of some other condition).

In the unified segmentation and normalization framework several processes are unified to improve the overall efficacy of each individual process. This process details how to classify

tissues in a structural image and then coregister it to a standard anatomical space. We first coregister (using linear methods) the structural image and a template structural image in a standard anatomic space (Montreal Neurological Institute or MNI space). We then perform an initial segmentation of the voxels into six tissue classes using a Gaussian mixture model weighted by a prior distribution. We can now iterate through this process to improve both and output a final segmentation (each voxel contains an individual probability for each tissue class) as well as a deformation field. The deformation field is a set of cosine bases that map each voxel into MNI space and is utilized to normalize most neuroimaging data by first coregistering an image to the native structural image and then applying the deformation field to warp it into MNI space (Penny *et al*, 2011).

3.1.4 Effects Of White Matter Disease On The Accuracy Of Automated Segmentation

This section describes previously published primary author work that investigated the effects of WMH on the accuracy of the automated segmentation in the statistical parametric mapping (SPM) toolbox (Karim *et al*, 2016b). We also investigated whether performing corrections resulted in any change in the segmentation. WMH are hyperintense regions (on T2-weighted images) that become more prevalent in late-life that are attributed to degenerative changes of long penetrating arteries, resulting in demyelination, gliosis, and axonal degeneration (Ovbiagele and Saver, 2006). It is associated with a wide variety of disorders (including depression) (Aizenstein *et al*, 2011; Sheline *et al*, 2010a; Taylor *et al*, 2003). WMH appear dark (hypointense) on the MPRAGE and typically look like gray matter intensities, thus using standard approaches the WMH are segmented as gray matter instead of white matter. This affects the overall accuracy of both the segmentation and normalization.

Two common methods for performing a correction are WMH filling (Battaglini *et al*, 2012; Eloyan *et al*, 2014) and multi-spectral segmentation (Ashburner *et al*, 2005). WMH filling involves first identifying voxels with WMH then filling them with normal appearing white matter values (NAWM). The WMH segmentation process is described later in section 3.5, but to fill them - all NAWM (white matter voxels that are not classified as WMH) are used to generate a distribution. Then each WMH voxel is filled with a random value from this distribution effectively disguising this region as not hyperintense thus correcting for the WMH (Eloyan *et al*, 2014). However, the main issue is that this is a brute force method that forces WMH to appear like normal white matter.

Multi-spectral segmentation methods instead rely on multiple tissue types (e.g., MPRAGE and FLAIR). The differences in distributions help better classify each tissue class. Consider that four separate tissues on the MPRAGE and FLAIR have differing properties. On the MPRAGE the increasing ordered rank of the mean intensities are: gray matter, WMH, caudate, and NAWM. Thus, WMH appear most like gray matter and caudate (a subcortical gray matter region). On the FLAIR, however, the increasing ordered rank of the mean intensities is: NAWM, gray matter, caudate, WMH. Thus, the properties of the intensities are altered in each image. Leveraging that information improves the Gaussian mixture model and helps better classify the tissues. Critically, we also set the number of Gaussians to fit for white matter to two instead of one, because there is a separate Gaussian distribution associated with the WMH.

We computed intraclass correlation coefficients (ICC) between the original uncorrected segmentation and each of the correction methods segmentations (McGraw and Wong, 1996). We found that both significantly altered the segmentation globally and locally. We found that the multi-spectral segmentation more greatly affected the overall segmentation. Further, it seemed to

more greatly affected the local segmentation/normalization as well – affecting subcortical structures like the caudate and amygdala most. This highlights a need to correct for WMH in studies where they are prevalent. The results from this experiment only suggest that the multi-spectral segmentation seems to more greatly affect the initial inaccurate segmentation – but not which is better. Multi-spectral segmentations utilize the full nature of multi-modal data acquisition in modern neuroimaging studies and seem to be an effective approach to perform correction without having to force WMH to appear like normal white matter.

3.1.5 Smoothing

Smoothing is a process where data is interpolated using some function to reduce the effects of large outliers by essentially blurring the data. Typically, a Gaussian smoothing kernel is used to interpolate the image that is described by its size or the full-width at half-maximum (FWHM). If for instance the FWHM is 8mm then the Gaussian distributions value at 4mm from the center is half the maximum value of the Gaussian distribution (e.g., if the max of the distribution is 1 then the value 4mm away from the center is 0.5). This describes the extent of the Gaussian distribution, and greater FWHM cause greater blurring/smoothing.

There are multiple reasons that this process is typically done, mainly: increased signal to noise ratio, to account for differences in structural anatomy, and finally this becomes critical for statistical group inference. Signal to noise increases because there can be (at the individual and group level) high amounts of noise (especially in ASL and fMRI) and smoothing removes large outliers (or noise). Further, as the structural coregistrations are not perfect there can be high variability between subjects structurally and functionally. Smoothing blurs functional/structural clusters – which increases the overlap between subjects (improving sensitivity). Finally, when

performing group statistical inference – neuroimaging methods rely heavily on Gaussian random field theory (described in detail in section 3.8). Smoothing decreases the number of independent statistical tests (by blurring neighboring voxels), and this increases the overall sensitivity (Mikl *et al*, 2008). Briefly, imagine performing four independent statistical tests in four neighboring voxels that generates four p-values. Bonferonni correction is a method to control for multiple comparisons (joint inference on all four voxels), which states that dividing the acceptable false positive rate by the number of *independent* statistical tests sufficiently controls the false positive rate across multiple tests. Using Bonferonni correction, this means that to control the false positive rate (alpha) at 0.05 then we need to divide by four. However, consider that two voxels are highly correlated (or smoothed) then the number to divide by should actually be three (as only three of them are truly independent). Similarly, smoothing reduces the severity of the multiple comparisons problem (described in greater detail in section 3.8).

3.2 LONGITUDINAL GRAY MATTER DENSITY ESTIMATION

Structural imaging data can be used to estimate gray matter density. This is a common variable of interest in neuroimaging studies as these are often associated with a wide range of disorders like depression severity. Depending on the disease being investigated the healthy brain templates used when segmenting and coregistering the brain may not be good proxies for diseased or aging brains. Thus utilizing study specific templates help improve the overall estimation of gray matter density.

After performing segmentation, a single probability map is output in a standard anatomic space. The probability maps can be coregistered in an iterative process where the mean of all the

probability maps is taken; they are then registered to the mean. The process is repeated, which increases the smoothness of the mean of each subsequent iteration. This generates a new study specific template in the same standard anatomical space that is better suited to the current sample (Ashburner, 2007). Consider that aging populations tend to have greater sized ventricles – however the template has much smaller ventricles affecting the overall normalization process. By creating a template, we circumvent to some degree as now we coregister to a template with larger ventricles. This normalizes each map into a standard anatomical space while relaxing the large deformations needed for certain structures (e.g., the ventricles).

Along with the segmentation, this process outputs what is known as the Jacobian (or the matrix of all first order partial derivatives of the deformation field, i.e., the gradient or local changes in deformation) (Ashburner and Friston, 2000). Without the Jacobian, the current segmentations represent the probability at each voxel that it is gray matter – however because the local tissues have been warped, the probabilities do not truly represent the *amount* of gray matter. By multiplying by the Jacobian, we can compute the gray matter density (Ashburner *et al*, 2000). Consider, a region that has to be shrunk to fit as part of the template – it then follows that we are forcing a certain amount of gray matter into a smaller region, thus it should have greater gray matter density. Conversely, a region that is increased to fit onto the template should have lower gray matter density.

We can improve this process even further longitudinally where we have multiple structural images for each participant across time. Similar to the previous process, we can generate a subject specific template prior to creating a study specific template. The subject specific template uses each of the tissue probability maps from each time point and creates a subject specific template. This subject specific template is then used to create a study specific

template. This further improves the overall estimation of gray matter density. Thus we compute a single gray matter density map for each time point.

To further improve the initial segmentation, a multi-spectral segmentation that utilizes multiple spectra/image types (e.g., MPRAGE and FLAIR) can be used (Ashburner *et al*, 2005). The Gaussian mixture model can thus consider two sets of distributions when trying to fit the model. The FLAIR significantly improves the classification of WMH as white matter (Karim *et al*, 2016b).

3.3 WMH SEGMENTATION

WMH burden (amount of WMH in the brain) has been shown to be associated with a wide range of neurological disorders and is thought to be a good marker of cerebrovascular disease (Alexopoulos *et al*, 1997; Sheline *et al*, 2010a). We describe the vascular depression hypothesis in a later section, however it states that WMH may be a driving factor late-onset depression due decreased cerebral blood flow and impaired cognitive function (Alexopoulos *et al*, 1997). Thus, segmenting and quantifying WMH burden is important for understand late-life depression.

Segmentation of WMH has been performed using a wide variety of methods – however one effective method uses the FLAIR image and an automated approach to select seeds and then grow them using fuzzy connectedness (Wu *et al*, 2006). The images are first intensity normalized, by calculating the mean and standard deviation of the cerebellum white matter (which is relatively devoid of WMH) and calculating a Z-score of the whole brain using that mean and standard deviation. We then choose voxels that are 3.5 standard deviations (chosen based on previous data at improving the segmentation) above the mean as seeds. These seeds are

then grown using a fuzzy connectedness algorithm (on the original data and not the Z-score data) (Wu *et al,* 2006). In this region-growing algorithm, the fuzzy adjacency and affinity are calculated between a seed and all voxels, which measure how strongly the seed and each corresponding voxel associate in space and intensity, respectively. This image is then threshold to generate a fuzzy segmentation for that seed – which is iterated through each seed and then combined across seeds. This generates a single WMH segmentation. We can compute the volume of the WMH (number of voxels x resolution in mm^3) and then divide by intracranial volume (ICV) to normalize the measure as a percentage of ICV (accounts for differences in head size).

3.4 DIFFUSION TENSOR IMAGING: FA AND MD

Diffusion weighted images are 4D and contain b_0 images (no diffusion) as well as diffusion images in different directions. Using this data we can compute fractional anisotropy and mean diffusivity, which are important neural correlates. Diffusion data is first eddy corrected – misalignment of the images due to the presence of eddy currents in the scanner. This is typically resolved using image registration methods. We then perform multiple linear regression to calculate diffusion tensor components from the set of diffusion images with differing directions/ orientations with respect to water diffusion. A tensor is a three by three matrix representing the diffusion of water in all three directions (3D) in a single voxel. After diagonalization of each corresponding matrix, we compute the eigenvalues and eigenvectors (the primary, secondary, and tertiary diffusion directions) (Behrens *et al*, 2003).

Two main measures are utilized commonly (although others exist): fractional anisotropy (FA) and mean diffusivity (MD) (Behrens *et al*, 2003). Consider the three eigenvalues and the mean eigenvalue then FA is computed as the square root of the sum of the squared difference between eigenvalues and the mean eigenvalue divided by the square root of the sum of squares of each eigenvalue multiplied by square root of 1.5. The higher FA translates to greater anisotropy or a more "oblong" shaped ellipsoid, and thus an FA of zero is perfectly spherical. This indicates the primary direction of diffusion. The MD is simply the mean of the eigenvalues, which represents the total diffusion within a voxel. FA is thought to be most sensitive to microstructural integrity in the white matter. MD is thought to be an inverse measure of membrane density sensitive to cellularity, edema, and necrosis. Thus, we compute a single voxel-wise FA and MD map. These maps can be coregistered to the structural image and then warped to MNI space using the deformation field.

3.5 ASL: PERFUSION

ASL data is first motion corrected and is performed as a two-stage process where tagged and untagged are motion corrected separately and then together. The images are then subsequently smoothed to improve estimation of perfusion. The perfusion is a measure of cerebral blood flow (CBF) normalized by volume of the each voxel (Detre *et al*, 1992). To calculate CBF we used an equation that is dependent on the following: difference between tagged/untagged images (deltaM), the blood/tissue water partition coefficient (lambda), longitudinal relaxation rate of blood (R), tagging efficiency (alpha), equilibrium magnetization of the brain (M_o usually

calculated from white matter), post-labeling delay (w), and duration of the labeling RF pulse train (tau). The following equation (Wang *et al*, 2008b) describes this relationship:

$$\text{Equation 1. } CBF = \frac{\delta M * \lambda * R * e^{w*R}}{2 * M_o * \alpha} * \frac{1}{1 - e^{-\tau * R}}.$$

This measure represents CBF at each voxel and can be used to estimate the perfusion. Typically, this is only calculated in voxels inside the brain. This measure is highly correlated with measures of metabolism in positron emission tomography as well as CBF as measured using actual contrast agents (gold standard for CBF measurements) (Chen *et al*, 2011). These maps can be coregistered to the structural image and then warped to MNI space using the deformation field.

3.6 RESTING STATE FMRI: EIGENVECTOR CENTRALITY

Resting state involves several stages of processing: slice-timing correction, motion correction, normalization to MNI space, smoothing, wavelet despiking, covariate regression and band-pass filtering, and then estimation of eigenvector centrality (Whitfield-Gabrieli and Nieto-Castanon, 2012). After performing slice-timing correction, motion correction, normalization, and smoothing (described in previous sections) the data are despiked, which removes large motion artifacts. The wavelet-despiking algorithm used identifies non-stationarity events across multiple frequencies using the following generalized steps: time-series decomposition into wavelet domain, identification of non-stationarity events, removal of those events, and reconstruction to the time domain (Patel *et al*, 2014). Spike artifacts are prioritized, as they are most likely due to motion (Patel *et al*, 2014).

In the next stage we removed several covariates that may corrupt our true resting state fluctuations, including: residual motion artifacts, low frequency noise, high frequency noise above a certain frequency, and global signal from the white matter and cerebrospinal fluid (Behzadi *et al*, 2007; Whitfield-Gabrieli *et al*, 2012). This is conducted by performing a mass-univariate regression between our observed data and a set of covariates. We then subsequently can analyze the residuals, which are essentially not 'corrupted' by these factors. Removing the motion parameters removes the effect of any residual motion not accounted for in the motion correction. Band-pass filtering removes non-resting state fluctuations, as it has been shown that resting state fluctuations seem to occur at a certain frequency band and other frequencies 'corrupt' this resting state signal. Finally, some regions are 'corrupted' by white matter of CSF signal, thus we can remove canonical signals from these tissues to account for their effects.

To remove motion artifacts we use the six motion parameters from the motion correction stage. As the data is discrete and has a low sampling rate (high TR, usually around 2 seconds), we can use a set of cosines of varying frequency to remove low and high frequencies that are of no interest. Mainly, we remove frequencies not in the band 0.008 to 0.15 Hz (Whitfield-Gabrieli *et al*, 2012). Thus, we can generate a set of cosines for the discrete valued signal that represent signals with frequencies above and below these bands and regress them against the observed data.

Previous studies included only two covariates: one for the mean white matter signal and one for the mean CSF signal, however current techniques utilize multiple signals to account for a larger proportion of the variance in signals in the each of these tissues. Segmentations of the white matter and CSF can be used to determine where to extract a matrix of time-series. Principal components analysis (PCA) can be conducted on these signals to generate several

principal (typically 5 components) time-series that represent a proportion of the variance within the matrix. Briefly, it is a method used to estimate a set of bases that are orthogonal (not correlated) and explain the variance in the data (i.e., it is a low dimensional feature space that the original data can be represented in). The observed matrix ('t' time points by 'v' voxels) can be represented as a set of scores (original data in the low dimensional space that is 't' by 'c' components) times as set of loadings ('c' by 'v'). The first five scores could then be used to represent the signals in the white matter and CSF and are regressed out.

After regressing out six motion parameters, a set of cosines that represents frequencies of no interest, five components from the white matter and CSF combined, as well as the mean of the time series, we can compute eigenvector centrality. Centrality is a graph theoretical measure that represents how important a voxel as a node, where higher centrality represents voxels that have greater connectedness (Wink *et al*, 2012). One way to compute centrality at a voxel is to correlate that voxel's time series with all other voxels and then compute an average. This is essentially mean centrality or the mean connectedness of a voxel, where if a voxel is highly correlated to most other voxels then its centrality will be greater (consider a region like dorsal anterior cingulate which has high involvement in many cognitive processes and enforces top-down control on many other brain structures). While this method is sound, it is computationally inefficient, because to calculate it voxel-wise a voxel-to-voxel correlation matrix must be computed, which is highly inefficient. However, we can utilize PCA to compute the eigenvariate (i.e., the scores of the PCA on the matrix of all time series) using the fast eigenvector centrality mapping (fastECM) algorithm (Wink *et al*, 2012). This allows for a similar measure (EVC) that is computationally efficient.

3.7 TASK-BASED FMRI ACTIVATION

Task-based fMRI is first slice-time corrected (if event related), motion corrected, normalized to MNI space, and spatially smoothed. A generalized linear model is used to estimate the effect of the tasks performed in the scanner (Wink *et al*, 2012). Similar to the resting state we remove several confounds, including: six motion parameters, a set of cosine terms that represent low frequency signals (effectively a high-pass filter, usually 1/128 Hz), and the mean. Note that we do not perform low-pass filtering or remove additional confounds from white matter or CSF. However, to model temporal auto-correlation due to aliased biorhythms and unmodelled signal an auto-regressive term with order one [AR(1)] is included, which includes a shifted (by one) signal of the time series. This models changes in heart rate and blood pressure from the previous time point to account for temporal auto-correlation, as an assumption of the regression is that the measurements (each time point) are independent.

To model the activation of the task performed in the scanner the onsets and durations are used to create boxcar functions (zero if not tasking and one if tasking) for each of the blocks. For example, in the face/shapes task there are two types of blocks (conditions): matching face emotions and matching shapes thus two boxcar functions would be generated representing the onsets and durations for those two tasks. Each boxcar for each condition is then convolved with a hemodynamic response function (HRF), which represents the canonical hemodynamic response in the BOLD response after a task. Convolution is a process where the HRF is translated in time along the boxcar and the integral of the element-wise multiplication is computed (i.e., sum of the multiplication). For each condition, an expected hemodynamic response is modeled for each task. We regress the expected hemodynamic responses of each condition against each voxel's time series (mass univariate regression) to estimate two parameter estimates representing the

activation during each condition (e.g., activation during faces vs. activation during shapes) (Wink *et al*, 2012). The greater the parameter estimate translates to a greater observed association with the expected response, thus the greater the actual activation (Penny *et al*, 2011). A contrast is essentially a difference in parameter estimates and is usually used to estimate the relative activation of one task while adjusting for another (e.g., faces minus shapes indicates activation during the faces while controlling for visual, motor, and matching aspects of the shapes). The output is a contrast value (difference in parameter estimates) at each voxel (a map) (Penny *et al*, 2011).

3.8 STATISTICAL GROUP INFERENCE

Voxel-wise group inference relies on several core principles: regression and random field theory. Regression can be used to determine the voxel-wise association between voxel-wise maps (gray matter density, FA/MD, EVC, or activation during a task) and a variable of interest (e.g., group differences or association with depression severity). A voxel-wise mass-univariate regression (similar to the previous section) can be used to determine the group level association between a variable of interest and a voxel-wise map. Each voxel has a corresponding parameter estimate, a degree of freedom, a statistical test, and a p-value. If we set our acceptable rate of false positive at alpha less than 0.05, then it does not suffice to threshold voxels whose p-value is below this threshold as we inflate our true rate with each independent statistical test.

This is the core of the multiple comparisons problem. Consider performing two independent statistical tests and setting our acceptable false positive rate alpha at 0.05. Then to make joint inference on both of those findings, we need to account for the inflation in the false

positive rate with each additional statistical test (i.e., adding more statistical tests increases the overall likelihood of false positives). The Bonferroni correction (a type of family-wise error correction) is one method that adjusts the false positive rate by dividing alpha by the number of statistical tests (e.g., 0.05/2 or 0.025). However, this is far too conservative of a measure when correcting for multiple comparisons in a voxel-wise analysis due to the high number of voxel-wise tests. Many voxel-wise data have statistical tests on the order of hundreds of thousands, thus for example if we adjust for multiple comparisons for 100,000 statistical tests we need a p-value of $5x10^{-7}$ using this correction method. However, Bonferroni correction requires correction for the number of *independent* t-tests. As the voxels are highly spatially and structurally correlated, and were spatially smoothed (which introduces even greater correlation between voxels) then the voxels are not necessarily independent statistical tests.

Random field theory can be used to estimate the number of independent statistical tests based on spatial correlation (smoothness) of the voxel-wise data (Poline *et al*, 1995). A resolution element (resel) is a unit of measure that determines the number of independent tests. Assuming we know that data that was 'x' by 'y' by 'z' number of voxels had 'V' smoothness then the number of resels is computed as the product of the number of voxels divided by the cubed smoothness (Poline *et al*, 1995). Consider the example with 100,000 voxels then if we assume a smoothness of 12mm (typical for functional neuroimaging data) then the number of resels is approximately 58 thus only a p-value less than $8.6x10^{-4}$ is needed to correct for multiple comparisons.

4.0 MACHINE LEARNING

This chapter introduces basic concepts of regression (logistic regression) and support vector machines (two commonly used algorithms), as well as introducing machine learning concepts, including: cross-validation, validation metrics (area under the curve, AUC), and permutation testing. After introducing core concepts, the common pitfalls and problems in machine learning are described: how high dimensionality causes over fitting and multi-collinearity issues; and then how these are overcome using kernel based machine learning methods. Finally, the keystone algorithm is described in the final two sections: multi-factor kernel based machine learning.

4.1 REGRESSION

Regression is the process of estimating a set of parameters that linearly model the association between an observed variable of interest and set of observed features. In the context of machine learning, we are often interested in how well the estimated parameters predict the observed variable of interest. Thus, after fitting a model we can predict new data and investigate how well the predicted variable matches the measured outcome. Consider the following problem:

Equation 2a. $y = X * \beta,$

where y is a vector that is length n subjects, X is a feature matrix that is n subjects by f features, and beta are the set of parameters that linearly model their association (vector length f) that we want to estimate. Then the following is the ordinary least squares solution:

Equation 2b. $\beta = (X^T X)^{-1}(X^T y)$.

This solution has several key assumptions: (1) linearity which states that the observed variable is a linear combination of the features; (2) homoscedasticity which states that the variables have constant variance (i.e., they do not have skewed variance along the full set of values) as well as normality which states that features must be normally distributed; (3) independence of response variables (more specifically their errors) which states the each measurement is independently measured; and (4) no multi-collinearity which states that features should not be highly correlated.

The two most important assumptions to consider are linearity and multi-collinearity. When fitting models with low number of subjects, it is often best to assume linearity as non-linear models may *over-fit* and not generalize well (i.e., model may be specific to the current data set and would not do well in a larger sample). Multi-collinearity occurs when features are highly correlated and thus there exists no unique solution to the parameter estimates. High dimensional data with large number of features tends to easily suffer from multi-collinearity as the probability that two features are highly correlated increases as the number of features increases.

4.1.1 Logistic Regression

Logistic regression is a scenario where the outcome variable is binary (or non-continuous). Unlike linear regression, logistic regression has no closed form solution for the parameter estimate, thus it is estimated through an iterative process. Parameters are usually estimated using maximum likelihood estimation where an initial solution is computed; it is updated using one of

several methods (e.g., Newton's method); and repeats until the solution maximizes the likelihood function.

To understand the mathematical formulation, consider a simple example where we predict a variable y (vector with length n subjects) that is either zero or one with a single feature x (vector with length n subjects). In this scenario, we are estimating a probability that y equals one or zero given x, which can be written as:

Equation 3a. $p(y=1|x)=\beta_0+\beta_1 x$

Equation 3b. $p(y=0|x)=1-p(y=1|x)$

where the beta terms model the mean and contribution of x, respectively. As we are trying to estimate a probability, the solution for the probability should be bounded by zero and one – however the right hand side of the equation is unbounded (negative to positive infinity). Thus, we can use what is known as the logit function (defined here as F), which is one at positive infinity and zero at negative infinity. This can be written as:

Equation 3c. $p(y=1|x)=F(\beta_0+\beta_1 x),$

Equation 3d. $p(y=0|x)=1-F(\beta_0+\beta_1 x),$

where these can be combined to a general form as:

Equation 3e. $p(y=y_i|x)=[F(\beta_0+\beta_1 x)]^{y_i}*[1-F(\beta_0+\beta_1 x)]^{1-y_i}.$

Note that if y_i is one then equation 3e becomes 3c and if it is one it becomes 3d (thus the generalization, or equation 3e, works). The likelihood (which is what we would like to maximize) is the product of these probabilities for all individuals:

Equation 3f. $L=\prod_{i=1}^{n}[F(\beta_0+\beta_1 x)]^{y_i}*[1-F(\beta_0+\beta_1 x)]^{1-y_i},$

thus to estimate the probability, we should compute the derivative of the log-likelihood and solve for the parameters when it equals zero (as this maximizes the likelihood). However, the derivative of a set of products is not easily computed. However, we utilize a well-known property that the solution using the derivative of the log-likelihood is the same as the derivative of the likelihood. Thus we define the log-likelihood:

Equation 3g.
$$l = \ln(L) = \ln\left(\prod_{i=1}^{n}[F(\beta_0+\beta_1 x)]^{y_i} * [1-F(\beta_0+\beta_1 x)]^{1-y_i}\right) = \sum_{i=1}^{n} y_i \ln\left(F(\beta_0+\beta_1 x)\right) + (1-y_i)\ln\left(1-F(\beta_0+\beta_1 x)\right)$$

Note that the log-likelihood reduces to a sum due to the property that the log of two products is the sum of the individual logs, and that the power terms are multiplicative due to the property that the log of a variable to a power is equal to the power times the log of the variable. At this moment, a derivative is computationally possible and we estimate the parameters such that the derivative with respect to each parameter of the log-likelihood is equal to zero. In reality, optimization algorithms that identify the maximum of the log likelihood function are able to estimate an initial approximation, then iteratively update the parameters (e.g., using Newton's method, gradient descent, etc.), and then define some convergence criteria for the iteration to halt. This process can be scaled for more than one feature.

We can fit a sparse logistic function (i.e. with fewer features included in the model) using step-wise regression. Step-wise regression is a model building approach that attempts to only include features that are highly predictive and generates the 'best' predictive model with the 'lowest' number of features (*attempts* to generate the most parsimonious model). This process involves first starting with an initial model that only includes modeling the mean. The algorithm then iterates over multiple steps where each step it tests all possible features to include and

includes those that are most predictive (or none), but simultaneously also tests for which features should be removed. Once there are no more features to add or remove – the procedure stops. This generates a sparser model, which can be useful when considering high dimensional feature sets.

4.2 SUPPORT VECTOR MACHINES (SVM)

Support vector machines (SVM) are machine-learning algorithms that (similar to regression) generate a model that separates two groups by attempting to find a hyperplane that best separates the data with the largest margin between both groups. The basic concept is that there exist some boundaries that separate the data into several classes and these boundaries confidently predict classes on both extreme ends of the boundary. This section is divided into the following: notation, definition of margins, optimizing margins, define the Lagrangian, we define the optimization problem using the Lagrangian, introduce the concept of kernels, and briefly describe the optimization algorithm sequential minimal optimization (SMO).

Let y be one of two class labels (-1 and 1), while x is a feature. Then, a classifier can be defined with the intercept being defined as b and parameters as w:

Equation 4a. $h(x) = g(w^T x + b)$.

4.2.1 Functional and Geometric Margins

The functional margin represents the confidence of the accuracy of the prediction, where a large functional margin represents high confidence of a correct prediction. We can define the functional margin as the following:

Equation 4b. $\hat{\gamma}^{(i)} = y^{(i)}(w^T x + b)$,

for a given training set $[x^{(i)},y^{(i)}]$, where given a large set of training data then we would minimize across functional margins. The functional margin suffers as it does give you a measure of confidence but not how close it is to the decision boundary and this is due to the scalability of w and b. This allows for the functional margin to be arbitrarily large without meaning, thus we can constrain it using the L2-norm:

Equation 4c. $\| w \| = 1$,

where the magnitude of w is constrained.

