Development of a Neuroinformatics Pipeline and its Application to Gene-environment Interaction in Neurodegenerative Disease

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Shauna Marie Overgaard

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Gyorgy Simon, Ph.D., Advisor, and Laël Gatewood, Ph.D., Co-advisor

December 2020

ACKNOWLEDGEMENTS

Dr. Gyorgy Simon, Dr. Prashanthi Vemuri, Dr. Laël Gatewood, and Dr. Weihua Guan, thank you. My dissertation committee saw my potential, believed in me and my ideas, and encouraged me to work at the cutting edge. Rather than face resistance for being “too big,” my vision was framed as innovative – never diminished but always seriously vetted – and for this support, I am grateful. A special appreciation to Dr. Vemuri whose expertise, collaborative nature, foundational work, example, laboratory, and tools sustained the project. Dr. Vemuri, Dr. Simon, and Dr. Guan thank you for being my tripartite advising team. You have invested your time, your intellect, your ideas, and the significance of your contributions will continue to increase the value of this work. Dr. Gatewood, you have set an example of excellence as a researcher, mentor, and role model. Thank you for being the anchor of my work and my light at the end of the tunnel. The committee members’ intellectual prowess, technical and domain expertise, autonomy support, and expectation of excellence made them my “dream team.” Importantly, I own all imperfections in this work and am extremely thankful for time these experts gave to me. From these mentors, I especially came to appreciate the following bits of wisdom: No model is perfect. In order to be skeptical, one must become disenchanted. Require documentation and reproducibility from other scientists. Accept when you are wrong, and you will be wrong. Always have reservations about causal claims and expect spurious correlations. Seek out experts. Embody Resilience and Grit. Regardless of circumstances, *never give up.*

My husband (Dr. Josh Overgaard) and daughters (Isabella, Geneviève, and Élise) continuously offer patience and unconditional love, which present daily as comfort, clarity, hope, laughter, and reality. They strap on my wings and give me life.

Mike McInerney (my father) and James McInerney (my brother) breathe confidence in me and stoke my inner fire and passion.

Ryan McInerney, my brother, and intellectual ally went through this journey first and redefined normalcy. He continues to inspire me with his brilliance.

Ranyah Aldekhyyel, my bosom friend and soul sister, has been my safe place for candor, empathy, laughter, and rejuvenation. Through graduate education and into the unknown – my colleague, sister, trusted companion.

I am grateful for the informatics community, the American Medical Informatics Association (AMIA), who continue to lay the groundwork and who share a similar vision. I am grateful for and acknowledge the Alzheimer’s Disease Neuroimaging Initiative (ADNI) scientists, particularly Dr. Arthur Toga, Dr. Clifford Jack, and Matthew Senjem, MS, for their work spearheading and developing neuroimaging and informatics protocols for automated analyses. I am grateful for Dr. Clifford Jack, Dr. Kejal Kantarci, and Dr. Prashanthi Vemuri’s leadership of the Mayo Clinic Aging and Dementia Imaging Research (ADIR) laboratory, and the ADIR programming and medical imaging staff for their work processing and analyzing ADNI data. The data used in preparation of this manuscript were approved by and obtained from ADNI. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the

Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

I also acknowledge and am thankful for the continued data analysis software development by the R-Project team, for the Tetrad team funded by of the Center for Causal Discovery, supported by grant U54HG008540, and for the pioneering work of Dr. Olaf Sporns and Dr. Danielle Bassett in their study and documentation of the structure and function of brain networks.

For years to come, the "current state" will continue to reflect only "the beginning" of artificial intelligence and precision medicine in neuroinformatics. There will always be so much more to be imagined and produced. The brain will be measured, and therapies enhanced well beyond the sophistication awarded to other organs. I shamelessly admit my excitement for the future of our field.

DEDICATION

This work is dedicated to my mother, Johanne (Anfossi) McInerney, who passed away during the completion of this work – but who sees me through to each end and continues to guide me.

This work is also dedicated to patients and to the loved ones of patients who suffer or have watched others suffer from a brain disorder or injury. There is no pain like losing yourself or the one you love.

ABSTRACT

Background

Alzheimer's disease is a neurodegenerative disease and the sixth-leading cause of death in the United States. While there has been a greater understanding of Alzheimer's disease (AD) processes in the last two decades, clinical trials in AD have not been successful, suggesting that further research is needed to understand key questions pertaining to the underpinnings of the disease. Alzheimer's disease is a complex disease with significant heterogeneity in the disease progression and expression of clinical symptoms. The presence of the ϵ4 allele of the Apolipoprotein (*APOE4*) increases the risk of Alzheimer's disease and is associated with earlier onset of Alzheimer's disease pathology. On the other hand, variability in "Resilience," i.e., the ability to cope with Alzheimer's disease pathology, is associated with the differences in the expression of clinical symptoms.

Objective

The work presented draws on the methodological strengths of health informatics, biostatistics, and neuroscience, to achieve two specific aims: (1) Deployment of a neuroinformatics pipeline for the replicable collection, manipulation, and analysis of AD data (2) Employment of the neuroinformatics pipeline to evaluate the potential impact of specific allele carriership on what is recognized as a resilience mechanisms in the context of AD. Specifically, the neuroscience questions are: (1) Does the effect of cognitive reserve on GMD differ by *APOE4* genotype? (2) Does *APOE4* carrier status impact clinical functioning, and is this effect mediated by global efficiency? (3) Do *APOE4* carriers as compared to non-carriers demonstrate differences in network recruitment (specifically, global efficiency of the default mode network)?

Methods

To evaluate the complex interplay of *gene-environment interactions in AD,* we investigated the impact of *APOE4* and education on brain structure in the first study. In our second study, we used a core construct from graph theory to compute global efficiency on single-subject 3T MRI scans and evaluated the interplay between pathology, *APOE4*, education, and global efficiency and their impact on clinical functioning. In our third study, we applied causal inference models to investigate the causal relationships between pathology, *APOE4*, education, and global efficiency, considered drivers of clinical functioning in AD.

Conclusion

This work uniquely contributes to health informatics through the construction of a neuroinformatics pipeline which combines multimodal biomedical data (neuroimaging, genomics, cognition, and clinical), employs database management, automated computing, graph theory, and biostatistics to answer clinical questions. This work contributes to science by proposing a method to measure and monitor brain health, providing additional insight into the mechanistic underpinnings of *APOE4* allele carriership underlying AD pathology.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

1.1 Background

In the United States, 5.7 million Americans suffer from Alzheimer's disease (AD), which is recognized as the most common cause of dementia among people over the age of 65. AD is a neurodegenerative disease and the sixth-leading cause of death in the United States. The economic burden of AD is estimated to be more than $250 billion yearly. The impact is more than financial, however: it manifests as a physical and emotional burden as well. More than 15 million Americans provide unpaid care for individuals with AD or another dementia.

1.2 Objective

While there has been a greater understanding of AD disease processes in the last two decades, clinical trials in AD have not been successful, suggesting a need for further research to understand key questions pertaining to the underpinnings of the disease.[1] The National Institutes of Health (NIH) recognizes this knowledge gap and supports the availability of databases containing relevant genomic, imaging, and proteomic data. Specifically, three national databases have been derived from this initiative: Alzheimer's Disease Neuroimaging Initiative (ADNI), Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL), and ADNI Department of Defense (ADNI-DOD). The wide availability of these databases has spurred international collaboration of scientists in the discovery of biomarkers and mechanisms for AD. To exploit these data, informatics tools built to merge and leverage disparate data sources, as well as the processing power to manipulate high-dimensional data, are necessary. There is a need for solid pipelines for the replicable collection, manipulation, and analysis of data to produce meaningful and verifiable information (Aim I).

The only established primary predictors of AD are the presence of the ϵ4 allele of the Apolipoprotein (*APOE4*) and age. The brain's ability to maximize performance and utilize alternative networks in the face of damage may be compromised due to the carriership of *APOE4*. There is compelling evidence showing an increased risk of AD in *APOE4* carriers versus carriers of other allelic variants (i.e., *APOE2*, *APOE3*). There is also significant heterogeneity in the risk of dementia in those with *APOE4* allele based on an individual's Resilience.

Inconsistencies in brain pathology such as the presence of Amyloid-β (Aβ) plaques, a major hallmark of AD[89,95], exist between healthy individuals with the same cognitive function. Deposition of Aβ into insoluble plaques is the earliest sign of AD, and aggregation depends on several factors, including the rate of production, clearance from the brain interstitial fluid (ISF) (where amyloid plaques are found), and the rate of fibrillation, all of which may be influenced by *APOE*. Such inconsistencies have been known to relate to lifelong enriching experiences such as education, occupational activities, bilingualism, and cognitive activity.

The theory of Reserve suggests the existence of a Resilience mechanism in the face of neurodegeneration and the presence of Aβ, and the data are supportive of such a mechanistic relation to gray matter structures based on clinical evidence of damage to the observable brain or to measures of cognition. The complex interplay between Reserve, *APOE4*, and their combined role in AD is an important neuroscience question that remains unsolved.

Cognitive Reserve (CR) is a manifestation of Resilience and a phenomenon which is used to explain individual variability in vulnerability to dementia. Education, a proxy for CR, has been hypothesized to generate protective effects against gray matter atrophy, enhance the expression of the plasticity gene, erg-1, and diminish *A*β deposition (the accumulation of Aβ is known to eventually cause neurodegeneration, a reduction of the brain's gray matter density (GMD)). We therefore aim to investigate these complexities through the employment of neuroinformatics tools, leveraging a unique position to construct and employ a neuroinformatics pipeline to address key neuroscience questions in this realm (Aim II): (1) Do education and APOE genotype differentially impact GMD? (2) Does *APOE4* carrier status impact clinical functioning, and is the effect mediated by global efficiency? (3) Do *APOE4* carriers as compared to non-carriers demonstrate differences in network recruitment (specifically, global efficiency of the default mode network)?

1.3 Methods

If the presence of *APOE4* leads to neurodegeneration (a reduction in GMD) at a rate higher than the absence of *APOE4*, in situations of equal CR, carriers may demonstrate comparative decreases in GMD. Objective 1 of this work was to test whether education and differentially impact GMD. Hypothesis 1: Education and *APOE* genotype differentially impact GMD.

As the brain is built on multi-scale interactions, the evaluation of cognition should include a measure of the density of brain matter combined with larger-scale network patterns (such as the default mode network) in order to observe activity and estimate the transfer of information. Network analysis allows for the evaluation of many regions as a pattern, rather than analyzing a single brain region of interest. The overall capacity for integrative processing may be associated with CR. Integrative processing can be evaluated using a core construct from graph theory known as global efficiency which is a measure of network analysis. Objective 2 of this work was to test whether *APOE4* carrier status affects clinical functioning, and if so, whether these effects are impacted by global efficiency. Hypothesis 2: *APOE4* carrier status affects clinical functioning, and the effect is mediated by global efficiency.

AD has been described as a disconnection syndrome that is characterized by disruptions in brain network that tend to overlap areas of known pathology.[2] The evaluation of CR as a contributor to a dynamic system that includes *APOE* ought to advance the understanding of CR's active resilience mechanisms and add to the groundwork for the development of innovative translational approaches and the evaluation of techniques for clinical intervention in AD. *APOE4* affects the biological drivers associated with neurodegeneration (e.g., amyloid). Neurodegeneration may increase the potential for reduced global brain network efficiency (integration). We aim to provide insight into the effect of *APOE4* on the neural basis of cognitive reserve, expecting that answers to our questions may contribute to identifying targets for intervention of neurodegeneration. *Objective 3 of this work was to test whether APOE4 carriers as compared to non-carriers demonstrate differences in network recruitment (specifically, global efficiency of the default mode network).* Hypothesis 3: *APOE4* carriers as compared to non-carriers will demonstrate differences in recruitment of the default mode network, as measured by interaction effects of global efficiency and *APOE4* carriership on functioning.

The specific neurophysiological mechanisms that facilitate the effective integration of experience and the development of neuroprotective intellectual abilities are unclear. The multifaceted nature of the disease construct points to the interaction of multiple contributing factors. In Objective 3, we discussed the statistical significance and variable combinations of the interdependencies of the biomarkers (i.e., *APOE4*, global efficiency, Aβ) evaluated.

Correlation does not imply causation; therefore, contributing factors (i.e., *APOE4*, global efficiency, Aβ) evaluated in Objective 3 were subjected to an objective search process, through the application of Fast Causal Inference (FCI), a reputable algorithm for causal discovery. We addressed the question of whether the relationships between variables evaluated for statistical significance and interactions in Objective 3 were causal (employing the FCI algorithm).

The overarching goal of this work was to create an informatics pipeline (Aim I) in order to strategically address complicated questions in neuroscience related to AD (Aim II). Neuroinformatics methods were employed to pinpoint whether individuals with the *APOE4* allele benefit equally from the publicized preventative strategy known as CR. The subsequently generated hypotheses operate under the premise that CR and brain organization are associated with each other (i.e., high CR may mean better network efficiency and vice versa) and each of the covariates (i.e., age, Aβ, CR, gender) have an impact on cognitive health. Under these assumptions, the use of neuroinformatics in a well-constructed combination of data pertaining to network efficiency, genetic predisposition (*APOE4*), pathology (Aβ, GMD), and demographics (age, gender) associated in an instance of injury positions us to carefully study the effect of CR in the presence (or absence) of *APOE4*.

1.4 Conclusion

The work presented draws on the methodological strengths of health informatics, biostatistics, and neuroscience, to evaluate the potential impact of specific allele carriership on what is recognized as a buffer to injury for the rest of the population. This work uniquely contributes to science through the construction of a neuroinformatics pipeline which combines multimodal biomedical data (neuroimaging, genomics, cognition, and clinical), and which employs database management, automated computing, graph theory, and biostatistics to answer complex clinical questions.

1.5 High-Level Overview of Chapter Contents

Chapter 2 describes both Aims in detail and provides the necessary background in terms of methods and neuroscience to understand the rest of the thesis. We provide an overview of our informatics pipeline (Aim I) through which a scientist can target the evaluation of challenging questions in neuroscience (Aim II).

Chapter 3 describes the informatics pipeline's structure through a visual framework and provides a summary of each step for ease of reference.

Chapter 4 provides a thorough description of the methodology specific to each of the three objectives and explains the links amongst the advancing hypotheses.

Chapter 5 provides a summary of accomplishments and contributions of the work to informatics and neuroscience through Aims I and II, comments on the generalizability of the work, proposes interpretations for consideration and provides conclusions based on findings and external studies.

Appendix provides the R code constructed for Aim I. Code is also available at the following GitHub web address: https://github.com/shaunaovergaard/neuroinformatics.git

.

Chapter 2 Background

2.1 Aim I: Development of a Neuroinformatics Pipeline

A. Neuroimaging Data Initiative

The current understanding of the brain's connectome is that it is a technically and mathematically complex network that requires detailed examination using powerful analytic tools. The recent advances in imaging and genetics have provided a unique opportunity to develop and apply neuroinformatics methods to improve the understanding of the brain. Although a wealth of genetic, imaging and cognitive data has become available through NIH funded initiatives, such as the ADNI and the Human Connectome Project (HCP), elucidation of the brain's network architecture and mechanisms still requires sophisticated computational and informatics tools to facilitate multimodal data integration. A challenge of this work was the interpretation of normative changes in brain structure that would occur as a result of environmental variations and demands, through substantiation by imaging measures and the integration of vast datasets. Further, the anticipated neuroinformatics challenges within this project included a) the integration of multiple data types (e.g., volume- and surface-based representations of the brain to which spatial coordinates are assigned to each voxel), b) linking network efficiency data to individual characteristics (e.g., demographics, genomics, cognitive performance, and behavior) and c) the employment of visualization platforms to accommodate multiple data types.[3-6]

The ADNI seeks to "improve clinical trials for the treatment and prevention of Alzheimer's disease" (http://adni.loni.usc.edu/). Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). This database was specifically selected for the present work given the wealth of data derived from a vast spectrum of modalities which suited hypothesis investigation by containing the variables of interest, the quality, and the validity of processed data. The contributions of the completed work may very well lead to insights for the improvement of clinical trials for the study of AD. ADNI is a multisite study that follows a reliable protocol, and the data have been used in hundreds of analyses chiefly focused on AD. The solid assessment and classification of patients into diagnostic categories has been at the forefront of protocol development, and the processing of samples has been closely monitored to ensure data integrity for the host of researchers who invest their time and techniques to advance science.

B. Neuroinformatics

Informatics is the science of how to use data, information, and knowledge to improve human health and the delivery of health care services. Explicating the disjunction between cognition and pathology is central to the success of preventative strategies targeted at neuropathology in cognitively healthy aging, and it is relevant for progressing our understanding of neurodegenerative disease. As summarized in Chapter 1, the significant risk presented by *APOE4* expression, its widespread prevalence, and its detrimental effects in normal aging to neuropathological disorders all unite to underscore the value of informatics methodologies for unraveling the nature of dynamic structural alterations. The study of cognitive function must be observed at multiple levels simultaneously and with traditional analytic techniques which rely on one outcome measure; for instance, the structure of a single brain area, rather than a network view, may insufficiently probe the driving force of function.[7] In his seminal paper, “The informatics core of the Alzheimer’s Disease Neuroimaging Initiative,”[8] informaticist Dr. Arthur Toga directed the future of informatics initiatives using ADNI data by stating, “Integrating a broader spectrum of ADNI data and providing tools for interrogating and visualizing those data will enable investigators to more easily and interactively investigate broader scientific questions.”[8] This is, in part, what the present work seeks to accomplish: through the use of ADNI data, continued informatics work in this realm will be key to understanding the characteristic ability of the brain to adapt and, ultimately, to improve individual lives by reducing suffering and advancing *individualized* therapies.[9-11] Neuroscience and informatics have both made significant advancements which have made their incorporation known as "*neuro*informatics" possible. In the late 1980s, the foundations of the informatics multidisciplinary field of neuroinformatics were laid. "Relating the complex structures and functions of the nervous system requires coordination among diverse domains of knowledge, integration across multiple levels of investigation, and fusion of seemingly disparate technical approaches, from molecules to behavior. The challenge of neuroinformatics specifically is to provide a unified computational information framework to enable, facilitate, and foster such an enterprise."[12] The innovation of the present work is the use of neuroinformatics for the integration of measurements that combine information from each mode of study in order to investigate the important dynamic of CR in *APOE4* carriers and whether the preventative measure of CR is equally effective in *APOE4* carriers as it is in *APOE4* non-carriers. This work combines multiple forms of healthcare data and employs neuroinformatics and data mining to create a signature that will be pivotal to the brain's ability to adapt. The evaluation of the expression of *APOE4* on these mechanisms will contribute to the advancement of individualized therapies. The neuroinformatics pipeline developed will allow for the substitution of input and output variables. The foundation may be used to estimate structural covariance using a conventional clinical imaging modality.