Consider now the geometric margin, which is the distance between a training point (A) and the decision boundary. Let's consider a point (B) that is orthogonal to A and lies on the decision boundary, while defining the distance between A and B as gamma. Thus, we can find the point B on the decision boundary using the following:

Equation 4d. $x^{(i)} - \gamma^{(i)} w / \|w\|$,

where the point A is by definition x, and gamma times the unit vector results in subtracting a value that is length gamma (distance from A to B) and direction orthogonal to the decision boundary. We can now use the value of B, which lies on the decision boundary, into the equation of the decision boundary:

Equation 4e. $$w^T\left(x^{(i)}-\gamma^{(i)}\frac{w}{\|w\|}\right)+b,$$

and then solving for gamma results in:

Equation 4f. $$\gamma^{(i)}=\left(\frac{w}{\|w\|}\right)^T x^{(i)}+\frac{b}{\|w\|},$$

which can be generalized to the definition of the geometric margin:

Equation 4g. $$\gamma^{(i)}=y^{(i)}\left(\left(\frac{w}{\|w\|}\right)^T x^{(i)}+\frac{b}{\|w\|}\right).$$

Given a large training set, we would then define the geometric margin as the smallest of the geometric margins across all the training data.

4.2.2 Defining the Minimization Problem

The goal is to find a decision boundary that maximizes the geometric margin, as this would reflect a high confidence set of predictions (results in a classifier or model that separates two classes with a margin or gap whose center is the decision boundary). This maximization problem can be written as:

Equation 5a. $$\begin{array}{l}\max_{\gamma,w,b}(\gamma)\\ st\\ y^{(i)}\left(w^T x^{(i)}+b\right)\geq\gamma, i\in\{1,\ldots,m\}\\ \|w\|=1\end{array}.$$

While we could attempt to solve this problem – it is not well constrained, mainly the L2-norm constraint is not convex (i.e., may have multiple local minima). However, we can alter this

problem using two ideas developed in section 4.2.1. The first is that the geometric margin is equal to the functional margin divided by the L2-norm:

$$\text{Equation 5b. } \gamma = \frac{\hat{\gamma}}{\|w\|},$$

and if we replace the geometric function in equation 5a then we get:

$$\max_{\hat{\gamma},w,b} \left(\frac{\hat{\gamma}}{\|w\|} \right)$$

$$\text{Equation 5c. } st \quad .$$

$$y^{(i)}\left(w^T x^{(i)} + b\right) \geq \hat{\gamma}, i \in \{1,\ldots,m\}$$

From the previous section, we know that the functional margin can be scaled without changing the solution, which we know take advantage of and scale the functional margin such that:

$$\text{Equation 5d. } \hat{\gamma} = 1,$$

and further since the maximization problem in 5c is reduced by equation 5d, then we are now maximizing one over ||w|| then we can instead define the problem as a minimization like this:

$$\min_{\gamma,w,b} \left(\frac{1}{2}\|w\|^2 \right)$$

$$\text{Equation 5e. } st \quad .$$

$$y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1, i \in \{1,\ldots,m\}$$

The current optimization problem is convex (squared term) and thus has an optimal solution. Note that minimizing the squared term is the same as maximizing the inverse term, and also that we have included a constant (1/2, which does not change the solution) but does simplify some of the computation in later stages. While this optimization problem is well constrained and solvable, we can define a better form (dual form) that performs much better and has some desirable properties regarding kernels (defined later).

4.2.3 Dual Form of the Minimization Problem

Using the Lagrangian we can define the dual, and so if we consider a general form of an optimization problem:

$$\text{Equation 6a. } \begin{aligned} &\min_w f(w) \\ &st \\ &h_i(w) = 0, i \in \{1,...,m\} \end{aligned} .$$

Then the Lagrangian is essentially the following:

$$\text{Equation 6b. } L(w,\beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w),$$

where beta are known as the Lagrangian multipliers, and we can calculate two partial derivatives to estimate the two parameters:

$$\text{Equation 6c. } \frac{\partial L}{\partial w_i} = 0; \frac{\partial L}{\partial \beta_i} = 0.$$

This is a well-defined construct that we do not review (except for equation 6a-6c). We can thus input these into the original equation to get the dual form. However, applying this concept to equation 5e, we can further simplify the minimization problem. Thus, the Lagrangian for this problem is:

$$\text{Equation 7a. } L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} \left(w^T x + b \right) - 1 \right],$$

where alpha are the Lagrangian multipliers. The partial derivatives are

$$\text{Equation 7b. } \begin{aligned} &\frac{\partial L}{\partial w} = 0 = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \\ &\therefore w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \end{aligned} ,$$

Equation 7c. $$\frac{\partial L}{\partial b} = 0 = \sum_{i=1}^{m} \alpha_i y^{(i)}.$$

Inserting these into the original problem (equation 5e) and after rearranging, we reach the dual form of the problem:

Equation 8a. $$\begin{aligned} &\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} > \\ &st \\ &\alpha_i \geq 0, i \in \{1, \ldots, m\} \\ &\sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \end{aligned}.$$

The dual form of this problem has several unique properties that can be utilized, and by utilizing kernels these algorithms can learn in high dimensional spaces.

4.2.4 Kernels

Consider that instead of learning using the feature set x, that it may be beneficial to learn using the squared and cubic terms as well. We can then define such a function:

Equation 9a. $$\phi(x) = \begin{bmatrix} x \\ x^2 \\ x^3 \end{bmatrix},$$

where phi is a mapping between an input feature and some new feature set that may be more informative than the original feature set. Equation 8a can be written entirely as dot products and so we can define a Kernel as:

Equation 9b. $$K(x,z) = \phi(x)^T \phi(z).$$

The dot products can thus be replaced by K(x,z) or the Kernel. In practice, we never compute phi, but rather the Kernel is often inexpensive to calculate. A particularly useful Kernel is known as the radial basis function (rbf):

Equation 9c. $$K(x,z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right),$$

where sigma is a parameter and "exp" refers to the exponential function. These functions can thus learn different boundaries (not just a straight line) using such an exponential function. The 'kernel trick' is that we can use linear algorithms to learn non-linear kernels (e.g., rbf).

4.2.5 Sequential Minimal Optimization (SMO)

The SMO algorithm is one of the most frequently used algorithms to solve the maximization (of alpha) problem in equation 8a. In this algorithm two alphas are chosen (based on a heuristic that picks alphas that progresses towards global maximum maximally), they are updated, then W is optimized with respect two updated alphas, and then this is repeated until convergence. Critically, two alphas are chosen because the sum of all alphas must equal one – thus by changing one, another must be updated as well.

The major reason this optimization scheme works well is that the update for alpha is computationally very efficient. Thus, even though the first step (which chooses alpha based on some heuristic) is not an optimized step, we can update alpha quite easily. Further, any update is bound to go towards the maxima (albeit some may have smaller step sizes). Finally, this method is *guaranteed* to converge.

4.3 MACHINE LEARNING MODEL BUILDING PROCEDURE

Several key concepts are needed to understand the basics of fitting machine learning models and testing their efficacy. Specifically, five core concepts are introduced: cross-validation, feature selection and parameter optimization, validation and the measures used, and permutation testing. The basic procedure for training and testing a model is: (1) divide data into a training set (used to fit the models) and a test set (used to validate its efficacy); (2) using only the training data choose features and optimize parameters of the machine learning model; (3) predict the testing set outcomes using the trained model; (4) compare the outcome of the test set to the predicted outcomes using validation measures.

Cross-Validation. The first step of most model building is determining which data will be used as training sets (i.e., data to fit the model) and subsequently which data will instead be the testing set (i.e., data to validate the final built model). If we fit the model and test it on the same dataset – we will inflate the accuracy thus it is critical to separate the data into training and test sets. If there are large samples, a good recommendation is to use approximately 70 percent of the data as training and 30 percent as testing. This is by far the most principled way of determining the accuracy and efficacy of the machine-learning model. However, if the sample size is small then this may not be possible, thus if the number of samples is *very* small then leave-one-out cross-validation may be a good option. In this approach, the model building process is repeated 'n' times (number of subjects) where in each fold one subject is left out as the test and the others are used as training data. The model is built then we predict the outcome on the test holdout. If we repeat this process then we will have built 'n' models and predicted each point (where on each fold that individual was left out of the model training). However, this approach can be highly biased towards the sample and often it is better to use a larger number of folds. Thus, we

can perform a similar iterative holdout procedure however instead of leaving just one individual out on each fold, we instead leave out a higher number (e.g., if N=50 then a 10-fold cross-validation would result in 10 folds with 5 individuals each, where the model training is done 10 times and the data of 5 individuals is left out as the test set).

Feature Selection and Parameter Optimization. In high dimensional feature sets, it is important to select features prior to fitting the machine learning models. If there are too many features, the model may over-fit (described in section 4.4.1). The most principled way to select features is to use some percentage (e.g., 15 percent) of the training data set to determine a set of relevant features that are then used in the model training. Again if the number of samples is small then this may have to be cross-validated as well. Often this can mean using another cross-validation inside the loop that performs the training. We can determine the best features inside this nested cross-validation. Similarly, we may want to optimize parameters of the machine learning models (e.g., box constraint in SVM) and we can allow for this to occur inside this nested cross-validation as well.

Notice that while allowing for a nested cross-validation allows for a good optimization of the parameters and good feature selection, it may be biased because the parameters chosen and the features selected may not be consistent for each fold of the original cross-validation. Thus, at best we can assume that the model training procedure is generalizable to a larger population – but not necessarily the specific model that we train.

Validation Metrics. For each fold of the cross-validation a model was fit and the outcome of the test set was predicted. Thus, after this procedure each subject has a single predicted outcome ($\hat{y}$), which is either a set of probabilities or a set of predicted labels, and a single *actual*

outcome (y). Two common measures to validate the efficacy of the trained model are: accuracy and area under the curve (AUC) of a receiver-operating characteristic (ROC) curve.

If a label is output, accuracy is a measure that determines percentage of accurately identified individuals. The following are needed to compute accuracy: true positives are the number of positives successfully identified and true negatives are the number of negatives successfully identified. Subsequently, false positives are the number of negatives identified as positives (type I error), and false negatives are the number of positives identified as negatives (type II error). Accuracy is the number of true positives and negatives combined divided by the total number of individuals. We can also identify two other important values: sensitivity or how accurately we identify positives (true positives divided by true positives plus false negatives) as well as specificity or how accurately we identify negatives (true negatives divided by true negatives plus false positives). While accuracy identifies how well we can identify labels, it does not identify how accurately we can identify positive and negative classes. In a scenario where there are a high number of negative classes, the model (for example) may just predict that all labels should be negative. This would result in a higher accuracy and specificity, but low sensitivity. A good model balances sensitivity and specificity.

If a probability is output (instead of a label), an ROC curve can be generated and AUC can be calculated. For a set of ‘m’ thresholds, we can threshold those probabilities and calculate the false positive rate (number of false positives divided by total negative) and plot it against the true positive rate (or sensitivity). This plot is known as the ROC curve, and the area under the ROC curve (the integral) is known as the AUC. An AUC of 1 indicates a perfect prediction, while an AUC of 0.5 indicates at chance prediction.

Permutation Testing. While a model may have high accuracy, we may want to test whether this accuracy is *significant*. A possible concern is that the model building procedure is *so* good that given any training labels, it is still able to achieve a high accuracy. A basic procedure that can be done is known as permutation testing, which involves establishing a distribution for some measure of interest (whose distribution is unknown, e.g., accuracy) and determine a significance based on this distribution.

This process involves: (1) randomly permuting (mixing) the labels in the outcome measure; (2) performing the entire model training process; (3) calculating the validation measure (e.g., accuracy). This process is repeated a large number of times (e.g., 5,000 times). For each permutation a single accuracy measure has been computed, which establishes the distribution of this specific measure. We can then investigate whether the accuracy measure we computed lies on the tails of the distribution (indicating its significance). To calculate a p-value directly, we can add the number of times the *permuted* validation measures are greater than the *actual* accuracy measure then divide by the number of permutations. Thus, if the accuracy is significant (defined as alpha < 0.05) it should be greater than the randomly permuted accuracy 95% of the time (as alpha is 5%). If not, then it means that our model building procedure is able to predict *any* given set of labels and thus the current model may be accurate but would not generalize to another sample.

4.4 PRACTICAL PROBLEMS AND SOLUTIONS

4.4.1 Common Machine Learning Problems

There are several common problems that plague machine learning, however the core problem is high dimensionality of features and low number of samples (especially in neuroimaging literature). This tends to also be the underlying cause of many other issues including: over fitting and biased feature selection and another separate problem is multi-collinearity.

High dimensional data (i.e., too many features) often suffers from generalizability issues, and models built in high dimensional feature spaces tend to perform well on the training data but worse on the test data. Often this may be because there are too many semi-random features that do not accurately and *reliably* model the data. One example where this may occur is that a single individual may be an outlier on some feature, and that feature may be chosen as predictive – which may bias the entire model. High dimensional problems tend to also be more complicated to understand, as there may be multiple interacting features. This tends to lead to over-fitting of the model as a high number of features (even random features) are able to generate a single model, but these tend to do well only on the training data. This can complicate feature selection as well; mainly if there are a lot of features it may be difficult to either know which features to include a priori or even determine algorithmically.

The number of samples needed to fit generalizable machine learning models should be considerably greater than the number of features. Similar to high dimensional data, this results in over-fitting and high variance in accuracy. Further, a small sample size is more likely to suffer from selection bias (i.e., sample may represent only a small subsample of the entire population) – thus even if the model is stable it may not generalize to a larger more general population or just

generalizes to specific subsample. If we select features algorithmically, because this process is usually done in either a separate sample or a nested loop – the number of samples becomes severely limited when trying to select the best features.

Multi-collinearity refers to the association between features. If two or more features are highly correlated, it may be difficult for the model to differentiate between them and many models often involve assumptions regarding feature independence. High dimensional data is more likely to contain features that are linearly correlated (based on central limit theorem). Often, a feature selection procedure is needed to remove highly collinear features.

As a result of multiple of these problems, another issue is that often many of these problems do not have an idealized solution. For example, the ordinary least squares solution for regression (equation 2b) is not an ideal solution when the number of features is greater than the number of samples or if the features are highly collinear. Thus, often we have to regularize the solution to penalize highly over-fit models (e.g., usually a penalty term with a corresponding parameter is used to reduce the number of features contributing to the overall model).

4.4.2 A Practical Solution: Principal Components Analysis

A powerful tool that is often utilized is known as principal components analysis (PCA), and is potentially able to resolve each of the problems in section 4.4.1 in a single step. PCA is a deterministic process that identifies a low dimensional feature space that can be used to represent higher dimensional data. To understand PCA intuitively, several critical concepts are needed.

Theoretically, it is a method that identifies an orthogonal (i.e., not correlated) basis set of the high dimensional feature space. Some examples of common orthogonal bases are the Cartesian coordinate system or the cosine basis. The Cartesian coordinate describes all points in

a 3D space by defining three points on three linear but orthogonal vectors (x, y, and z). Similarly, the cosine basis can define all signals in time as a set of cosine functions of varying frequency and amplitude. Thus the discrete cosine basis can define *any* signal as a linear combination of many cosines and so it forms a basis or a low dimensional feature space (cosines) that defines a very high dimensional feature space (all discrete signals). In the same way, PCA attempts to identify orthogonal vectors (called eigenvectors) that represent the entirety of the data in a low dimensional space.

Another intuition important in PCA and basis sets is that of rank, which represents the nondegenerateness of the feature space. Consider the Cartesian coordinate system again, to define a 3D data set a minimum of three vectors are needed to identify all points in space, however if we added another vector it would be rank deficient. These four vectors are not orthogonal when defining a 3D space, as a maximum of 3 vectors is needed (i.e., only three numbers are needed to understand where in a 3D space a point is located while the fourth adds no new information). Similar to the high dimensionality issues described earlier, having a fourth dimension or vector is similar to having too many features.

Singular value decomposition (SVD) is a method that can be used to decompose any matrix into a combination of three matrices: two unitary matrices (i.e., conjugate transpose is also its inverse) and a rectangular diagonal matrix. Using SVD, we can perform PCA on any given matrix. Given a matrix 'X' ('n' subjects by 'f' features) then SVD can be represented as:

Equation 10a. $X = U\Sigma V^T$,

where U (n by n) and V (f by f) are square unitary (e.g., $U*U^T=I$) matrices and Σ (n by f) is a diagonal matrix. Critically, U and V represent the left and right eigenvectors while Σ contains the eigenvalues along its diagonal. While PCA is usually described as decomposition on the

covariance matrix, SVD allows for a simpler decomposition (i.e., U and V of X are the same as the eigenvectors of $X*X^T$). The UΣ is also known as the scores while V is known as the coefficients or loadings, where the scores represent the original data in the principal component bases while the coefficients are the transformation between the original and principal component bases.

The scores is a square matrix that is n by n, where the n-rows represent each individual subject and the n-columns represent the n-bases (or principal components). Further, UΣ*(UΣ) is the identity matrix, thus the n-bases are not correlated. We have now constructed a feature space that is not collinear, has as many features as number of subjects (thus not rank deficient), and because of this has an idealized solution in regression problems. This is the basis of principal components regression or kernel (since the scores from PCA are a kernel) based machine learning. The basic approach is to first reduce the feature matrix using PCA to get scores, fit a machine-learning model between the scores and the outcome, and transform the parameters back to the original space using the coefficients. Section 4.4.3 describes the general process of kernel-based regression.

4.4.3 Kernel-Based Regression

Consider a feature matrix X (n by f) and a vector of outcomes y (length n) then while we could solve the general regression problem (y=XB), it may suffer from some of the problems described previously. An alternative is to first perform PCA on X (which we first center by removing the mean of each column):

Equation 11a. $X = U\Sigma V^T = WV^T$,

where W is the scores and V is the coefficients. Then we can solve the following problem:

Equation 11b. $y = W\beta_{pcr}$,

where we now instead solve for a new set of regression parameters (one for each component in the PCA), which does not suffer some of the typical pitfalls described in previous sections. We can then back-project these parameters into the original space (if we are interested in understanding the model) using the coefficients:

Equation 11c. $\beta = V\beta_{pcr}$,

which is an ideal solution for the original problem and has several unique properties, but most importantly it addresses the multi-collinearity problem and has a regularization effect. Thus, by utilizing this 'kernel trick' we are able to fit a well-defined and efficient model.

Choosing Number of Principal Components: This model is improved further by choosing a set of principal components instead of utilizing the entire feature space. There are several methods for choosing the number of components, the most common being to investigate a scree plot (cumulative variance explained by number of components) and either choosing the number of components that explain a certain percentage of variance or by finding the 'elbow' of the scree plot (i.e., find the maximum of the second derivative of the scree plot also known as the inflection point). However, another method, which is one of the most recommended but most under utilized, is called Horn's parallel analysis. This method utilizes permutation testing and permutes the data matrix into uncorrelated normal variables and performs PCA getting a set of eigenvalues over several permutations. For each eigenvalue we can compute a p-value by adding the number of times the random eigenvalue is greater than the actual eigenvalue and dividing by the total number of permutations. Eigenvalues with p-values less than some pre-determined alpha (e.g., alpha < 0.05) are considered significant and are retained.

4.4.4 Multi-Factor Analysis

Multi-factor analysis (MFA) is an extension of PCA that seeks to balance several feature sets (Abdi *et al*, 2013). Consider an example where some outcome (e.g., depression severity) is dependent on two feature sets (e.g., clinical surveys and genetic data), where we assume that the number of clinical surveys (f) is much less than the number of genetic features (g). We could investigate the association between depression severity and each clinical survey or each genetic variable using principal components regression (PCR). While we could perform these analyses separately, it may be useful to understand their joint association with depression (i.e., the association between depression severity and both feature sets). We could combine these feature sets into one dataset (which has f plus g features) and then perform PCR. However, the surveys suffer greatly because the kernel that is generated using PCA is influenced by the genetic data solely due to its size (as it has a much larger number of features, g).

MFA extends upon this by performing PCA twice. We compute a set of scores for both feature sets (i.e., one kernel for the surveys and one for the genetic data) and then these scores (which are identical in size) are then input into another PCA where another set of scores is generated. This assumes that the genetic data and clinical surveys both serve an important role in understanding the depression severity and does not bias the kernel towards either feature set. This becomes essential as in neuroimaging studies we often collect multiple neuroimaging data (structural and functional that are very high dimensional on the order of hundreds of thousands of voxels), but also collect clinical measures (with much lower number of features) that may be just as important. Thus, MFA can be used to balance the predictive capabilities of each set of features without the loss of balance between features. Such an approach will enable a unique combination of multiple feature sets and will utilize the full set of features within a study.

4.5 MULTI-FACTOR KERNEL BASED MACHINE LEARNING

Two major algorithms are used in this work: (1) single feature set learning and (2) multi-feature set learning. Single feature set learning uses kernel-based (PCA) learning while the multi-feature set uses an extension using another kernel based on the MFA instead of PCA.

4.5.1 Single Feature: Kernel Based Learning

Given a single feature set X (n subjects by f features, which can be images, surveys, or any other data) and a vector of outcomes (y, length n subjects) then the implemented algorithm (PCR) is as follows:

1. Horn's Parallel Analysis is conducted on X to determine the number of components (or c, where c < f) to keep in the PCA.
2. Perform PCA on X to calculate scores (W, n by c) and coefficients (V, c by f).
3. Conduct a 10-fold cross-validation dividing data into training and test sets. Then for each fold perform the following:
    a. Fit a model (using either SVM, Logistic Regression, or step-wise Logistic Regression) between the training scores (W_{train}) and the outcomes (y).
        i. Optimize the models using the training data (if needed).
    b. Using the model predict a set of outcomes ($\hat{y}$) using the test scores (W_{test}).
4. After we iterate through all 10 folds, the predicted outcomes ($\hat{y}$) are compared to the actual outcomes (y). We compute AUC and accuracy.
5. As the cross-validation in step 3 is non-exhaustive (i.e., it does not learn and predict on all possible ways to divide the data into training and test sets), it benefits to generate a

large set of cross-validations. Thus, we repeat steps 3 and 4 several (50) times where on each repetition we generate another cross-validation scheme. This will output a distribution on the validation measure (i.e., 50 AUC and accuracy measures are generated).

6. We fit a single model on all W and y to get a set of parameters (β_{pcr}, length c) that are projected back into the original space using the coefficients ($\beta=V\beta_{pcr}$, length f). We perform permutation testing (step 7) to determine which parameters in the original space significantly contribute to the prediction model.
7. Generate a set of 1,000 random permutations of y. For each permutation:
    a. Fit a single model between W and $y_{permuted}$ to get a set of ($\beta_{pcr\_permuted}$) that are projected back into the original space using the coefficients ($\beta_{permuted}=V\beta_{pcr\_permuted}$).
8. For each parameter (β), add the number of times its absolute value is less than the absolute value of the permuted parameter ($\beta_{permuted}$) then divide by the number of permutations (1,000) to generate a single p-value for each parameter. This determines whether the parameter significantly contributes to the overall model.

4.5.2 Multiple Features: Multi-Factor Kernel Based Learning

Consider two feature sets X (n subjects by f features) and S (n subjects by g features) and a vector of outcomes (y, length n subjects). Note that this process extends the previous algorithm to have *two* instead of one PCA (hence multi-factor), thus fit parameters are projected twice. The implemented algorithm (MFA) is as follows:

1. PCA is conducted on both X and S to generate scores (W and S, both n by n matrices) and coefficients (V, length n by f; and N, length n by g). We generate a single scores matrix from all scores (R, n by 2n).
2. Horn's Parallel Analysis is conducted on R to determine the number of components (or c) to keep in the PCA.
3. Perform PCA on R to calculate scores (Q which is n by c) and coefficients (P which is c by 2n).
4. Conduct a 10-fold cross-validation dividing data into training and test sets. Then for each fold perform the following:
    a. Fit a model (using either SVM, Logistic Regression, or step-wise Logistic Regression) between the training scores (Q_{train}) and the outcomes (y).
        i. Optimize the models using the training data (if needed).
    b. Using the model predict a set of outcomes ($\hat{y}$) using the test scores (Q_{test}).
5. After we iterate through all 10 folds, the predicted outcomes ($\hat{y}$) are compared to the actual outcomes (y). We compute AUC and accuracy.
6. As the cross-validation in step 3 is non-exhaustive (i.e., it does not learn and predict on all possible ways to divide the data into training and test sets), it benefits to generate a large set of cross-validations. Thus, we repeat steps 4 and 5 several (50) times where on each repetition we generate another cross-validation scheme. This will output a distribution on the validation measure (i.e., 50 AUC and accuracy measures are generated).
7. We fit a single model on all R and y to get a set of parameters (β_{MFA}, length c) that are projected back into the MFA space using the coefficients ($\beta_{pcr}=P\beta_{MFA}$, length 2n).

Another projection allows the parameters to be projected to the first PCA space ($\beta_X=V\beta_{pcr}$, length f; or $\beta_S=N\beta_{pcr}$, length g). We perform permutation testing (step 7) to determine which parameters in the original space significantly contribute to the prediction model.

8. Generate a set of 1,000 random permutations of y. For each permutation:
    a. Fit a single model between Q and $y_{permuted}$ to get a set of ($\beta_{MFA\_permuted}$) that are (twice) projected back into the original space using the coefficients ($\beta_{permuted}$).
9. For each parameter (β), add the number of times its absolute value is less than the absolute value of the permuted parameter ($\beta_{permuted}$) then divide by the number of permutations (1,000) to generate a single p-value for each parameter. This determines whether the parameter significantly contributes to the overall model.

5.0 NEURAL CORRELATES OF LATE-LIFE DEPRESSION

This chapter introduces major depressive disorder (MDD) and late-life depression (LLD) as well as the structural and functional neural correlates of pharmacotherapy. The neuroimaging prediction literature is also reviewed briefly. The introduction sections of chapters 6 to 8 cover the resting state, emotion reactivity, and prediction of response literature, respectively.

5.1 MAJOR DEPRESSIVE DISORDER

Depression is characterized by several core symptoms: low/depressed mood, anhedonia (inability to feel pleasure), low energy or fatigue as well as disturbed sleep, pessimism, feelings of guilt, loss/gain of weight, and suicidal tendencies. It is a complex disorder dependent on genetic, environmental, and neural factors. Depression has a high prevalence (16.6% of individuals will meet criteria for MDD at least once in the US (Kessler *et al*, 2005)) and is associated with high medical comorbidity and mortality resulting in more years lived with disability than any other disease (Alexopoulos and Kelly, 2009; Moussavi *et al*, 2007). Depression ranks fourth in disability-adjusted life years (Moussavi *et al*, 2007). It is projected (by 2020) that it will only be second to heart disease in its contribution to global disease burden (Hinrichsen and Hernandez, 1993). Further, individuals with MDD have worse cognitive functioning, greater prevalence and severity of chronic medical conditions such as arthritis, hypertension, and diabetes, as well as

increased utilization of medical services and greater health care costs (Bruce *et al*, 2004; Charney *et al*, 2003; Stevens *et al*, 1999). Despite significant strides in our understanding of MDD etiology, pathophysiology, and mechanisms for treatment, it has proven challenging to prevent, diagnose, and treat depression effectively.

While MDD has a strong genetic susceptibility component (first-degree relatives carry a threefold increase in risk compared to the general population (Sullivan *et al*, 2000)) it is also associated with a wide variety of neurobiological factors. The most well-known and well-characterized hypothesis is the dysfunction of monoamine systems, specifically serotonin, dopamine, or norepinephrine (Bunney and Davis, 1965; Schildkraut *et al*, 1965). This was largely supported in studies that showed that decreased synaptic concentrations of these neurotransmitters could cause depression symptoms (Bunney *et al*, 1965; Schildkraut *et al*, 1965). Further, treatment with drugs that increased synaptic serotonin and norepinephrine would ameliorate depressive symptoms (Charney, 1998; Delgado *et al*, 1990; Miller *et al*, 1996). Other systems are also affected such as the corticotropin-releasing hormone, which affects the hypothalamus-pituitary-adrenal axis, and substance P, which is involved in the response to stress (Gold *et al*, 1984; Holsboer *et al*, 1984; Kramer *et al*, 1998; Nemeroff *et al*, 1984). Circadian dysregulation are also described in MDD (Kupfer *et al*, 1982), where sleep deprivation can result in a short-lived remission to depression.

Other environmental and demographic factors are also associated with the susceptibility to depression (reviewed in (Vink *et al*, 2008; Wong and Licinio, 2001)). Women and older individuals are more susceptible to depression. Prior depression is also a significant predictor, though it is unclear whether this is due to genetic susceptibility or chronicity. Social aspects like marital status (unmarried), lower socio-economic status, living alone or without a support

structure, and recent bereavement are also associated with greater risk. Current or past health is also strongly associated with depression, including disability, current poor health, new medical illness, and a history of medical illness. Cognitive impairment is also a strong predictor of risk for depression. Thus, a variety of factors increase the risk of depression and likely remission as well.

5.2 LATE-LIFE DEPRESSION (LLD)

There is a second peak of incidence of depression in late-life (first peak in youth) that is associated with greater risk of suicide, medical comorbidity, disability, and family caregiving burden (Katon *et al*, 2010; Mulsant *et al*, 2006; Nelson *et al*, 2013). In late-life, there are several age-related factors that may further influence symptoms. For example, comorbid anxious-depression is more prevalent in late-life (Chou, 2009). Aging individuals have greater illness burden as well cognitive decline or dementia - both of which increase susceptibility to depression.

Additionally, in late-life a novel mechanism may drive some subtypes of depression. The 'vascular depression hypothesis' states that cerebrovascular disease may further predispose, precipitate, or perpetuate depressive symptoms (Taylor *et al*, 2013). Cerebral perfusion deficits induce microbleeds and infarcts in the white matter tracts, and consequently result in the dysconnectivity of various brain regions, thus worsening both cognitive function and mood symptoms (Taylor *et al*, 2013). The white matter lesions are noticed as hyperintense regions on T2-weighted MRI images.

5.2.1 Treatment of LLD

The most common treatment of LLD is antidepressant pharmacotherapy. Treatment of depression often involves a trial and error process of multiple antidepressants before an effective regimen is found. Approximately 40-50% of patients fail to respond to initial pharmacologic treatments (Andreescu *et al*, 2011). Typically, for midlife MDD, a clinician needs 3-4 weeks to determine whether the current regimen will be effective, an interval which increases in LLD to approximately 6-8 weeks (Andreescu *et al*, 2011; Reynolds *et al*, 2006). This interval is associated with increased risk of suicide and dropping from care. This is one of the most challenging features of treatment in LLD, thus it is particularly important to detect early treatment markers (prior to the behavioral response period) that indicate future clinical improvement (Aizenstein *et al*, 2014).

While current treatments often improve symptom severity, achieving full remission and maintaining remission is more difficult and likely explains why depression has more years lived with disability than other disorders. Although approximately two-thirds of patients eventually respond to some antidepressant therapy, relapse rates are high (especially in late-life) (Andreescu *et al*, 2011). This is further complicated by increased side effects from antidepressants, which reduces compliance with treatment and thus likelihood to achieve remission of symptoms (Andreescu *et al*, 2011).

Previous studies have identified several biomarkers of treatment response (reviewed in (Aizenstein *et al*, 2014; Breitenstein *et al*, 2014)). Two biological predictors include the serotonin transporter gene (S allele) and the decreased rapid eye movement sleep latency which are both associated with poor response. Further, glucose metabolism in the subgenual and the dorsal anterior cingulate is associated with better response. There are a host of clinical variables

that are predictive of better response such as low medical burden, early symptom improvement, early age of onset, no sleep disturbance, lower pre-treatment depression and anxiety severity, low suicidal ideation, and previous response to antidepressants. Thus, a broad category of social, genetic, and neural markers are associated with remission. In the next section, we describe in greater detail the neural changes associated with treatment response in LLD.

5.3 NEURAL PREDICTORS OF RESPONSE TO PHARMACOTHERAPY

Currently, MRI is only used clinically to screen whether depression symptoms are related to structural atrophy or cerebrovascular disease (Botteron *et al*, 2012; Gelenberg *et al*, 2010). However, several studies have investigated the pre-treatment structural and functional MRI predictors of response to antidepressants in LLD or the MRI changes associated with successful pharmacotherapy.

There are several common structural neuroimaging features associated with resistance to treatment. The most common is the white matter hyperintensities (WMH) burden, which correlates with the overall vascular burden described in the vascular depression hypothesis (Taylor *et al*, 2013; Taylor *et al*, 2003). High pre-treatment WMH burden has been associated with poor response to antidepressant pharmacotherapy. This is one of the most consistent findings in the neuroimaging literature in LLD and further supports the notion that LLD, especially the vascular subtype, may have a different mechanism to remission. To further support this, some studies have also reported a trend increase in WMH burden in non-responders during the course of a trial (Sheline *et al*, 2010a). These results have been interpreted as a product of the

vicious pathophysiologic circle (vascular lesions trigger depression which increases inflammation and worsens the vascular burden).

A smaller literature implicates several other pre-treatment structural markers (reviewed in (Aizenstein *et al*, 2014; Breitenstein *et al*, 2014)). Low pre-treatment dorsal and rostral anterior cingulate volumes predict poor response to antidepressants, which have been functionally implicated in depression and are related to emotion reactivity and regulation. Similarly, lower dorsolateral prefrontal cortex volumes also correlate with poor response to pharmacotherapy and are thought to be associated with lowered cognitive control over emotions. Lower hippocampus volumes have also been shown to be associated with poor response, however its relation to depressive symptoms is less clear. While amygdala (emotion reactivity and memory) volume and activity has also been shown to be associated with depressive symptoms and response to antidepressants, this result is less robust in LLD. Fractional anisotropy (FA, which measures microstructural integrity of the white matter) has also been implicated in response, mainly that low FA in the frontal cortex and anterior cingulate predicts poor response – however it is unclear whether this is driven by WMH (since WMH burden tends to accumulate in the anterior and posterior cingulum). Further, much of the literature regarding pre-treatment structural predictors of response in LLD is mixed and it is unclear how specific these predictors are to individual antidepressants (since each of these studies varied in the course and antidepressant used).

The literature directly investigating LLD and functional neural markers is more limited, however some of the functional changes in mid-life may be important in understanding the context of changes in late-life. Resting state studies indicated that high pre-treatment amygdala-cingulate and insula-cingulate connectivity was predictive of poor response (Lui *et al*, 2011). Mainly limbic structures are strongly implicated. High pre-treatment default mode network and

low executive control network connectivity was associated with better response to antidepressants (McGrath *et al*, 2013). While presenting visual images of sadness, previous studies found that high anterior cingulate activation was predictive of better response and that during presentation of happy images, higher hippocampal activation was predictive of better response (Chen *et al*, 2007; Fu *et al*, 2008; Fu *et al*, 2007; Langenecker *et al*, 2007; Lemogne *et al*, 2010; Lisiecka *et al*, 2011). Limbic reactivity during emotional face matching has also been shown to predict response to antidepressants and greater orbitofrontal-cerebellar connectivity was predictive of better response during this task (Lisiecka *et al*, 2011). A meta-analysis of positron emission tomography using fluorodeoxyglucose further implicated the following: high subgenual cingulate and medial prefrontal cortex metabolism and low putamen as well as insula/inferior frontal gyrus metabolism were associated with better response (Fu *et al*, 2013). The amygdala however showed high heterogeneity – some showing high and others showing low reactivity in responders (Fu *et al*, 2013). These changes reflect altered resting state default mode network and executive control network connectivity (hypothetically related to ruminative aspects of depressive symptoms), altered limbic reactivity (high emotional response), and low activation in cognitive regions (low emotional control).

LLD studies have shown a similar pattern of altered cognitive and limbic networks. LLD studies showed alterations in pre-treatment resting state executive control network connectivity and default mode network connectivity (Aizenstein *et al*, 2009; Alexopoulos *et al*, 2012; Andreescu *et al*, 2013; Brassen *et al*, 2008; Wang *et al*, 2008a). One study reported decreased posterior cingulate-striatum connectivity following treatment response (Andreescu *et al*, 2013). Several studies during a wide set of tasks reported low prefrontal cortex activation that normalized after treatment.

The treatment of depression and depressive symptoms is complicated by individual differences in etiology, pathophysiology, and treatment response. Machine learning approaches use high volume data to create a computational model that considers a large number of features outside of what is typically considered. A clinician typically considers a host of features, including time of onset, single vs. recurrent vs. chronic depressive symptoms, severity of symptoms, whether the individual is in partial or full remission, presents with/without psychotic, catatonic or melancholic features, seasonal patterns, and whether the symptoms are associated with other medical illnesses. However, even the most experienced clinicians may draw from past and learned experiences but may not be able to interpolate from the most current research. A machine-learning model may be able to help identify and summarize a much larger host of features (including genetic, neuroimaging, surveys, as well as exercise and diet information) from a large dataset. These models can also incorporate new technological advances and their use in modern society (e.g., internet usage as a predictor). Producing such summaries may help further guide clinicians when making decisions on the best course of treatment. However, most models fail to produce highly desirable and/or reproducible effects, as they were not designed for use in prediction models.

We can design studies that are biased to high number of samples instead of towards non-useful markers. While structural MR markers are useful and may be predictive – they are not cost-effective, as it would require an MRI, which is expensive. However, functional measurements can be made using a wide variety of techniques that are more cost-effective (e.g., electroencephalogram or EEG). Surveys are a cheap and effective method to determine the current depressive state as well as a host of important variables (e.g., age, gender, socioeconomic status, education, etc.). Genetic sampling can be done and is cost-effective mainly as it need only

be done once, while other clinical measures can be performed easily (e.g., blood pressure as well as blood biomarkers). Thus using such designs, it may be possible to generate sophisticated models of response. In the next sections we present results of a large multi-modal LLD cohort.

6.0 INTRINSIC FUNCTIONAL CONNECTIVITY IN LATE-LIFE DEPRESSION (LLD): TRAJECTORIES OVER THE COURSE OF PHARMACOTHERAPY IN REMITTERS AND NON-REMITTERS

This chapter is a modified version of work that has been previously published in Molecular Psychiatry (Karim *et al*, 2016a). This work (including only a subset of the full data as this is what was available) was intended to fulfill the first half of aim 1 to investigate resting state connectivity changes in the brain following antidepressant treatment, how this differed between remitters and non-remitters, and if any acute changes (e.g., following a single dose of antidepressants) was observed. The paper is reprinted here (with permission from Nature Publishing Group).

6.1 ABSTRACT

Previous studies in late-life depression (LLD) have found that patients have altered intrinsic functional connectivity in the dorsal default mode network (DMN) and executive control network (ECN). We aimed to detect connectivity differences across a treatment trial among LLD patients as a function of remission status. LLD patients (N=37) were enrolled into a 12-week trial of venlafaxine and underwent five functional magnetic resonance imaging (fMRI) resting state scans during treatment. Patients had no history of drug abuse, psychosis,

dementia/neurodegenerative diseases, or medical conditions with known effects on mood. We investigated whether there were differences in three networks: DMN, ECN, and ASN (anterior salience network) connectivity as well as a whole brain centrality measure (eigenvector centrality, EVC). We found that remitters showed increases in ECN connectivity in the right precentral gyrus as well as decreases in DMN connectivity in the right inferior frontal gyrus and supramarginal gyrus. The ECN and DMN had regions (middle temporal gyrus and bilateral middle/inferior temporal/fusiform gyrus, respectively) that showed reversed effects (decreased ECN and increased DMN, respectively). Early changes in functional connectivity can occur after initial medication exposure. This study offers new data indicating that functional connectivity changes differ depending on treatment response and can occur shortly after exposure to anti-depressant medication.

6.2 INTRODUCTION

Treatment of major depression often requires multiple trials of medications before identifying an effective regimen. Forty percent of patients drop from care within the first month of treatment (Gaynes *et al*, 2009; Holtzheimer and Mayberg, 2011) (an important risk of incomplete response (Warden *et al*, 2007)), and for those who remain in treatment; over half do not respond (Trivedi *et al*, 2006). Although conventional methods of increasing dose and using augmentation strategies increase overall response rates (Trivedi *et al*, 2006), these trials require patients to endure prolonged episodes of depression. Failure to respond to treatment can increase suicide risk, contribute to worsening of medical co-morbidities, disability, cognitive impairment, and death (Katon *et al*, 2010; Mulsant *et al*, 2006; Nelson *et al*, 2013). Because depressed older

adults are at increased risk for all of these negative heath consequences, shortening the window from clinical presentation to effective treatment is particularly important.

Several prior functional MRI (fMRI) studies have identified potential biological correlates, or markers of mid- and late-life depression (Aizenstein *et al*, 2014). They suggest that depression is associated with changes spanning multiple resting state networks. Specifically, depression has been linked to changes within the executive control network (ECN), default mode network (DMN), and anterior salience network (ASN) (Aizenstein *et al*, 2014). We have defined these networks based on previous work by Greicius (Shirer *et al*, 2012). We utilized a region of interest (ROI) based connectivity approach.

Late-life depression (LLD) has been associated with decreased functional connectivity in the ECN (Alexopoulos *et al*, 2012). The left dorsolateral prefrontal cortex (dlPFC) is highly correlated with emotion regulation and often used as the ROI for ECN (Banks *et al*, 2007; Ochsner *et al*, 2012). The ECN is important for goal-directed behaviors and complex cognitive tasks such as working memory, cognitive control, and decision-making (Menon and Uddin, 2010). In LLD, poor cognitive control is often reported (Aizenstein *et al*, 2009; Alexopoulos, 2002) and ECN connectivity has been associated with certain features of executive dysfunction, including rigidity in processing information/learning (Aizenstein *et al*, 2006; Aizenstein *et al*, 2005), deficits in working memory, and attention and cognitive inhibition (Alexopoulos *et al*, 2012; Carter and van Veen, 2007).

Several studies in mid-life depression and LLD suggest that depression is associated with greater connectivity in the DMN (Andreescu *et al*, 2013; Lui *et al*, 2011). The midline posterior cingulate cortex (PCC) has been used extensively as a central node of the DMN (Damoiseaux *et al*, 2008; Fransson and Marrelec, 2008; Leech *et al*, 2012). Previous studies have shown that

greater DMN activity is associated with negative bias, increased self-referential thoughts, and rumination (Alexopoulos *et al*, 2012; Greicius *et al*, 2007; Gusnard *et al*, 2001; Hamilton *et al*, 2011; Marchetti *et al*, 2012; Sheline *et al*, 2010b). In mid-life depression it has been shown that PCC and ventro-medial prefrontal cortex (vmPFC) connectivity predicted rumination severity (Berman *et al*, 2011). Further, therapeutic effects of antidepressants are associated with decreased neural response to negative self-referential stimuli (Nejad *et al*, 2013).

Finally, greater functional connectivity in the ASN is associated with increased anxiety and somatization (Andreescu *et al*, 2015; Paulus and Stein, 2006). The right anterior insula (RAI) is a central node of the ASN, and has been shown to be more greatly activated (relative to the left) in studies of emotion reactivity and regulation (Feinstein *et al*, 2006; Klumpp *et al*, 2012; Paulus *et al*, 2003). The ASN is extensively connected with regions involved in motivation, reward, as well as salience (cognitive, homeostatic, or emotional) (Craig, 2009; Menon *et al*, 2010). Increased ASN connectivity has also been associated with interoceptive hijacking, which may represent the neural basis of increased anxiety and somatization described in LLD (Paulus *et al*, 2006; Simmons *et al*, 2013).

Whole brain networks were examined using eigenvector centrality (EVC), which identifies important nodes that are densely connected (Zuo *et al*, 2012). These nodes may play an important compensatory role in damaged networks (Binnewijzend *et al*, 2014), and they provide a measure of how central a node is within the brain (summarizing the number of connections and their relative strength). This metric is particularly responsive to acute exposure to selective serotonin reuptake inhibitor (SSRI) (Schaefer *et al*, 2014). Early changes in these networks might signal whether a treatment is likely to succeed.

By pairing fMRI scans with a pharmacological challenge, it is now possible to track whether/how brain activity changes in response to particular medications by looking at changes in functional connectivity after a single dose (Bourke and Wall, 2015). It is possible that early markers of circuit engagement, in response to LLD treatment, will help identify remitters with greater accuracy than pre-treatment imaging alone. This dynamic fMRI approach that can help refine current hypotheses regarding the correlation between treatment response and activity in functional circuits. Furthermore, by using early changes in brain activity, this early change can help predict clinical outcomes for individual patients.

The feasibility of fMRI markers is supported by recent studies showing functional imaging changes as early as 1–7 days after starting a new medication (Godlewska *et al*, 2012; Takahashi *et al*, 2005). Positron emission tomography (PET) studies have indicated similar potential: increases in monoaminergic occupancy rates are detectable after a single dose of an SSRI (Meyer *et al*, 2001; Parsey *et al*, 2006). However, no longitudinal study has examined dynamic functional connectivity changes that occur during an LLD treatment trial.

We investigated how changes in functional brain connectivity over a 12-week trial of venlafaxine differed between remitters and non-remitters. Patients underwent five resting state fMRI scans. We would expect that early in the treatment trial that the DMN and ASN would decrease in connectivity, while the ECN would increase (decreased rumination and anxiety, and increased cognitive control, respectively). We hypothesized that these early changes would be sustained until the end of the treatment trial.

6.3 METHODS

6.3.1 Study Design and Subjects

This project was part of a five-year multi-site study of treatment of LLD, which used venlafaxine in the first phase and then followed up with aripiprazole in non-remitters in the second phase. This was based on a study that found that augmentation of venlafaxine with aripiprazole improved treatment outcomes in treatment resistant patients (Rutherford *et al*, 2007). It was also chosen due to its dual mechanism of action (at low versus high doses). Participants were included if they were $\geq$65 years of age, meeting DSM-IV criteria for major depressive episode (non-bipolar, non-psychotic), with Montgomery-Asberg Depression Rating Scale (MADRS)$\geq$15 (Montgomery and Asberg, 1979). Exclusion criteria: history of mania/psychosis, alcohol/substance abuse within the last 3 months, dementia/neurodegenerative disease, and conditions with known effects on mood (e.g. stroke, multiple sclerosis, vasculitis, significant head trauma, and unstable hypertension and hypothyroidism). After informed consent, five MRI scans were performed: baseline, following the placebo lead-in (placebo), after first exposure to venlafaxine (day one), a week after beginning treatment (week one), and at the end (figure 1).

A total of 37 participants signed consent, but four were excluded due to venlafaxine side effects (N=2), non-adherence to protocol (N=1), and an inaccurate diagnosis of major depressive disorder (N=1). Thus 33 subjects were included in this analysis. All subjects completed the first four scans, but six failed to complete the fifth scan (but were included). Nine participants were on benzodiazepines (12 hour exclusion period prior to scanning) during the study (mean lorazepam dose=0.61 mg). There were no significant differences (p=0.19) of lorazepam dose

between remitters (N=4, 0.5mg) and non-remitters (N=5, 0.7mg). Four participants were on anti-hypertensive medications throughout the study.

Detailed dosage information has been published (Joel *et al*, 2014) and are available in the supplement. Patients were designated as remitters at 12 weeks if they had a MADRS≤10 for 2 consecutive weeks during the trial (Joel *et al*, 2014; Riso *et al*, 1997).

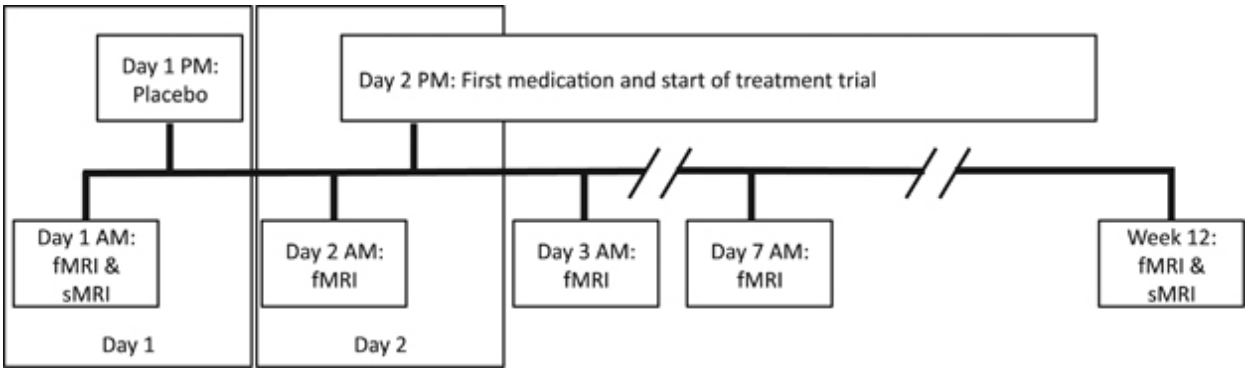

Figure 1. The study design protocol. Functional and structural MRI (fMRI and sMRI, respectively) was performed throughout the treatment period. All scanning was done in the morning. On day one, participants came in for an fMRI scan (Baseline) and then were given a placebo following the scan. On day two (~12 hours after placebo) they returned for another fMRI scan (Placebo) and then were started on venlafaxine following the scan. They returned the next day (~12 hours later) for another fMRI scan (Day One, i.e. day one of treatment). They continued on their medication as normal and came in for scans on week one (Week One) and at the end of the trial (End).

6.3.2 MRI Data Collection

Scanning was conducted using a 3T Siemens Trio TIM scanner located at the MR Research Center at the University of Pittsburgh. A high-resolution T1-weighted sequence was collected (TR=2300ms, TI=900ms, FA=9°) with a field of view 256x224 with 176 slices. $T2^*$-weighted BOLD acquisition using gradient-echo echoplanar imaging (EPI) was also collected (TR=2000ms, TE=34ms, in-plane resolution=128x128, 28 slices, voxel size=2x2x4mm^3). During resting scans, subjects (while awake, eyes open) observed a cross-hair.

6.3.3 Preprocessing

Data were preprocessed using statistical parametric mapping software (SPM12) (Penny *et al*, 2007). Functional volumes were first slice-time corrected then motion corrected. There were no significant differences between groups/time in mean relative motion and max absolute motion (see Supplement for descriptive statistics). Manual skull stripping was done, using ITK-SNAP (Yushkevich *et al*, 2006), to improve functional to structural coregistration. The stripped structural image was then co-registered to the mean functional volume.

The structural image was segmented using six spatial priors (including gray/white matter). This generated a deformation field that was applied to the functional images (Ashburner *et al*, 2005). Smoothing was applied using a Gaussian kernel with full-width half-maximum (FWHM) of 8mm.

6.3.4 Eigen-Vector Centrality (EVC) and ROI to Voxel Maps

Analyses were performed using in-house MatLab code.

Processing in both EVC and ROI to voxel analyses

We extracted a principal time-series from the white matter (WM) and cerebrospinal fluid (CSF) using singular value decomposition (SVD). We used these two signals and the motion parameters from the preprocessing in a multiple linear regression at each voxel. We extracted the residual time-series from each voxel, which represents the time-series not accounted for by WM, CSF, or motion. A band-pass filter (0.01-0.1Hz Butterworth) was applied. This pipeline was adapted from Whitfield-Gabrieli et al (Whitfield-Gabrieli *et al*, 2012).

Eigen-Vector Centrality (EVC)

A whole brain connectivity measure was calculated (EVC) (Binnewijzend *et al*, 2014; Joyce *et al*, 2010; Lohmann *et al*, 2010; Wink *et al*, 2012; Zuo *et al*, 2012). The matrix of covariate removed, band-pass filtered residuals across all voxels was put through an SVD. The principal Eigen-vector is the EVC measure. The matrix was centered then weighted by the inverse of the variance of each signal. In doing so, the SVD is done on the correlation rather than the covariance matrix. Z-scores were generated (mean zero and standard deviation one), smoothed, then masked for only gray matter.

ROI to Voxel

The signal within the ROI was correlated to each voxel. An SVD was performed to generate a principal time-series for the ROI. We computed the correlation between the ROI and all other voxels.

The Z-score map for these correlations was smoothed, then masked for only gray matter. This analysis was done for three separate ROI's. The PCC seed (DMN) was extracted from the posterior cingulate (eroded by hand in ITK-SNAP) from the Automated Anatomical Labeling (AAL) (Wu *et al*, 2011). The RAI seed (ASN) is extracted from the right insular cortex defined in the Automated Anatomical Labeling (AAL) atlas in the WFU Pick-Atlas. The left dlPFC (ECN) is defined as the left Brodmann area (BA) 46 in the Talairach Daemon database from the WFU Pick-Atlas. The network terminology used will reflect the terminology used in another study that performed an independent components analysis (ICA) (Shirer *et al*, 2012).

6.3.5 Statistical and Cluster Analysis

Statistical analyses were performed using SPM12 for each ROI connectivity and EVC maps. A repeated-measures ANOVA was performed containing the factors: group (response to treatment, 2 levels), time (5 levels, during treatment), an interaction between group and time, and a subject effect (models variability due to differences in average response of each subject).

In this study we assessed the significance of group, time, and group by time interaction effects. Permutation methods for peak-cluster level error correction (AlphaSim, http://afni.nimh.nih.gov/afni/) were applied for this whole-brain analysis by taking into account the significance of the peak voxel (p-value<0.005), thereby controlling for multiple comparisons (returning a minimum of 195 voxels). If the F-test was significant, we extracted the mean of each

significant cluster (as well as 99% confidence intervals, CI) and plotted that across the five time points for each group to examine trends within these significant clusters.

To show regional changes in connectivity, we performed four change score analyses for each of the significant interactions. We subtracted baseline connectivity from placebo, day one, week one, and end connectivity and performed a regression with two coefficients: a constant, and a grouping variable. Parameter estimate means (tests whether there is a significant difference in group) and 99%CI were extracted for each significant ROI and plotted.

6.4 RESULTS

Table 1 shows the clinical and demographic characteristics by group (remitters [N=20 (16F)] and non-remitters [N=13 (7F)]). We found no significant differences in any of the demographic or clinical measures (in the table) except for follow-up MADRS. We found no differences in white-matter hyperintensity (WMH) burden by group either at baseline or follow-up (see supplement for information on WMH segmentation/quantification(Wu *et al*, 2006)). The average venlafaxine dose (mean, 99%CI) in non-remitters was 263mg (227.3, 298.7), which was significantly greater (as expected; see supplement for titration information) than in remitters, which was 181.3mg (153.9, 208.7). There were no significant group/time or interaction effects in duration of depression and anxiety as measured by a single item in MADRS (see supplement).

Table 1. Clinical/demographic differences between groups. As designed, MADRS at end of trial differed between remitters and non-remitters. If the number of subjects is fewer in the analysis then the total, it is listed in parentheses. NOTE: MDE-Major Depressive Episode, CIRSG-Cumulative Illness Rating Scale for Geriatrics, MMSE-Mini-Mental State Examination, MADRS-Montgomery-Asberg Depression Rating Scale, WMH-White Matter Hyperintensity

	Non-Remitters (N = 13)	Remitters (N = 20)	Group Comparison (X/W,p)
Age (median, IQR)	65,6	66, 11	W = 126.5, p = 0.906
Gender (F)	7	16	Fisher's exact p= 0.139
Education (median, IQR)	15, 4	14, 5.25	W = 130.5, p = 0.992
Age at first MDE (median, IQR)	29, 15.25 (N=12)	29.5, 33.50 (N=18)	W = 109, p = 0.975
CIRSG Heart (0/1/2/3)	9/2/1/1	14/2/0/4	Fisher's exact p= 0.518
CIRSG Vascular (0/1/2)	4/0/9	4/1/15	Fisher's exact p= 0.810
MMSE Baseline (median, IQR)	29, 1	30, 2	W = 101, p = 0.273
MADRS Baseline (median, IQR)	26, 9	22, 8.75	W = 181.5, p = 0.058
MADRS End (median, IQR)	19.5, 10.5 (N = 12)	3, 5.5 (N = 19)	W = 211, p < 0.05 **
WMH Baseline (median, IQR)	0.0008, 0.0006	0.0011, 0.0015	W = 133, p = 0.9277
WMH End (median, IQR)	0.0011, 0.0012 (N = 12)	0.0013, 0.0017 (N = 19)	W = 100, p = 0.589

Only the ECN and DMN had significant group-by-time interaction effects. ASN and EVC had only significant group effects (remitters vs. non-remitters). All neuroimaging results are summarized in table 2. These results are robust to Benzodiazepine use and baseline MADRS. We demonstrate the associations of connectivity and features of clinical response and medication. There were group differences independent of time (excluding areas with significant interactions) in DMN and ECN connectivity.