2.2 Aim II: Application of a Neuroinformatics Pipeline

A. Measures of Alzheimer's Disease Pathology

1. Cognitive Reserve (CR)

Evidence for CR comes from neurological observations that have shown a disconnect between the degree of brain pathology and the clinical manifestation of that damage. The initial study of reserve against injury stems from multiple observations in neurology and clinical psychology research, bridging numerous disorders, revealing a disjunction between the degree of brain pathology and the clinical manifestation of that damage.[13] For instance, in early work by Katzman et al.,1989, advanced AD pathology was identified at the time of death in the extracted brains of individuals categorized as "cognitively normal."[14] Similarly, there is ample and recent evidence that a stroke of equal magnitude can produce catastrophic impairment in the cognitive outcome of one patient while marginally influencing another.[15] This disjunction has been attributed to CR: that is, brain resilience to injury gained through *prior* enrichment.

2. Amyloid-β (Aβ)

In more recent examples, including the "Nun Study," which is an ongoing longitudinal report of aging and dementia amongst sisters who have experienced a uniform adult lifestyle, it has been shown that individuals who do not develop AD are most often those with higher levels of education.[16] Inconsistencies in brain pathology such as the presence of Amyloid-β (Aβ) plaques, a major hallmark of AD[89,95], exist between healthy individuals with the same cognitive function. Deposition of Aβ into insoluble plaques is the earliest sign of AD, and aggregation depends on several factors, including the rate of production, clearance from the brain interstitial fluid (ISF) (where amyloid plaques are found), and the rate of fibrillation, all of which may be influenced by *APOE*. Such inconsistencies have been known to relate to lifelong enriching experiences such as education, occupational activities, bilingualism, and cognitive activity. To further complement what has been confirmed post-mortem, a body of literature describes the modifying influence of CR on a spectrum of neurological and psychiatric disorders in living patients (Please see Figure 2.1. below).[17-21]

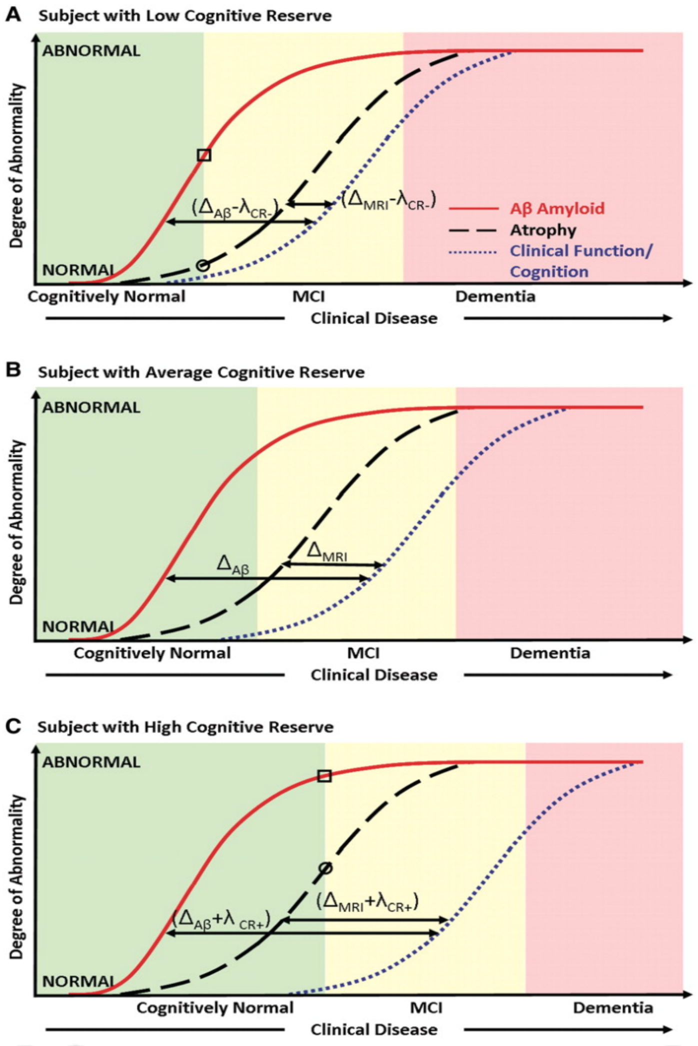


Figure 2.1 Cognitive reserve and Alzheimer's disease biomarkers are independent determinants of cognition.

Cognitive reserve and Alzheimer's disease biomarkers are independent determinants of cognition Image and description used with permission. Brain. 2011 May; 134(5): 1479–1492. Published online 2011 Apr 7. doi: 10.1093/brain/awr049. Description as originally published, "Model illustrating the independent effect of cognitive reserve on the relationship between biomarkers of pathology and cognition in subjects with (A) low, (B) average, and (C) high cognitive reserve. Clinical disease stage is indicated on the horizontal axis and the magnitude of biomarker abnormalities (from normal to maximally abnormal) on the vertical axis. The biomarker curve labels are indicated in A. In A and C, the levels of amyloid-b are indicated by a square and the levels of atrophy are indicated by a circle at the point where cognitively normal subjects progress to mild cognitive impairment. This illustrates that at an equivalent clinical diagnostic threshold, subjects with high cognitive reserve have greater biomarker abnormalities than low cognitive reserve subjects. MCI = mild cognitive impairment."[17]

What is also evident, however, is that elucidating these mechanisms by breaking down components in order to more broadly illuminate the effect of CR on brain function is central to the role of healthy aging and is crucial for all who will experience neurodegeneration.

3. Apolipoprotein E4 (APOE4)

While enrichment through CR in the form of education has been shown to buffer the effects of injury and pathology, there are known detrimental effects of *APOE4* that have been shown to remain unrestored by environmental enrichment.[22] The expression of isoform *APOE4* can be directly injurious to neuronal cells. Indeed, *APOE4* is considered the most established genetic risk factor for AD and has been associated with 1) increased neuron lysosomal leakage, 2) impaired dendritic remodeling, 3) severe loss of neurons and synapses in AD, regarded as a failure of plastic neuronal response to injury, and 4) decreased Amyloid-β (Aβ) clearance.

The majority of *APOE* in the central nervous system is produced by astrocytes, which play a central role in the cellular clearance of Aβ. The clearance of Aβ appears to be impaired more by *APOE ϵ4*, relative to other *APOE* isoforms (i.e., *ϵ2* and *ϵ3*). The adverse effects of the *APOE4* allele, both independent and dependent of Aβ, may impact normal brain functions and are strongly predictive of brain atrophy rates, known to synergistically activate the neurotoxic pathogenesis of synaptic degeneration in AD and Lewy Body Disease. In contrast, education, a proxy for cognitive reserve, may generate protective effects against gray matter atrophy, enhance the expression of the plasticity gene, erg-1, and diminish Aβ deposition. The mechanisms by which *APOE4* are expressed are in fact initiated by injury or stress.[23] It has been demonstrated that the onset of neurodegenerative disease and the negative outcomes of *APOE4* are associated with the perception of stress.[24-29] Recent (2015) reports by the U.S. Department of Veterans Affairs highlight the significantly increased risk of *APOE4* allele carriers developing PTSD, in conjunction with high levels of combat exposure, when compared to fellow veterans who do not carry the *E4* variant.[29] While this is an important consideration, the severity of the environment x gene interaction is made even more evident in other studies that exemplify detrimental influences on health with very little “stress” in comparison to “high levels of combat.” For instance, simply living in a psychosocially hazardous neighborhood is shown to be associated with significantly worse cognitive function in individuals with the *APOE4* genotype.[30]

Vulnerability or resilience to stress, broadly defined by the actual or anticipated threat to the well-being of an individual or the disruption of organism homeostasis, are influenced by gender, personality traits, or early life experiences, and are determined by genetic and epigenetic (environment x gene) interactions.[31] Various processes of adaptation to changing conditions can be measured through an organism’s physiological response to stress (termed “allostasis” in 1988 by Sterling & Eyer).[32] One measure often employed in research laboratories is “allostatic load or overload,” the release of glucocorticoid (measured as cortisol), which is secreted from the adrenal glands following stimulation by the anterior pituitary hormone adrenocorticotropic hormone (ACTH). Allostatic overload is, more precisely, exposure to or perception of too much stress and its subsequent inefficient management.[33] The detrimental release of cortisol in stress response is significantly increased in *APOE4* carriers,[31] and exposure to “real life” difficulties are also shown to cause memory loss in elderly *APOE4* carriers far beyond the loss in non-carriers.[34]

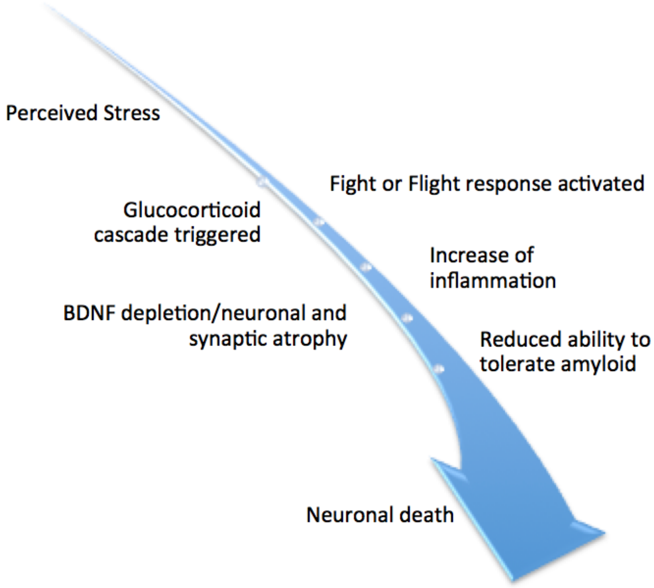

Perceived Stress
Fight or Flight response activated
Glucocorticoid cascade triggered
Increase of inflammation
BDNF depletion/neuronal and synaptic atrophy
Reduced ability to tolerate amyloid
Neuronal death

Figure 2.2 The downward dart of perceived stress: A conceptual model of the effects of stress. Each of the figure components are described in the section below. Amidst the perception of stress, the "fight or flight" response is activated, and a host of biological events (named the "glucocorticoid cascade") are set in motion.

Figure 2.2. serves as a representation of the pro-inflammatory effects of chronic exposure to stress. The profound impact of environmental and social stress alters brain structure through the release of glucocorticoid hormones into the central nervous system, in preparation for harm or deprivation of basic needs.[35] This stress response further catalyzes hippocampal neuronal death through an increase of inflammation and a decrease in glucose.[36] As the brain ages, it becomes increasingly more vulnerable to hippocampal neuronal death triggered by stressful life events, and the ability to regenerate neurons progressively worsens as the exposure to stress leads to an excess of glucocorticoids.[37] This process thereby accelerates the degeneration of the hippocampus.[38] The pathological manifestations of stress are evidenced to cause neuronal and synaptic atrophy/malfunction as well as immunosuppression.[39] Stress causes a depletion of BDNF, a necessary protein for synaptic plasticity and has also been shown to influence the brain's ability to tolerate AB toxicity.[40-41] Again, these events are shown to impact cognitive decline and clinical outcome more severely in carriers of the *APOE4* allele.[42-43]

All of this suggests that *APOE4* is a risk (not necessarily causal) factor that can influence human health through *multiple* pathways. Given the increased likelihood of the worst prognosis, close clinical monitoring and a detailed investigation into the effects of *APOE4* on preventative and therapeutic actions is clinically warranted. How gray matter density and network architecture, composed of genetic influence ─ specifically the carriership of *APOE4* ─ and brain regions, interact with our behavioral and cognitive capacities is still under investigation.

4. Gray Matter Density (GMD)

It is understood that the integrity and density of the brain's structure and function is controlled by genetic factors,[44] but the degree to which genetic factors influence brain connectivity and GMD requires further investigation. Figure 2.3. proposes a conceptual model of the *APOE4* toxic cycle, ending in neurodegeneration – that is, a reduction in GMD. Each element is described in the subsequent section.

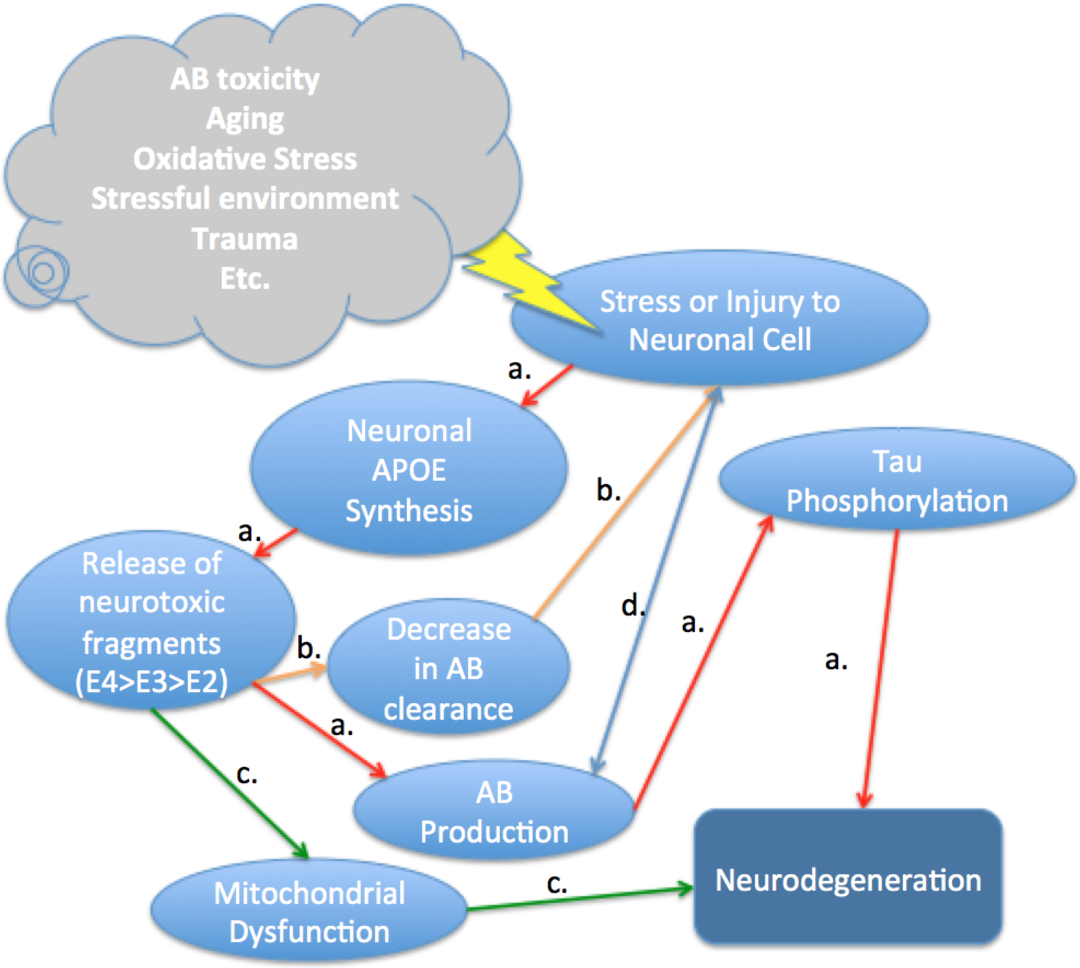


Figure 2.3 The toxic cycle: Influence of APOE on brain architecture.