Table 2. Resting state results summary table. X, Y, Z are the locations in MNI space. F is the maximum F-statistic within the cluster. Voxels is the size of the cluster. NS refers to Non-Significant results. If the Group x Time interaction is significant, then the main effects cannot be interpreted by themselves regardless of their significance. Since an interaction term is present (reaching statistical significance) that means that the relationship between the outcome variable and time is not the same for both groups. NOTE: ECN-Executive Control Network, DMN-Default Mode Network, ASN-Anterior Salience Network, EVC-EigenVector Centrality, NS-Not significant, NA-Not Applicable, BA-Brodmann Area

Network	GroupxTime Interactions	Time	Group	X	Y	Z	F	Voxels
ECN	Right Precentral/ Postcentral	NA	NA	63	0	12	16.5	251
	Right Middle Temporal/Occipital	NA	NA	48	-80	24	16.3	246
DMN	Right Inferior/Middle Frontal	NA	NA	44	24	18	16.3	670
	Left inferior/middle temporal gyrus/fusiform	NA	NA	-52	-62	-10	19.8	392
	Right inferior/middle temporal gyrus/fusiform	NA	NA	48	-36	-16	22	1407
	Right Supramarginal	NA	NA	60	-58	36	22.3	297
ASN	NS	NS	Left Inferior Frontal Gyrus	-38	6	24	15.9	240
	NS	NS	Left Middle Frontal Gyrus	-32	54	20	12.6	240
EVC	NS	NS	Left Inferior Frontal Gyrus	-56	6	28	13.2	221
	NS	NS	Right Inferior Frontal Gyrus	52	34	-12	16.1	203
	NS	NS	Medial Frontal Gyrus/BA 10	2	64	-8	15.5	713

6.4.1 Executive Control Network (ECN)

The regions with a significant group-by-time interaction (after multiple comparison correction) were the right precentral/postcentral gyrii (PCG) and the right middle temporal/occipital gyrii (rMTG/MOG), $p<0.05$ (corrected), see table 2 and figure 2A. The 99%CIs suggest no differences between remitters/non-remitters (figure 2A). The change score analysis (figure 2C, left) illustrates, relative to baseline, a larger change in connectivity following treatment than placebo. Across time rPCG increased in connectivity while rMTG/MOG decreased.

6.4.2 Default Mode Network (DMN)

Four clusters had significant interactions, they were the right inferior/middle frontal gyrus (rIFG/MFG), bilateral inferior/middle temporal gyrus/fusiform gyrii (bITG/MTG), and right supramarginal gyrus (rSMG), $p<0.05$ (corrected), see table 2 and figure 2B. Much like the ECN, the 99%CI suggests no differences between remitters and non-remitters at any time point. The 99%CI suggests that, relative to baseline, there is a larger change in connectivity following treatment than placebo (figure 2C, right). Across time bITG/MTG increased in connectivity while rIFG and rSMG decreased in connectivity in remitters.

6.4.3 Anterior Salience Network (ASN)

After applying the multiple comparison correction, no regions had significant group-by-time interaction effects. We then ran a model without the interaction effect and tested whether there were significant group and time effects. There was no significant time effect, but there were

significant group effects in the left inferior frontal gyrus (lIFG) and left middle frontal gyrus (lMFG), $p<0.05$ (corrected), table 2 and figure 3A.Non-remitters had greater ASN connectivity in both regions.

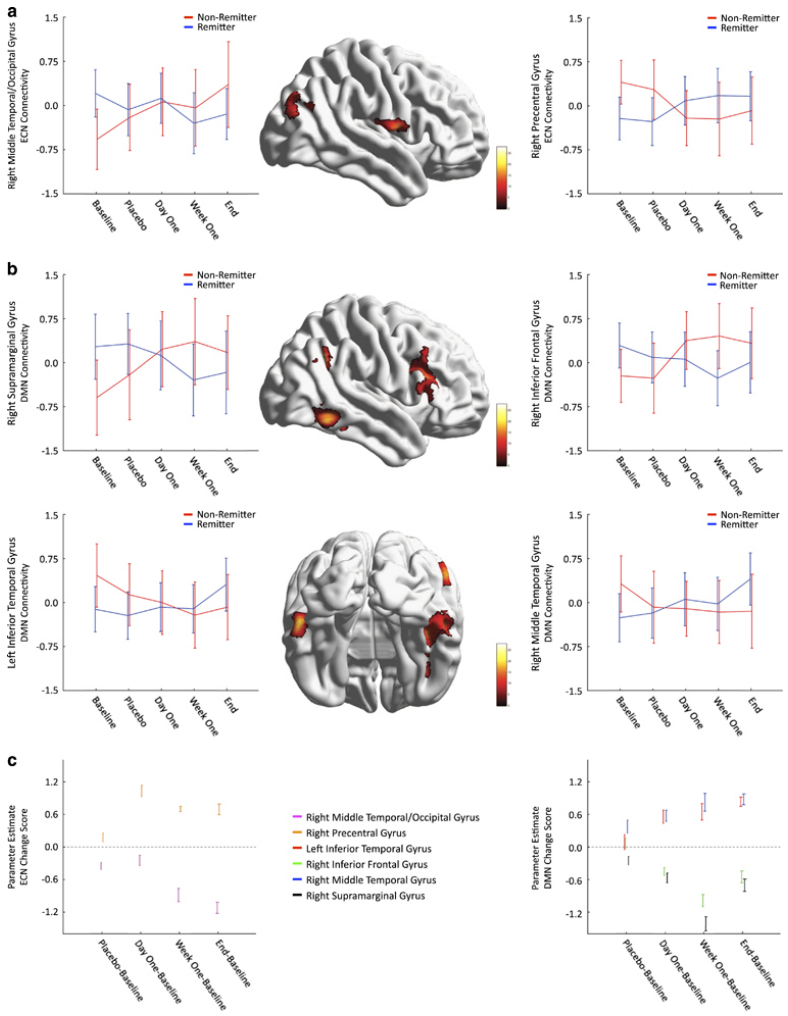

Figure 2. Connectivity changes where the interaction (group x time) was significant. (A) ECN connectivity changes that were significant. (B) DMN connectivity changes that were significant. For parts A and B, Non-remitters are shown in red and remitters are shown in blue. The color bar indicates the value of the F-statistic. Error bars represent the 99%CI. (C) Change score analysis results. Different regions are shown as different colors. The values represent mean and 99%CI for the parameter estimate that tested whether there was a significant difference between remitters/non-remitters in the change scores (placebo/day one/week one/end – baseline). Dotted line represents beta estimate of zero.

6.4.4 Eigen-Vector Centrality (EVC)

Eigen-vector centrality is a summary measure of the influence of a node (voxel) in a network. No interaction between group and time was found for the EVC. However, there was a significant effect of group (but not time) in the bilateral inferior frontal gyrus (bIFG) and the medial frontal gyrus (MeFG), $p<0.05$ (corrected), table 2 and figure 3B. Non-remitters had greater EVC in the bIFG but lower EVC in the MeFG compared to remitters.

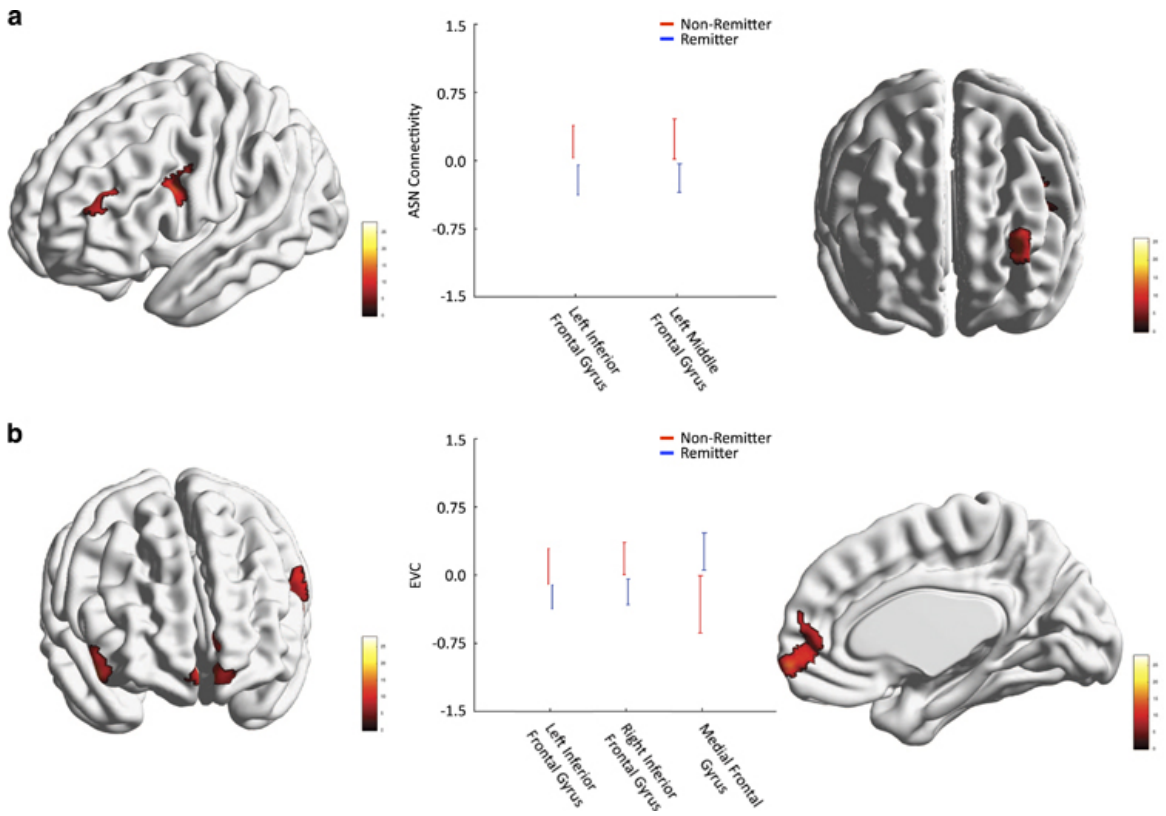

Figure 3. Group differences in connectivity. Analyses where the interaction (group x time) was not significant, but where the group effect alone (not the time effect) was significant. (A) Regions where the ASN connectivity differed between remitters (blue) and non-remitters (red). (B) Regions where the EVC measure differed between groups. The color bar indicates the value of the F-statistic. Error bars represent the 99%CI.

6.5 DISCUSSION

This is the first study reporting early dynamic fMRI markers of treatment response variability in LLD. We evaluated changes in three functional networks and in EVC at five time-points. Two networks (ECN and DMN) showed significant group-by-time effects (increased ECN-rPCG and DMN-bMTG/ITG as well as decreased ECN-rMTG and DMN-rIFG and rSMG in remitters across the trial compared to non-remitters). Only significant group (but not time) effects were found in the ASN (left IFG and MFG greater in non-remitters compared to remitters) and EVC (MeFG greater in remitters compared to non-remitters but lower in the bIFG).

Previous LLD research suggests patients, compared with controls, have a hyperactive DMN and a hypoactive ECN (Aizenstein *et al*, 2014; Alexopoulos *et al*, 2012; Andreescu *et al*, 2013). These may reflect clinical features of LLD such as increased rumination (hyperactive DMN) and cognitive impairment indicating low cognitive control of limbic regions associated with emotional response (hypoactive ECN). A meta-analysis in mid-life depression found that DMN connectivity was predictive of treatment response (Nejad *et al*, 2013; Pizzagalli, 2011). Another study found that DMN connectivity was positively associated with treatment response while dlPFC connectivity was negatively correlated (Aizenstein *et al*, 2014). Other studies have found a normalization of task-based response following successful treatment (Aizenstein *et al*, 2014). Our novel findings demonstrate, for the first time, that these effects are seen early following treatment, and appeared larger in magnitude than placebo.

In remitters we observed increased ECN-rPCG connectivity and decreased ECN- rMTG connectivity relative to non-remitters. While there is an effect of placebo, there appears to be an even greater effect following administration and continued treatment with venlafaxine. This suggests that the change in connectivity is related to the administration of venlafaxine and not to

placebo. ECN-rPCG increases in remitters may reflect an improvement in cognitive control as a predictor of successful treatment. ECN-rMTG decreases did not show a large change following first exposure (day one) to venlafaxine; rather this change is seen at week one. ECN-MTG (outside the ECN) connectivity changes may indicate increased and dispersed effort in the non-remitters.

In remitters we observed decreased DMN-rSMG and rIFG connectivity and an increase DMN-rMTG/lITG/MTG and left fusiform gyrii connectivity relative to non-remitters. Like the ECN, the magnitude of the connectivity change appeared to be greater following treatment than following placebo. Decreased DMN-rSMG/rIFG connectivity may reflect an improvement in future ruminative thought processes in remitters, as suggested previously (Aizenstein *et al*, 2014; Alexopoulos *et al*, 2012). Increased DMN-rMTG/lMTG connectivity suggests that clinical correlates of neural changes (rumination–hyperactive DMN) are actually related to connectivity changes between specific nodes (PCC-prefrontal cortex). Thus, we may witness a "rebalance" of the DMN in remitters, with a decrease in the "damaged" PCC-prefrontal connectivity and an increase in the connectivity between the other nodes.

Of note, the supramarginal gyrus has been involved (together with other sensory processing/associative brain regions, such as the fusiform gyrus), in the disrupted DMN connectivity in mid-life depression (Chen *et al*, 2015; Peng *et al*, 2015). This may reflect disruptions in social interaction processes such as empathy (Shamay-Tsoory, 2011) and social engagement (Li *et al*, 2014), which may ameliorate with improvement in depression symptoms. With regard to changes in PCC-IFG connectivity, we may speculate that given the recent reports regarding the role of right IFG in cognitive control but also in emotional appraisal and alexithymia and verbalization of emotional responses/states (Khalaf *et al*, 2016), we may infer

IFG, as a key region in the emotion-cognition interplay (Okon-Singer *et al*, 2015) becomes less involved during resting state, once depressive symptoms remits.

Alternatively, these results could be interpreted as increased intra-network coupling (increased ECN-rPCG and DMN-bMTG/ITG) and decreased inter-network coupling (decreased ECN-rMTG and DMN-rIFG and rSMG). In healthy individuals, ECN and DMN have inverse activations during tasks and this is disrupted in depression (Chen *et al*, 2013; Menon *et al*, 2010; Sridharan *et al*, 2008). This may reflect an important rebalancing of this association in remitters. These temporal regions are not nodes of the dorsal but rather ventral DMN and rSMG is part of the right ECN.

Recent evidence that shows that changes in DMN/ECN connectivity as well as other functional brain activation can be achieved through meditation, trans-cranial magnetic stimulation (TMS), cognitive behavioral therapy, and psychotherapy (Brewer *et al*, 2011; Farb *et al*, 2012; Goldapple *et al*, 2004; Jang *et al*, 2011; Linden, 2006; Liston *et al*, 2014). These different therapies target different symptoms of depression and by targeting affected symptoms (e.g., high rumination) then it might be possible to achieve these changes through alternative means.

The early interaction may reflect a network engagement due to the increase in synaptic serotonin that seems to be consistently engaged (relative to the end scan). Thus, it seems that the network changes occur at a much earlier stage and these may be correlated with future changes in depression severity, rumination, and cognitive control (though we do not demonstrate that here).

ASN-left IFG and MFG connectivity was higher in non-remitters than remitters. Previous studies reported higher ASN connectivity in LLD participants compared with non-depressed

elderly (Aizenstein *et al*, 2014), a possible marker of increased anxiety and somatization (Andreescu *et al*, 2015). Given the lack of time differences, this may represent a trait, rather than a state marker in LLD.

Using EVC (measures node importance), we found only group effects where EVC was higher in remitters than non-remitters in the MeFG, but lower in the bIFG. These findings suggest a potential neurobiological profile indicating positive response to treatment. Thus, participants who start with high connectivity in the DMN (and increased EVC in the MeFG) are more likely to respond to treatment. This will require further empirical testing.

Several limitations should be noted. This study had a relatively small sample size, unequal group sizes, and tested treatment response using only one medication. This result may not generalize well to other patient groups, including mid-life depression. Our definition of remitter, while established, has important limitations especially in borderline cases. A well-known observation in LLD is that white-matter hyperintensity burden differs between remitters and non-remitters (Taylor *et al*, 2013), which we failed to replicate, possibly due to the clinical and neurobiological heterogeneity of LLD (Taylor *et al*, 2013). This study utilized ROI based connectivity whereas others have utilized data-driven approaches. Importantly, there is a strong correspondence between the two methods (Rosazza *et al*, 2012). We limited our analyses to three ROI's that represented core nodes of the default mode, executive control, and salience networks-however each of these networks has multiple nodes that we did not explore. All participants had similar dosages of venlafaxine at all measurements except the final, where non-remitters had significantly greater mean dose than remitters. This was not controlled for in this analysis, and may account for some differences at the final time-point between remitters/non-remitters. Importantly, the dosage was equivalent over the course of the early changes (early interactions).

While there exists a literature that associates DMN/ECN connectivity with rumination/cognitive control measures, we did not specifically test this, and so future studies should perform these direct associations to validate these interpretations.

These group differences in trajectory of treatment may be important in predicting changes in depression symptoms, however group differences do not necessarily give the ability to distinguish individual subjects.

Despite these limitations, we validate previous findings of pre- and post- treatment effects. Further, we found that there were early changes in the DMN and ECN, but not ASN during the treatment trial and that the treatment was associated with greater magnitude of change than placebo. Future studies should test if an inter-network interaction between ECN and DMN exists, and investigate other nodes of each of these networks, as well as investigate the structural changes that may occur during the entire treatment trial.

7.0 FUNCTIONAL BRAIN ACTIVATION DURING EMOTION REACTIVITY FOLLOWING PHARMACOTHERAPY IN LATE-LIFE DEPRESSION: MARKERS OF REMISSION

This chapter reports unpublished results using the full cohort of the same study described earlier. This chapter is intended to fulfill the second half of Aim 1 to characterize the changes in emotion reactivity in the brain following pharmacotherapy in remitters/non-remitters to depression, whether any acute changes (e.g., following a single dose) occur. To better understand the results in this chapter, we also investigated how these results related to other neuroimaging data collected in the same sample (structural and functional changes).

7.1 ABSTRACT

Major depressive disorder has a second peak of incidence in late-life (LLD), which is associated with an additional host of negative health outcomes. Despite the vast number of studies on depression, there is currently no accepted biomarker for the diagnosis, prevention, or treatment of depression. Neuroimaging data has shown that there exist small changes in functional activation/connectivity following acute pharmacotherapy, which may be associated with eventual response. We investigated changes in functional activation during an emotion reactivity task following acute pharmacotherapy as well as over the entire course of a treatment trial. We

recruited LLD (N=51) participants into a treatment trial and collected functional magnetic resonance imaging (fMRI) data at five time points: baseline, following a placebo lead-in, following a single dose of venlafaxine, following a week of pharmacotherapy, and the end of the trial (12 weeks). We found that there existed baseline differences in activation, but more importantly that there existed acute increases following only a single dose in the left insula in remitters while non-remitters showed only decreases in activation. Further, we found that the parahippocampus increased in activation following a single dose in remitters, but decreased in non-remitters and this pattern persisted through the entire trial. We found similar changes during an explicit emotion regulation task – thus these changes may reflect an early change in implicit processing and regulation during the emotion reactivity task, but a chronic change in explicit regulation. Emotion regulation may represent a mechanism for remission in LLD.

7.2 INTRODUCTION

Major depressive disorder (MDD) is a leading cause of disability and global disease burden (Alexopoulos *et al*, 2009; Moussavi *et al*, 2007). Depression has a second peak of incidence in late-life, which carries additional risk of suicide, medical comorbidity, disability, and caregiver burden (Katon *et al*, 2010; Mulsant *et al*, 2006; Nelson *et al*, 2013). Despite significant improvements in our understanding of course, prognosis, and the neurobiology of depression, new biomarkers and treatment developments have lagged. There are currently no widely accepted neural or genetic biomarkers to aid in the diagnosis, treatment, or its long-term management. This likely reflects the notable heterogeneity underlying both depression pathophysiology and remission mechanisms. Currently, clinicians match patients with specific

treatments through a prolonged trial and error process that delays improvement, and significantly increases the overall burden of illness. This delay is even longer in late-life depression (LLD) and is further associated with a host of negative health outcomes, including higher risk of suicide, cardiovascular disease and cognitive deterioration (Andreescu *et al*, 2011). Previous work using functional magnetic resonance imaging (fMRI) has identified some possible biomarkers that act as predictors of treatment response. To better understand this process, researchers have investigated different neural circuitry related to MDD.

7.2.1 MDD: Disruption of Emotional Face Processing

Some of the core neural changes in depression are associated with the emotion-reactivity and regulation neural circuitry. Low mood and high anxiety have been consistently associated with emotion dysregulation. The processing of emotional faces in MDD has been often used to explore different aspects of emotion dysregulation. This general process involves three major stages, each subserved by fairly distinct neural structures: visual processing [primary (visual area 1/2) and secondary (fusiform, superior temporal gyrus)], emotion encoding and recognition [amygdala, insula] and response/appraisal [insula, orbitofrontal cortex, ventral striatum] as well as integration [insula and anterior cingulate], and finally monitoring affective state and emotion regulation [anterior cingulate, dorsomedial prefrontal cortex, dorsolateral prefrontal cortex] (Leppanen, 2006).

7.2.2 High Emotion Reactivity

Previous studies have identified changes in emotional face processing in MDD, including a bias towards negative stimuli in the initial appraisal stage. However, a meta-analysis that included 44 studies (with a total of 795 MDD and 792 never-depressed individuals) revealed hyperactivation beyond the amygdala and the insula, in regions like the parahippocampus, putamen, insula, and fusiform gyrus, while the dorsolateral prefrontal cortex has been consistently hypoactive (Groenewold *et al,* 2013).

Each of these regions may have a distinct role in further amplifying the neural dysfunction. The amygdala's role is to direct attention at emotional information, facilitate emotional memory, and generate responses to emotionally salient information. The fusiform's early role in the visual processing stream and its hyperactivation indicates that the negative bias may be encoded semi-automatically (i.e. biased visual-limbic feedback loop). Anterior cingulate and anterior insula may be involved in working alongside the amygdala to generate the relationship between external/internal stimuli and the self, where the anterior cingulate specifically is part of the dorsal default mode network which is involved in the ruminative aspect in MDD. Generally, they could be involved in the attendance to negative stimuli (as opposed to neutral stimuli). Insula lesions following stroke have been shown to be consistently associated with post-stroke depression implicating them even further (Sprengelmeyer *et al*, 2011). The putamen, however, may be related to the automaticity of the learned emotional response to the visual stimuli. Parahippocampus may be involved in the contextual information related to the stimuli as past studies have identified that the amygdala's activation can be modulated by presenting contextual information alongside an emotional face (e.g. decreased activation achieved by presenting an angry face with the sentence "they just had a bad day and are not upset

at you") (Bar and Aminoff, 2003; Groenewold *et al*, 2013). This may be also be related to inaccurately contextualizing non-salient information (e.g. picture of an angry face in the MR scanning environment).

7.2.3 Impaired Emotion Regulation

Another aspect that is disrupted in MDD is the *monitoring* and *regulation* of the emotional response, a feature associated with changes in the insula, anterior cingulate and dorsolateral prefrontal cortex activation. These changes are implicated in the inability to detect and then down-regulate the amygdala's response. Emotion regulation is thought to have two basic forms (though there is no clear boundary): explicit (or effortful) and implicit (or automatic) regulation (Etkin *et al*, 2015; Gyurak *et al*, 2011). Regulation is a goal-directed process that influences the intensity, duration, or type of experienced emotion and this can be done with conscious effort or without monitoring, insight, and awareness (Etkin *et al*, 2015). Explicit regulation involves the activation of dorso-/ventro-lateral prefrontal cortex as well as supramarginal gyrus, insula, supplemental motor, and pre-supplemental motor areas (Etkin *et al*, 2015). While behaviorally, it has been shown that MDD individuals are able to explicitly regulate their emotional response (i.e. feeling less negative after regulation); there is a clear difference in the neural response. Mainly, while both groups have decreased amygdala activation following regulation – the level of dorsolateral prefrontal cortex activation is severely reduced in MDD (Erk *et al*, 2010). Critically, amygdala down-regulation was associated with lower depression severity (i.e. severity of depression influences the *intensity* of the down-regulation in amygdala activation, which may be due to greater emotional reactivity or decreased cognitive regulation) (Erk *et al*, 2010). This may reflect this notion that individuals with MDD are still able to explicitly regulate.

More recently, implicit emotion regulation has been implicated in MDD. Even without instructing someone to explicitly and actively regulate emotions – there is an automatic underlying process that occurs with the presentation of any new stimuli (Gyurak *et al*, 2011). Several studies have shown that when presenting a Stroop like emotional face viewing paradigm (e.g. matching expressions when an angry face with either the word "angry" or "happy" presented simultaneously) showed that there is increased ventral anterior cingulate and dorsolateral prefrontal cortex activity and lower amygdala activation on incongruently presented faces (e.g. angry face with the word "happy" written over it) – suggesting an implicit regulation of emotional responses (Etkin *et al*, 2006; Etkin *et al*, 2010). To further strengthen this, when matching genders of neutral faces (with conflicting gender text written over the face) – there is an increase in dorsolateral prefrontal cortex activation coupled with fusiform activation, but no change in ventral anterior cingulate or amygdala (Egner *et al*, 2008). This suggests a dissociable neural pathway for implicit emotion regulation.

Habitual emotion regulation is a form of implicit regulation that is performed daily and involves regulating small aspects of daily life (Gyurak *et al*, 2011). This has been demonstrated in studies that found that individuals with higher reappraisal tendencies (ability to reimagine a scenario in a more positive way) were able to better regulate negative emotions during anger provocation (Mauss *et al*, 2007). The ventral anterior cingulate has also been implicated in implicit regulation – however the anterior insula is thought to play a major role. Previous studies have found that during a task where participants are presented with fair or unfair offers the insula's activation is associated with the degree to which individuals apply reappraisal strategies daily (Gross and John, 2003). This type of regulation is likely to be associated with greater affect, better interpersonal and social functioning, and overall better well-being (Gyurak *et al*,

2011). This is a possible mechanism through which individuals with MDD are impaired and a possible mechanism of remission in MDD.

7.2.4 Functional Changes Following Treatment

Past studies have shown that a variety of regions are normalized during emotional face processing following successful remission. Following successful remission, there is reduced activation of the amygdala (normalization) (Arce *et al*, 2008; Bigos *et al*, 2008; Godlewska *et al*, 2012). Elevated baseline anterior cingulate activity has been associated with greater response to antidepressant medications (Davidson *et al*, 2003; Wang *et al*, 2012). One study showed an increase in activation of the middle frontal gyrus following fluoxetine treatment (Wang *et al*, 2012). Another study showed that there was an increased insula activation following treatment with venlafaxine during a negative image viewing task and further replicated the anterior cingulate finding at baseline (Davidson *et al*, 2003). They showed that the change in insula activation occurred after only two weeks of treatment and sustained the activation – however the anterior cingulate changes occurred at a later stage (8 weeks) (Davidson *et al*, 2003). Another study showed normalization of amygdala reactivity following escitalopram treatment after only 7 days (Godlewska *et al*, 2012).

In this study we sought to investigate whether this change could be detected as early as following a single dose of medication. Positron emission tomography has shown that there is an increase in monoaminergic occupancy in the brain following a single dose of selective serotonin reuptake inhibitors (SSRI) (Meyer *et al*, 2001; Parsey *et al*, 2006). Several studies have indicated the early functional changes occurring following acute drug administration, including our recent work that showed single dose connectivity changes at rest following a single dose of venlafaxine

(Karim *et al*, 2016a). This study indicated a possible early engagement effect that is either sustained or diluted. In a previous study in never-depressed individuals using a double blind balanced crossover design, single dose or acute administration of citalopram (following a 30 minute injection) compared to saline resulted in *increased* activation of the amygdala during emotional face viewing (Bigos *et al*, 2008). While studies investigating the more chronic effects in never-depressed individuals showed a clear decrease in activation in the amygdala and insula, this may reflect an early engagement effect that is down regulated chronically.

We investigated early effects of venlafaxine in LLD as well as long-term changes in remitters (N=26) and non-remitters (N=25) to depression. We used an emotional face-viewing paradigm, and investigated activation differences at baseline (between remitters and non-remitters), the early changes following a single dose as well as after just a week of treatment, and then at the end of the trial. The face/shapes task is an emotion reactivity task, which likely has an implicit regulation component that involves regulating to some degree the negative faces presented. We investigated activation differences between remitters and non-remitters at five time-points: at baseline, following a placebo lead-in, following a single dose of venlafaxine, after a week of treatment, and at the end of the trial. We hypothesize that the depression-specific changes in activation during emotion reactivity have correspondence in changes associated with emotion regulation as well as with regional structural and CBF changes in the same regions involved in emotional face processing.

7.3 METHODS

7.3.1 Study Design and Participants

We collected data as part of a larger 5-year multi-site study of treatment in LLD that collected neuroimaging data at one site (Pittsburgh, USA). Participants were recruited and were given Venlafaxine due to its dual mechanism of action (at high doses it is both a serotonin *and* norepinephrine reuptake inhibitor). Participants were included if they were at least 55 years old, met *Diagnostic and Statistical Manual of Mental Disorders IV* criteria for MDD and had a Montgomery-Asberg depression rating scale (MADRS) score of 15 or higher at baseline. Participants were excluded if they had a history of mania or psychosis, alcohol or substance abuse (within last 3 months), dementia or neurodegenerative disease as well as conditions with known effects on mood (e.g. stroke, multiple sclerosis, vasculitis, significant head trauma, and/or unstable hypertension). After informed consent approved by the University of Pittsburgh institutional review board, five MRI scans during the treatment trial were collected.

All MRI scanning was conducted in the morning. Participants came in on the first day for a baseline scan (no medication). In the evening of that scan they were given a placebo, after which they returned the next day for another scan (placebo scan). The evening of that scan, they were given their first dose of Venlafaxine (35 mg), after which they returned the next day for another scan (single dose scan). They continued their medication for approximately one week and returned for another scan (week one scan). They returned a final time after the end of the treatment trial (12 weeks, end scan).

During the trial, participants returned for weekly or bi-weekly clinical visits and the Venlafaxine dosage was increased as necessary (up to a maximum of 175 mg). Participants who

did not show signs of response by week 6 had venlafaxine increased up to a maximum of 350 mg. At the end of the study, participants were classified as remitters if they had a MADRS less than 10 for at least two weeks during the trial (and remained so until the end of the trial).

A total of 62 participants signed consent. Eleven were excluded due to: side effects of medication (N=2), non-adherence to protocol (N=2), inaccurate diagnosis of MDD (N=1), and inability to determine remission status due to lost/missing data (N=6). Among the remaining data (N=51), two participants did not complete all MRI scanning but did complete the treatment trial. Wherever possible we included the data from these two participants.

7.3.2 MRI Data Collection

All scanning was conducted at the University of Pittsburgh Medical Research Center on a 3T Siemens Trio TIM scanner (Munich, Germany) on a 12-channel head coil. Baseline and end scans collected both a structural and functional image, while other scans collected only functional sequences. While this analysis focuses solely on the emotion reactivity task (face/shapes), our hypothesis involves understanding the structural as well as other functional changes that occur concurrently in those regions. Thus we also describe the collection and processing of the following: resting state as well as explicit emotion regulation task (IAPS) BOLD sequences, a resting pseudo-continuous arterial spin labeling (pCASL) sequence, and a diffusion weighted imaging (DTI) sequence.

An axial, whole brain 3D magnetization prepared rapid gradient echo (MPRAGE) was collected with repetition time (TR)=2300ms, echo time (TE)=3.43ms, flip angle (FA)=9 degrees, inversion time (TI)=900ms, field of view (FOV)=256x224, 176 slices, 1mm isotropic resolution and with GeneRalized Autocalibrating Partial Parallel Acquisition (GRAPPA) factor=2. An

axial, whole brain 2D fluid attenuated inversion recovery (FLAIR) was collected with TR=9160ms, TE=90ms, FA=150 degrees, TI=2500ms, FOV=256x212, 48 slices, and 1x1x3 mm resolution.

An axial, whole brain (excluding cerebellum) echo planar (EPI) T2*-weighted functional image was collected to measure the blood oxygen level dependent (BOLD) response with TR=2000ms, TE=34ms, FA=90 degrees, FOV=128x128, 28 slices, 2x2x4 mm resolution. The face/shapes task had 117 volumes, the explicit emotion regulation task had 270 volumes, and the resting state had 150 volumes. Due to variability in placement by MR technicians the coverage of the functional scan was in general limited to above the cerebellum and below the top aspect of the motor cortex (though this varied slightly between functional sequences).

An axial, whole brain (excluding cerebellum) pCASL sequence was collected at rest to measure perfusion in the brain with TR=4ms, TE=13ms, FA=90degrees, FOV=64x64, 32 slice, 4mm isotropic resolution, and 80 volumes. Finally an axial, whole brain DTI sequence was collected with TR=5300ms, TE=88ms, FA=90degrees, FOV=128x128, 40 slices, 2x2x3mm resolution, 12 directions, and 4 b_0 images.

7.3.3 Functional Tasks

Emotion Reactivity (Faces-Shapes Task)

This task is widely used and has been tested to robustly activate the amygdala (Hariri *et al*, 2002). Participants were instructed to match either a face cue or a shapes cue. A cue was shown on the center of the screen and they were instructed to respond with an MR-compatible glove (left or right index finger) by matching to one of two simultaneously presented faces. The facial expressions shown were either angry or fearful. During the shapes, they match a shape to one of

two simultaneously presented shapes. The shapes task (5 blocks) was interleaved with the faces task (4 blocks) and each block lasted 24 seconds containing 6 trials (4 seconds each). Before the beginning of each block participants are instructed visually to “match emotion” or “match form” (2 seconds). The faces images are presented from a set 12 different images (six per block, three of each gender) and are all derived from a standard set of pictures of facial affect. Stimulus presentation and responses were controlled using E-prime software (Psychology Software Tools, Inc., Pittsburgh).

Explicit Emotion Regulation Task (IAPS)

The main results of this task have been previously published in a smaller subset. Participants were shown emotionally neutral or negative images from the standardized International Affective Picture System (IAPS) and were instructed to either “Look” or “Decrease.” During the look instruction, participants were to view content naturally. During the decrease instruction, participants were instructed to reappraise the image to actively alter the elicited emotion. After each image they were asked to rate how negatively they felt from 1 to 5. The neutral (11 events), negative (15 events), negative regulate (15 events) conditions were interleaved and each event lasted 6 seconds. The images are presented from a set of images and stimulus presentation and responses were controlled using E-prime software (Psychology Software Tools, Inc., Pittsburgh). A master level instructor instructed participants on how to reappraise prior to entering the scanner.

Instructions for Resting State during pCASL and BOLD

The following data was used to further understand the changes occurring during the emotional reactivity task. The results of these data have been previously published in a smaller subset. During resting state perfusion and BOLD, participants were instructed to lie awake in the scanner while viewing a white cross hair.

7.3.4 Structural Processing

All processing was conducted using statistical parametric mapping (SPM12) (Penny *et al*, 2011). Interpolation was conducted using 4^{th} degree B-spline interpolation, normalized mutual information similarity metric for coregistration between images of different types, and mutual information similarity metric for motion correction unless otherwise stated. The FLAIR was coregistered to the MPRAGE (affine transform). Both images were input into a multi-spectral segmentation, which (after bias correction) segmented them into gray, white matter, cerebrospinal fluid, air, soft-tissue, and air. Due to high white matter hyperintensity burden the number of Gaussians used to identify white matter was two (which improved the segmentation) (Karim *et al*, 2016c). This process generates a deformation field that can be used to normalize other images to a standard anatomic space (Montreal Neurological Institute, MNI). An automatic mask for the intracranial volume was generated by thresholding the intracranial tissues with a probability of 0.1, filling the mask (imfill), and then performing a morphological closing operation (imclose, sphere of one voxel) in MatLab (MATLAB2016b, The MathWorks Inc., Natick, MA, 2000). This mask (intracranial volume, ICV) was applied to the MPRAGE to remove non-brain tissues (which improves functional-structural coregistration). The skull-

stripped MPRAGE was normalized to MNI space. An average of all baseline structural images was generated to overlay all functional imaging results.

To generate gray matter density images, we used DARTEL (Diffeomorphic Anatomical Registration using Exponentiated Lie algebra) (Ashburner, 2007). This leveraged the longitudinal data by first creating a subject specific template and then a study specific template and has been previously described in detail. After segmentation, we created a single template for each subject using DARTEL, which improved the coregistration between baseline and end structural data *within* a single subject. Those templates were then used to generate a study specific template *across* subjects. Briefly, this method iteratively creates averaged templates that slightly improve the coregistration process. This is thought to be important when calculating gray matter density maps especially in late-life studies that have greater gray/white matter deformations. The Jacobian of the transformations is multiplied by the final probability to generate a gray matter density image (instead of a probability). This is because the warping of tissue increases or decreases the actual density and needs to be adjusted for (e.g. thin cortical regions that are expanded to a larger template will have *lower* density as the gray matter was stretched) (Ashburner, 2007). The gray matter density images were smoothed using a Gaussian kernel of full-width at half-maximum (FWHM) of 6mm. The mean gray matter density was extracted from regions that we found in voxel-wise analyses of the face/shapes task– no voxel-wise analyses of this data was conducted. This was used to demonstrate structural differences between groups in certain regions.

7.3.5 BOLD Pre-Processing

The IAPS task and the resting state data were slice time corrected (temporally middle slice was used as reference) prior to performing motion correction. All functional BOLD data was motion corrected (rigid coregistration to the mean), coregistered to the skull-stripped MPRAGE (mean functional image used to calculate affine transformation), normalized to MNI space using the deformation field calculated previously (2mm isotropic resolution), and smoothed using a Gaussian kernel with FWHM of 8mm. All images were investigated by human eye to confirm that coregistration and normalization steps were accurate. Functional data from the first four scans utilized the baseline MPRAGE, while the end scan utilized its MPRAGE.

Motion was evaluated using ArtRepair toolbox (Mazaika *et al*, 2007). During the emotional faces reactivity task, participants had low maximum translations [mean=1.26mm (std=1.21)], low root mean squared (RMS) [1.11mm (0.81)], and low percentage of volumes displaying head jerks above 0.5mm [6.2% (10.7%)]. During the resting state, participants had low maximum translations [1.27mm (1.26)], low root mean squared (RMS) [1.04mm (0.85)], and slightly higher percentage of volumes displaying head jerks above 0.5mm [10.9% (19.9%)] that were corrected for using wavelet-despiking in later stages. During the explicit emotion regulation task, participants had low maximum translations [1.87mm (1.91)], low root mean squared (RMS) [1.40mm (1.08)], and low percentage of volumes displaying head jerks above 0.5mm [9.4% (30.8%)], except for a few particularly bad cases that were removed.

For resting state BOLD, spike artifacts were removed using a previously established method that uses wavelets to filter spike artifacts (Patel *et al*, 2014). Five principal components of white matter and cerebrospinal fluid were extracted as well as 6 motion parameters and a vector to model the mean of the time series. Band-pass filtering was conducted by including

several regressors that represented cosines with all discrete frequencies except those within the standard expected resting state frequencies (0.008 to 0.15 Hz).

7.3.6 Modeling Task Activation: Face/Shapes and IAPS

Mass-univariate general linear modeling (i.e. each voxel is independently modeled) was performed to model the mean of each signal, faces task, shapes task, and six parameters of motion (from motion correction). The canonical hemodynamic response function was used to convolve the faces and shapes tasks to expected hemodynamic responses. A high-pass filter of 1/128 Hz was utilized to account for low frequency noise. An autoregressive [AR(1)] filter was used to account for serial correlations due to aliased biorhythms and unmodelled activation. The contrast faces minus shapes was used to perform all voxel-wise group level analyses (i.e. regions that are active during faces relative to shapes and vice versa). Our voxel-wise analyses utilized only data from this contrast.

Similarly, the IAPS task included similar parameters however it modeled the activation during the neutral and negative viewing tasks as well as the reappraisal task (during viewing of some negative images). The contrast of interest was negative reappraise minus negative viewing, which modeled the activation during reappraisal adjusting for activation during the negative viewing task. The activation of specific regions during explicit emotion regulation were extracted from regions that we found in voxel-wise analyses of the face/shapes task– no voxel-wise analyses of this contrast was conducted. This was used to show how activation during explicit emotion regulation changed across the treatment trial.

7.3.7 Resting State BOLD: Eigenvector Centrality (EVC)

Eigenvector centrality was calculated using the fastECM toolbox (Lohmann *et al*, 2010). Briefly, centrality is a measure of connectedness of a voxel or region. FastECM uses singular value decomposition to circumvent the calculation of large correlation matrices. The centrality at rest was extracted from regions that we found in voxel-wise analyses of the face/shapes task– no voxel-wise analyses of this data was conducted. This was used to show how centrality changed across the treatment trial.

7.3.8 Pre-processing pCASL and Perfusion Calculation

After performing motion correction and spatial smoothing of the pCASL data, we coregistered the skull-stripped MPRAGE to the mean ASL image and applied the transformation to the ICV mask and white matter segmentation. White matter segmentation was used to calculate the M_0 magnetization in the white matter while the ICV allowed for calculation only within the brain. The following parameters were used to calculate perfusion using ASL toolbox (Wang *et al*, 2008b): label time = 1.1, delay time = $3.6x10^{-4}$, slice time = 37.25, and labeling efficiency = 0.85. The mean perfusion image was calculated across the entire time series for each voxel then coregistered to the skull-stripped MPRAGE and normalized to MNI space using the standard deformation field (4 mm isotropic resolution). The perfusion at rest was extracted from regions that we found in voxel-wise analyses of the face/shapes task– no voxel-wise analyses of this data was conducted. This was used to show how perfusion changed across the treatment trial.

7.3.9 DTI Preprocessing and Mean Diffusivity

After performing eddy correction via FSL, we coregistered the skull-stripped MPRAGE to the first b_0 image and applied the transformation to the ICV mask (used to calculate mean diffusivity only in the brain). FSL (Jenkinson *et al*, 2012) was used to calculate mean diffusivity (an inverse measure of membrane density) and is calculated by adding the first three eigenvalues and dividing by three. The MD image was coregistered to the skull-stripped MPRAGE and normalized to MNI space (2 mm isotropic resolution). The mean diffusivity was extracted from regions that we found in voxel-wise analyses of the face/shapes task– no voxel-wise analyses of this data was conducted. This was used to demonstrate structural differences between groups in certain regions.

7.3.10 Statistical Analysis

Group differences in demographic and clinical variables were tested using the Statistical Package for Social Science (IBM Corp. Released 2013. IBM SPSS Statistics for Mac, Version 24.0. Armonk, NY: IBM Corp.). Independent t-tests (continuous data) or Fischer's exact p-value (categorical data) was conducted where appropriate.

Statistical non-parametric toolbox (SnPM12) was used to perform all voxel-wise statistical analyses, which computes non-parametric p-values which are then corrected using a cluster-wise inference method (cluster forming threshold of $p<0.001$) that controls the family wise error rate (FWE) at $\alpha=0.05$ (Nichols and Holmes, 2002).

We conducted a one-sample t-test to test for the main effect of the task (areas activated more during faces than during shapes independent of group) to show that this task robustly activated the amygdala.

We investigated whether group differences (independent t-test) existed at baseline, following a placebo, and at the end of the trial during the face/shapes task. As we sought to understand baseline differences in emotional reactivity, in regions that were significantly different between groups we extracted baseline gray matter density, mean diffusivity, perfusion, EVC, and activation during emotion reappraisal. We then tested (via SPSS) whether there existed group differences in any of these regions within each measure and controlled the false discovery rate (FDR) at alpha less than 0.05 using the Benjamini-Hochberg procedure.

We conducted paired t-tests to investigate *within* group differences between baseline and end (as well as following a placebo and end). This reveals total changes across the entire treatment trial (i.e. effect of the medication).

In the next set of analyses, we wanted to investigate whether any acute changes depended on group (interaction: independent t-test on the difference). We investigated whether groups differed on how activation changed acutely (between placebo and a single dose) as well as sub-acutely (between placebo and following a week of treatment).

We subsequently also extracted mean perfusion, EVC, and activation during emotion reappraisal in the clusters that showed significant change during the faces-shapes task across the entire treatment trial and those that showed early changes during faces-shapes after a single dose of anti-depressant. In these analyses, we tested whether there were significant differences between baseline and end as well as whether there existed associations between the acute changes in the emotional reactivity (face/shapes baseline minus single does) and total changes

(baseline minus end) in each of the other measures. We extracted mean diffusivity and gray matter density in these regions to test whether the changes were dependent on baseline structural measures.

Functional imaging results were generated using xjview (Cui *et al*, 2011). Table 2 was generated by dividing significant clusters into regions in the automatic anatomic labeling (AAL) template (Tzourio-Mazoyer *et al*, 2002) and Brodmann areas were reported if they overlap with at least 30 percent of the cluster. The same principal was used to determine if a network [predetermined from an established set of resting state networks (Smith *et al*, 2009)] label should be assigned to that cluster.

7.4 RESULTS

7.4.1 Clinical Group Differences

Remitters had significantly lower baseline depression severity, which we adjusted for in several analyses (see table 3). As expected, remitters had even lower depression severity by the end of the study as well as lower serum venla/des-venla levels (by design as non-responders had dosage increased). We found no differences at baseline between remitters/non-remitters in WMH burden.

Table 3. Group differences in clinical/demographic features (full sample). NOTE: CIRSG-Cumulative illness rating scale for geriatrics; MMSE-Mini-mental state examination; MDD-major depressive disorder; MADRS-Montgomery-Asberg Depression rating scale; WMH-white matter hyperintensities.

	Non-Remitter (N = 25)	Remitter (N = 26)	t-statistic, p-value
	Mean (Standard Deviation) or Number of Subjects		
Age	65 (6)	67 (7)	t(49)=-1.1, p=0.297
Gender	11 F	7 F	p=0.249
Race	21 CC	22 CC	p=1.000
Education	15 (3)	15 (3)	t(49)=1.2, p=0.255
Depression Type (Single/Recurrent)	10 single [N=24]	8 single [N=24]	p=0.565
CIRSG	9 (5) [N=24]	10 (4)	t(48)=-0.7, p=0.494
MMSE	29 (1) [N=24]	29 (2)	t(48)=-0.01, p=0.987
Serum Venla/Des-Venla End	333.5 (122.9) [N=22]	238.3 (101.4) [N=24]	t(42)=2.8, p<0.01*
MADRS Baseline	27 (5)	23 (8)	t(49)=2.2, p<0.05*
MADRS End	18 (7)	5 (4)	t(47)=8.0, p<0.0001**
WMH	3.04 (1.59)	2.56 (1.42)	t(49)=1.2, p=0.256

7.4.2 Faces-Shapes Task: Robust Activation of Emotional Circuits

Independent of group, the task significantly activated the bilateral amygdala, visual cortex and secondary visual processing areas (including parietal cortex, precuneus, fusiform gyrus), hippocampus, parahippocampus, insula, as well as inferior and middle frontal.

Table 4. Results of all statistical analyses on emotion reactivity task. The analysis conducted and effect tested are reported as well as the significant regions (including hemisphere and BA/network, if applicable), number of voxels, the max value of the statistical test, and x, y, and z coordinates in MNI space. Regions are labeled with a BA or network if at least 30 percent of that cluster overlaps with the structural BA or functional network definitions.

Analysis	Effect	Region	Side	Network	BA	# Voxels	Max	x, y, z
Interaction: Time by Group	Baseline/Placebo by Group	Not Significant						
	Placebo/Single Dose by Group	Parahippocampus	L	vDMN		191	5.5	-20, -38, -8
	Placebo/Week One by Group	Not Significant						
	Placebo/End by Group	Not Significant						
Group Differences (Independent T-test) at Baseline, Placebo, and End	Baseline Non-Remitter > Remitter (Adjusting for Baseline Depression Severity)	Caudate	L			67	4.3	-18, 22, 0
			R			99	3.9	22, 24, 4
		Cerebellum Declive	R		37	105	4.8	28, -56, -20
		Anterior Cingulate	L	dDMN, RECN	32	111	4.0	-2, 42, 16
			R		32	92	4.2	14, 50, 20
		Inferior Frontal (Orb)	L	RECN	47	269	4.8	-26, 34, -12
			R		47	106	4.0	34, 38, -6
		Inferior Frontal (Tri)	L	RECN	45, 47	271	4.1	-44, 28, 0
		Inferior Frontal (Orb)	L		11, 47	74	4.9	-26, 36, -12
		Middle Frontal	L	ASN, RECN	46	547	4.8	-28, 44, 30
		Superior Frontal	L		9	135	4.3	-18, 44, 30
		Superior Medial Frontal	L		10	79	3.8	-10, 56, 26
		Heschl Gyrus	R		48	50	4.4	40, -24, 18
		Hippocampus	R		20	100	4.2	40, -22, -8
		Insula	L	ASN	48	104	4.1	-42, 8, -2
			R		48	226	4.5	38, 12, -12
		Rolandic Operculum	R		48	120	5.3	44, -24, 20
		Inferior Temporal	R		20	68	4.6	54, -4, -28
		Middle Temporal	L		20, 21	226	5.4	58, -4, -22
		Superior Temporal Pole	L	ASN	38	82	5.1	-50, 12, -12
		Superior Temporal	L		48	98	4.1	-50, 6, -12
			R		48	568	4.8	58, -8, 2
		Thalamus	L			79	4.0	-2, -14, 6
	Placebo Group Differences	Not Significant						
	End Group Differences	Not Significant						

Table 4 (continued)

Time Differences (Paired T-test) in Non-Remitters and Remitters	Non-Remitters: Baseline > End	Calcarine	L		17, 18	194	4.4	2, -96, 8
			R		18	79	4.4	22, -94, -4
		Cerebellum Culmen	R	vDMN	37	60	4.4	20, -46, -14
		Cerebellum Declive	R		18, 19	385	6.0	24, -78, -18
		Cerebellum Declive	R		18	98	5.6	34, -76, -22
		Cuneus	L		18	161	4.6	-10, -96, 18
			R		18	144	5.3	8, -92, 26
		Fusiform	R		18, 19	159	5.6	24, -82, -16
		Insula	L		48	85	4.6	-46, 2, 2
			L		19	122	5.0	-18, -52, -8
		Lingual	R		18	275	5.1	22, -84, -14
		Inferior Occipital	R		18, 19	106	4.4	30, -82, -16
		Middle Occipital	L		18	129	4.6	-28, -94, 12
			R		18	88	4.8	26, -92, 12
		Superior Occipital	L		17, 18	158	4.9	-12, -96, 20
			R		18	163	5.3	22, -92, 20
		Inferior Parietal	L	RECN	2, 3	76	4.8	-56, -22, 46
		Rolandic Operculum	L		48	101	4.8	-50, 2, 4
		Thalamus	L			76	4.9	-6, -8, 6
			R			50	4.7	6, -12, 6
	Remitters: End > Baseline	Insula	L	ASN	13	132	5.4	-40, 14, -4
	Non-Remitters: Placebo > End	Parahippocampus	L	vDMN		136	4.8	-16, -24, -10
	Remitters: Placebo = End	Not Significant						

7.4.3 Baseline Hyperactivation in Non-Remitters Relative to Remitters

We found that non-remitters had greater activation than remitters (even after adjusting for baseline depression severity) in the: bilateral caudate, anterior cingulate, inferior frontal (orbital), superior temporal, and insula; as well as the left thalamus, inferior, middle, and superior frontal; and the right hippocampus, rolandic operculum, and inferior temporal gyrus (figure 4 and table 4). We found no group differences at placebo or the end of the treatment trial.

In each of these regions, we extracted mean gray matter density, mean diffusivity, resting perfusion, resting EVC, and activation during explicit emotion regulation and tested for group differences at baseline (adjusted for multiple comparisons by controlling the FDR). We found that in the right inferior orbital gyrus, remitters had lower mean diffusivity [$t(55)=-2.9$, $p_{unc}=0.0056$, FDR=0.028] than non-remitters. Thus, the group differences in activation may be partially driven by group differences in diffusivity in the right inferior orbital gyrus.

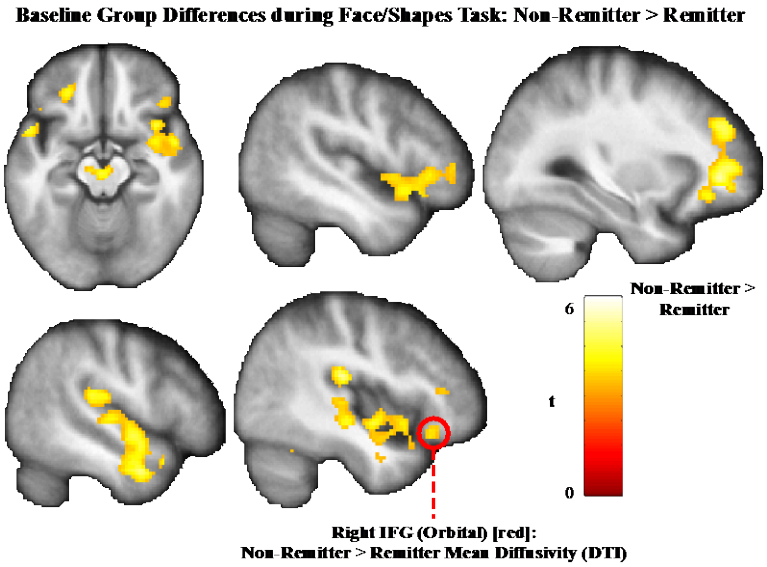

Figure 4. Group differences in emotion reactivity at baseline. Regions that were significantly more active during face/shapes task in non-remitters relative to remitters at baseline (adjusting for baseline depression severity). Colors indicate the value of the t-statistic (independent t-test), where lighter values indicate regions where non-remitters have greater activation than remitters (no regions in reverse direction). Non-remitters also showed group differences in right inferior frontal gyrus (orbital) mean diffusivity, but showed no other group differences in structural (gray matter density and mean diffusivity) or functional (resting perfusion, resting EVC, and activation during IAPS) measures in these regions.

7.4.4 Decreased Activation Following Ineffective Pharmacotherapy (Non-Remitters)

Non-remitters showed significant decreases in activation between baseline and the end of the trial in visual cortex and secondary visual processing areas (fusiform, inferior parietal), thalamus,

and bilateral insula (figure 5 and table 4). However, we only found that the left parahippocampus significantly decreased between placebo and the end of the trial in the non-remitters (table 4).

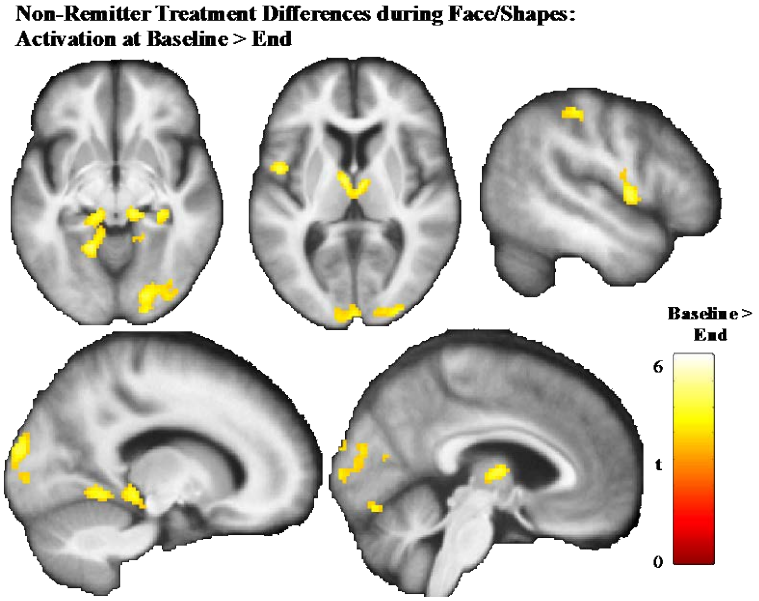

Figure 5. Baseline vs. end emotion reactivity in non-remitters. Several clusters showed significant decreases from baseline to end (but no increases) in activation in non-remitters. Colors on the brain show the t-statistic for the paired t-test between baseline and the end of the trial.

7.4.5 Increased Insula Activation Following Effective Pharmacotherapy (Remitters)

Remitters showed a significant *increase* in the left anterior insula between baseline and the end of the trial (figure 6 and table 4). To better understand this finding in the remitters we extracted resting perfusion, resting EVC, and activation during explicit emotion regulation in the left AI. We tested for total changes across the entire trial as well as whether total changes associated with acute changes in emotion reactivity.