Path a: Neurons synthesize APOE to assist in the repair or remodeling of neurons. In doing so, however, APOE4 releases neurotoxic fragments leading to tau phosphorylation45 by Amyloid-β production, a known disease agent to AD. Amyloid-β (Aβ) regulation and its potential roles in normal neuronal biology are still under investigation, however, it is known that Aβ production can be stimulated by injury to neurons through oxidative stress.46-48 Path b: It is also known that APOE4 can harmfully impact Aβ clearance and deposition, leading to increased neuronal stress.49-51 Path c: The generation of neurotoxic fragments leads to mitochondrial dysfunction, which contributes further to the onset of neurodegenerative disease.52 Path d: Further, Aβ itself injures neurons; neuronal injury stimulates APOE production, which induces APOE neurotoxic fragment formation, and which thereby further perpetuates the toxic cycle.[23]

It is possible that upon initiation of this toxic cycle (Figure 2.3) *APOE4* carriers lose reserves built by education at a faster rate. An opportunity to gauge the toxic cycle may be to evaluate amyloid load. The carriership of a single *APOE4* allele in healthy younger and older adults is associated with changes in neural activation, throughout the working memory encoding phase.[53] Filbey et al. investigated failing compensatory mechanisms and concluded that the *APOE4* may be associated with "early" compensatory mechanisms compared to non-carriers and that these compensatory mechanisms may fail earlier in older *APOE4* carriers than in non-carriers.[53]

Our work on *APOE4* and cognitive reserve imply that carriers and non-carriers of the *APOE4* allelic variant may a) differ in their response to environmental enrichment, and/or b) may be further along in the deterioration process of aging, such that the presumed benefits are undetectable (Figure 2.4.).[54]

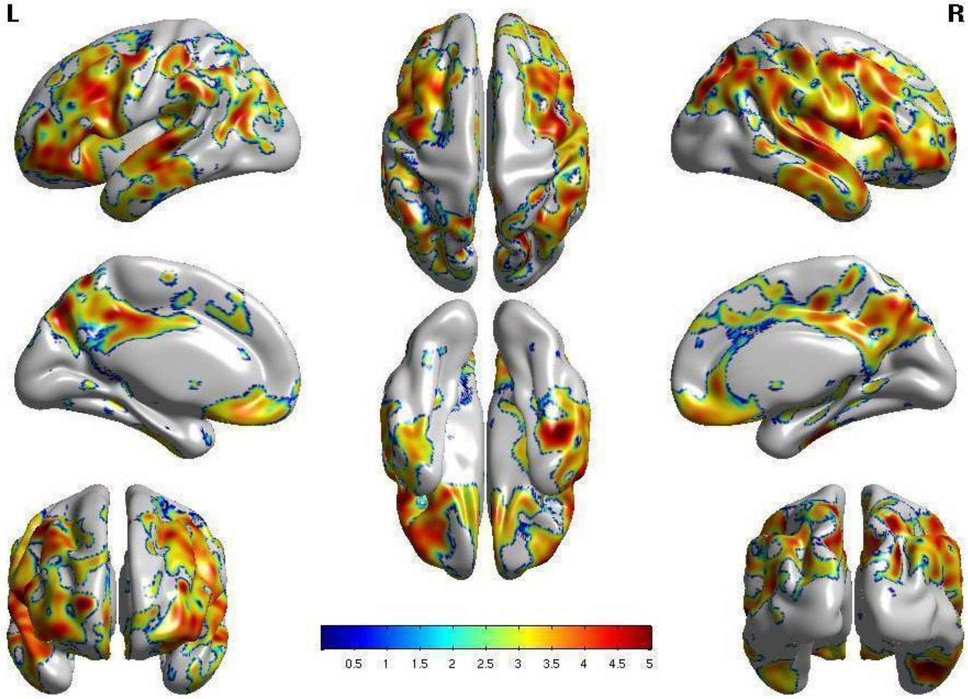


Figure 2.4 Education effects on regional GMD in APOE4 non-carriers.

This image displays only regional GMD in APOE4 non-carriers (n=207) associated with years of education after adjusting for age, sex, and Aβ retention (FDR p<0.01). No effect of education on regional GMD was found in APOE4 carriers. The first row from left to right is the lateral view of the left hemisphere, top side, lateral view of the right hemisphere; the second row from left to right are the medial view of the left hemisphere, bottom side, medial view of right hemisphere; and the third row are frontal side and back side. The diverging scale represents the difference from zero in positive volume mapping.

In this work, while education may help *APOE4* non-carriers as well as *APOE4* carriers, the neuroplasticity mechanisms through which education aids in delaying AD differs by *APOE* genotype, and non-carriage of the *ϵ4* allele may serve as a developmental benefit. To evaluate this further, a younger cohort ought to be observed, as well as the possibility that environmental enrichment may involve a subjective component, rather than act solely as an objective measure of achievement. Given what we know of the mechanisms of *APOE4* (please refer to Chapter 2, Figure 2.3.), it is also plausible that *APOE4* carriers with elevated stress or neuronal injury might experience an increasingly detrimental effect of injury,[55-56] as compared to non-carriers, regardless of their achievement status.

While education has been associated with neurogenesis and increased neuroplasticity, *APOE4* has been associated with impaired dendritic remodeling and failure of the plastic neuronal response to injury. The adverse effects of *APOE4* on neuronal plasticity are markedly heightened upon exposure to stress, reducing amyloid β clearance, and increasing the rate of brain atrophy.[57-60] In other words, it is worth evaluating whether the neurocognitive mechanisms through which synaptic plasticity is regulated are in fact promoted by CR but hampered by *APOE4* carriership,[57, 61-62] given that both appear to target the brain's ability to withstand injury and modulate network structure.

Given that evaluation of *APOE4* carrier status provides evidence to support differences in GMD, as well as a lack of success in therapy development in trials of reducing plaque deposition and excessive tau phosphorylation,[63-64] there is logical backing to focus efforts on understanding the impact of *APOE4* on CR. Neuroinformatics-based approaches and multimodal data will be essential as tools for the development of novel diagnostic and therapeutic strategies. The advancement toward mapping patterns of structural changes in individuals and the elucidation of the structurally complex and dynamic functions of the neural reserve can be catalyzed by the strategic neuroinformatics integration of genomics, cognitive data, and imaging approaches.[64]

Based on this fundamental premise, the present work aims to develop neuroinformatics methodologies to understand the neural substrates of cognitive reserve and sought to construct a neuroinformatics integration of genomic, imaging, and cognitive data to understand cognitive reserve. It accomplishes this by leveraging the collaborative sphere of available training and informatics resources. The project stands at the intersection of informatics, neuroscience, and imaging in order to answer these fundamental questions investigating cognitive reserve. While the focus of the work isolates the effects of *APOE4* carriership vs. non-carriership, the more generalizable neuroinformatics pipeline is structured for expansion to other genetic risk factors. How network architecture, composed of genetic influence and brain regions (further integrated with the dynamic activity of neurons) interacts with our behavioral and cognitive capacities, are under investigation. Elucidating these mechanisms by breaking down components in order to more broadly illuminate the effect of CR on brain function is central to the role of healthy aging; however, it is crucial for those who will experience neurological diseases of aging or neurodegeneration by any trigger.

5. Global Efficiency (Eglobal)

The brain's network architecture, influenced by cognitive reserve, shapes behavioral and cognitive capacities. The context of the present investigation is centered on evidence indicating that CR is based on the efficient utilization of brain networks when challenged with demands, and the ability to maximize performance and employ alternative networks in the face of brain damage. This may be accomplished by employing models for brain networks that would otherwise be unengaged by a specific function during the absence of brain damage. CR is thought to play a large role in healthy aging and in the reduction of clinical manifestation of damage in neuropathology. From the study of "enriched rodents", it is known that a stimulating environment, a component of CR, fosters the growth of new neurons (in the form of neurogenesis), and upregulates Brain-Derived Neurotrophic Factor (BDNF), thereby fostering neural plasticity.[65-67] Notably, in humans, a stimulating environment is considered the experience of learning, formal education, occupational attainment, and engagement in leisure activities – and there is indeed a vast body of literature corroborating the effects of experience-dependent plasticity, (also known as environmental enrichment (EE)),[9, 68] on the human brain.[17, 19, 20, 69, 70] The fact that neural plasticity is conditional on experience suggests a combined structural and functional basis for individual differences in performance.[71] Therefore, experience-dependent plasticity may be a major contributor to the copious variations of the brain's structure and connectivity existing across individual connectomes and extending into adulthood. This background literature has prompted speculation that higher CR will be associated with higher network efficiency and will involve specialized use of neural processing. *How* precisely CR may modulate connectivity patterns and network structure within the aging brain is fundamental information that has yet to be revealed.

6. Pfeffer Functional Activities Questionnaire

The Pfeffer Functional Activities Questionnaire (FAQ),[72] assesses 10 common activities of complex cognitive and social functioning and serves as the measure of neuropsychological performance. Based on the ADNI study, performed by Ritter et al., 2015, which utilized all available modalities for feature selection and classification, the FAQ was identified as the best performing single feature of MCI conversion to AD at 3-year follow-up.[73]

2.3 Analytics Methods Employed

A. Multiple Regression

Multiple regression models are built to understand the association between multiple regressor variables X on a single outcome variable Y. Interactions and the strength contributions the variance of the outcome measure can serve as indicators of association and interdependence. Importantly, regression does not provide insight into the cause of X on Y, and nor does it provide information pertaining to the directionality of the relationship. Prior work on ADNI data includes the employment of multiple regression analysis in the evaluation of stage-specific associations of biomarkers to neurodegenerative phenotypes.[74-75]

B. Fast Causal Inference (FCI)

Although widely employed for the purpose of causal inference, regression is ill-suited for causal discovery.[76] While regression models are often endorsed to estimate the influence of regressor Xi on an outcome measure Y, regression measures correlation, not causation. It would be an error to accept the significance of regression estimate Xi if Xi and outcome measure Y has one or more unmeasured common causes. Spirtes et al., 2000, highlight another important consideration: the statistical dependence of a regressor Xi on an outcome variable Y may be biased by an unmeasured common cause of other regressors (Xk), which could be variables that do not actually influence Y.[76] Unfortunately, in an observational study one cannot confidently measure all common causes of the outcome variable and regressors.

The Fast Causal Inference (FCI) algorithm, which does not make assumptions about latent variables, was used to state whether variable Xa *directly* influences Y, *may* influence Y, *does not* influence Y, or *is undetermined* in its influence of Y. FCI works in two stages: 1) "skeleton identification," which works to identify conditional independence between each pair of variables X, Y, and in which process, where X is conditionally dependent on Y given Z, Z is then stored; and 2) "orientations stage," which uses the stored conditioning sets to orient the edges. The FCI output forms a partial ancestral graph,[76] wherein each variable has a node (or vertex). An edge (line) is drawn from node X to node Y to depict causal relationship. Methods for direct causal discovery that exist (e.g., PC, FCI, FGES) typically assume faithfulness and a lack of unobserved confounders (Please see Figure 2.5.).

Within the Tetrad software used for this work, one may apply prior knowledge concerning temporal ordering, and there is no limit to the input assumptions (the data may be continuous, discrete, or mixed). The cutoff for p-values (alpha) was set to 0.05, indicating that conditional independence tests with p-values greater than 0.05 will be stored as "independent." Although the FCI algorithm employed in this work has the ability to identify all arrowheads within the model, a limitation of the Tetrad program is that it does not have the ability to accurately identify all tails. Within a causal graph, each variable has a node (or vertex). A line is drawn from A to B when there is a hypothesized response of B when A changes.

V-structure B is considered a collider variable given that it is causally influenced by two or more variables (in this case, A and C). Conditioning on the collider B may open a path between A and C; however, this will introduce bias into the estimate of cause between A and C, which may then name associations where none truly exist (such is the case in regression). Note, this is different from a confounding variable which ought to be controlled in the estimation of regression.

An FCI output would state whether variable Xa *directly* influences Y, *may* influence Y, *does not* influence Y, or *is undetermined* in its influence of Y. FCI has the ability to discover latent confounding. Notably, this is quite different from a regression model which may err through the denotation of “significant” variables biased by an unmeasured common cause of other regressors (which could be variables that do not actually influence Y).

Existing causal discovery methods fall into three broad categories: constraint-based, score-based, and hybrid. FCI uses a constraint-based method to estimate conditional independence, eliminating graphs that are inconsistent with the constraints set. Notably, a limitation of the FCI is that it often performs poorly on small sample sizes and large sets of variables (in order to remedy this, Greedy Fast Causal Inference (GFCI) combines multiple causal inference algorithms[77] and performs well on small sample sizes).

Unlike regression, causal structure discovery infers directionality when possible (Please see Figure 2.5 below). Further, while causal structure discovery works to evaluate the conditional independence of variable pairs and provides information about whether a causal pathway exists, causal inference is used to extrapolate based on causal structure discovery (what has been interpolated). The methods in this dissertation employ both multiple regression and FCI.

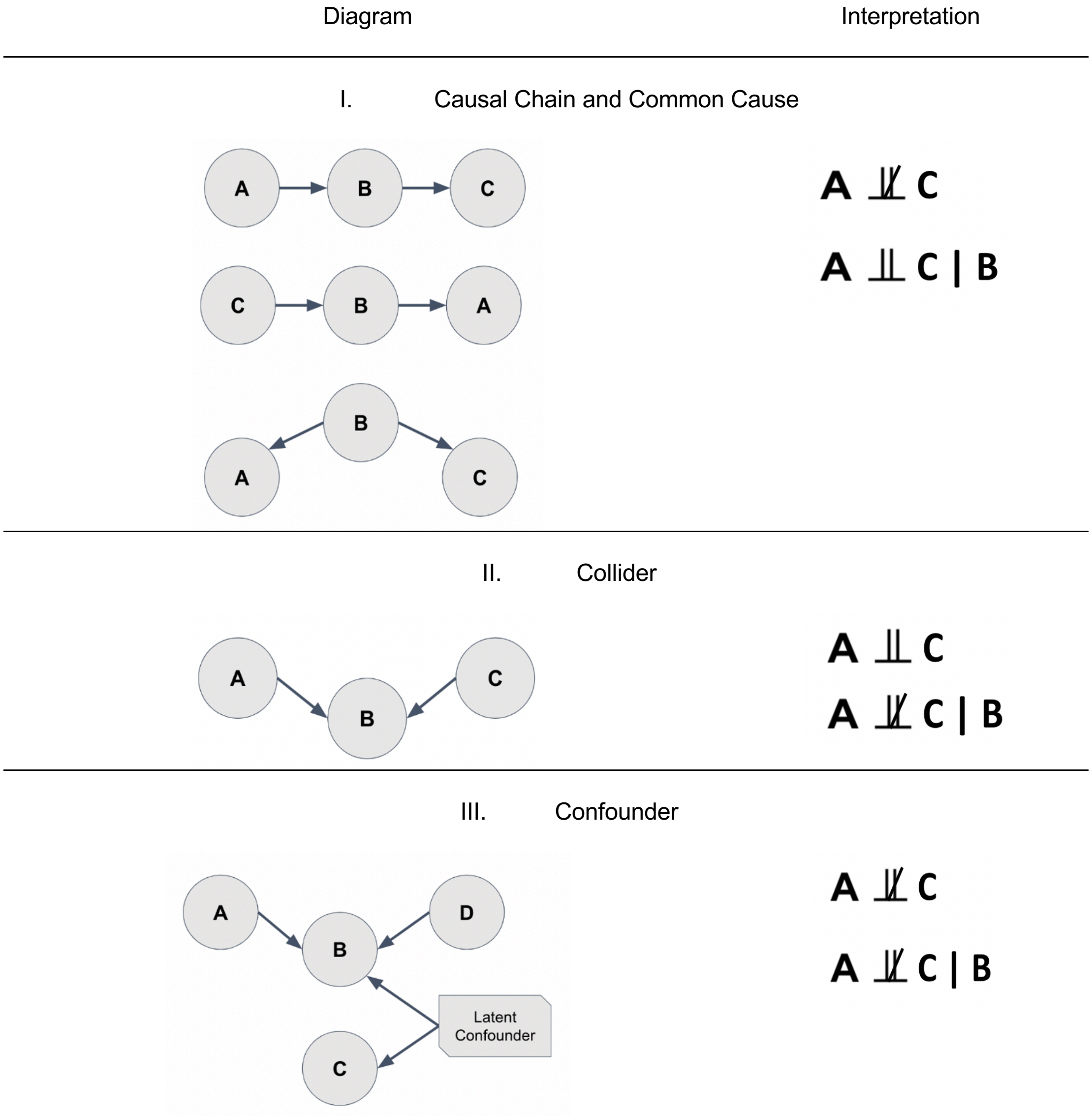


Diagram
Interpretation
I. Causal Chain and Common Cause
A
B
C
C
B
A
B
A
C
$A \not\perp\!\!\!\perp C$
$A \perp\!\!\!\perp C \mid B$
II. Collider
A
B
C
$A \perp\!\!\!\perp C$
$A \not\perp\!\!\!\perp C \mid B$
III. Confounder
A
D
B
Latent Confounder
C
$A \not\perp\!\!\!\perp C$
$A \not\perp\!\!\!\perp C \mid B$

Figure 2.5 Intuition for causal discovery.

(I) Causal Chain and Common Cause. An arrow from A to B represents that A is a direct cause of B. Roughly, this indicates that the value of A makes some causal difference in the value of B, and that A influences B through a process not mediated by any other variable in the set of variables represented (rows 1 and 2). B is a common cause of A and C (row 3). (II) Collider. V-structure B is considered a collider variable given that it is causally influenced by two or more variables (in this case, A and C). (III) Conditioning on the collider B may open a path between A and C; however, this will introduce bias into the estimate of cause between A and C, which may then name associations where none truly exist (such is the case in regression).

C. Graph Theory

Graph theory is the study of the way in which elements interact with one another in a system.[7] Mathematics model elements and their connections as nodes and edges, respectively. For example, statistical relationships may be measured as correlations between cortical thickness distribution, and physical relationships may be representative of axons between neurons.[88] A graph could indicate the strength of covariance results in a symmetric covariance matrix C (i, j) where each row i would represent the edge that goes out from node i, to arrive at each node j represented by column j. The connections between nodes can be tested over time, in relationship with subnetworks, behavior, or a range of other metrics.[85]

Chapter 3 Development of A Neuroinformatics Pipeline (Aim I)

3.1 Chapter Overview

The pipeline developed (Please see Figure 3.1 below) for these studies adheres to the emerging neuroinformatics compact to produce open source and replicable methods. This is achieved through: (1) intentional limitation of the number of platforms used (2) employment of open-source computing software (R-Project[96]) (3) brevity in scripting yet thorough documentation of code (4) community storage allowing for open use of, and comment on, the application and supporting materials (5) proof of concept established using publicly available data.

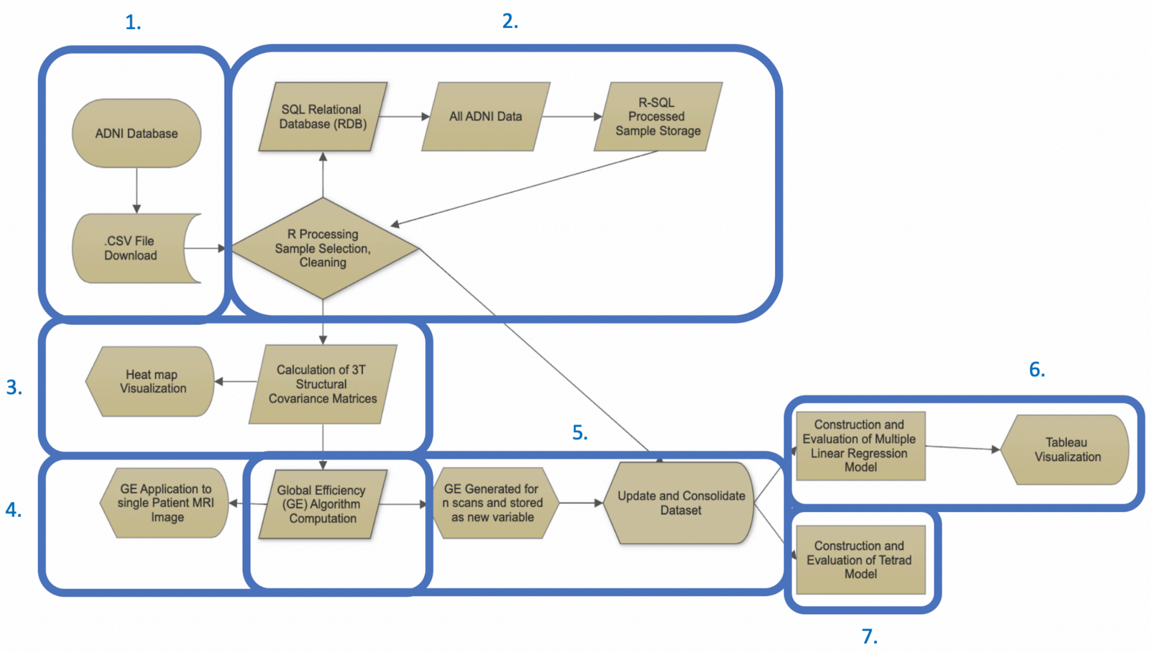

1.
2.
ADNI Database
.CSV File Download
SQL Relational Database (RDB)
All ADNI Data
R-SQL Processed Sample Storage
R Processing Sample Selection, Cleaning
3.
Heat map Visualization
Calculation of 3T Structural Covariance Matrices
4.
GE Application to single Patient MRI Image
Global Efficiency (GE) Algorithm Computation
5.
GE Generated for n scans and stored as new variable
Update and Consolidate Dataset
6.
Construction and Evaluation of Multiple Linear Regression Model
Tableau Visualization
Constuction and Evaluation of Tetrad Model
7.

Figure 3.1 Neuroinformatics pipeline in sectioned format.

The visualization follows the construction of a neuroinformatics pipeline and the consideration of tools and analysis techniques that are both generalizable and replicable. The method is described through a sectioned process map: (1) Data acquisition (2) Sample processing and storage (3) Computation and visualization of brain structural covariance (4) New variable generation through the statistical computation of graph theoretical metrics (5) Merging of datasets (6) Variable manipulation and construction of regression model (7) Objective validation of the model performed using the FCI search algorithm. The Tetrad software output provided visualizations of graphs. Additionally, Tableau was used to create a visualization of interaction findings. Processes corresponding to sections 2-6 are automated (using a single R script, provided in the appendix).

3.2 Description of Neuroinformatics Pipeline

A. Sectional overview of neuroinformatics pipeline

References to the neuroinformatics pipeline (Please see Figure 3.1.) and corresponding sections used to detail the work are presented in Table 3.1. The R scripts for pipeline sections 2-6 are provided in the Appendix. Pipeline Sections 1 and 7 were not included in the R script at the time of analysis.

Table 3.1

Key Sections in Neuroinformatics Pipeline

Pipeline Section	Key Steps
1. Data Acquisition	• Following hypothesis generation, relevant datasets are downloaded from the public repository.
2. Sample Processing and Storage	• Disparate datasets are prepared and loaded into a relational database for sample construction. • Local Structured Query Language (SQL) phpMyAdmin Database is created. • R is used to load and prepare the data, which are subsequently uploaded to the SQL database.
3. Computation and Visualization of Brain Structural Covariance	• Algorithms are applied to individual 3T MRI data and covariance matrices are produced. • Heat maps employing a diverging color scheme are generated to describe the intensity of covariance between nodes.
4. New variable generation through graph theory	• Global efficiency (E_{global}) calculated based on the covariance matrix. • E_{global} output is generated for a single subject and corresponds to a heat map. • E_{global} is generated for the entire sample.
5. Merging of Datasets	• Newly generated data are merged with a foundational data set based on subject ID, creating a new variable in the dataset.
6. Variable manipulation and construction of regression model	• Variables are manipulated through grouping and classification. • Regression models are prepared to directly test the original hypotheses. • Data are visualized in Tableau and a user-friendly hypothesis evaluation interface is produced.
7. Employment of Fast Causal Inference (FCI) Algorithm	• Data are tiered in Tetrad software and data are stratified based on preliminary results. • Processes are mapped and validated.

1. Data Acquisition

Alzheimer's Disease Neuroimaging Initiative (ADNI) data (http://adni.loni.usc.edu) was downloaded and stored in a local instance of phpMyAdmin (http://localhost/phpmyadmin/index.php) to facilitate reproducibility and to maintain data integrity. Four databases from the central ADNI data repository were downloaded to a local drive, original folder names preserved.

2. Sample Processing and Storage

The ADNI databases were uploaded to a local instance (http://localhost/phpmyadmin/index.php) of the open source MySQL tool, phpMyAdmin (https://www.phpmyadmin.net), a stable Relational Database Management System (RDMS) which was initially released in 1998 by The phpMyAdmin Project. The tool is written in PHP and JavaScript and provides a web hosting service. Tables were written into the phpMyAdmin RDMS using Structured Query Language (SQL) commands which were annotated and stored as reproducible scripts written in R and subsequently published to GitHub (https://github.com), a collaborative development platform for software development. The R Project for Statistical Computing96 (https://www.r-project.org) is an open-source software environment that provides a platform for statistical computing and graphic construction.

3. Computation and Visualization of Brain Structural Covariance

a. ADNI Magnetic Resonance Imaging Processing

Vertex-wise cortical thickness measurements were determined by a FreeSurfer algorithm which computes the distances, at any given point, between highly accurate models of gray and white matter plial surfaces.[80] Values of mean cortical thickness and standard deviation were determined from FreeSurfer *.aparc.stats file output. Anatomic regions-of-interest (ROIs) affiliated with the default mode network (DMN)[81] were selected as a subset of 16 from the 68 cortical ROI Desikan-Killiany Atlas.[82]

b. Structural Covariance Calculation

To compute a structural covariance matrix for each subject, the cortical thickness data for each nodal region within the default mode network was extracted. Mapping individual brain networks using statistical similarity in regional morphology from MRI was proposed by Xiang-zhen Kong et al. in 2015.[83] In 2016, Hee-Jong Kim et al. proposed a structural covariance network for single subject scans using MATLAB.[84] The theory[85-87] was studied, and the constructs applied in the present work on non-accelerated T1 scans using R.

For each pair (i, j) of ROIs, the Z-score of the cortical thickness value was calculated using the mean and standard deviation of the i-th and j-th ROI, where Z (i, j) represents the average deviation of the i-th ROI from the j-ith ROI (eq.1).

$$Z(\,i, j) = \frac{\mu\,(i) - \,\mu\,(j)}{\sigma(j)}$$

(1)

Using the same logic, the Z-scores of the cortical thickness value to signify the deviation of the j-th from the i-th ROI was calculated (eq.2).

$$Z(\,j, i) = \frac{\left(\mu\,(j)\; - \;\mu\,(i)\right)}{\sigma\,(i)}$$

(2)

From these values, a weighted undirected graph indicating the strength of the covariance results in a symmetric covariance matrix C (i, j) (eq.3), where each row i represents the edge that goes out from node i, to arrive at each node j represented by column j.

$$C(\,i, j) = \frac{\left((|\,Z\,(i, j)\;|\; + \;|\,Z\,(j, i)\;|)\right)}{2}$$

(3)

Using these definitions, strength of structural covariance was represented as the mean of the absolute values of Z(i,j) and Z(j,i), where C(i,j) measures the similarity in cortical thickness distribution between nodes i and j (eq.3). The structural covariance values represented in the adjacency matrix, thereby quantify the influence generated by each node (region of interest in the brain) on each individual node within the network. A function was created in R which computed the structural covariance adjacency matrix by selecting the mean and standard deviation of node i, and the mean and standard deviation of node j. From these variables, the function computed the z score of ij, and the z score of ji and manipulated these values into the structural covariance matrix using the formula in eq.3. Using R, these covariance values were represented in a heat map where default mode network regions were listed as 16 nodes on the X and the Y axes, and a colored scale gradient was applied to the adjacency cells. Using this visualization technique, areas of greatest covariance in thickness distribution between nodes emerge as a pattern unique to each individual. Using the 16 x 16 matrix, nodal efficiency and global efficiency were calculated for each subject's T1 scan.

Default mode network regions were listed as 16 regions of interest (8 nodes bilaterally: 1) rostral anterior cingulate, 2) precuneus, 3) parahippocampal, 4) posterior cingulate, 5) medial orbitofrontal, 6) lateral orbitofrontal, 7) isthmus cingulate, 8) caudal anterior cingulate) on the X and the Y axes and a colored scale gradient was applied to the adjacency cells. Using this visualization technique, areas of greatest covariance in thickness distribution between nodes emerge as a pattern unique to each individual. Using the 16 x 16 matrix, nodal efficiency and global efficiency were calculated for each subject's T1 scan.

c. Calculation of Nodal Efficiency

The structural covariance adjacency matrix was used to calculate the efficiency of each node and the entire global efficiency of a network. Nodal efficiency (eq.4) represents the influence of one node within a given network and how well that node is integrated within the network, given its shortest paths.

$$E_{nodal}(j) = \frac{1}{N-1}\sum_{i}\frac{1}{l_{ij}}$$

(4)

Nodal and global efficiency of the brain's structural default mode network were calculated using the structural covariance network, first computed through the application of a function written in R and saved as 'adjacency_matrix', which is provided in the appendix A.3. Local and Global Efficiency Calculation.

d. Calculation of Global Efficiency (Eglobal)

After the covariance adjacency matrix was computed, the threshold for distinguishing a true or false connection between nodes was set. For instance, where the threshold was set to 0.3 or below, all cells in the adjacency matrix with a structural covariance value greater than 0.3 were set to 1, and all others were set to 0. Thus, a binary matrix was formed according to the presence or absence of a true connection, as determined by the preset threshold. Following the construction of the binary matrix, a breadth-first search algorithm was applied to the data in order to determine the shortest distance between nodes, based on the presence or absence of an edge. Using this algorithm, node i, if not immediately sharing an edge with node j, would pass through one or multiple neighboring nodes in order to eventually reach node j. Employing the breadth-first search algorithm allowed for the identification of the shortest path links between nodes. This generated an adjacency matrix of shortest paths between nodes.

First, the distance of the shortest path between nodes of this undirected graph was calculated using Dijkstra's algorithm[93], an iterative process that identifies the shortest path from one node (source node) to all other nodes in the graph. This created an additional 16 x 16 adjacency matrix of shortest path from node i to j. After accounting for the removal of the matrix diagonal, the column sum of the inverse of the shortest path length for each node i to j, divided by the number of rows, was used to calculate nodal efficiency (eq. 4). Nodal efficiency for each of the 16 default mode network nodes was stored as a separate variable within the dataset. From the computation of nodal efficiency, global efficiency was calculated as the average of the inverse shortest path length and quantified as the efficient exchange of information where each node is capable of exchanging information using the shortest path length. Applying the formula proposed by Latora et al. (eq.5)[88] generated one global efficiency value for the default mode network of an individual subject scan.

$$E_{global} = \frac{1}{L'} = \frac{1}{N(\,N-1)} \sum_{i \neq j} \frac{1}{l_{ij}}$$

(5)

This global efficiency value was stored for each subject matrix and included in the dataset as a new variable. It is important to note that in the case where a node cannot be reached at a set threshold (where the threshold is too high, and a node appears isolated) and no degree links are available, the global efficiency of the network cannot be equally compared between scans and should therefore be penalized in subsequent analyses. Conversely, the threshold could be limited from excluding any nodes within the predefined network, thereby forcing the system to include all network nodes. To address the issue of the absent edge, in the binary matrix, the proportion of connections made by each node was calculated. That is, in the network A, B, C, D, if an edge was present between nodes A-C and A-D but not between A-B, the proportion for that node would be 2/4, or 0.5. In this way, the proportion of true edges leaving the node became the nodal efficiency for that node and the process of computing global efficiency.

4. Merging of Data Sets

a. Interfacing with R

The analysis dataset was reduced using R and “sqldf,” an R package for running SQL statements on R data frames (https://github.com/ggrothendieck/sqldf, https://cran.r-project.org/web/packages/sqldf/index.html).

b. Sample Selection

Of priority for sample selection was the robustness of the dataset. Given the need for power in setting out to define an abstract concept, maximization of the dataset, while maintaining the integrity of the scientific rationale, was of great priority. Thus, the primary variables of interest were Processed 3T MRI non-accelerated structural scan thickness data for default mode network cortical ROIs within the Desikan-Killiany Atlas[82], carrier status of the *APOE4* allele, and Education level attainment data, which served as a proxy for cognitive reserve.[94]

All thickness data was retrieved online via ADNI MR Image Analysis, UCSF Cross-Sectional FreeSurfer (5.1), as produced by the ADNI funded MRI core, Center for Imaging of Neurodegenerative Diseases, UCSF, led by Norbert Schuff. Specifically, the UCSF – Cross-Sectional FreeSurfer (5.1) [ADNI1, GO, 2] downloadable file, UCSFFSX51_08_01_16, contained 1051 unique RIDs (Research Identification). Notably, although another file of the sample publication date, named UCSF – Longitudinal FreeSurfer (5.1) – All Available Baseline Image [ADNIGO, 2], was available for download, this file contained only 470 unique cases. Of the 1051 cases listed within the UCSFFSX51_08_01_16 file, four subjects (RID 2117, 2118, 2154, 2281) were removed given that processed non-accelerated T1 scan data were unavailable, reducing the analysis sample to 1047 subjects. Of subjects listed within the UCSFFSX51_08_01_16 file, only four (RID 1072 (exam date 2010-03-18), 1131(2010-03-04), 1169(2010-01-11), 1241(2010-02-16)) were linked to the ADNI1 collection protocol. Given that the data from these four subjects were collected in visit m36 of the ADNI1 protocol, with exam dates between 2012 and 2014 (note: all data within the dataset range from 2010-01-11 to 2016-02-23), they were included in the sample. Of these 1047 cases, 986 RIDs linked to apolipoprotein E4 specimen results reporting a binary apolipoprotein E4 carrier status (via 'Key ADNI tables merged into one table 'adnimerge', downloaded file: ADNIMERGE.csv), and 975 RIDs linked genetic load of the apolipoprotein E4 allele for each subject (via 'ApoE – Results [ADNI1, GO, 2]', table name 'APOERES', downloaded file: APOERES.csv). Of the 986 cases within the scored ADNI data, 973 had processed data within both the ADNIMERGE and APOERES files. A complete table of 973 cases, named 'MR_Image_Analysis.datanoacc1047allcols', which excluded all accelerated T1 scans, 'MR_Image_Analysis.UCSFFSX51_08_01_16', 'Detached.adnimerge', and 'Biospecimen_Results.APOERES' was created within the local ADNI data repository via the R package 'sqldf' (https://cran.r-project.org/web/packages/sqldf/sqldf.pdf). Given that the program creates rownames as an automatic index, rownames within the merged datasets were removed to avoid conflicting key columns. Using R, the dataset was unlisted, and a function was applied so that data were grouped by RID, then a new column was assigned systematic values according to chronological order of visit date corresponding to the adnimerge table exam date (column, ‘EXAMDATE.1’). For example, the first and second time points within a set of RIDs received an ‘order.by.group’ value of 1 and 2, respectively. A subset of the data was pulled from the complete file, which selected only ‘order.by.group’ values of 1 (i.e. first recorded visit), generating one row of data for each of the 973 cases, thereby producing a row x column matrix of dimension 973 x 484. This file was saved to the MR_Image_Analysis database and named ‘ADNI_973_ALL_ordergrp1’. The script for the 973-sample selection is commented and stored in Appendix 1 for replicability (‘ADNI Sample Selection’)

5.Variable Extraction

Variables extracted for analysis were APOE status, Pfeffer Functional Activities Questionnaire, Gray Matter Density, Amyloid-β, Cognitive Reserve, and Global Efficiency. Table 3.2 below provides additional information about the variables and their relevance throughout the dissertation.

Table 3.2

Variables Evaluated

Name	Acronym	Variable in R	Relevance
Amyloid-β	Aβ	AV45_bl	• A brain protein that accumulates (forming tau) and eventually disrupts communication between brain cells resulting in the death of the cell. • Throughout this work appears as a chief contributor to variation in statistical models.
Apolipoprotein E	APOE4	APOE4	• Carriers of the apolipoprotein E4 allele (*APOE4)* are at increased risk of developing Alzheimer's Disease. • In this work, used to stratify the sample.
Cognitive Reserve	CR	PTEDUCAT EduOrdNum	• The brain's resistance to injury and ability to maintain cognitive functions in the face of stress or injury.
Diagnostic Classification	Dx	DxOrd	• Participant placement on the spectrum of illness according to predetermined criteria established by (National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's Disease and Related Disorders Association (ADRDA). • Progressing by order of severity Control (CN) > Mild Cognitive Impairment (MCI) > Alzheimer's Disease (AD).
Global Efficiency	E_{global}	globeff	• In this work, measure is computed using structural covariance matrices constructed from 3T MRI data.