In the remitters, we found that the left anterior insula EVC significantly increased from baseline [$t(23)=-2.3$, $p=0.0325$], while the acute increase (following single dose relative to baseline) in activation was associated with total increase (end relative to baseline) in perfusion [$r(23)=0.42$, $p=0.0385$]. Of note, we found that the IAPS activation increased during the long-term treatment period (end relative to a week after beginning treatment), but not significantly [$t(23)=-1.83$, $p=0.0804$]. We did not find any associations between structural features at baseline and either the acute or long-term changes in activation. Thus, the increase in left anterior insula activation during the emotion reactivity task was coupled with a more chronic change in explicit regulation, activation was associated with change in perfusion, and changes were independent of structural influence.

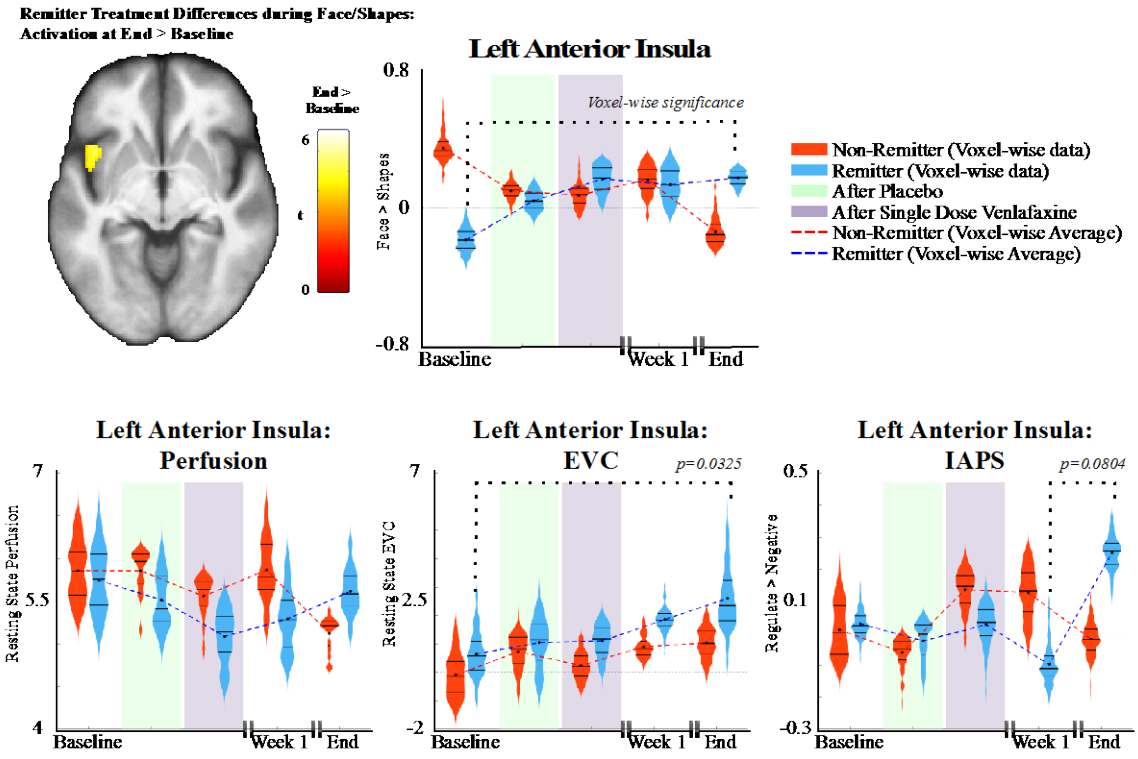

Figure 6. Baseline vs. end emotion reactivity in remitters. The left *anterior* insula showed significant increases in activation during face/shapes task in remitters. Colors on the brain show the t-statistic for the paired t-test between baseline and the end of the trial. While the entire course is plotted, the test was only done on two scans (baseline and end) in remitters and while we plotted the *non-remitter* data there was no test done (only plotted for reference). Violin plots show the voxel-wise variance for this cluster by plotting a mirrored histogram (i.e. a vertically mirrored histogram for the voxel-wise data in the cluster). We plot average changes in dotted lines. We found that the acute change in activation during face/shapes was associated with increased total perfusion and that the EVC of this region also significantly increased. While we found no significant baseline to end differences during IAPS, we did find a marginal effect showing differences between the end of the trial and a week after beginning treatment.

7.4.6 Single Dose Engagement in Parahippocampus

We found that following a single dose of Venlafaxine relative to placebo there was an increase in activation in remitters that significantly differed than the decrease in activation in non-remitters. This increase was detected in the left parahippocampus (figure 7 and table 4). While we found no such changes (that depended on group) between placebo and after a week of treatment or at the end of the trial, we did find that the left parahippocampus *qualitatively* showed some effect in both analyses (i.e. a cluster appeared that did *not* pass multiple comparisons correction, $p<0.001$ uncorrected).

To better understand this effect in the remitters we extracted mean resting perfusion, resting EVC, and activation during IAPS in the parahippocampus and tested for significant changes across the entire trial. We found that only the IAPS task significantly increased from baseline in the remitters [$t(23)=-3.9$, $p=0.0008$]. We did not find any associations between structural features at baseline and either the acute or long-term changes in activation. Thus, the acute increase in activation in the left parahippocampus was coupled with a more chronic change in explicit regulation, but was independent of structural influence.

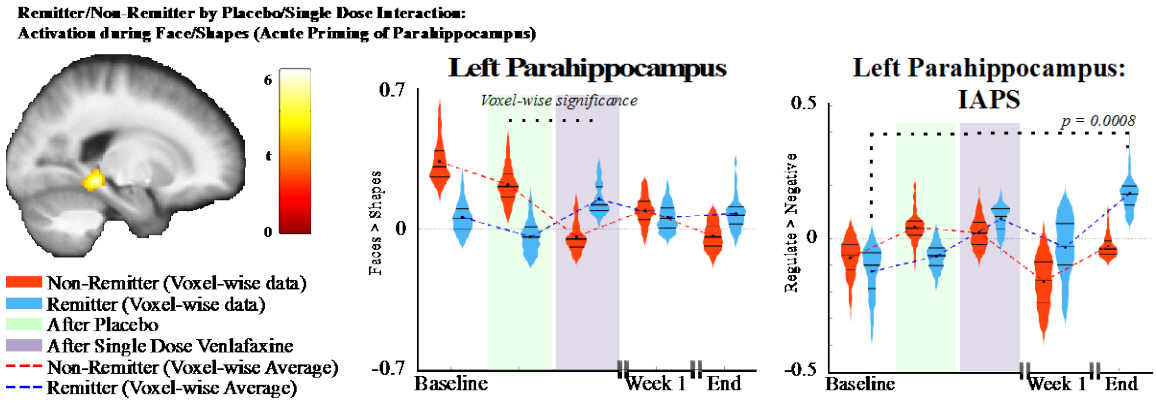

Figure 7. Acute single-dose changes in activation of the emotion reactivity task. The left parahippocampus showed a significant interaction between group (remitters and non-remitters) and time (after placebo and after single dose). While remitters and non-remitters showed a slight decrease following a placebo, they showed opposite changes following the initial dose (which seems to remain). Colors on the brain show the t-statistic for the independent t-test between the difference (placebo minus initial dose). While the entire course is plotted, the test was only done on two scans (placebo and single dose). Violin plots show the voxel-wise variance for this cluster by plotting a mirrored histogram (i.e. a vertically mirrored histogram for the voxel-wise data in the cluster). We plot average changes in dotted lines. We also found that the left parahippocampus showed a significant increase from baseline (relative to end) in this region.

7.5 DISCUSSION

As expected, the face/shapes task robustly activated the amygdala as well as supporting structures during the emotion reactivity task, including structures such as the parahippocampus, insula, and executive structures (Groenewold *et al*, 2013).

7.5.1 Baseline Hyperactivation

At baseline, we found that non-remitters exhibited a heightened reactivity (as measured by the level of activation that was independent of their baseline depression severity) in a wide set of regions that did *not* include the amygdala. As hyperactivation of the amygdala is seen in MDD relative to never-depressed individuals, this may reflect a similar level of automatic *reactivity* – but a possible difference in either appraisal and/or regulation. Regions such as the caudate, insula, anterior cingulate, and frontal structures were hyperactive. Thus, this may reflect a greater level of the secondary appraisal processing that follows the initial amygdala reactivity (hence hyperactivation of the insula, caudate, anterior cingulate). Frontal structures (including orbitofrontral) and the anterior cingulate may reflect a greater *need* for emotional regulation as a result of hyperactivation in these other limbic structures. Interestingly, we found a lower mean diffusivity in the right inferior orbital gyrus in the remitters compared to non-remitters, and as this region is critical for both implicit and explicit aspects of regulation may reflect a difference in ability to regulate, but also a difference in remission capacity (at least to this regimen). Interestingly, there was no detected difference at placebo – which may reflect a difference in reactivity that is quickly tempered.

7.5.2 Decreased Activation in Non-Remitters

Interestingly, non-remitters displayed a decrease in activation in visual processing areas (primary and fusiform/parahippocampus), insula, inferior parietal lobe, and thalamus. Notably, all participants in the study improved to *some* degree following treatment. Thus, non-remitters do display improvement to *some* degree in symptom severity and this decrease in limbic reactivity may be a possible mechanism of these changes. Several of these changes (including the insula and parahippocampus) occur early as well, which suggests that a decrease in these limbic structures may provide relief possibly through a change in reactivity. However, it does not allow for total remission of symptoms.

7.5.3 Increased Left Anterior Insula Activation

In contrast to non-remitters, remitters showed *only* a significant (steady) *increase* in the left anterior insula activation. This suggests that remission (at least to venlafaxine) has a specific neural signature and replicates similar previous findings but extends the changes to activation over a more acute period (Arce *et al*, 2008; Davidson *et al*, 2003; Wang *et al*, 2012). While there was not a significant change in perfusion, the increased activation was associated with an increase in perfusion. This supports the notion that there are meaningful changes occurring within this region in those who experience improvement in symptom severity. We argue that these changes reflect an improvement in the insula's core function to relate external and internal stimuli to the self. Specifically, as this region has been associated with the amount of daily habitual implicit regulation there may be greater implicit regulation of emotional responses (Gross *et al*, 2003). This is further supported by significant changes in centrality (EVC) that

reflects a greater whole brain connectedness within this region, which may be involved in overall improvement even at rest. Notably, we did not find any changes in amygdala activation throughout the study in remitters. Thus, we can conclude that there *is* no change in reactivity but rather a change in the processes that follow. To further implicate these changes, we also found that there is a total change in *explicit* emotion regulation in a separate task that occurs over a later period (no early changes). As these changes were not related to baseline structural features, we can thus conclude that while there were no changes in reactivity there were important changes in how those emotions were regulated (implicit regulation acutely and explicit regulation chronically).

7.5.4 Acute Parahippocampal Engagement

Unlike the left anterior insula that showed a steady increase in activation, the left parahippocampus exhibited a strong acute (single dose) increase in the remitters (decrease in non-remitters) that remained for the entirety of the trial. The change occurred following only a single dose, thus it is a strong candidate as an early biomarker. As with the left anterior insula, this region also showed a significant increase in activation during the explicit regulation task in the remitters. Again, this supports a change in regulatory strategies and may reflect a change in how the stimuli are perceived contextually.

7.5.5 Chronic Behavioral Changes and Implicit Improvement

A well-known aspect of the treatment process is that while the anti-depressants are known to modulate (increase) serotonin levels acutely (within hours), the behavioral changes do not

present until a much later period (3-5 weeks in mid-life and 6-8 weeks in late-life) (Andreescu *et al*, 2011). These results may provide some insight, specifically that while there is a change in processing (neural processing of implicit regulation) there are no changes behaviorally after a single dose. Similarly, the changes in explicit emotion regulation, which requires conscious continuous effort, do not change either at the neural or behavioral level until a much later period. These results provide a possible mechanistic explanation (as these changes occur following one dose) of the changes occurring following treatment: modulation of serotonin changes the neural processing involved in implicit emotion regulation that is a marker of eventual changes in explicit emotion regulation (which may actually be an effect of remission).

7.5.6 Relevance to Late-Life and Limitations

We recruited a late-life sample thus while it is not clear whether these results would generalize to a mid-life sample there exists previous studies in mid-life that showed similar changes in the anterior insula. Further in this sample, we did not find any differences in white-matter hyperintensity burden between remitters and non-remitters even though we have in the past. Non-remitters displayed lower mean diffusivity than remitters in the right inferior orbital gyrus, which may reflect differences in remission capacity to this particular anti-depressant. While compared to other neuroimaging studies we have a good sample size, a larger study would help us better understand the generalizability of these results and more importantly their reliability. These changes may reflect changes following therapy to only this particular anti-depressant and may not generalize to other anti-depressants.

Several mid-life studies have shown changes in amygdala reactivity following successful pharmacotherapy, but there are fewer studies that have shown a change in amygdala reactivity in

LLD. We could speculate that this is related to the difference in etiology – many older individuals have not (as youth or in mid-life) experienced any depressive symptoms, thus it may that older individuals have impairments in another process of the fearful/angry viewing paradigm, mainly the appraisal or implicit regulation steps. A lower ability to regulate implicitly could explain the worsened mood as well as hyperactivity of the amygdala. This is supported by studies that have found (in never depressed individuals) similar amygdala activation in young and elderly individuals but altered fusiform and insula activation (Wright *et al*, 2006).

As the amount of time needed to determine whether patients are responding to a particular therapy is longer in late-life (6-8 weeks) and is associated with an increased risk of suicide, finding biomarkers of remission is of utmost importance. These changes may be an important step towards this goal.

7.5.7 Conclusion

We have identified two possible regions of interest: the left parahippocampus and the left anterior insula. We argue that these changes occur and are specific to the implicit emotion regulation neural circuitry that translates to eventual changes in the explicit emotion regulation neural circuitry. The left anterior insula in particular could be a target for transcranial magnetic stimulation (TMS) as with deep TMS we are now able to reach deeper and deeper structures. Overall, these may explain the chronic behavioral changes that occur overall a longer time scale compared to the acute neural changes in implicit regulation.

8.0 PREDICTING REMISSION IN LLD: MULTI-FACTOR KERNEL BASED MACHINE LEARNING

This chapter focuses on the results of several models that attempt to predict remission. The models utilize the same dataset described in the last two chapters. Several parameters of the model fitting process are explored, including: single vs. multiple features, PCA vs. MFA for feature reduction (or kernel), and different combinations of feature sets (mainly demographic and clinical data, baseline structural imaging, baseline functional imaging, pharmacological change in functional imaging).

8.1 INTRODUCTION

Increased interest in machine learning approaches have resulted in many studies that have attempted to generate a model for predicting remission to depression. This has proven to be complicated with many models suffering from over-fitting and low generalizability due to high-dimensional features and low sample sizes. A common problem is combining several feature sets in an intuitive fashion (e.g., neuroimaging and clinical data). One approach that has been widely utilized uses kernel-based machine learning models. These approaches typically reduce the feature set into a single kernel that can be used to model the observed outcomes.

Principal components analysis (PCA) reduces high-dimensional features into low-dimensional vectors (or eigenvectors), which explain a certain proportion of variance within the data (related to the eigenvalue). PCA reduces the matrix of features into a set of scores (which represent the original data in the low-dimensional feature space). These scores are then used to fit a model with the observed outcome. These models have several desirable properties. Consider a linear regression model that is either solved using the standard approach (ordinary least squares, OLS) or using the PCA approach. The first property is that any linear form of the principal components method has a lower variance than the OLS solution. The covariance matrix of the scores (from PCA) is identity, which means that none of the features are collinear. This completely resolves the multi-collinearity problem in regression. This method can be considered a regularized solution and is also an optimal regularized solution.

While the PCA kernel can be used to reduce a single feature, the problem of multiple feature sets is not resolved. Multi-factor analysis is an extension of this approach. Combining all the data into a single matrix then performing PCA is undesirable, as the components will be dominated by the matrices with the greatest number of features (e.g., including neuroimaging data will introduce a large number of voxels). By first performing a PCA on each individual data set (clinical/demographic and neuroimaging data separately) then using the scores to perform another PCA, the scores will be equally weighted on each individual data set rather than by the number of features within each individual set.

In this study, a cohort (N=51) of LLD individuals was recruited into an open-label trial of venlafaxine (a serotonin-norepinephrine reuptake inhibitor). Neuroimaging data was collected at baseline and following a single dose of venlafaxine (among 3 other time points – however they

are not utilized in this analysis). We used kernel-based machine learning approaches to predict treatment outcomes (remitters/non-remitters) at 12 weeks.

8.2 METHODS

The following sections have already been described in the previous chapter (7.3.1-7.3.9): Study Design and Participants, MRI Data Collection, Functional Tasks, Structural Processing, BOLD Pre-Processing, Modeling Task Activation: Face/Shapes and IAPS, Resting State BOLD: Eigenvector Centrality (EVC), Pre-processing pCASL and Perfusion Calculation, DTI Preprocessing and Mean Diffusivity. The only exception is in 7.3.9, where both mean diffusivity (MD) and fractional anisotropy (FA) are calculated. The following sections are split into: single feature set and multiple features set learning (mainly PCA vs. MFA as a feature reduction method). The theory behind these models is detailed in chapter 4.

8.2.1 Single Feature Set: Principal Components Learning

This method utilizes principal components analysis to reduce features and works well with 'single feature sets.' The following single feature sets were evaluated: (1) demographic and clinical data (e.g., age, gender, education, MADRS, WMH burden, etc.); (2) baseline functional neuroimaging (each of the following was independently used: emotion reactivity, emotion regulation, EVC, and perfusion); (3) baseline structural imaging (each of the following was independently used: gray matter density, MD, FA); (4) difference in functional neuroimaging between baseline and placebo or first dose in each of the functional tasks.

These models assume that a single feature matrix exists (X, n subjects by f features where f can be the number of clinical/demographic variables or the number of voxels depending on the feature set) and this is used to predict the outcomes (y, binary vector length n subjects). The model building process is reviewed graphically in figure 8. After determining the number of principal components (using Horn's parallel analysis, HPA), then we perform PCA on the feature matrix to get a set of scores (U, n subjects by c components) and a set of coefficients (λ, c components by f features).

First we fit a single model between the outcomes (y) and the scores (U) using either logistic regression, step-wise logistic regression, or support vector machines. Support vector machines optimize the box constraint and the kernel scale using a random 25 percent of the samples. This step outputs a set of parameter estimates (β, length c components) that are projected back into the original space using the coefficients (B, length f features). The next two steps test the generalizability of the model (by computing area under the curve, AUC) and which features are most significantly predictive (using permutation testing).

A 10-fold cross-validation was utilized, where for each fold a set of training data (U_{train} and y_{train}) was used to fit a single model, which was used to predict on the test data (U_{test}) to fit a set of predicted outcomes ($yhat_{test}$). After the cross-validation, a set of predicted outcomes (yhat) are output that can be compared to the actual outcomes (y) by calculating AUC of a receiver operating characteristic curve. However, it is possible that the cross-validation we have utilized is negatively/positively biased, thus we repeat this cross-validation a total of 50 times to get a range on the AUC statistic. This statistic tests for the generalizability of the full model, i.e., how well this model will perform given entirely new data.

Permutation testing is performed (1000) to determine the significance of the parameters in original model (i.e., which features significantly contribute to the final model). For each permutation, the outcomes are permuted or shuffled (y_p) then a model is fit (using U) to get a set of parameters (β_p) that are projected into the original feature space (B_p). This is repeated for each permutation to get a set of 1000 values for each feature (B_p is f by 1000) that represents a distribution for each feature that can be used to compute a p-value. The p-value is the number of times the absolute value of the actual parameter estimates (B) are less than the absolute value of the permuted parameter estimates (B_p) divided by 1000.

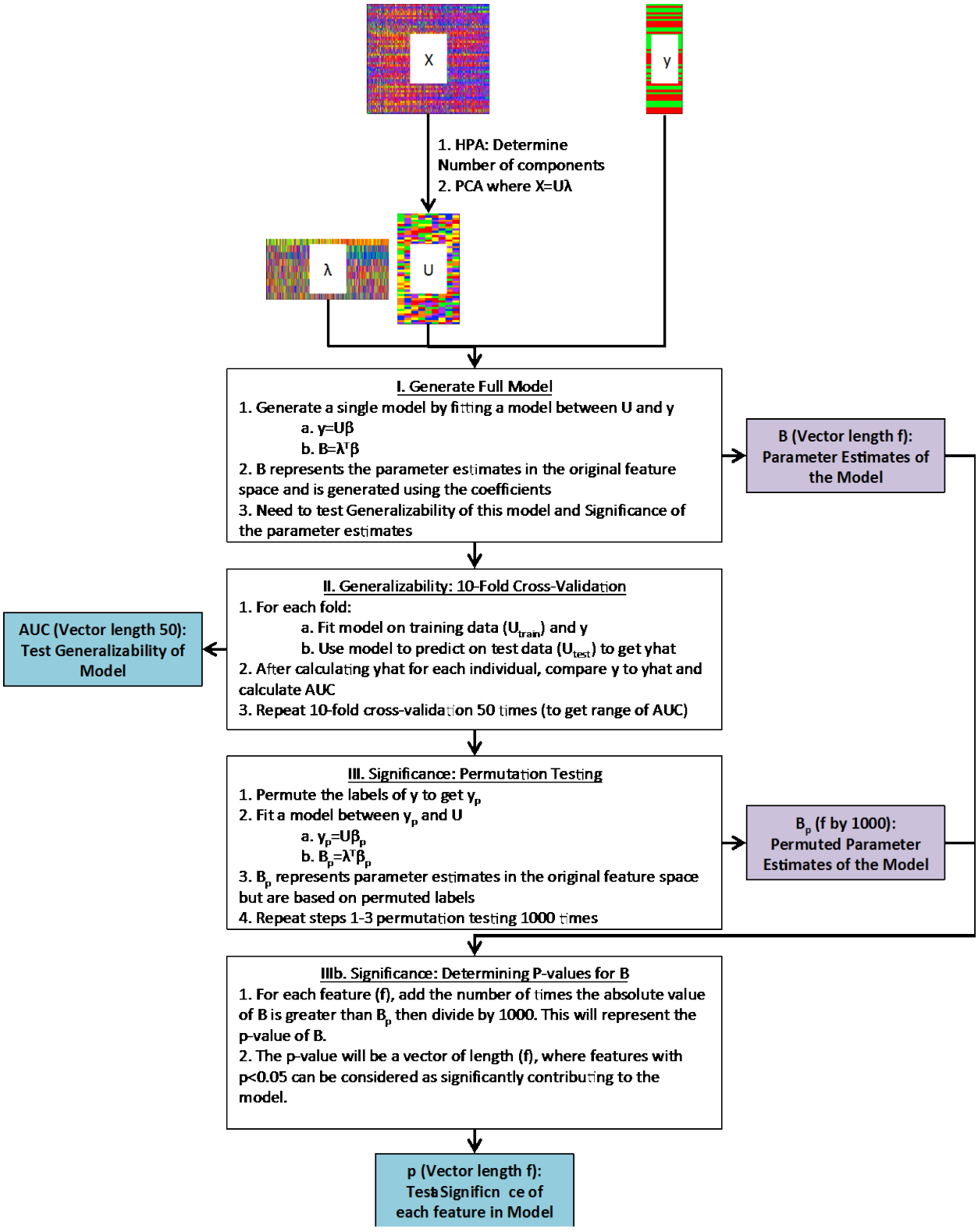

Figure 8. Model building procedure for single feature sets.

8.2.2 Multiple Feature Sets: Multi-Factor Learning

This method utilizes multiple factor analysis (MFA) to reduce features and works well with 'multiple feature sets.' Consider two sets of predictive features (e.g., demographic data and genetics data) where one feature set is particularly larger than the other (genetics data has a high number of features). Combining these two sets together will 'wash out' the effect of the demographic data – though that feature set may be highly predictive. MFA attempts to ameliorate this by performing a PCA on each set of data independently then using the scores from each to perform another PCA (hence the name "multi"-factor). This allows for each feature set to weight the scores matrix equally. The following multiple feature sets were evaluated: the demographic and clinical data along with (1) each of the functional neuroimaging features (independently then all together); (2) each of the structural neuroimaging features (independently then all together); (3) each of the functional neuroimaging features along with the pharmacological difference in the neuroimaging data; and (4) each of the functional neuroimaging features along with the pharmacological difference in the neuroimaging data as well as all the structural neuroimaging data.

These models assume greater than one feature matrix exists (e.g., X, n by f and S, n by v) and this is used to predict the outcomes (y, binary vector length n subjects). The model building process is reviewed graphically in figure 9. The process is identical to the previous algorithm, with a few important changes. Most importantly the first step is to perform PCA on both feature sets independently to extract a set of scores (U and V) as well as coefficients (λ and δ).

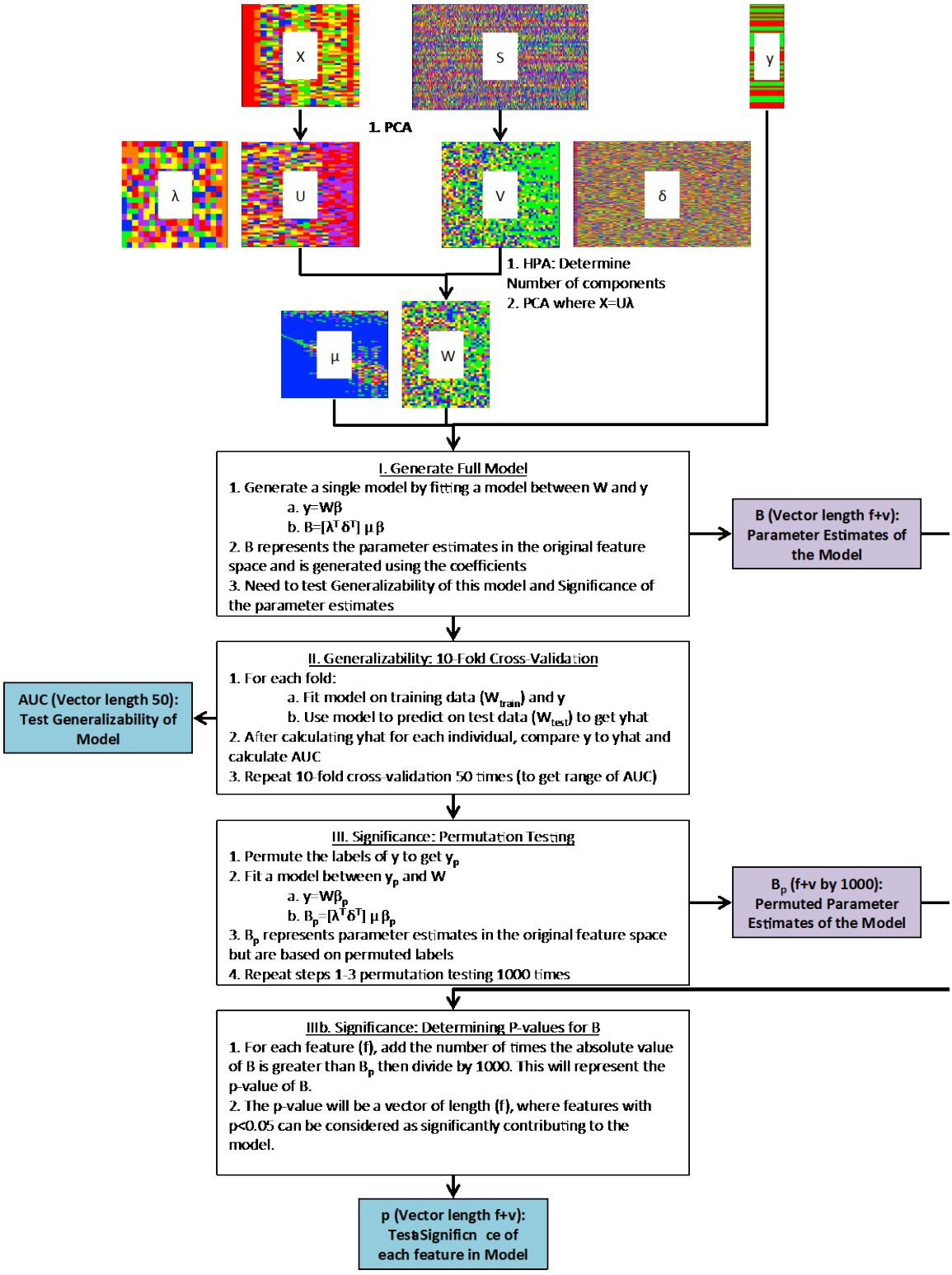

Figure 9. Model building procedure for multiple feature sets.