Table 3.2

Variables Evaluated

Name	Acronym	Variable in R	Relevance
Gray Matter Density	GMD	GMD	• Gray matter consists of neuronal cell bodies, neuropil, glial cells, synapses, and capillaries. • Neurodegeneration equates to a decrease in GMD. • Decrease in GMD is considered a pathological feature of AD. • Used as outcome measure throughout this work.
Pfeffer Functional Activities Questionnaire	FAQ	FAQ	• Assesses 10 common activities of complex cognitive and social functioning. • Often used as a primary outcome measure throughout this work.

1. Amyloid-β (Aβ)

Amyloid-β (A*β*) load was evaluated as a continuous variable, a higher value indicating increased pathology. Cases were also categorized for analysis as “amyloid positive” and “amyloid negative.” Using Landau’s[89] April 2018 Neurology paper as a reference, the cutoff was set to an amyloid load of 1.11, in order to categorize cases as amyloid negative. In this subset n=348 (AD=17, CN=125, MCI=206), the amyloid load ranged from 0.8385 to 1.1081. Amyloid positive was classified as cases where amyloid load was equal to or greater than 1.11.

2. Apolipoprotein E4 (APOE4)

APOE genotyping was performed using the Illumina HumanOmniExpress BeadChip, the ADNI genetic data protocol can be found in the paper, “Alzheimer’s Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans” (Saykin et al., 2010). Each participant’s *APOE* genotype was coded in two forms: first as a 3-level exposure variable to carriership of the E4 risk allele: 0=not present (E3/E3, E3/E2, E2/E2), 1=1

E4 allele (E3/E4, E2/E4), 2=2 E4 alleles (E4/E4), then as a binary variable (0=not present (E3/E3, E2/E3, E2/E2) or 1=present (E2/E4, E3/E4, E4/E4).

3. Cognitive Reserve (CR)

Completion of formal education in total years was derived from the ‘adnimerge’ file. Following sample selection, the data were restructured into two additional variables: a) to represent four ordinal levels of educational attainment (‘No Highschool Diploma’ (<12 years), ‘At Least Highschool’ (>=12 years), ‘At Least Undergrad’ (>16 years), ‘At Least Gradschool’ (>= 20)); and b) the bottom 25% and top 75% of educational attainment (‘Low Education’ (<= 14 years), and ‘High Education’ (>=18 years)).

4. Diagnostic Classification (Dx)

Diagnostic status was derived from the ‘adnimerge’ file and transformed into a three-level ordinal variable: ‘CN’ (Control subjects), ‘MCI’ (Mildly Cognitively Impaired subjects which combined late (LMCI) and early MCI (EMCI)), and ‘AD’ (Alzheimer’s Disease subjects). Each individual underwent clinician verification of diagnostic status only after reviewing the 3T MRI Radiology report and Clinical Read. The screening procedures for new participants ensured that control participants were free of memory complaints, with normal memory function as documented by scoring above education adjusted cutoffs on the Logical Memory II subscale (maximum score is 25) (>=9 where 16 or more years of education, >=5 where 8-15 years of education, >=3 where 0-7 years of education), Clinical Dementia Rating scale (symptom free with a rating of 0), and a score between 24 and 30 on the Mini-Mental State Exam. For inclusion, MCI participants (combined EMCI and LMCI) must have had a subjective memory concern as reported by the participant themselves or the clinician, abnormal memory function documented by the Logical Memory II subscale (score of <= 8 where 16 or more years of education, score of <= 4 where 8-15 years of education, score of <=2 where 0-7 years of education), Clinical Dementia Rating of at least 0.5, and having sufficiently preserved functional performance such that a diagnosis of Alzheimer’s disease could not be made. Participants who were categorized as having Alzheimer’s disease must have had a subjective memory concern, abnormal memory functioning as documented by the Logical Memory II subscale (same as LMCI), a Mini-Mental State Exam score of between 20 and 26, a Clinical Dementia Rating of 0.5 or 1.0, and meeting criteria for probable AD as determined by the NINCDS-ADRDA criteria[79](National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's Disease and Related Disorders Association (ADRDA).

5. Global Efficiency (Eglobal)

Global efficiency of the brain's structural default mode network was calculated using the structural covariance network was first computed through the application of a function written in R saved as 'adjacency_matrix' which is provided in the appendix A.3. Local and Global Efficiency Calculation. A global efficiency value was attributed to each scan id.

6. Gray Matter Density (GMD)

Gray matter density (GMD) of the bilateral middle temporal region, hippocampus, whole brain, entorhinal, and fusiform were selected and entered into statistical models. Notably, middle temporal and hippocampal regions are included within the default mode network. This dissertation work used these regions within the computation of the measure of global efficiency in the default mode network.

7. Pfeffer Functional Activities Questionnaire (FAQ)

The ADNI protocol included several widely used and efficient, standardized neuropsychological tests to represent the domains of interest in the control, MCI, and AD population. Seven of the ADNI cognitive measures are used as part of a collective Uniform Data Set (UDS) facilitated by the National Alzheimer's Coordinating Center (NACC). Performance of cognitive functioning such as memory, processing speed, and executive functioning was assessed at screening and baseline, the timing coinciding with the structural MRI scan and F-AV-45 amyloid imaging. The Pfeffer Functional Activities Questionnaire (FAQ), which assesses 10 common activities of complex cognitive and social functioning, was selected as the measure of neuropsychological performance. Based on an ADNI study which utilized all available modalities for feature selection and classification, the FAQ was identified as the best performing single feature of MCI conversion to AD at 3-year follow-up .[73]

F. Construction of statistical models and data visualization

1. Multiple Regression

Multiple regression, which evaluates the association of more than one predictor variable X with a single outcome variable Y, can be employed within mediation analysis. Multiple models were built to understand the association between multiple regressor variables X on a single outcome variable Y. Interactions and the strength contributions to the variance of the outcome measure can serve as indicators of association and interdependence. Importantly, regression does not provide insight into the cause of X on Y, and nor does it provide information pertaining to the directionality of the relationship.

Regression models were built on the entire sample to first determine whether interactions with *APOE4* carrier status existed. Namely, AD pathology, global efficiency, and cognitive reserve were tested, as were the interactions' contributions to functioning as evaluated by functional activities of daily living. Where interactions existed, regression models were run on stratified sets of variables (e.g., *APOE4* carriers vs. non-carriers).

2. Data Visualization

Tableau was used to allow for interactive visualization of the dissertation findings. The dashboard was made publicly available on Tableau Public: https://public.tableau.com/profile/shauna.overgaard1810#!/vizhome/GlobalEfficiencyEvaluation/Dashboard2.

G. Employment of Fast Causal Inference (FCI) Algorithm

The FCI algorithm, which does not make assumptions about latent variables, was used to state whether variable Xa directly influences Y, may influence Y, does not influence Y, or is undetermined in its influence of Y. As previously stated, FCI uses a constraint-based method to estimate conditional independence, eliminating graphs that are inconsistent with the constraints set. FCI works in two stages: 1) “skeleton identification”, which works to identify conditional independence between each pair of variables X, Y, and where X is conditionally dependent on Y given Z, Z is then stored; and 2) “orientations stage,” which uses the stored conditioning sets to orient the edges.

Chapter 4 Application of Neuroinformatics Pipeline (Aim II)

4.1 Chapter Overview

Aim II, Neuroscience questions proposed using the neuroinformatics pipeline (Appendix A), was addressed by developing three objectives. Objectives 1-3 directly tested our original pathophysiology-related hypotheses: (1) Effect of CR on GMD will differ by *APOE* genotype (2) A higher cognitive reserve will be associated with higher network efficiency (3) *APOE4* carriers as compared to non-carriers will demonstrate differences in network recruitment. We further addressed the statistical significance and variable combinations of the interdependencies evaluated within Objectives 1-3 and addressed the question of whether the relationship between variables evaluated for statistical significance and interactions in Objective 3 were causal (employing the FCI algorithm). The work presented below first describes Methods in Section 4.2, which details sample selection, variable processing, and statistical analysis applied across all studies. Note, these sections were provided in Chapter 3 as they describe the neuroinformatics pipeline (Aim I). Following 4.2, 4.3-4.5 describe Objectives 1-3, respectively, with dedicated headings that state Hypothesis, Models, Results, Conclusion, and Discussion for each objective.

4.2 Methods

A. Data Source

Alzheimer's Disease Neuroimaging Initiative (ADNI) data (http://adni.loni.usc.edu) was downloaded, preserving variable names. The detailed clinical description of the ADNI cohort has been published.[78] AD diagnoses were made according to the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association criteria.[79]

B. Sample Inclusions and Exclusions

Thus, the primary variables of interest were Processed 3T MRI non-accelerated structural scan thickness data for default mode network cortical ROIs within the Desikan-Killiany Atlas[82], carrier status of the *APOE4* allele, and Education level attainment data which served as a proxy for cognitive reserve. [94] All thickness data was retrieved online via ADNI MR Image Analysis, UCSF Cross-Sectional FreeSurfer (5.1) (https://ida.loni.usc.edu/) as produced by the ADNI funded MRI core, Center for Imaging of Neurodegenerative Diseases, UCSF, led by Norbert Schuff. Specifically, the UCSF – Cross-Sectional FreeSurfer (5.1) [ADNI1, GO, 2] downloadable file, UCSFFSX51_08_01_16, contained 1051 unique RIDs (Research Identification). Notably, although another file of the sample publication date, named UCSF – Longitudinal FreeSurfer (5.1) – All Available Baseline Image [ADNIGO, 2], was available for download, this file contained only 470 unique cases. Of the 1051 cases listed within the UCSFFSX51_08_01_16 file, four subjects (RID 2117, 2118, 2154, 2281) were removed given that processed non-accelerated T1 scan data were unavailable, reducing the analysis sample to 1047 subjects.

Of subjects listed within the UCSFFSX51_08_01_16 file, only four (RID 1072 (exam date 2010-03-18), 1131(2010-03-04), 1169(2010-01-11), 1241(2010-02-16)) were linked to the ADNI1 collection protocol. Given that the data from these four subjects were collected in visit m36 of the ADNI1 protocol, with exam dates between 2012 and 2014 (note: all data within the dataset range from 2010-01-11 to 2016-02-23), they were included in the sample. Of these 1047 cases, 986 RIDs linked to apolipoprotein E4 specimen results reporting a binary apolipoprotein E4 carrier status (via 'Key ADNI tables merged into one table 'adnimerge', downloaded file: ADNIMERGE.csv), and 975 RIDs linked genetic load of the apolipoprotein E4 allele for each subject (via 'ApoE – Results [ADNI1, GO, 2]', table name 'APOERES', downloaded file: APOERES.csv). Of the 986 cases within the scored ADNI data, 973 had processed data within both the ADNIMERGE and APOERES files. A complete table of 973 cases, named 'MR_Image_Analysis.datanoacc1047allcols' which excluded all accelerated T1 scans, 'MR_Image_Analysis.UCSFFSX51_08_01_16', 'Detached.adnimerge', and 'Biospecimen_Results.APOERES' was created within the local ADNI data repository via the R package 'sqldf' (https://cran.r-project.org/web/packages/sqldf/sqldf.pdf). Given that the program creates rownames as an automatic index, rownames within the merged datasets were removed to avoid conflicting key columns.

Using R, the dataset was unlisted, and a function was applied so that data were grouped by RID, after which a new column was assigned systematic values according to chronological order of visit date corresponding to the adnimerge table exam date (column, 'EXAMDATE.1'). For example, the first and second time points within a set of RIDs received an ‘order.by.group’ value of 1 and 2, respectively. A subset of the data was pulled from the complete file, which selected only ‘order.by.group’ values of ‘1’ (i.e. first recorded visit), generating one row of data for each of the 973 cases, thereby producing a row x column matrix of dimension 973 x 484. This file was saved to the MR_Image_Analysis database and named ‘ADNI_973_ALL_ordergrp1’. The script for the 973 sample selection is commented and stored in Appendix 1 for replicability (‘ADNI Sample Selection’).

C. ADNI Magnetic Resonance Imaging Processing

Vertex-wise cortical thickness measurements were determined by a FreeSurfer algorithm which computes the distances, at any given point, between highly accurate models of gray and white matter plial surfaces.[80] Values of mean cortical thickness and standard deviation were determined from FreeSurfer *.aparc.stats file output. Anatomic regions-of-interest (ROIs) affiliated with the default mode network[81] were selected as a subset of 16 from the 68 cortical ROI Desikan-Killiany Atlas.[82]

D. Computation of Global Efficiency

1. Structural Covariance Calculation

To compute a structural covariance matrix for each subject, the cortical thickness data for each nodal region within the default mode network was extracted. Mapping individual brain networks using statistical similarity in regional morphology from MRI was proposed by Xiang-zhen Kong et al. in 2015.[83] In 2016, Hee-Jong Kim et al. proposed a structural covariance network for single subject scans using MATLAB.[84] The theory[85-87] was studied, and the constructs applied in the present work on non-accelerated T1 scans using R.

For each pair (i, j) of ROIs, the Z-score of the cortical thickness value was calculated using the mean and standard deviation of the i-th and j-th ROI, where Z(i,j) represents the average deviation of the i-th ROI from the j-ith ROI (eq.1).

$$Z(\,i,j) = \frac{\mu\,(i) - \,\mu\,(j)}{\sigma(j)}$$

(1)

Using the same logic, the Z-scores of the cortical thickness value to signify the deviation of the j-th from the i-th ROI was calculated (eq.2).

$$Z(\,j, i) = \frac{\left(\mu\,(j)\;-\;\mu\,(i)\right)}{\sigma\,(i)}$$

(2)

From these values, a weighted undirected graph indicating the strength of the covariance results in a symmetric covariance matrix C (i, j) (eq.3), where each row i represents the edge that goes out from node i, to arrive at each node j represented by column j.

$$C(\,i, j) = \frac{\left((|\,Z\,(i,j)\,|\;+\;|\,Z\,(j,i)\,|)\right)}{2}$$

(3)

Using these definitions, strength of structural covariance was represented as the mean of the absolute values of Z (i, j) and Z (j, i), where C (i, j) measures the similarity in cortical thickness distribution between nodes i and j (eq.3). The structural covariance values represented in the adjacency matrix thereby quantify the node-to-node influence generated by within the network. A function was created in R which computed the structural covariance adjacency matrix by selecting the mean and standard deviation of node i, and the mean and standard deviation of node j. From these variables, the function computed the z score of ij, and the z score of ji and manipulated these values into the structural covariance matrix using the formula in eq.3. Using R, these covariance values were represented within a heat map whereby default mode network regions were listed as 16 regions of interest (8 nodes bilaterally: 1) rostral anterior cingulate, 2) precuneus, 3) parahippocampal, 4) posterior cingulate, 5) medial orbitofrontal, 6) lateral orbitofrontal, 7) isthmus cingulate, 8) caudal anterior cingulate) on the X and the Y axes and a colored scale gradient was applied to the adjacency cells. Using this visualization technique, areas of greatest covariance in thickness distribution between nodes emerge as a pattern unique to each individual. Using the 16 x 16 matrix, nodal efficiency and global efficiency were calculated for each subject’s T1 scan.

2. Calculation of Nodal Efficiency

The structural covariance adjacency matrix was used to calculate the efficiency of each node and the entire global efficiency of a network. Nodal efficiency (eq.4) represents the influence of one node within a given network and how well that node is integrated within the network, given its shortest paths.

$$E_{nodal}(j) = \frac{1}{N-1}\sum_{i}\frac{1}{l_{ij}}$$

(4)

Nodal and global efficiency of the brain's structural default mode network were calculated using the structural covariance network was first computed through the application of a function written in R and saved as 'adjacency_matrix', which is provided in the appendix A.3. Local and Global Efficiency Calculation

3. Calculation of Global Efficiency

After the covariance adjacency matrix was computed, the threshold for distinguishing a true or false connection between nodes was set. For instance, where the threshold was set to 0.3 or below, all cells in the adjacency matrix with a structural covariance value greater than 0.3 were set to 1, and all others were set to 0. Thus, a binary matrix was formed according to the presence or absence of a true connection, as determined by the preset threshold. Following the construction of the binary matrix, a breadth-first search algorithm was applied to the data in order to determine the shortest distance between nodes, based on the presence or absence of an edge. Using this algorithm, node i, if not immediately sharing an edge with node j, would pass through one or multiple neighboring nodes in order to eventually reach node j. Employing the breadth-first search algorithm allowed for the identification of the shortest path links between nodes. This generated an adjacency matrix of shortest paths between nodes.

First, the distance of the shortest path between nodes of this undirected graph was calculated using Dijkstra's algorithm[93], an iterative process that identifies the shortest path from one node (source node) to all other nodes in the graph. This created an additional 16 x 16 adjacency matrix of shortest path from node i to j. After accounting for the removal of the matrix diagonal, the column sum of the inverse of the shortest path length for each node i to j, divided by the number of rows, was used to calculate nodal efficiency (eq. 4). Nodal efficiency for each of the 16 default mode network nodes was stored as a separate variable within the dataset. From the computation of nodal efficiency, global efficiency was calculated as the average of the inverse shortest path length and quantified as the efficient exchange of information where each node is capable of exchanging information using the shortest path length. Applying the formula proposed by Latora et al. (eq.5)[88] generated one global efficiency value for the default mode network of an individual subject scan.

$$E_{global} = \frac{1}{L'} = \frac{1}{N(\,N-1)} \sum_{i \neq j} \frac{1}{l_{ij}}$$

(5)

This global efficiency value was stored for each subject matrix and included in the dataset as a new variable. It is important to note that in the case where a node cannot be reached at a set threshold (where the threshold is too high, and a node appears isolated) and no degree links are available, the global efficiency of the network cannot be equally compared between scans and should therefore be penalized in subsequent analyses. Conversely, the threshold could be limited from excluding any nodes within the predefined network, thereby forcing the system to include all network nodes. To address the issue of the absent edge, in the binary matrix, the proportion of connections made by each node was calculated. That is, in the network A, B, C, D, if an edge was present between nodes A-C and A-D, but not between A-B, the proportion for that node would be 2/4, or 0.5. This way, the proportion of true edges leaving the node became the nodal efficiency for that node and the process of computing global efficiency.