The next step uses the scores from each feature set (U and V), combines them into a single matrix and performs yet another HPA analysis to determine the number of necessary components then a PCA on that matrix to get a set of scores (W) and coefficients (μ). Next, the model building procedure is exactly the same as before, however the new scores are used to fit the models. Another change is that the transformation back into the original feature space requires two transformations (using μ first then λ and δ to transform into X and S feature space, respectively).

8.3 RESULTS

The single feature models produced several models with sufficiently high AUC (reported in table 5), specifically the demographic/clinical features with a step-wise logistic model resulted in the highest AUC (median 0.735). Two other features that performed well were: baseline emotion reactivity using a step-wise logistic model (median AUC of 0.766) and baseline MD using a step-wise logistic model (median AUC of 0.658). The ROC curves for each of these models are plotted in figure 10. The most significant predictors in the demographic and clinical model are reported in table 6 (ordered by most predictive to least). The voxels that were most predictive in the emotion reactivity model are reported in figure 11, while the voxels that were most predictive in the MD model are reported in figure 12. The emotion reactivity task showed the following regions to be predictive: bilateral visual cortex (angular, calcarine, cuneus, middle occipital), fusiform, precentral, supramarginal, middle and superior temporal, caudate, thalamus, putamen, insula, anterior and middle cingulate, inferior orbital and inferior triangular gyrii, middle and superior frontal, and superior medial frontal gyrii; as well as right amygdala, heschl gyrus,

hippocampus, precuneus, inferior temporal; and left parahippocampus, and inferior and superior parietal as well as parts of the cerebellum. The MD showed the following regions to be predictive: bilateral middle cingulate, superior orbital, insula, inferior parietal, pre-/post-central, and middle temporal; as well as right angular, inferior frontal, middle frontal, superior parietal, supplemental motor, and superior temporal; and left calcarine, anterior cingulate, hippocampus, lingual, precuneus, and supramarginal.

Table 5. AUC of each of the single feature models. NOTE: SVM=support vector machine; swLogistic=step-wise logistic model; AUC=area under the ROC curve; Hypothesize Mean=T-test tests the null hypothesis that the AUC is greater than the hypothesized mean.

Feature	Model	AUC median (IQR)	One Sample T-test (df=49)	Hypothesized Mean	p-value
Demographic and Clinical Features	SVM	0.655 (0.079)	7.2	0.6	1.60E-09
	Logistic	0.682 (0.029)	26.9	0.6	0
	swLogistic	0.735 (0.018)	15.6	0.7	0
Baseline Emotion Reactivity	SVM	0.618 (0.079)	1.8	0.6	0.0850
	Logistic	0.655 (0.040)	14.9	0.6	0
	swLogistic	0.766 (0.040)	17.1	0.7	0
Baseline Emotion Regulation	SVM	0.517 (0.045)			
	Logistic	0.530 (0.046)			
	swLogistic	0.575 (0.066)			
Baseline EVC	SVM	0.426 (0.078)			
	Logistic	0.325 (0.065)			
	swLogistic	0.354 (0.062)			
Baseline Perfusion	SVM	0.347 (0.058)			
	Logistic	0.530 (0.058)			
	swLogistic	0.547 (0.074)			
Baseline Gray Matter Density	SVM	0.531 (0.063)			
	Logistic	0.431 (0.060)			
	swLogistic	0.263 (0.063)			
Baseline FA	SVM	0.332 (0.062)			
	Logistic	0.495 (0.043)			
	swLogistic	0.442 (0.085)			
Baseline MD	SVM	0.510 (0.058)			
	Logistic	0.578 (0.045)			
	swLogistic	0.658 (0.055)	11.8	0.6	2.77E-16
phMRI Emotion Reactivity	SVM	0.583 (0.063)			
	Logistic	0.411 (0.052)			
	swLogistic	0.275 (0.075)			
phMRI Emotion Regulation	SVM	0.331 (0.069)			
	Logistic	0.503 (0.071)			
	swLogistic	0.297 (0.088)			
phMRI EVC	SVM	0.501 (0.041)			
	Logistic	0.571 (0.038)			
	swLogistic	0.586 (0.088)			

Table 5 (continued)

phMRI Perfusion	SVM	0.323 (0.0610)			
	Logistic	0.313 (0.032)			
	swLogistic	0.281 (0.066)			

Table 6. Features predictive in model that utilized clinical/demographic features. Features of the model that utilized demographic and clinical features and a step-wise logistic regression ordered by contribution to overall model.

Feature	Parameter Estimate (Z)	p-value
Negative Affect	-2.936	0.004
MADRS8	-0.280	0.004
MADRS9	-0.268	0.004
Education	-0.266	0.004
Gender	0.076	0.004
MADRS1	-0.192	0.005
MMSE	-0.045	0.005
MADRS	-1.959	0.008
MADRS2	-0.125	0.008
Positive Affect	2.651	0.009
MADRS3	-0.268	0.017
MADRS7	-0.185	0.031
MADRS4	-0.170	0.042
CIRSG	0.284	0.047
MADRS6	-0.190	0.047
MADRS5	-0.183	0.047
Age	0.695	0.048
MADRS10	-0.098	0.048
WMH	0.072	0.048
Race	-0.016	0.048
Depression Type	-0.004	0.048

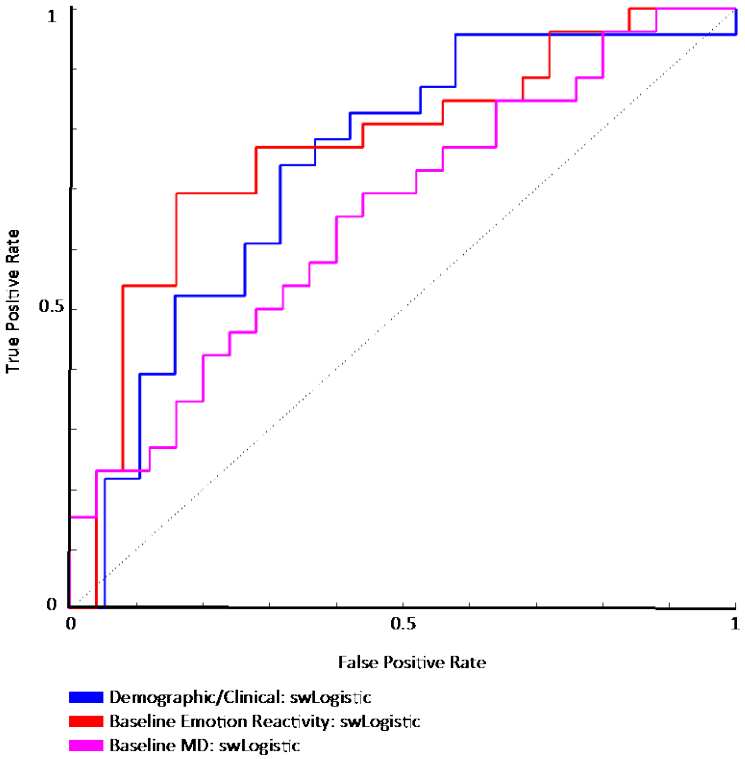

Figure 10. ROC curves for the most accurate models.

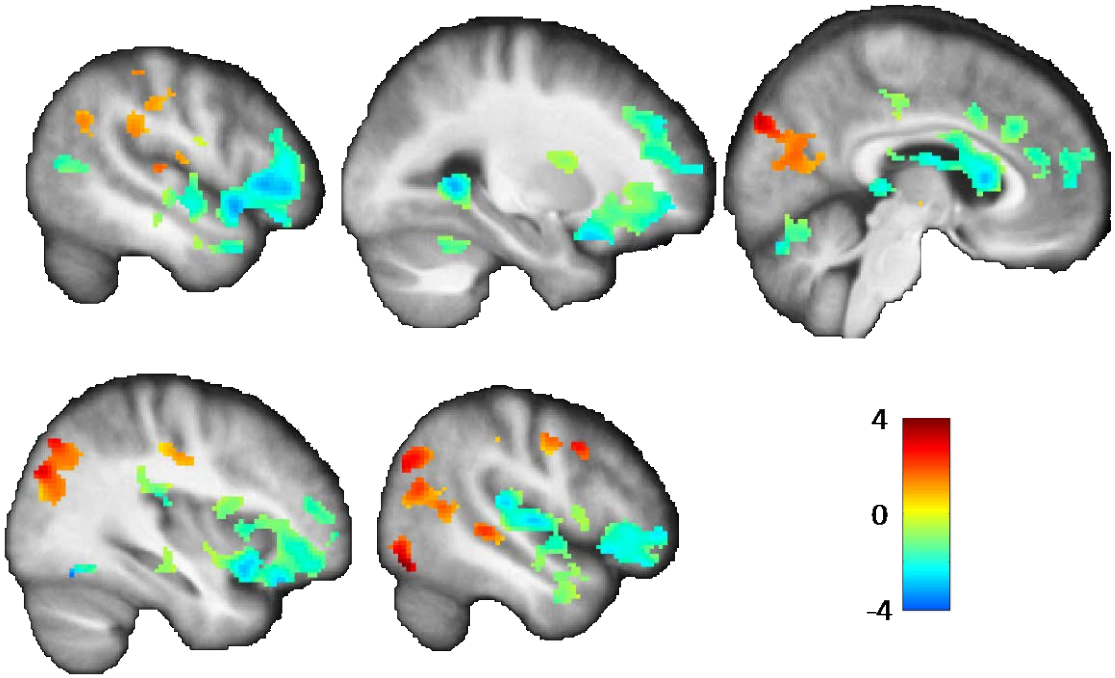

Figure 11. Most predictive voxels in emotion reactivity at baseline model. Regions during emotion reactivity (baseline) significantly contributing to the prediction of remission. Colorbar indicates value of the parameter estimate (Z-score).

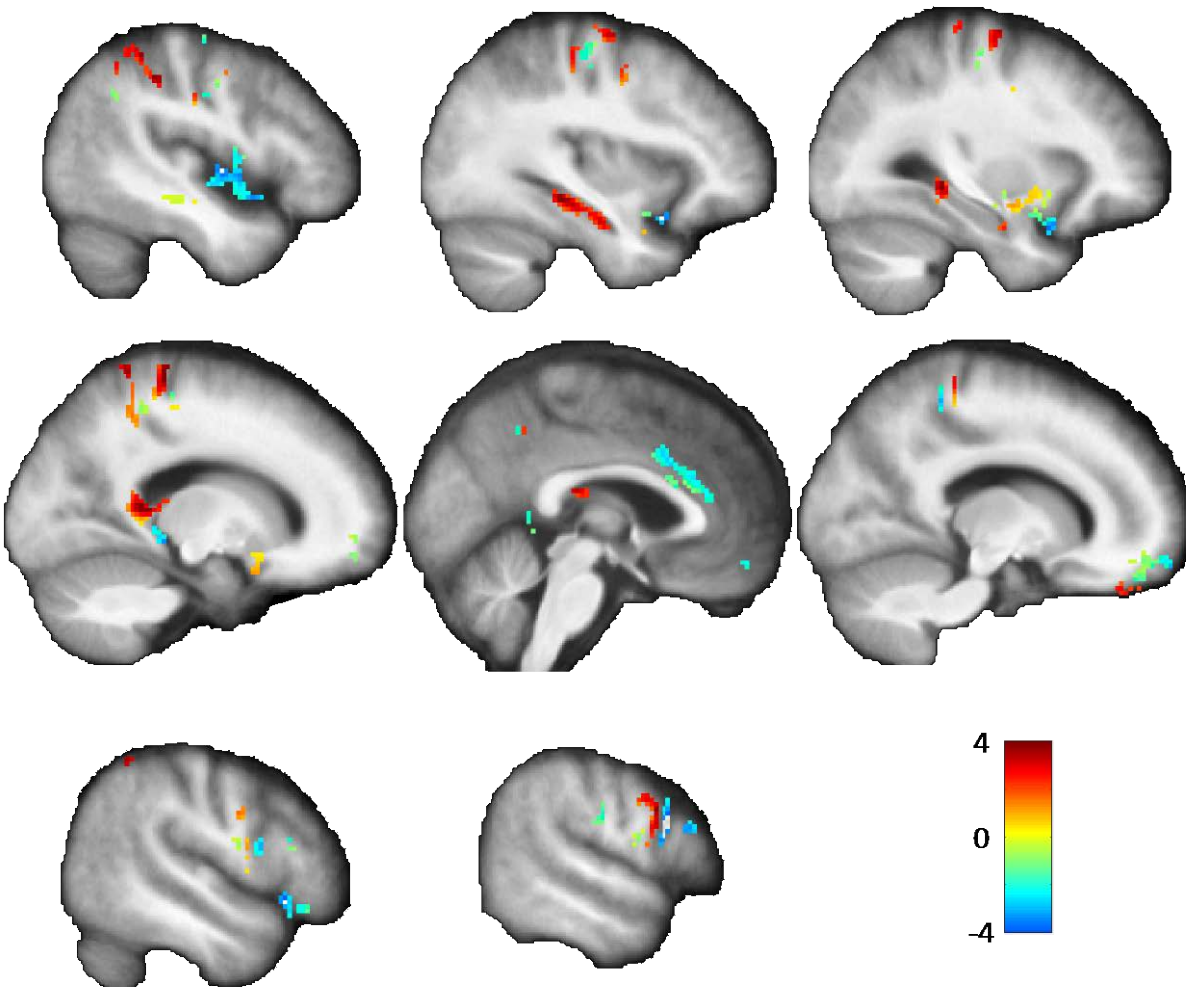

Figure 12. Most predictive voxels in mean diffusivity at baseline model. MD (at baseline) of regions significantly contributing to the prediction of remission. Colorbar indicates value of the parameter estimate (Z-score).

While some of the multiple feature models performed similarly (reported in table 7), they performed worse than their single feature counterparts. The best performing model amongst them combined the demographic/clinical features and the baseline emotion reactivity (median AUC of 0.661), however this is worse than either of the models separately. Similarly, the model that combined the demographic/clinical features and the baseline MD (median AUC of 0.602)

also performed worse than either model separately. We investigated the AUC of the full models (i.e., fitting a model on the full set of data and then predicting on that set as well) to understand whether this was possible due to over-fitting. We found that the AUC was highest for the MFA model combining demographic/clinical features and baseline emotion reactivity (AUC=1), but was much lower for both demographic/clinical features (AUC=0.785) and the emotion reactivity (AUC=0.834). This suggests that the models may have over-fit, which can occur when too many features are entered into the model.

Table 7. AUC of each of the multiple feature models. SVM=support vector machine; swLogistic=step-wise logistic model; AUC=area under the ROC curve.

Feature	Model	AUC median (IQR)
Demographic/Clinical and Baseline Emotion Reactivity	SVM	0.609 (0.79)
	Logistic	0.661 (0.058)
	swLogistic	0.581 (0.082)
Demographic/Clinical and Baseline Emotion Regulation	SVM	0.500 (0.096)
	Logistic	0.483 (0.056)
	swLogistic	0.582 (0.077)
Demographic/Clinical and Baseline EVC	SVM	0.490 (0.096)
	Logistic	0.495 (0.076)
	swLogistic	0.510 (0.088)
Demographic/Clinical and Baseline Perfusion	SVM	0.489 (0.042)
	Logistic	0.510 (0.075)
	swLogistic	0.471 (0.082)
Demographic/Clinical and Baseline Gray Matter Density	SVM	0.343 (0.061)
	Logistic	0.439 (0.051)
	swLogistic	0.602 (0.080)
Demographic/Clinical and Baseline FA	SVM	0.386 (0.095)
	Logistic	0.500 (0.064)
	swLogistic	0.543 (0.091)
Demographic/Clinical and Baseline MD	SVM	0.430 (0.058)
	Logistic	0.464 (0.063)
	swLogistic	0.564 (0.083)
Demographic/Clinical and Baseline Functional Neuroimaging	SVM	0.342 (0.059)
	Logistic	0.465 (0.074)
	swLogistic	0.423 (0.093)
Demographic/Clinical and Baseline and phMRI Emotion Reactivity	SVM	0.447 (0.102)
	Logistic	0.568 (0.072)
	swLogistic	0.573 (0.082)
Demographic/Clinical and Baseline and phMRI Emotion Regulation	SVM	0.500 (0.000)
	Logistic	0.495 (0.058)
	swLogistic	0.517 (0.092)
Demographic/Clinical and Baseline and phMRI EVC	SVM	0.380 (0.060)
	Logistic	0.458 (0.070)
	swLogistic	0.262 (0.067)
Demographic/Clinical and Baseline and phMRI Perfusion	SVM	0.320 (0.060)
	Logistic	0.530 (0.066)
	swLogistic	0.540 (0.054)

8.4 DISCUSSION

Kernel-based regression approaches are a promising approach to fitting models with a high number of features and limited sample size. These approaches reduce the feature space into a low-dimensional vector space (kernel), perform a model fitting approach, and then project back into the original space. Due to the 'kernel' trick, this fits models that behave non-linearly and by constraining the number of principal components used – only the vectors that explain the highest variance are used (which improves the overall generalizability of the models).

The single feature kernels performed well on both neuroimaging and demographic and clinical data, where a step-wise logistic regression model on the scores of the demographic and clinical data resulted in the best performing model (median AUC 0.735). The emotion reactivity task (at baseline) also performed well (median AUC 0.766). While some of the multiple feature sets performed well – they did not perform as well as either of their individual parts. This may either indicate a core flaw in the multi-factor based kernels or may be a result of the limited sample size. It is possible that the effectiveness of these kernels improves with a greater sample size as we can more robustly estimate the MFA kernel. A simulation-based study may help in better understanding the characteristics of such a machine learning approach and some of the possible assumptions.

Surprisingly, while the statistical results (chapter 6 and 7) suggested acute changes following a single dose of venlafaxine (i.e., the pharmacological effect), the models that utilized these neuroimaging data did not perform as well. This presents a clear difference between some of the statistical results and the models used to predict remission. While this may be a result of the poor performance of the MFA models, it is unclear why these features alone (using the single feature models) did not perform well. The difference between baseline and following single dose

neuroimaging was calculated and it is possible that a better measure relating these two imaging data can be computed (e.g., slope between baseline and single dose). Another limiting factor is that the kernel used may be diluted by non-specific (to the treatment effect) voxels as only a few regions indicated a specific response (i.e., insula or parahippocampus). A priori feature selection can help improve the specificity of the model.

Some of the most predictive clinical features were negative affect, MADRS items 8 (inability to feel) and 9 (pessimistic thoughts), and education (among others). The model suggests that non-remitters have worse negative affect, inability to feel, or pessimistic thoughts. Further, non-remitters tended to have higher education. In figure 11 and 12, negative parameter estimates suggest that greater activation or MD at baseline (in those regions) is predictive of worse outcome. Subsequently, positive parameter estimates suggest that greater activation or MD (in these regions) at baseline is predictive of better outcome.

While kernel based approaches offer a promising approach to fitting high-dimensional models, combining multiple feature sets may require greater samples to sufficiently estimate the MFA kernels and to avoid over-fitting. Simulation-based studies may help elucidate the assumptions and characteristics of such models. Several promising models utilized clinical/demographic data or baseline emotion reactivity activation, and leveraging the predictability of both models as well as managing to efficiently use the phMRI effects may result in the best models.

9.0 SUMMARY AND CONCLUSIONS

The current approach to treatment is an often-prolonged trial and error process that matches patients to a working regimen. A single trial may take several weeks before any sign or indication that the current regimen is working, and even then if the regimen is not working – patients are taken off that medication (which may take some time) and then another medication is used. This process is often lengthy and associated with patient worsening in a host of negative health outcomes (even sometimes their severity of depression). In LLD, this window is much longer and is associated with a greater risk of suicide. Thus, identifying useful biomarkers and generating machine learning models is critical for the future of personalized psychiatry.

We have identified a unique pattern of resting state connectivity and emotion reactivity changes that occur following only a single dose of medication. These changes reflect the future change in functional connectivity and reactivity and thus may reflect an early engagement of the implicit networks via increased synaptic monoamine occupancy. It is well-known that there monoamine occupancy increases within hours of receiving medication – and while the behavioral change may occur much later, the change in functional connectivity and activation changes acutely. This may serve as an important biomarker for remission.

Machine learning models have the unique capability of learning high dimensional ‘rules’ that are predictive of remission by using complex non-linear kernels. Our models have identified several key predictors of remission (in this sample), and future work can aim to increase the

sample sizes to properly fit more complex models. Ultimately, non-linear models can be generated with appropriately sized studies and these models are likely to be far better at predicting remission.

Future studies should aim to collect: genetic, demographic, behavioral and psychological assessments, cognitive batteries, past history of treatment, actigraphy (sleep and activity assessment), as well as neuroimaging data. A true multi-modal approach can be used to fit more complex and robust models of remission.

9.1 ACKNOWLEDGEMENTS

This work was supported by the NIH grants: NIMH R01 MH076079, K23 MH086686, and NIMH R01 MH111265.

BIBLIOGRAPHY

Abdi H, Williams LJ, Valentin D (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics* **5**(2): 149-179.

Aizenstein HJ, Andreescu C, Edelman KL, Cochran JL, Price J, Butters MA, *et al* (2011). fMRI correlates of white matter hyperintensities in late-life depression. *Am J Psychiatry* **168**(10): 1075-1082.

Aizenstein HJ, Butters MA, Clark KA, Figurski JL, Andrew Stenger V, Nebes RD, *et al* (2006). Prefrontal and striatal activation in elderly subjects during concurrent implicit and explicit sequence learning. *Neurobiology of aging* **27**(5): 741-751.

Aizenstein HJ, Butters MA, Figurski JL, Stenger VA, Reynolds CF, 3rd, Carter CS (2005). Prefrontal and striatal activation during sequence learning in geriatric depression. *Biological psychiatry* **58**(4): 290-296.

Aizenstein HJ, Butters MA, Wu M, Mazurkewicz LM, Stenger VA, Gianaros PJ, *et al* (2009). Altered functioning of the executive control circuit in late-life depression: episodic and persistent phenomena. *Am J Geriatr Psychiatry* **17**(1): 30-42.

Aizenstein HJ, Khalaf A, Walker SE, Andreescu C (2014). Magnetic resonance imaging predictors of treatment response in late-life depression. *J Geriatr Psychiatry Neurol* **27**(1): 24-32.

Alexopoulos GS (2002). Frontostriatal and limbic dysfunction in late-life depression. *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry* **10**(6): 687-695.

Alexopoulos GS, Hoptman MJ, Kanellopoulos D, Murphy CF, Lim KO, Gunning FM (2012). Functional connectivity in the cognitive control network and the default mode network in late-life depression. *J Affect Disord* **139**(1): 56-65.

Alexopoulos GS, Kelly RE, Jr. (2009). Research advances in geriatric depression. *World Psychiatry* **8**(3): 140-149.

Alexopoulos GS, Meyers BS, Young RC, Campbell S, Silbersweig D, Charlson M (1997). 'Vascular depression' hypothesis. *Arch Gen Psychiatry* **54**(10): 915-922.

Andreescu C, Reynolds CF, 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.

Andreescu C, Sheu LK, Tudorascu D, Gross JJ, Walker S, Banihashemi L, *et al* (2015). Emotion reactivity and regulation in late-life generalized anxiety disorder: functional connectivity at baseline and post-treatment. *Am J Geriatr Psychiatry* **23**(2): 200-214.

Andreescu C, Tudorascu DL, Butters MA, Tamburo E, Patel M, Price J, *et al* (2013). Resting state functional connectivity and treatment response in late-life depression. *Psychiatry Res* **214**(3): 313-321.

Arce E, Simmons AN, Lovero KL, Stein MB, Paulus MP (2008). Escitalopram effects on insula and amygdala BOLD activation during emotional processing. *Psychopharmacology (Berl)* **196**(4): 661-672.

Ashburner J (2007). A fast diffeomorphic image registration algorithm. *Neuroimage* **38**(1): 95-113.

Ashburner J, Andersson JL, Friston KJ (1999). High-dimensional image registration using symmetric priors. *Neuroimage* **9**(6 Pt 1): 619-628.

Ashburner J, Friston KJ (2000). Voxel-based morphometry--the methods. *Neuroimage* **11**(6 Pt 1): 805-821.

Ashburner J, Friston KJ (2005). Unified segmentation. *Neuroimage* **26**(3): 839-851.

Attwell D, Iadecola C (2002). The neural basis of functional brain imaging signals. *Trends Neurosci* **25**(12): 621-625.

Bammer R (2003). Basic principles of diffusion-weighted imaging. *Eur J Radiol* **45**(3): 169-184.

Banks SJ, Eddy KT, Angstadt M, Nathan PJ, Phan KL (2007). Amygdala-frontal connectivity during emotion regulation. *Social cognitive and affective neuroscience* **2**(4): 303-312.

Bar M, Aminoff E (2003). Cortical analysis of visual context. *Neuron* **38**(2): 347-358.

Battaglini M, Jenkinson M, De Stefano N (2012). Evaluating and reducing the impact of white matter lesions on brain volume measurements. *Hum Brain Mapp* **33**(9): 2062-2071.

Behrens TE, Woolrich MW, Jenkinson M, Johansen-Berg H, Nunes RG, Clare S, *et al* (2003). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magn Reson Med* **50**(5): 1077-1088.

Behzadi Y, Restom K, Liau J, Liu TT (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage* **37**(1): 90-101.

Berman MG, Peltier S, Nee DE, Kross E, Deldin PJ, Jonides J (2011). Depression, rumination and the default network. *Social cognitive and affective neuroscience* **6**(5): 548-555.

Bigos KL, Pollock BG, Aizenstein HJ, Fisher PM, Bies RR, Hariri AR (2008). Acute 5-HT reuptake blockade potentiates human amygdala reactivity. *Neuropsychopharmacology* **33**(13): 3221-3225.

Binnewijzend MA, Adriaanse SM, Van der Flier WM, Teunissen CE, de Munck JC, Stam CJ, *et al* (2014). Brain network alterations in Alzheimer's disease measured by eigenvector centrality in fMRI are related to cognition and CSF biomarkers. *Human brain mapping* **35**(5): 2383-2393.

Botteron K, Carter C, Castellanos FX, Dickstein DP, Drevets W, Kim KL, *et al* (2012). Consensus report of the APA work group on neuroimaging markers of psychiatric disorders. *American Psychiatric Association*.

Bourke JH, Wall MB (2015). phMRI: methodological considerations for mitigating potential confounding factors. *Front Neurosci* **9**: 167.

Bracewell RN (1989). The Fourier transform. *Sci Am* **260**(6): 86-89, 92-85.

Brassen S, Kalisch R, Weber-Fahr W, Braus DF, 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.

Breitenstein B, Scheuer S, Holsboer F (2014). Are there meaningful biomarkers of treatment response for depression? *Drug Discov Today* **19**(5): 539-561.

Brewer JA, Worhunsky PD, Gray JR, Tang YY, Weber J, Kober H (2011). Meditation experience is associated with differences in default mode network activity and connectivity. *Proc Natl Acad Sci U S A* **108**(50): 20254-20259.

Bruce ML, Ten Have TR, Reynolds CF, 3rd, Katz, II, Schulberg HC, Mulsant BH, *et al* (2004). Reducing suicidal ideation and depressive symptoms in depressed older primary care patients: a randomized controlled trial. *JAMA* **291**(9): 1081-1091.

Bruhl AB, Kaffenberger T, Herwig U (2010). Serotonergic and noradrenergic modulation of emotion processing by single dose antidepressants. *Neuropsychopharmacology* **35**(2): 521-533.

Bunney WE, Jr., Davis JM (1965). Norepinephrine in depressive reactions. A review. *Arch Gen Psychiatry* **13**(6): 483-494.

Carter CS, van Veen V (2007). Anterior cingulate cortex and conflict detection: an update of theory and data. *Cognitive, affective & behavioral neuroscience* **7**(4): 367-379.

Charney DS (1998). Monoamine dysfunction and the pathophysiology and treatment of depression. *J Clin Psychiatry* **59 Suppl 14**: 11-14.