E. Variable Preparation for Statistical Evaluation

1. Amyloid-β (Aβ)

Amyloid-β load (AV45_bl) was evaluated as a continuous variable, a higher value indicating increased pathology. Cases were also categorized for analysis as “Amyloid-β positive” and “Amyloid-β negative." Using Landau’s[89] April 2018 Neurology paper as a reference, the cutoff was set to an Amyloid-β load of 1.11, in order to categorize cases as Amyloid-β negative. In this subset n=348 (AD=17, CN=125, MCI=206), the Amyloid-β load ranged from 0.8385 to 1.1081. “Amyloid-β positive” was classified as cases where Amyloid-β load was equal to or greater than 1.11.

2. Apolipoprotein E4 (APOE4)

APOE genotyping was performed using the Illumina HumanOmniExpress BeadChip. The ADNI genetic data protocol can be found in the paper, “Alzheimer’s Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans” (Saykin et al., 2010). Each participant’s *APOE* genotype was coded in two forms: first as a 3-level exposure variable to carriership of the E4 risk allele (where 0=not present (E3/E3, E3/E2, E2/E2), 1=1 E4 allele (E3/E4, E2/E4), 2=2 E4 alleles (E4/E4)), then as a binary variable (where 0=not present (E3/E3, E2/E3, E2/E2) and 1=present (E2/E4, E3/E4, E4/E4)).

3. Cognitive Reserve (CR)

Completion of formal education in total years was derived from the ‘adnimerge’ file. Following sample selection, the data were restructured into two additional ordinal variables: a) to represent four ordinal levels of educational attainment (‘No Highschool Diploma’ (<12 years), ‘At Least Highschool’ (>=12 years), ‘At Least Undergrad’ (>16 years), ‘At Least Gradschool’ (>= 20)); and b) the bottom 25% and top 75% of educational attainment (‘Low Education’ (<= 14 years), and ‘High Education’ (>=18 years)).

4. Alzheimer’s Disease Diagnostic Classification

Diagnostic status was derived from the ‘adnimerge’ file and transformed into a three-level ordinal variable: ‘CN’ (Control subjects), ‘MCI’ (Mildly Cognitively Impaired subjects which combined late (LMCI) and early MCI (EMCI)), and ‘AD’ (Alzheimer’s Disease subjects). Each individual underwent clinician verification of diagnostic status only after reviewing the 3T MRI Radiology report and Clinical Read. The screening procedures for new participants ensured that control participants were free of memory complaints, with normal memory function as documented by scoring above education adjusted cutoffs on the Logical Memory II subscale (maximum score is 25) (>=9 where 16 or more years of education, >=5 where 8-15 years of education, >=3 where 0-7 years of education), Clinical Dementia Rating scale (symptom free with a rating of 0), and a score between 24 and 30 on the Mini-Mental State Exam. For inclusion, Mildly Cognitively Impaired participants (combined EMCI and LMCI) must have had a subjective memory concern as reported by the participant themselves or the clinician, abnormal memory function documented by the Logical Memory II subscale (score of <= 8 where 16 or more years of education, score of <= 4 where 8-15 years of education, score of <=2 where 0-7 years of education), Clinical Dementia Rating of at least 0.5, and having sufficiently preserved functional performance such that a diagnosis of Alzheimer's disease could not be made. Participants who were categorized as having Alzheimer's disease must have had a subjective memory concern, abnormal memory functioning as documented by the Logical Memory II subscale (same as LMCI), a Mini-Mental State Exam score of between 20 and 26, a Clinical Dementia Rating of 0.5 or 1.0, and meeting criteria for probable AD as determined by the NINCDS-ADRDA criteria[79] (National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's Disease and Related Disorders Association (ADRDA)).

5. Global Efficiency (Eglobal)

Global efficiency of the brain's structural default mode network was calculated using the structural covariance network which was first computed through the application of a function written in R and saved as 'adjacency_matrix', which is provided in the appendix A.3. Local and Global Efficiency Calculation. A global efficiency value was attributed to each scan id.

6. Gray Matter Density (GMD)

Gray matter density of the bilateral middle temporal region, hippocampus, whole brain, entorhinal, and fusiform were selected and entered into statistical models. Middle temporal and hippocampal regions are included within the default mode network. This dissertation used these regions within the computation of the measure of global efficiency in the default mode network (DMN).

7. Pfeffer Functional Activities Questionnaire (FAQ)

The ADNI protocol included several widely used and efficient, standardized neuropsychological tests to represent the domains of interest in the control, MCI, and AD population. Seven of the ADNI cognitive measures are used as part of a collective Uniform Data

Set (UDS) facilitated by the National Alzheimer's Coordinating Center (NACC, https://www.alz.washington.edu). Performance of cognitive functioning such as memory, processing speed, and executive functioning was assessed at screening and baseline, the timing coinciding with the structural MRI scan and F-AV-45 Amyloid-β imaging (https://adni.loni.usc.edu/wp-content/uploads/2008/07/adni2-procedures-manual.pdf). The Pfeffer Functional Activities Questionnaire (FAQ), which assesses 10 common activities of complex cognitive and social functioning, was selected as the measure of neuropsychological performance. Based on an ADNI study which utilized all available modalities for feature selection and classification, the FAQ was identified as the best performing single feature of MCI conversion to AD at 3-year follow-up) [73]

4.3 Objective 1: Investigate Brain Structure

A. Models

The hypothesis for Objective 1 was that education and APOE genotype differentially impact GMD. The Objective 1 Model (Table 4.1) was run on the total sample ((CN=221, MCI=501, AD=145); n=867, the breakdown of *APOE4*Cat (non-carriers = 0 vs carriers = 1) was as follows: CN0=159, CN1=62, MCI0=263, MCI1=238, AD0=47, AD1=98). The hypothesis of the present study, "Education and *APOE* genotype differentially impact GMD," was evaluated through the employment of multivariate regression, setting GMD in five key regions associated with AD pathology, 1) middle temporal, 2) hippocampus, 3) whole brain, 4) entorhinal, and 5) fusiform as dependent variables, and CR as the predictor, and diagnostic category, Amyloid-β, age, and gender, as covariates.

In each equation, contributions to GMD of these contributing factors were tested within the entire population using a multiple regression analysis.

Table 4.1

Objective 1 Model Development

Model Notation in R	Model Purpose and Description
GMD ~ CR + DxOrd + Age + Gender + APOE4 status + (CR:APOE4 status)	• Evaluation of CR independent contributions to bilateral GMD ROIs • Evaluation of APOE Carrier Status independent contributions to bilateral GMD ROIs • Evaluation of CR x APOE Carrier Status Interaction • Adjustments for diagnosis category (CN = Control, MCI = Mild Cognitive Impairment, AD = Alzheimer's Disease), Age, and Gender

APOE4 Status = Apolipoprotein E4 Carriership (binary: E4+ vs. E4-), CR = Cognitive Reserve, DxOrd = Diagnostic Status (ordinal), FCI = Fast Causal Inference, GMD = Gray Matter Density, ROIs = Regions of Interest, GMD ROIs = Middle Temporal, Hippocampus, Whole Brain, Entorhinal, and Fusiform.
Note. The models depicted use R notation: Dependent Variable ~ Independent Variable + Covariates + (Independent Variable: Independent Variable), where "~ "means regressed on, and ":" is an interaction term.

Table 4.1 describes the approach to investigate the interplay of *APOE4* genotype and cognitive reserve on GMD we regressed GMD on a model to determine a) whether there were independent effects of cognitive reserve and *APOE4* status which sustained after age and gender adjustments, and b) whether the interaction of CR and *APOE4* status significantly contributed to GMD in our five regions of interest.

B. Results

The model notation and key findings are presented in Table 4.2 below.

a) Education (i.e., CR) showed significant independent effects on middle temporal GMD, whole brain GMD, entorhinal GMD, and fusiform GMD

b) *APOE4* status showed independent effects in the middle temporal GMD and whole brain GMD.

c) Significant interactions between CR and *APOE4* were identified in the middle temporal GMD and in the whole brain GMD.

Table 4.2

Objective 1 Results

	CR:APOE4Cat
Outcome Measure	Linear Model t value Pr(>|t|)
Gray Matter Density (GMD)	
Middle Temporal	coeff. -134, t -2.108 (p<0.05) *
Hippocampus	coeff. -36.4, t 7.321 (p=0.2)
Whole Brain	coeff. -5095, t -2.17 (p<0.05) *
Entorhinal	coeff. 2.305, t -0.23 (p=0.8)
Fusiform	coeff. -104.28, t 1.40 (p=0.09)

Multivariate analysis was performed on gray matter density ROI. This table represents the results of Objective 1 Model.

APOE4 Status = Apolipoprotein E4 Carriership (binary: E4+ vs. E4-), CR = Cognitive Reserve, DxOrd = Diagnostic Status (ordinal), FCI = Fast Causal Inference, GMD = Gray Matter Density, ROI = Region of Interest, GMD ROIs = Middle Temporal, Hippocampus, Whole Brain, Entorhinal, and Fusiform.

Note. The models depicted use R notation: Dependent Variable ~ Independent Variable + Covariates + (Independent Variable : Independent Variable); where "~" means regressed on, and ":" is an interaction term. **p* < .05. ***p* < .01. ****p* < .001

C. Conclusion

Education and *APOE4* differentially impact GMD in the middle temporal and whole brain.

D. Discussion

We looked at these 5 regions and found that CR and *APOE4* do affect GMD, as shown in prior literature. We found significant independent effects of CR and *APOE4* status on the middle temporal region as well as in the whole brain measure. We identified interaction effects between CR and *APOE4* carrier status on functioning in the middle temporal and whole brain measures of GMD (Table 4.2). We wanted to study an integrative picture of neurodegeneration, to reduce the number of singular regions investigated. In doing this, we bring strength to our findings by reducing multiple comparisons. Therefore, rather than evaluating a single brain region of interest, we decided to develop and utilize a metric that better represents the covarying nature of neurodegeneration and would summarize the structural deficits in the brain.

While controlling for Amyloid-β, CR continued to show independent effects on GMD in the middle temporal, whole brain, and fusiform. CR and Amyloid-β independently predicted GMD in the whole brain, middle temporal, and fusiform measures (the association with entorhinal GMD was lost upon adjustment of Amyloid-β status). There was a significant interaction between CR and *APOE4* carriership in middle temporal lobe GMD and whole brain GMD measures.

Our prior work evaluating the effects of *APOE4* and cognitive reserve support the likelihood of an *APOE* x CR interaction.[54] Specifically, carriers and non-carriers of the *APOE4* allelic variant a) may differ in their response to environmental enrichment, and/or b) may be further along in the deterioration process of aging, such that the presumed benefits are undetectable.

4.4 Objective 2: Create Composite Measure for Neurodegeneration

A. Models

We hypothesized that *APOE4* impacts clinical functioning, and that the effect is mediated by neurodegeneration ─ which we computed as global efficiency of the default mode network. We regressed clinical functioning on global efficiency, *APOE4* status, and cognitive reserve, adjusting for age and gender, and we added an interaction term of global efficiency and *APOE4* carriership. In doing this, we sought to answer the question whether the effect of global efficiency on clinical functioning varies by *APOE4* carriership. In other words, if an individual is a carrier of the *APOE4* allele, might neurodegeneration have more of an effect on clinical functioning? Thus, in Objective 2 Primary Models (Table 4.3), we sought to understand whether any effect would be sustained upon controlling for Amyloid-β, which is known as a “driver” of neurodegeneration.

Table 4.3

Objective 2 Model Development

MLR Model*	MLR Model Purpose
1. FAQ ~ globeff + APOE4 + (globeff : APOE4) + CR + AGE + PTGENDER	- Evaluate interaction effect of E_{global} and *APOE4* on Clinical Functioning - Evaluate E_{global} independent effects on Clinical Functioning - Evaluate *APOE4* Carrier Status independent effects on Clinical Functioning - Evaluate Independent Effects of CR - Adjustments for Age and Gender
2. FAQ.bl ~ globeff + APOE4 + (globeff : APOE4) + CR + AGE + PTGENDER + Amyloid	- Model 1 repeated, controlling for Amyloid-β

APOE4 Status = Apolipoprotein E4 Carriership (binary: E4+ vs. E4-), CR = Cognitive Reserve, FAQ = Pfeiffer's Functional Activities Questionnaire, globeff = Eglobal = Global Efficiency (composite measure of gray matter density), Amyloid = Amyloid-β. The models depicted use R notation: Dependent Variable ~ Independent Variable + Covariates + (Independent Variable: Independent Variable); where "~" means "regressed on."

Now, recall that in our preliminary findings in the MCSA we observed an influence of CR on GMD in non-carriers but not in carriers. Thus, in the Objective 2 Dichotomized Model (Table 4.4) we looked at the influence of CR and global efficiency within a dichotomized model in order to more closely evaluate the impact of *APOE*.

Table 4.4

Objective 2 Dichotomized Model Development

APOE4 Subset	MLR Model*
Non-Carriers (E4 -) n=469 Education: mean = 16.23, sd = 2.69, yrs. range = 7-20	**1.** FAQ ~ CR + AGE + GENDER + Amyloid + globeff
Carriers (E4 +) n=398 Education: mean = 15.95, sd = 2.75, yrs. range = 6-20	**2.** FAQ ~ CR + AGE + GENDER + Amyloid + globeff

APOE4 Status = Apolipoprotein E4 Carriership (binary: E4+ vs. E4-), CR = Cognitive Reserve, FAQ = Pfeiffer’s Functional Activities Questionnaire, Amyloid = Amyloid-β, globeff = Eglobal = Global Efficiency (composite measure of gray matter density).

Note. The models depicted use R notation: Dependent Variable ~ Independent Variable + Covariates + (Independent Variable: Independent Variable); where “~“ means “regressed on."

B. Results

We found that the interaction between global efficiency and *APOE4* carriership neared significance, and that each of the individual terms in the model contributed to functioning (Table 4.6, Model 1). Amyloid-β appeared to account for the influence of *APOE4* in the effect on functioning (Table 4.6, Model 2).

Table 4.5

Objective 2 Model Results

MLR Model*	Key Findings
1. FAQ.bl ~ globeff + APOE4 + (globeff : APOE4) + CR + AGE + PTGENDER	• *APOE4* and E_{global} interaction neared significance (coeff. 0.20, p=0.054) Independent Effects on Functioning • E_{global} (coeff. -0.25, p<0.005)** • CR (coeff. -0.88, p<0.005)** • *APOE4* status (coeff. -4.01, p<0.0005)*** • Age (coeff. 0.09, p<0.05)* • Gender (coeff. 1.09, p<0.0005)***
2. FAQ.bl ~ globeff + APOE4 + (globeff : APOE4) + CR + AGE + PTGENDER + Amyloid	• *APOE4* and E_{global} interaction non-significant (coeff. 0.11, p=0.28) Independent Effects on Functioning • E_{global} (coeff. -0.18, p<0.05)* • CR (coeff. -0.95, p<0.0005)*** • *APOE4* status (coeff. -1.23, p=0.28) • Age (coeff. 0.06, p<0.05)* • Gender (coeff. 1.40, p<0.005)** • Amyloid (coeff. 9.15, p<2e-16)***

APOE4 Status = Apolipoprotein E4 Carriership (binary: E4+ vs. E4-), CR = Cognitive Reserve, FAQ = Pfeiffer's Functional Activities Questionnaire, Amyloid = Amyloid-β. This table displays the results of Objective 2 Primary Models (Table 4.4).

Note. The models depicted use R notation: Dependent Variable ~ Independent Variable + Covariates + (Independent Variable : Independent Variable); where "~ " means "regressed on".

**p* < .05. ***p* < .01. ****p* < .001

Results of the dichotomized models indicated a lack of effect of global efficiency on functioning in *APOE4* non-carriers (Table 4.6, Model 1), however, global efficiency was significant in *APOE4* carriers (Table 4.6, Model 2). Amyloid-β was significant in both dichotomized models. Cognitive reserve was significant in *APOE4* non-carriers, but not in *APOE4* carriers (Table 4.6)

Table 4.6

Objective 2 Dichotomized Model Results

MLR Model*	Key Findings
Non-Carriers (E4-) 1. FAQ.bl ~ CR + AGE + GENDER + Amyloid + globeff	• CR (coeff. -0.30, p<0.005)** Age (coeff. -0.08, p<0.05)* • Gender (coeff. 2.02, p<0.0005)*** • Amyloid (coeff. 7.62, p<1.29e-08)*** • E_{global} (coeff. -0.084, p=0.195)
Carriers (E4+) 2. FAQ.bl ~ CR + AGE + GENDER + Amyloid + globeff	• CR (coeff. -0.23, p=0.051) Age (coeff. 0.04, p=0.433) • Gender (coeff. 0.83, p=0.21) • Amyloid (coeff. 10.66, p<2.66e-12)*** • E_{global} (coeff. -0.17, p<0.05)*

APOE4 Status = Apolipoprotein E4 Carriership (binary: E4+ vs. E4-), CR = Cognitive Reserve, FAQ = Pfeiffer's Functional Activities Questionnaire.

Note. The models depicted use R notation: Dependent Variable ~ Independent Variable + Covariates + (Independent Variable: Independent Variable); where "~" means "regressed on." This table displays the results of Objective 2 Dichotomized Models (Table 4.4).