Charney DS, Reynolds CF, 3rd, Lewis L, Lebowitz BD, Sunderland T, Alexopoulos GS, *et al* (2003). Depression and Bipolar Support Alliance consensus statement on the unmet needs in diagnosis and treatment of mood disorders in late life. *Arch Gen Psychiatry* **60**(7): 664-672.

Chen AC, Oathes DJ, Chang C, Bradley T, Zhou ZW, Williams LM, *et al* (2013). Causal interactions between fronto-parietal central executive and default-mode networks in humans. *Proceedings of the National Academy of Sciences of the United States of America* **110**(49): 19944-19949.

Chen CH, Ridler K, Suckling J, Williams S, Fu CH, Merlo-Pich E, *et al* (2007). Brain imaging correlates of depressive symptom severity and predictors of symptom improvement after antidepressant treatment. *Biol Psychiatry* **62**(5): 407-414.

Chen Y, Wang C, Zhu X, Tan Y, Zhong Y (2015). Aberrant connectivity within the default mode network in first-episode, treatment-naive major depressive disorder. *J Affect Disord* **183**: 49-56.

Chen Y, Wolk DA, Reddin JS, Korczykowski M, Martinez PM, Musiek ES, *et al* (2011). Voxel-level comparison of arterial spin-labeled perfusion MRI and FDG-PET in Alzheimer disease. *Neurology* **77**(22): 1977-1985.

Chou KL (2009). Age at onset of generalized anxiety disorder in older adults. *Am J Geriatr Psychiatry* **17**(6): 455-464.

Craig AD (2009). How do you feel--now? The anterior insula and human awareness. *Nature reviews Neuroscience* **10**(1): 59-70.

Cui X, Li J, Song X (2011). xjview: a viewing program for SPM. *Retreived from www alivelearn net/xjview8*.

Dai W, Garcia D, de Bazelaire C, Alsop DC (2008). Continuous flow-driven inversion for arterial spin labeling using pulsed radio frequency and gradient fields. *Magn Reson Med* **60**(6): 1488-1497.

Damoiseaux JS, Beckmann CF, Arigita EJ, Barkhof F, Scheltens P, Stam CJ, *et al* (2008). Reduced resting-state brain activity in the "default network" in normal aging. *Cerebral cortex* **18**(8): 1856-1864.

Davidson RJ, Irwin W, Anderle MJ, Kalin NH (2003). The neural substrates of affective processing in depressed patients treated with venlafaxine. *Am J Psychiatry* **160**(1): 64-75.

Delgado PL, Charney DS, Price LH, Aghajanian GK, Landis H, Heninger GR (1990). Serotonin function and the mechanism of antidepressant action. Reversal of antidepressant-induced remission by rapid depletion of plasma tryptophan. *Arch Gen Psychiatry* **47**(5): 411-418.

Detre JA, Leigh JS, Williams DS, Koretsky AP (1992). Perfusion imaging. *Magn Reson Med* **23**(1): 37-45.

Egner T, Etkin A, Gale S, Hirsch J (2008). Dissociable neural systems resolve conflict from emotional versus nonemotional distracters. *Cereb Cortex* **18**(6): 1475-1484.

Eloyan A, Shou H, Shinohara RT, Sweeney EM, Nebel MB, Cuzzocreo JL, *et al* (2014). Health effects of lesion localization in multiple sclerosis: spatial registration and confounding adjustment. *PLoS One* **9**(9): e107263.

Erk S, Mikschl A, Stier S, Ciaramidaro A, Gapp V, Weber B, *et al* (2010). Acute and sustained effects of cognitive emotion regulation in major depression. *J Neurosci* **30**(47): 15726-15734.

Etkin A, Buchel C, Gross JJ (2015). The neural bases of emotion regulation. *Nat Rev Neurosci* **16**(11): 693-700.

Etkin A, Egner T, Peraza DM, Kandel ER, Hirsch J (2006). Resolving emotional conflict: a role for the rostral anterior cingulate cortex in modulating activity in the amygdala. *Neuron* **51**(6): 871-882.

Etkin A, Prater KE, Hoeft F, Menon V, Schatzberg AF (2010). Failure of anterior cingulate activation and connectivity with the amygdala during implicit regulation of emotional processing in generalized anxiety disorder. *Am J Psychiatry* **167**(5): 545-554.

Farb NA, Anderson AK, Segal ZV (2012). The mindful brain and emotion regulation in mood disorders. *Can J Psychiatry* **57**(2): 70-77.

Feinstein JS, Stein MB, Paulus MP (2006). Anterior insula reactivity during certain decisions is associated with neuroticism. *Social cognitive and affective neuroscience* **1**(2): 136-142.

Fox MD, Snyder AZ, Vincent JL, Corbetta M, Van Essen DC, Raichle ME (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc Natl Acad Sci U S A* **102**(27): 9673-9678.

Fransson P, Marrelec G (2008). The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: Evidence from a partial correlation network analysis. *NeuroImage* **42**(3): 1178-1184.

Fu CH, Mourao-Miranda J, Costafreda SG, Khanna A, Marquand AF, Williams SC, *et al* (2008). Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. *Biol Psychiatry* **63**(7): 656-662.

Fu CH, Steiner H, Costafreda SG (2013). Predictive neural biomarkers of clinical response in depression: a meta-analysis of functional and structural neuroimaging studies of pharmacological and psychological therapies. *Neurobiol Dis* **52**: 75-83.

Fu CH, Williams SC, Brammer MJ, Suckling J, Kim J, Cleare AJ, *et al* (2007). Neural responses to happy facial expressions in major depression following antidepressant treatment. *Am J Psychiatry* **164**(4): 599-607.

Gaynes BN, Warden D, Trivedi MH, Wisniewski SR, Fava M, Rush AJ (2009). What did STAR*D teach us? Results from a large-scale, practical, clinical trial for patients with depression. *Psychiatric services* **60**(11): 1439-1445.

Gelenberg AJ, Freeman MP, Markowitz JC, Rosenbaum JF, Thase ME, Trivedi MH, *et al* (2010). Practice guideline for the treatment of patients with major depressive disorder third edition. *The American Journal of Psychiatry* **167**(10): 1.

Godlewska BR, Norbury R, Selvaraj S, Cowen PJ, Harmer CJ (2012). Short-term SSRI treatment normalises amygdala hyperactivity in depressed patients. *Psychol Med* **42**(12): 2609-2617.

Gold PW, Chrousos G, Kellner C, Post R, Roy A, Augerinos P, *et al* (1984). Psychiatric implications of basic and clinical studies with corticotropin-releasing factor. *Am J Psychiatry* **141**(5): 619-627.

Goldapple K, Segal Z, Garson C, Lau M, Bieling P, Kennedy S, *et al* (2004). Modulation of cortical-limbic pathways in major depression: treatment-specific effects of cognitive behavior therapy. *Arch Gen Psychiatry* **61**(1): 34-41.

Greicius MD, Flores BH, Menon V, Glover GH, Solvason HB, Kenna H, *et al* (2007). Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biological psychiatry* **62**(5): 429-437.

Groenewold NA, Opmeer EM, de Jonge P, Aleman A, Costafreda SG (2013). Emotional valence modulates brain functional abnormalities in depression: evidence from a meta-analysis of fMRI studies. *Neurosci Biobehav Rev* **37**(2): 152-163.

Gross JJ, John OP (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *J Pers Soc Psychol* **85**(2): 348-362.

Gusnard DA, Akbudak E, Shulman GL, Raichle ME (2001). Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America* **98**(7): 4259-4264.

Gyurak A, Gross JJ, Etkin A (2011). Explicit and implicit emotion regulation: a dual-process framework. *Cogn Emot* **25**(3): 400-412.

Hajnal JV, De Coene B, Lewis PD, Baudouin CJ, Cowan FM, Pennock JM, *et al* (1992). High signal regions in normal white matter shown by heavily T2-weighted CSF nulled IR sequences. *J Comput Assist Tomogr* **16**(4): 506-513.

Hamilton JP, Furman DJ, Chang C, Thomason ME, Dennis E, Gotlib IH (2011). Default-mode and task-positive network activity in major depressive disorder: implications for adaptive and maladaptive rumination. *Biol Psychiatry* **70**(4): 327-333.

Hariri AR, Mattay VS, Tessitore A, Fera F, Weinberger DR (2003). Neocortical modulation of the amygdala response to fearful stimuli. *Biol Psychiatry* **53**(6): 494-501.

Hariri AR, Tessitore A, Mattay VS, Fera F, Weinberger DR (2002). The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage* **17**(1): 317-323.

Hinrichsen GA, Hernandez NA (1993). Factors associated with recovery from and relapse into major depressive disorder in the elderly. *Am J Psychiatry* **150**(12): 1820-1825.

Holsboer F, Von Bardeleben U, Gerken A, Stalla GK, Muller OA (1984). Blunted corticotropin and normal cortisol response to human corticotropin-releasing factor in depression. *N Engl J Med* **311**(17): 1127.

Holtzheimer PE, Mayberg HS (2011). Stuck in a rut: rethinking depression and its treatment. *Trends in neurosciences* **34**(1): 1-9.

Jang JH, Jung WH, Kang DH, Byun MS, Kwon SJ, Choi CH, *et al* (2011). Increased default mode network connectivity associated with meditation. *Neurosci Lett* **487**(3): 358-362.

Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM (2012). Fsl. *Neuroimage* **62**(2): 782-790.

Joel I, Begley AE, Mulsant BH, Lenze EJ, Mazumdar S, Dew MA, *et al* (2014). Dynamic prediction of treatment response in late-life depression. *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry* **22**(2): 167-176.

Joyce KE, Laurienti PJ, Burdette JH, Hayasaka S (2010). A new measure of centrality for brain networks. *PloS one* **5**(8): e12200.

Kanal E, Borgstede JP, Barkovich AJ, Bell C, Bradley WG, Felmlee JP, *et al* (2002). American College of Radiology White Paper on MR Safety. *AJR Am J Roentgenol* **178**(6): 1335-1347.

Karim H, Andreescu C, Tudorascu D, Smagula S, Butters M, Karp J, *et al* (2016a). Intrinsic functional connectivity in late-life depression: trajectories over the course of pharmacotherapy in remitters and non-remitters. *Molecular psychiatry*.

Karim HT, Andreescu C, MacCloud RL, Butters MA, Reynolds CF, 3rd, Aizenstein HJ, *et al* (2016b). The effects of white matter disease on the accuracy of automated segmentation. *Psychiatry Res* **253**: 7-14.

Karim HT, Andreescu C, MacCloud RL, Butters MA, Reynolds CF, Aizenstein HJ, *et al* (2016c). The effects of white matter disease on the accuracy of automated segmentation. *Psychiatry Research: Neuroimaging* **253**: 7-14.

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.

Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry* **62**(6): 593-602.

Khalaf A, Karim H, Berkout OV, Andreescu C, Tudorascu D, Reynolds CF, *et al* (2016). Altered Functional Magnetic Resonance Imaging Markers of Affective Processing During Treatment of Late-Life Depression. *The American Journal of Geriatric Psychiatry* **24**(10): 791-801.

Klumpp H, Angstadt M, Phan KL (2012). Insula reactivity and connectivity to anterior cingulate cortex when processing threat in generalized social anxiety disorder. *Biological psychology* **89**(1): 273-276.

Kramer MS, Cutler N, Feighner J, Shrivastava R, Carman J, Sramek JJ, *et al* (1998). Distinct mechanism for antidepressant activity by blockade of central substance P receptors. *Science* **281**(5383): 1640-1645.

Kupfer DJ, Shaw DH, Ulrich R, Coble PA, Spiker DG (1982). Application of automated REM analysis in depression. *Arch Gen Psychiatry* **39**(5): 569-573.

Lang PJ, Bradley MM, Cuthbert BN (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. *Technical report A-8*.

Langenecker SA, Kennedy SE, Guidotti LM, Briceno EM, Own LS, Hooven T, *et al* (2007). Frontal and limbic activation during inhibitory control predicts treatment response in major depressive disorder. *Biol Psychiatry* **62**(11): 1272-1280.

Leech R, Braga R, Sharp DJ (2012). Echoes of the brain within the posterior cingulate cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **32**(1): 215-222.

Lemogne C, Mayberg H, Bergouignan L, Volle E, Delaveau P, Lehericy S, *et al* (2010). Self-referential processing and the prefrontal cortex over the course of depression: a pilot study. *J Affect Disord* **124**(1-2): 196-201.

Leppanen JM (2006). Emotional information processing in mood disorders: a review of behavioral and neuroimaging findings. *Curr Opin Psychiatry* **19**(1): 34-39.

Li W, Mai X, Liu C (2014). The default mode network and social understanding of others: what do brain connectivity studies tell us. *Front Hum Neurosci* **8**: 74.

Linden DE (2006). How psychotherapy changes the brain--the contribution of functional neuroimaging. *Mol Psychiatry* **11**(6): 528-538.

Lisiecka D, Meisenzahl E, Scheuerecker J, Schoepf V, Whitty P, Chaney A, *et al* (2011). Neural correlates of treatment outcome in major depression. *Int J Neuropsychopharmacol* **14**(4): 521-534.

Liston C, Chen AC, Zebley BD, Drysdale AT, Gordon R, Leuchter B, *et al* (2014). Default mode network mechanisms of transcranial magnetic stimulation in depression. *Biol Psychiatry* **76**(7): 517-526.

Lohmann G, Margulies DS, Horstmann A, Pleger B, Lepsien J, Goldhahn D, *et al* (2010). Eigenvector centrality mapping for analyzing connectivity patterns in fMRI data of the human brain. *PLoS One* **5**(4): e10232.

Loubinoux I, Pariente J, Boulanouar K, Carel C, Manelfe C, Rascol O, *et al* (2002). A single dose of the serotonin neurotransmission agonist paroxetine enhances motor output: double-blind, placebo-controlled, fMRI study in healthy subjects. *Neuroimage* **15**(1): 26-36.

Lui S, Wu Q, Qiu L, Yang X, Kuang W, Chan RC, *et al* (2011). Resting-state functional connectivity in treatment-resistant depression. *Am J Psychiatry* **168**(6): 642-648.

Marchetti I, Koster EH, Sonuga-Barke EJ, De Raedt R (2012). The default mode network and recurrent depression: a neurobiological model of cognitive risk factors. *Neuropsychology review* **22**(3): 229-251.

Mauss IB, Cook CL, Cheng JY, Gross JJ (2007). Individual differences in cognitive reappraisal: experiential and physiological responses to an anger provocation. *Int J Psychophysiol* **66**(2): 116-124.

Mazaika P, Whitfield-Gabrieli S, Reiss A, Glover G (2007). Artifact repair for fMRI data from high motion clinical subjects. *Human Brain Mapping*: 2007.

McGrath CL, Kelley ME, Holtzheimer PE, Dunlop BW, Craighead WE, Franco AR, *et al* (2013). Toward a neuroimaging treatment selection biomarker for major depressive disorder. *JAMA Psychiatry* **70**(8): 821-829.

McGraw KO, Wong SP (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods* **1**(1): 30.

Menon V, Uddin LQ (2010). Saliency, switching, attention and control: a network model of insula function. *Brain structure & function* **214**(5-6): 655-667.

Meyer JH, Wilson AA, Ginovart N, Goulding V, Hussey D, Hood K, *et al* (2001). Occupancy of serotonin transporters by paroxetine and citalopram during treatment of depression: a [(11)C]DASB PET imaging study. *Am J Psychiatry* **158**(11): 1843-1849.

Mikl M, Marecek R, Hlustik P, Pavlicova M, Drastich A, Chlebus P, *et al* (2008). Effects of spatial smoothing on fMRI group inferences. *Magn Reson Imaging* **26**(4): 490-503.

Miller HL, Delgado PL, Salomon RM, Berman R, Krystal JH, Heninger GR, *et al* (1996). Clinical and biochemical effects of catecholamine depletion on antidepressant-induced remission of depression. *Arch Gen Psychiatry* **53**(2): 117-128.

Miskowiak K, Papadatou-Pastou M, Cowen PJ, Goodwin GM, Norbury R, Harmer CJ (2007). Single dose antidepressant administration modulates the neural processing of self-referent personality trait words. *Neuroimage* **37**(3): 904-911.

Montgomery SA, Asberg M (1979). A new depression scale designed to be sensitive to change. *The British journal of psychiatry : the journal of mental science* **134**: 382-389.

Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B (2007). Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *Lancet* **370**(9590): 851-858.

Mugler JP, 3rd, Brookeman JR (1990). Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). *Magn Reson Med* **15**(1): 152-157.

Mulsant BH, Houck PR, Gildengers AG, Andreescu C, Dew MA, Pollock BG, *et al* (2006). What is the optimal duration of a short-term antidepressant trial when treating geriatric depression? *J Clin Psychopharmacol* **26**(2): 113-120.

Murphy SE, Norbury R, O'Sullivan U, Cowen PJ, Harmer CJ (2009). Effect of a single dose of citalopram on amygdala response to emotional faces. *Br J Psychiatry* **194**(6): 535-540.

Nejad AB, Fossati P, Lemogne C (2013). Self-referential processing, rumination, and cortical midline structures in major depression. *Frontiers in human neuroscience* **7**: 666.

Nelson JC, Delucchi KL, Schneider LS (2013). Moderators of outcome in late-life depression: a patient-level meta-analysis. *Am J Psychiatry* **170**(6): 651-659.

Nemeroff CB, Widerlov E, Bissette G, Walleus H, Karlsson I, Eklund K, *et al* (1984). Elevated concentrations of CSF corticotropin-releasing factor-like immunoreactivity in depressed patients. *Science* **226**(4680): 1342-1344.

Nichols TE, Holmes AP (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping* **15**(1): 1-25.

Ochsner KN, Bunge SA, Gross JJ, Gabrieli JD (2002). Rethinking feelings: an FMRI study of the cognitive regulation of emotion. *J Cogn Neurosci* **14**(8): 1215-1229.

Ochsner KN, Silvers JA, Buhle JT (2012). Functional imaging studies of emotion regulation: a synthetic review and evolving model of the cognitive control of emotion. *Annals of the New York Academy of Sciences* **1251**: E1-24.

Ogawa S, Lee TM (1990a). Magnetic resonance imaging of blood vessels at high fields: in vivo and in vitro measurements and image simulation. *Magn Reson Med* **16**(1): 9-18.

Ogawa S, Lee TM, Nayak AS, Glynn P (1990b). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn Reson Med* **14**(1): 68-78.

Okon-Singer H, Hendler T, Pessoa L, Shackman AJ (2015). The neurobiology of emotion-cognition interactions: fundamental questions and strategies for future research. *Front Hum Neurosci* **9**: 58.

Ovbiagele B, Saver JL (2006). Cerebral white matter hyperintensities on MRI: Current concepts and therapeutic implications. *Cerebrovasc Dis* **22**(2-3): 83-90.

Parsey RV, Kent JM, Oquendo MA, Richards MC, Pratap M, Cooper TB, *et al* (2006). Acute occupancy of brain serotonin transporter by sertraline as measured by [11C]DASB and positron emission tomography. *Biol Psychiatry* **59**(9): 821-828.

Patel AX, Kundu P, Rubinov M, Jones PS, Vertes PE, Ersche KD, *et al* (2014). A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. *Neuroimage* **95**: 287-304.

Patel K, Abdool PS, Rajji TK, Mulsant BH (2017). Pharmacotherapy of major depression in late life: what is the role of new agents? *Expert Opin Pharmacother* **18**(6): 599-609.

Paulus MP, Rogalsky C, Simmons A, Feinstein JS, Stein MB (2003). Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. *NeuroImage* **19**(4): 1439-1448.

Paulus MP, Stein MB (2006). An insular view of anxiety. *Biol Psychiatry* **60**(4): 383-387.

Peng D, Liddle EB, Iwabuchi SJ, Zhang C, Wu Z, Liu J, *et al* (2015). Dissociated large-scale functional connectivity networks of the precuneus in medication-naive first-episode depression. *Psychiatry Res* **232**(3): 250-256.

Penny W, Friston K, Ashburner J, Kiebel S (2007). *Statistical parametric mapping: the analysis of functional brain images*.

Penny WD, Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE (2011). *Statistical parametric mapping: the analysis of functional brain images* Academic press.

Pizzagalli DA (2011). Frontocingulate dysfunction in depression: toward biomarkers of treatment response. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* **36**(1): 183-206.

Poline JB, Worsley KJ, Holmes AP, Frackowiak RS, Friston KJ (1995). Estimating smoothness in statistical parametric maps: variability of p values. *J Comput Assist Tomogr* **19**(5): 788-796.

Rawlings NB, Norbury R, Cowen PJ, Harmer CJ (2010). A single dose of mirtazapine modulates neural responses to emotional faces in healthy people. *Psychopharmacology (Berl)* **212**(4): 625-634.

Reynolds CF, 3rd, Dew MA, Pollock BG, Mulsant BH, Frank E, Miller MD, *et al* (2006). Maintenance treatment of major depression in old age. *N Engl J Med* **354**(11): 1130-1138.

Riso LP, Thase ME, Howland RH, Friedman ES, Simons AD, Tu XM (1997). A prospective test of criteria for response, remission, relapse, recovery, and recurrence in depressed patients treated with cognitive behavior therapy. *Journal of affective disorders* **43**(2): 131-142.

Rosazza C, Minati L, Ghielmetti F, Mandelli ML, Bruzzone MG (2012). Functional connectivity during resting-state functional MR imaging: study of the correspondence between independent component analysis and region-of-interest-based methods. *AJNR American journal of neuroradiology* **33**(1): 180-187.

Rutherford B, Sneed J, Miyazaki M, Eisenstadt R, Devanand D, Sackeim H, *et al* (2007). An open trial of aripiprazole augmentation for SSRI non-remitters with late-life depression. *Int J Geriatr Psychiatry* **22**(10): 986-991.

Schaefer A, Burmann I, Regenthal R, Arelin K, Barth C, Pampel A, *et al* (2014). Serotonergic modulation of intrinsic functional connectivity. *Curr Biol* **24**(19): 2314-2318.

Schildkraut JJ, Gordon EK, Durell J (1965). Catecholamine metabolism in affective disorders. I. Normetanephrine and VMA excretion in depressed patients treated with imipramine. *J Psychiatr Res* **3**(4): 213-228.

Shamay-Tsoory SG (2011). The neural bases for empathy. *Neuroscientist* **17**(1): 18-24.

Sheline YI, Pieper CF, Barch DM, Welsh-Bohmer K, McKinstry RC, MacFall JR, *et al* (2010a). Support for the vascular depression hypothesis in late-life depression: results of a 2-site, prospective, antidepressant treatment trial. *Arch Gen Psychiatry* **67**(3): 277-285.

Sheline YI, Price JL, Yan Z, Mintun MA (2010b). Resting-state functional MRI in depression unmasks increased connectivity between networks via the dorsal nexus. *Proc Natl Acad Sci U S A* **107**(24): 11020-11025.

Shirer WR, Ryali S, Rykhlevskaia E, Menon V, Greicius MD (2012). Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cereb Cortex* **22**(1): 158-165.

Simmons WK, Avery JA, Barcalow JC, Bodurka J, Drevets WC, Bellgowan P (2013). Keeping the body in mind: insula functional organization and functional connectivity integrate interoceptive, exteroceptive, and emotional awareness. *Human brain mapping* **34**(11): 2944-2958.

Sladky R, Friston KJ, Trostl J, Cunnington R, Moser E, Windischberger C (2011). Slice-timing effects and their correction in functional MRI. *Neuroimage* **58**(2): 588-594.

Smith SM, Fox PT, Miller KL, Glahn DC, Fox PM, Mackay CE, *et al* (2009). Correspondence of the brain's functional architecture during activation and rest. *Proc Natl Acad Sci U S A* **106**(31): 13040-13045.

Sperling RA, Bates JF, Cocchiarella AJ, Schacter DL, Rosen BR, Albert MS (2001). Encoding novel face-name associations: a functional MRI study. *Hum Brain Mapp* **14**(3): 129-139.

Sprengelmeyer R, Steele JD, Mwangi B, Kumar P, Christmas D, Milders M, *et al* (2011). The insular cortex and the neuroanatomy of major depression. *J Affect Disord* **133**(1-2): 120-127.

Sridharan D, Levitin DJ, Menon V (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proceedings of the National Academy of Sciences of the United States of America* **105**(34): 12569-12574.

Stevens JA, Hasbrouck LM, Durant TM, Dellinger AM, Batabyal PK, Crosby AE, *et al* (1999). Surveillance for injuries and violence among older adults. *MMWR CDC Surveill Summ* **48**(8): 27-50.

Sullivan PF, Neale MC, Kendler KS (2000). Genetic epidemiology of major depression: review and meta-analysis. *Am J Psychiatry* **157**(10): 1552-1562.

Takahashi H, Yahata N, Koeda M, Takano A, Asai K, Suhara T, *et al* (2005). Effects of dopaminergic and serotonergic manipulation on emotional processing: a pharmacological fMRI study. *NeuroImage* **27**(4): 991-1001.

Taylor WD, Aizenstein HJ, Alexopoulos GS (2013). The vascular depression hypothesis: mechanisms linking vascular disease with depression. *Mol Psychiatry* **18**(9): 963-974.

Taylor WD, Steffens DC, MacFall JR, McQuoid DR, Payne ME, Provenzale JM, *et al* (2003). White matter hyperintensity progression and late-life depression outcomes. *Arch Gen Psychiatry* **60**(11): 1090-1096.

Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L, *et al* (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *The American journal of psychiatry* **163**(1): 28-40.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, *et al* (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15**(1): 273-289.

Vink D, Aartsen MJ, Schoevers RA (2008). Risk factors for anxiety and depression in the elderly: a review. *J Affect Disord* **106**(1-2): 29-44.

Wang L, Krishnan KR, Steffens DC, Potter GG, Dolcos F, McCarthy G (2008a). Depressive state- and disease-related alterations in neural responses to affective and executive challenges in geriatric depression. *Am J Psychiatry* **165**(7): 863-871.

Wang Y, Xu C, Cao X, Gao Q, Li J, Liu Z, *et al* (2012). Effects of an antidepressant on neural correlates of emotional processing in patients with major depression. *Neurosci Lett* **527**(1): 55-59.

Wang Z, Aguirre GK, Rao H, Wang J, Fernandez-Seara MA, Childress AR, *et al* (2008b). Empirical optimization of ASL data analysis using an ASL data processing toolbox: ASLtbx. *Magn Reson Imaging* **26**(2): 261-269.

Warden D, Trivedi MH, Wisniewski SR, Davis L, Nierenberg AA, Gaynes BN, *et al* (2007). Predictors of attrition during initial (citalopram) treatment for depression: a STAR*D report. *The American journal of psychiatry* **164**(8): 1189-1197.

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

Wink AM, de Munck JC, van der Werf YD, van den Heuvel OA, Barkhof F (2012). Fast eigenvector centrality mapping of voxel-wise connectivity in functional magnetic resonance imaging: implementation, validation, and interpretation. *Brain Connect* **2**(5): 265-274.

Wong ML, Licinio J (2001). Research and treatment approaches to depression. *Nat Rev Neurosci* **2**(5): 343-351.

Wright CI, Wedig MM, Williams D, Rauch SL, Albert MS (2006). Novel fearful faces activate the amygdala in healthy young and elderly adults. *Neurobiol Aging* **27**(2): 361-374.

Wu M, Andreescu C, Butters MA, Tamburo R, Reynolds CF, 3rd, Aizenstein H (2011). Default-mode network connectivity and white matter burden in late-life depression. *Psychiatry research* **194**(1): 39-46.

Wu M, Rosano C, Butters M, Whyte E, Nable M, Crooks R, *et al* (2006). A fully automated method for quantifying and localizing white matter hyperintensities on MR images. *Psychiatry Res* **148**(2-3): 133-142.

Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, *et al* (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage* **31**(3): 1116-1128.

Zhu X, Wang X, Xiao J, Liao J, Zhong M, Wang W, *et al* (2012). Evidence of a dissociation pattern in resting-state default mode network connectivity in first-episode, treatment-naive major depression patients. *Biol Psychiatry* **71**(7): 611-617.

Zuo XN, Ehmke R, Mennes M, Imperati D, Castellanos FX, Sporns O, *et al* (2012). Network centrality in the human functional connectome. *Cerebral cortex* **22**(8): 1862-1875.