**p* < .05. ***p* < .01. ****p* < .001

C. Conclusion

APOE4 carrier status impacted clinical functioning, and the effect was mediated by global efficiency. When accounting for Amyloid-β in this model, the effect of global efficiency was slightly reduced. The effect of global efficiency and Amyloid-β on clinical functioning appear to be greater in *APOE4* Carriers.

D. Discussion

The evaluation of cognitive reserve as a dynamic system ought to advance the understanding of the active resilience mechanism and add to the groundwork for innovative, translational approaches to prompt and evaluate techniques for clinical intervention in AD. Our work supports prior literature indicating that a neural system may operate differently given *APOE4* carrier status. We found that upon the addition of Amyloid-β into our model, we completely lost the interaction effects of *APOE4* status and global efficiency, as well as the independent effects of *APOE4* status on functioning. We observed a decrease in the effect of global efficiency and gender on functioning but observed an increase of CR effects on functioning. The most important part of these findings was that Amyloid-β appears to mediate the effect of *APOE* on functioning. We made the following three key observations in our second objective:

1. Amyloid-β is a driver and effects of *APOE* and CR on clinical functioning are captured through Amyloid-β.
2. Global efficiency captures the upstream impact of deterioration.
3. To understand how clinical functioning is impaired, we must understand the complex interplay between genetics, cognitive reserve, pathological changes as measured by Amyloid-β and their impact on brain efficiency.

In investigating the impact of *APOE* on clinical functioning (please see Figure 4.1 and 4.2 below), we found from the regression models that *APOE* had an effect on clinical functioning; however, this relationship was masked by Amyloid-β. Global efficiency, our measure of neurodegeneration, was observed to affect clinical functioning, as did Amyloid-β.

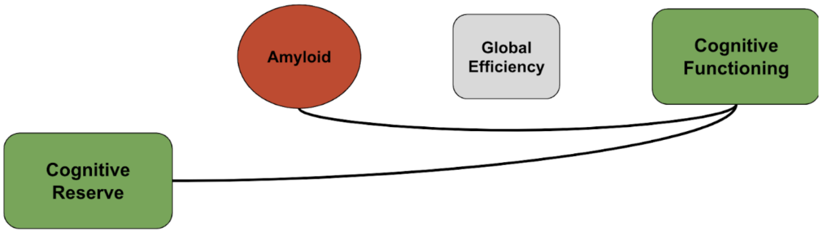


Figure 4.1 Investigation of impact on clinical functioning in APOE4 non-carriers.

Investigation of impact on clinical functioning in APOE4 Non-Carriers. This figure is a visual representation of the Objective 2, Dichotomized Model 1 (Table 4.6), illustrating that Cognitive Reserve appeared to have an effect on Clinical/Cognitive Functioning in APOE4 non-carriers, however, Global Efficiency did not.

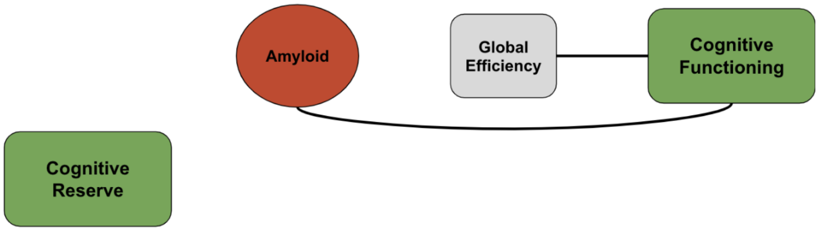


Figure 4.2 Investigation of impact on clinical functioning in APOE4 carriers.

This figure is a visual representation of the Objective 2, Dichotomized Model 2 (Table 4.6), illustrating that in contrast to Figure 4.1, Cognitive Reserve did not appear to have an effect on Cognitive functioning in *APOE4* carriers. Global Efficiency appeared to effect Clinical/Cognitive Functioning in *APOE4* carriers.

Our findings indicate that cognitive reserve appears to significantly affect functioning in the non-carrier subset, but not in the *APOE4* carrier subset. We observed similarities with our preliminary study where we see education effects, but only within the *APOE4* non-carriers. In our model, global efficiency only appeared as a significant contributor to functioning in the *APOE4* carriers.

Multiple regression models are built to understand the association between multiple regressor variables on a single outcome variable. Interactions and the strength contributions of the variance of the outcome measure can serve as indicators of association and interdependence. Regression does not provide insight into the cause of the predictor or outcome variable, nor does it provide information pertaining to the directionality of the relationship. To understand the causal relationship between variables, we sought to complement our regression findings by putting our effort into using causal inference. For this we used algorithms tailored to take into account latent relationships to discover causal associations.

4.5 Objective 3: Apply Causal Inference

A. Models

The hypothesis of Objective 3 was that APOE4 carriers as compared to non-carriers will demonstrate differences in network recruitment. Causal inference was applied to the data to complement the regression models by inferring directionality and independence of variables. Inferred associations from regression modeling can be better depicted by removing conditional independence and summarizing findings in a causative graph through Fast Causal Inference (FCI) (Please see Table 4.7 below).

Table 4.7

Objective 3 Fast Causal Inference (FCI)

Fast Causal Inference (FCI) Model Purpose	FCI Input Variables	FCI Results	FCI Alignment with MLR: Key Findings
Test of Objective 2 *APOE4* and CR differentially impact clinical functioning and the effect is mediated by Global Efficiency.	• Functioning • CR • Global Efficiency • Amyloid • Age • Gender • *APOE4* Carriership	• Age, Amyloid-β, and global efficiency are directly, causally related to functioning. • Age is directly, causally related to Amyloid-β. • There is an unrecorded cause of Global Efficiency and Amyloid-β	• Aligned • CR is not associated with global efficiency • Amyloid-β and global efficiency interact to affect functioning • Global efficiency, CR, Amyloid-β, and age, independently affect functioning • Unaligned • Gender associations to functioning may be due to a latent variable (FCI) • Gender independent predictor of functioning (MLR)
APOE4 non-carriers (*E4* -) evaluated to understand the interplay of variables.	• Functioning • CR • Global Efficiency • Amyloid • Age • Gender	• Amyloid-β is directly, causally related to functioning. • There is an unrecorded common cause between age and functioning • There is a relationship between gender and functioning which may be influenced by a confounding/latent variable	• Aligned • CR affects Amyloid-β • Amyloid-β affects functioning • Age affects functioning • Unaligned • Education does not directly affect functioning (FCI) • Education directly affects functioning (MLR)
APOE4 carriers (*E4* +) evaluated to understand the interplay of variables.	• Functioning • CR • Global Efficiency • Amyloid • Age • Gender	• Direct causal relationship of Amyloid-β to functioning • Relationship between global efficiency and Amyloid-β may be influenced by a confounding/latent variable • Relationship between age and Amyloid-β may be influenced by a confounding/latent variable	• Aligned • Global efficiency independently affects functioning • Education does not significantly affect functioning • Global efficiency is directly causally related to functioning. • Age is related to Amyloid-β • Unaligned • Amyloid-β and Age association may be due to a latent variable (FCI) • Amyloid-β and Age are associated (MLR)

APOE4 Status = Apolipoprotein E4 Carriership (binary: E4+ vs. E4-), CR = Cognitive Reserve, DxOrd = Diagnostic Status (ordinal), FCI = Fast Causal Inference, GMD = Gray Matter Density, MLR = Multivariate Linear Regression.

Below, Table 4.8 depicts the Tetrad FCI Graph Symbols. "An edge between two variables in the output, however the ends of edge are marked, indicates that there is a causal pathway (a direct cause in one direction or the other, or a common cause) connecting the two variables, that does not contain any other observed variable."

Table 4.8

Tetrad FCI Graph Interpretation

Tetrad FCI Symbol Depiction	Indicator	Implication
X --Y	An edge from X to Y that is unmarked	● X is a cause of Y. X may not, however, be a direct cause of Y. ● Similarly, Y is a cause of X. Y may not, however, be a direct cause of X.
X-->Y	An edge from X to Y that has an arrowhead directly into Y	● X is a cause of Y. ● Y is not a cause (not an ancestor) of X.
X<-->Y	An edge with two arrowheads connecting X and Y	● There exists an unrecorded common cause of X and Y
Xo-->Y	An edge end is marked with "o"	● The algorithm cannot determine whether there should or should not be an arrowhead at that edge end.

This table depicts the Tetrad FCI graph symbols, their indicators, and implications for interpretation.

B. Results

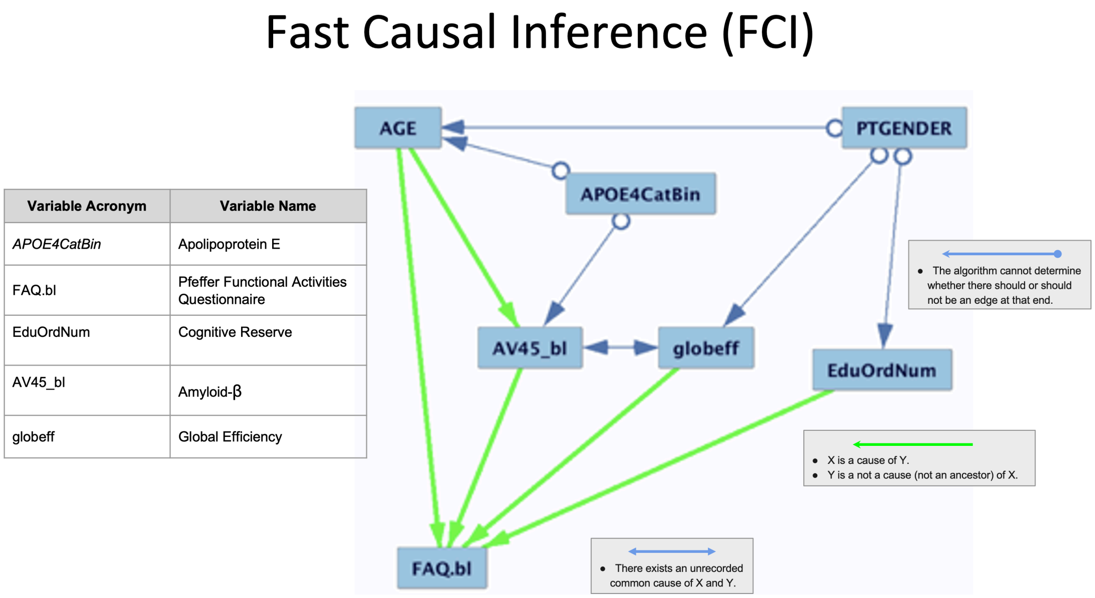
In the FCI graph, the cutoff for p-values (alpha) was set to 0.05, meaning that conditional independence tests with p-values greater than 0.05 would be stored as "independent." At this significance value of 0.05, the search algorithm identified a direct relationship with FAQ.bl and Age, FAQ.bl and amyloid, FAQ.bl and global efficiency, FAQ.bl and CR (level of education (ordinal)), and amyloid and age, as indicated by a bright line in the Tetrad visual output (Please see Figure 4.3). The association between global efficiency and amyloid (thickened line) implied that there is a relationship with no latent confounder. *APOE4* carriership was identified as either the cause of amyloidosis, and/or that there was an unmeasured confounder of *APOE4* carriership and amyloidosis. The unmeasured confounder (e.g., Latent Confounder) (Figure 2.5) of B and C would indicate that there may be unmeasured variables along the causal pathway from B to C.

Global efficiency and gender were related in that gender affected global efficiency or there was an unmeasured confounder. There also appeared to be a potentially unmeasured confounder between gender and age, *APOE4* carriership and age, and gender and CR (ordinal, level of education). Please refer to Tetrad FCI Graph Symbols (Table 4.8) for a table of FCI path symbols and descriptions, and to Biomarkers Evaluated in this Dissertation (Table 3.2) for a tabular review of biomarkers and corresponding variable names. *APOE4* carriership (*APOE4*CatBin) does not appear to be causally related to global efficiency in effects on functioning (FAQ.bl). Global efficiency directly contributes to functioning and Amyloid-β load (AV45_bl). Amyloid-β load directly affects functioning. Gender affects global efficiency and level of education. Age directly affects functioning. Education directly affects functioning. Age directly affects Amyloid-β load. AV45_bl = Amyloid-β. FCI Search Algorithm, alpha = 0.05.

C. Conclusion

Amyloid-β load masks the effect of *APOE4* on functioning. The relationship between *APOE4* carriership and global efficiency, as well as *APOE4* carriership and functioning, arises in the absence of Amyloid-β. Education has a direct causal relationship on Amyloid-β, which then has a direct causal relationship on functioning, only in *APOE4* non-carriers.

Amyloid-β has a direct causal effect on functioning in *APOE4* carriers, however, there exists no relationship of education on Amyloid-β (unlike what is demonstrated in the *APOE4* non-carrier sample. Further, only in the *APOE4* carriers is there a relationship between global efficiency and functioning, and global efficiency and Amyloid-β.



Fast Causal Inference (FCI)

Variable Acronym	Variable Name
APOE4CatBin	Apolipoprotein E
FAQ.bl	Pfeffer Functional Activities Questionnaire
EduOrdNum	Cognitive Reserve
AV45_bl	Amyloid-β
globeff	Global Efficiency

- The algorithm cannot determine whether there should or should not be an edge at that end.

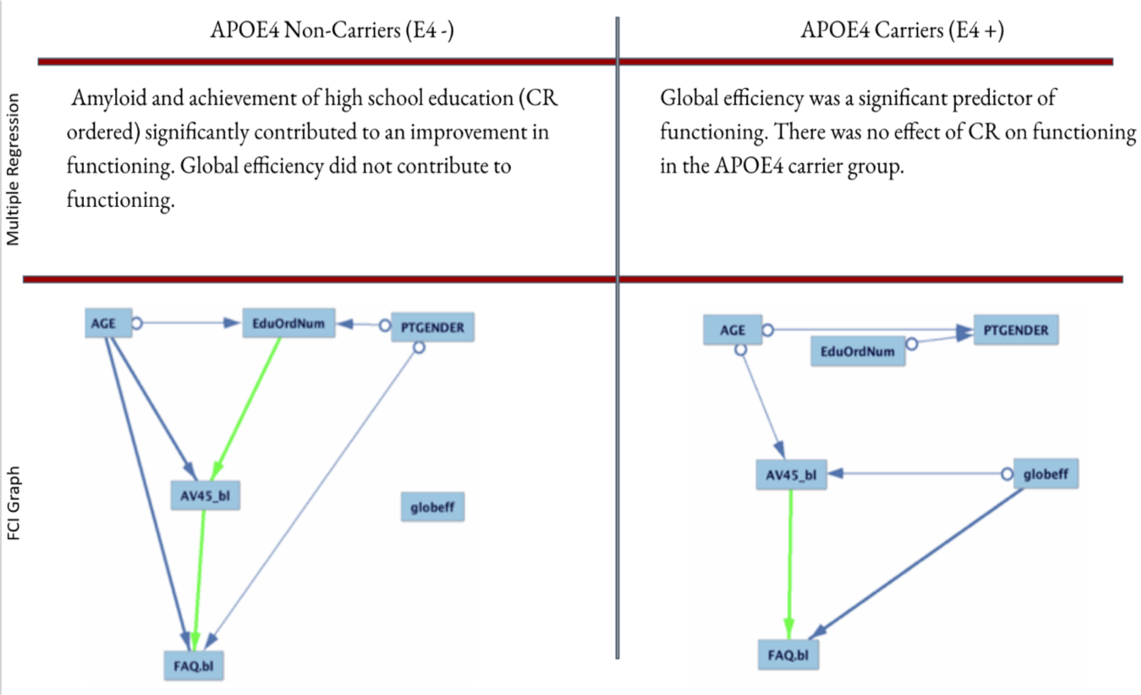
- X is a cause of Y.
- Y is a not a cause (not an ancestor) of X.

- There exists an unrecorded common cause of X and Y.

Figure 4.3 Overarching Biomarker FCI Graph.

This is the output of the Tetrad FCI graph, referencing the Objective 3 question: Do APOE4 carriers as compared to non-carriers demonstrate differences in network recruitment (specifically, global efficiency of the default mode network)? APOE4CatBin = APOE4 Carrier Status, AV45_bl = Amyloid-β, globeff = Global Efficiency, EduOrdNum = CR (ordinal), FAQ = Functioning, PTGENDER = Gender. Fast Causal Inference. Alpha was set to 0.05, implicating conditional independence tests with p-values greater than 0.05 were stored as “independent."

In the dichotomization of *APOE4* carriership (Figure 4.4), we observed that in *APOE4* non-carriers, there does appear to be a direct causal relationship of education on amyloid that may result in a direct relationship from amyloid to functioning, which also appears to be influenced by age. As shown in our previous work, the effect of education on functioning does not appear to be direct. Meanwhile, the effect of global efficiency on functioning is direct, and the algorithm appears uncertain as to whether amyloid directly affects global efficiency.



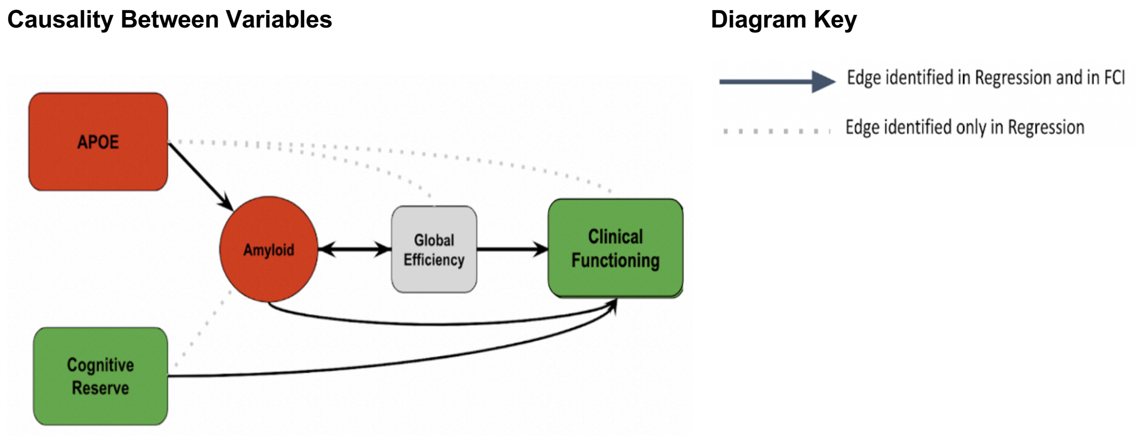
Amyloid and achievement of high school education (CR ordered) significantly contributed to an improvement in functioning. Global efficiency did not contribute to functioning.

Global efficiency was a significant predictor of functioning. There was no effect of CR on functioning in the APOE4 carrier group.

Figure 4.4 APOE4 Effects: Regression and FCI Comparison.

This figure presents comparison of the multiple regression analysis and the FCI analysis. In APOE4 non-carriers, the impact of Cognitive Reserve appears to contribute significantly to Amyloid-β, which then impacts Clinical Functioning. Global Efficiency does not appear to influence Clinical Functioning (Column 1). Conversely, in APOE4 carriers, the effect of Cognitive Reserve on amyloid is absent, and global efficiency effects both Amyloid-β and Clinical Functioning.

In the below image, the bold represent findings identified using both FCI and regression. Figure 4.5 explains the study findings in terms of causality. The bold lines represent findings generated by regression modeling and confirmed by FCI. Although we see similarities in edges between nodes, we see the elimination of relationships that may be due to conditional independence. For instance, the direct relationship of *APOE* with functioning identified in the regression model is eliminated and we see in the causal model, that the effect of *APOE* may be passing through a causal chain initiated by a direct relationship with amyloid. On the other hand, cognitive reserve (Education) does not appear to be directly contributing to amyloid. Thus, while *APOE* and CR appear to be directly influencing amyloid, it is only *APOE* that directly influences amyloid in the overall sample.



Causality Between Variables

Diagram Key

APOE
Amyloid
Global Efficiency
Clinical Functioning
Cognitive Reserve
Edge identified in Regression and in FCI
Edge identified only in Regression

Figure 4.5 Application of causal inference

Application of causal inference to understand the complex interplay of variable pathway and impact (Objective 3). Amyloid is the driver of neurodegeneration and cognitive decline. *APOE* directly impacts Amyloid.

Global efficiency is a composite of GDM, which we are using to evaluate neurodegeneration. Given that the relationship of CR to amyloid disappeared when the FCI model was run to test for independence, it is possible that *APOE4* was a latent confounder on the relationship between cognitive reserve and amyloid. Also, *APOE4* no longer appears to directly contribute to cognitive functioning, but may instead work through amyloid. This confirms that amyloid is the driver of neurodegeneration.

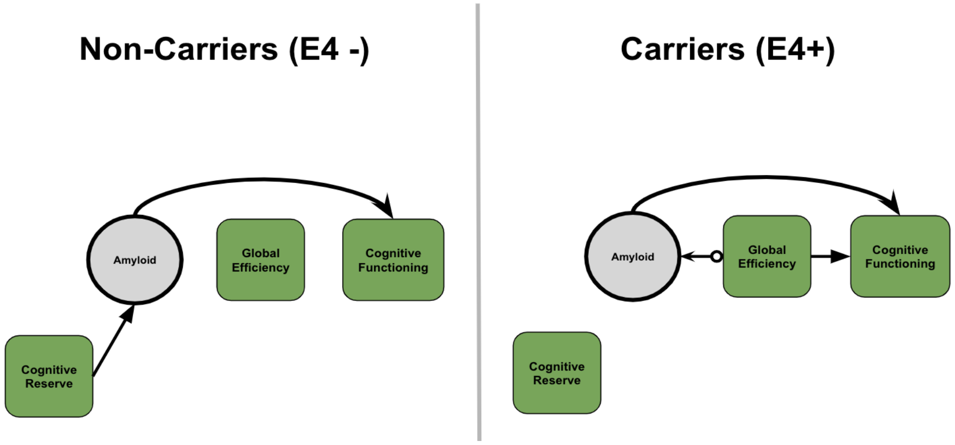


Figure 4.6 Application of causal Inference dichotomized by APOE4 carrier-status.

While Amyloid-β appears to impact Clinical/Cognitive functioning in carriers and non-carriers, global efficiency only appears to directly influence Clinical/Cognitive functioning in carriers of the APOE4 allele.

As shown in Figure 4.6, repeated in the dichotomization of *APOE4* carrier-status is the observation of CR effects within *APOE4* non-carriers which remain absent in the E4 carrier model.

D. Discussion

The specific neurophysiological mechanisms that facilitate the effective integration of experience and the development of neuroprotective intellectual abilities are unclear. The multifaceted nature of the disease construct points to the interaction of multiple contributing factors. The literature states that the adverse effects of *APOE4* on neuronal plasticity are markedly heightened upon exposure to stress, reduce amyloid-β clearance[95], and increase the rate of brain atrophy. While education may help *APOE4* non-carriers and *APOE4* carriers, the neuroplasticity mechanisms through which education aids in delaying AD differs by *APOE* genotype, and non-carriage of the ϵ4 allele, and rather carriage of the ϵ2 allele, and may serve as a developmental benefit.[90] Thus, given what we know of the mechanisms of *APOE4* (Figure 2.3), it is also plausible that *APOE4* carriers with neuronal injury and/or elevated stress might experience an increasingly detrimental effect of the toxic cycle,[55-56] as compared to non-carriers, regardless of their achievement status. Future research ought to consider the inclusion of stress as a measure of possible contribution to the interdependencies of the surveyed AD biomarkers.

Using the Tetrad knowledge input, *APOE4* carriership, age, and gender, were set to the first tier, Amyloid-β, CR (level of education), and global efficiency were set to the second tier, and functioning was set to the third tier.

In a second evaluation, the total sample was stratified by *APOE4* carriership (n_{E+}=398; n_{E-}=469). Note that multiple regression models containing identical sets of input variables were previously run in R for both carrier groups (FAQ.bl~globeff + Amyloid + (globeff:Amyloid) + CR + AGE + PTGENDER). In the Tetrad FCI search model evaluating only *APOE* carriers, Amyloid-β and global efficiency were both found to be associated with functioning.

Correlation does not imply causation; therefore, contributing factors (i.e., *APOE4*, global efficiency, Aβ) evaluated in Objective 2 were subjected to an objective search process, through the application of FCI, causal discovery. We therefore asked whether the relationships between variables evaluated for statistical significance and interactions in Objective 3 are causal (employing the FCI algorithm). In line with Objective 2, CR and global efficiency are not causally related in FCI graphs. The results indicate that a) Amyloid-β and global efficiency interact to affect functioning, and b) global efficiency, CR, Amyloid-β, and age, showed significant independent effects on functioning, and the Amyloid-β and global efficiency interaction was significant. Note that gender associations with functioning identified in Objective 3 may be due to a latent variable.

As described in Objective 2, the removal of the Amyloid-β term appears to be causally related to differences in network recruitment (DMN global efficiency). There exists the possibility of a latent variable affecting *APOE4* carriership and Amyloid-β, as well as the possibility of a latent variable affecting *APOE4* carriership in relationship with global efficiency, where Amyloid-β is absent from the model (Figure 4.4).

Figure 4.5 A is a comparison of regression and FCI tests comparing the presence of Amyloid-β in the model versus an absence of Amyloid-β from the model. In the Figure 4.5 B, FCI graph, *APOE4* carriership is related to Amyloid-β with possibility of a latent variable associated. Meanwhile, in Figure 4.5 C, *APOE4* carriership is directly related to functioning where Amyloid-β is not included as a term in the model.

It is possible that although prior knowledge inserted in Tetrad correctly represents biological functioning (*APOE4* precedes Amyloid-β), this incorrectly represents the graphical representation of the process. In other words, perhaps Amyloid-β must begin to accumulate before the effects of Amyloid-β enter as an identifiable causal contributor to functioning. This would be in line with the biochemical cascade theorized in Chapter 2 (Figure 2.3), specifically paths b where *APOE4* harmfully impacts Amyloid-β clearance and deposition, leading to increased neuronal stress and d, where Amyloid-β itself injures neurons; neuronal injury stimulates *APOE* production which induces *APOE* neurotoxic fragment formation, which thereby further perpetuates the toxic cycle.

The *APOE4* non-carrier group showed a direct effect of Amyloid-β, a relationship between age and functioning without a latent confounder, and the relationship between gender and functioning, which may be influenced by a confounding/latent variable. These relationships identified by the FCI algorithm confirmed those found in the regression model, with the exception of the direct effects of education. Figure 4.5 identifies a direct relationship of education on Amyloid-β, and a direct relationship of Amyloid-β on functioning. Education has a direct causal relationship on Amyloid-β, which then has a direct causal relationship on functioning, in *APOE4* non-carriers.

Causal influence of education on Amyloid-β identified in *APOE4* non-carriers is not reflected in the *APOE4* carriers FCI model. Conversely, causal influence of global efficiency on functioning and of global efficiency influence on Amyloid-β within the *APOE4* carrier FCI model is not reflected in *APOE4* non-carrier FCI model.

Recall Objective 2 findings: a) global efficiency is a significant predictor of functioning only in the *APOE4* carrier group (where higher global efficiency is related to better performance (lower scores) on the FAQ); and b) education significantly affects functioning only in the non-carrier group.

Recall Objective 3 findings: In the *APOE4* carrier group, global efficiency is causally related to functioning. Global efficiency and Amyloid-β are related, with a possible latent variable contributing. Age is related to Amyloid-β and not directly to functioning.

The *APOE4* non-carrier group showed a direct effect of Amyloid-β, a relationship between age and functioning without a latent confounder, and the relationship between gender and functioning which may be influenced by a confounding/latent variable. The relationships identified by the FCI algorithm confirmed those found in the regression model, with the exception of the direct effects of education found within the regression model (Figure 4.5 identifies a direct relationship of education on Amyloid-β, and a direct relationship of Amyloid-β on functioning). Education has a direct causal relationship on Amyloid-β, which then has a direct causal relationship on functioning, in *APOE4* non-carriers. CR and global efficiency are not causally related in the FCI graph.

Global efficiency of the DMN is associated with Amyloid-β and functioning. The relationship between *APOE4* carriership and global efficiency arises in the absence of Amyloid-β, where *APOE4* is marked as either the cause of global efficiency variance, or as affected by an unmeasured confounder of *APOE4* carriership and global efficiency (Figure 4.5). The relationship between *APOE4* carriership and functioning also arises in the absence of Amyloid-β. These findings were identified within the regression models described in Objective 3, with the exception of age effects on functioning (the regression models showed direct effects of age on functioning, while the FCI model implied that the age effects were targeted toward Amyloid-β).

Amyloid-β has a direct causal effect on functioning in *APOE4* carriers; however, there exists no relationship of education on Amyloid-β (unlike what is demonstrated in the *APOE4* non-carrier sample). Further, only in the *APOE4* carriers is there a relationship between global efficiency and functioning, and global efficiency and Amyloid-β (Global efficiency does not appear as a predictor of functioning in the *APOE4* non-carriers, nor is there an interaction between global efficiency and Amyloid-β in this sample).

Results of the Tetrad FCI search model for the *APOE4* non-carrier group showed a direct effect of Amyloid-β (bright arrow), a relationship between age and functioning without a latent confounder (thick blue line), and the relationship between gender and functioning which may be influenced by a confounding/latent variable (arrow anchored by circle). The relationships identified by the FCI algorithm confirmed those found in the regression model, with the exception of the direct effects of education on functioning found within the regression models; rather, Figure. 4.5 identifies a direct relationship of education on Amyloid-β, and a subsequent direct relationship of Amyloid-β on functioning.

Chapter 5 Conclusion

5.1 Summary of accomplishments and contributions

A. Contribution to Neuroinformatics

This dissertation research uniquely contributes to health informatics through the construction of a neuroinformatics pipeline (Appendix A) by employing and combining multimodal biomedical data (neuroimaging, genomics, cognition, and clinical), database management, automated computing, graph theory, and biostatistics. The application of the pipeline demonstrates that it can be used to successfully address neuroscience questions.

The work presented drew on the methodological strengths of health informatics, biostatistics, and neuroscience to evaluate the potential impact of specific allele carriership on what is recognized as a resilience mechanism for the rest of the population. While there has been a greater understanding of Alzheimer's disease (AD) processes in the last two decades, clinical trials in AD have not been successful, suggesting the need for further research to understand key questions pertaining to the underpinnings of the disease.[1] The NIH-funded Alzheimer's Disease Neuroimaging Initiative (ADNI) recognized this knowledge gap and continues to fund this database so that it contains relevant genomic, imaging, and proteomic data. Brain study generates high-dimensional data, such that combining disparate data sources requires solid and replicable processing pipelines if one seeks to advance science through expediting collaboration and leveraging prior work. To exploit these data, we have responded to action calls by informaticists Dr. Arthur Toga and Dr. Ivo Dinov and have built a neuroinformatics pipeline (Appendix A) for the replicable collection, manipulation, and analysis of data to produce meaningful and verifiable information (Aim I).

B. Contribution to Neuroscience

We set out to investigate the complex interplay between genetics, cognitive reserve, pathological changes via Amyloid-β, and their impact on brain efficiency, which influences functioning. We did indeed find complex relationships:

1. *APOE* appears to affect functioning and global efficiency through a direct relationship with Amyloid-β.

2. The effect of education on functioning is different between carriers and non-carriers.
    a. Education may not necessarily interact with *APOE*, however:
        i. In the case of *E4* non carriers, CR affects functioning through Amyloid-β.
        ii. In the case of *APOE4* carriers, education does not appear to contribute to Amyloid-β, nor does it appear to contribute directly to functioning.

Questions about such relationships were rooted in the literature, including some of our own prior work, which indicates that there exists an increased risk of AD in *APOE4* carriers versus carriers of other *APOE* allelic variants (i.e., *APOE2, APOE3*), and that GMD in AD carriers of the *APOE4* variant shows increased atrophy. Heterogeneity in the risk of dementia appears to be based on an individual's Reserve and Resilience. CR has been hypothesized to generate protective effects against gray matter atrophy (a decrease in GMD), enhance the plasticity of gray matter, and diminish the accumulation of Amyloid-β in the brain (Amyloid-β is known to eventually cause neurodegeneration). CR may have a mechanistic relation to GMD and structural networks. We incorporated CR because we were interested in studying the existence of a resilience mechanism in the face of neural injury as a function of *APOE4* carriership. The complex interplay between cognitive reserve, *APOE4*, and their combined role in AD is compelling in neuroscience. Therefore, as we aimed to investigate these complexities through the employment of neuroinformatics tools, we leveraged our unique position to construct and employ a replicable neuroinformatics pipeline to address key questions in this realm (Aim II): (1) Do education and *APOE* genotype differentially impact GMD?, (2) Does *APOE4* carrier status impact clinical functioning, and is the effect mediated by global efficiency?, and (3) Do APOE4 carriers as compared to non-carriers demonstrate differences in network recruitment (specifically, global efficiency of the default mode network)?

5.2 Generalizability of the results

In our work, the effect of cognitive reserve on GMD did appear to differ by *APOE* genotype. Our logic was that if the presence of *APOE4* leads to neurodegeneration (a reduction in GMD) at a rate higher than the absence of *APOE4*, and if individuals who carry two *APOE2* alleles or one *APOE2* allele and one *APOE3* allele are less likely to develop AD,[91] in situations of equal CR, carriers and non-carriers of the *APOE4* allele may demonstrate differences in GMD. We found that the effect of cognitive reserve on GMD differed by carriership of the *APOE4* genotype, and specifically that achievement of a high school diploma appeared protective from degeneration in the middle temporal and whole brain measures of GMD, but only in those who were not carriers of the *APOE4* allele.

Global efficiency did not appear to be significantly influenced by CR. Further, we found that CR does not appear within the model as a direct causal variable in *APOE4* non-carrier graphs (Figure 4.5); and similarly, in *APOE4* carrier graphs, CR does not appear within the model as a direct causal variable (however, importantly, it arises in relation to gender with the possibility of a latent confounder). As shown in Figure.4.7, a causal influence of education on Amyloid-β, identified by FCI within *APOE4* non-carriers is not identified in *APOE4* carriers FCI model. Conversely, causal influence of global efficiency on functioning and of global efficiency influence on Amyloid-β within the *APOE4* carrier FCI model is not reflected in the *APOE4* non-carrier FCI model.

Network analysis allows for the computation of many brain regions of interest (in this case, nodes) as a pattern. Although the default mode network has been associated with decreased resting state activity and increased hypometabolism in AD, it is possible that our measure of global efficiency is not a sensitive measure to detect a higher overall capacity for integrative processing. It is also possible that the measure would be better suited in the evaluation of whole-brain functional networks, or by using a different neuroimaging basis for network measurement (e.g., diffusion tensor imaging (DTI), functional magnetic resonance imaging (fMRI)). Notably, as our measure does appear sensitive to other variances in AD pathology (e.g., Amyloid-β, *APOE4* carriership, neurodegeneration) it is possible that global efficiency is not well-suited to capture effects of CR. We maintain that the evaluation of CR as a dynamic system ought to advance the understanding of the active resilience mechanism and add to the groundwork for innovative translational approaches to prompt and evaluate techniques for clinical intervention in AD. Therefore, we encourage the undertaking of future studies to evaluate the biological mechanism representing the underpinnings of CR contributions to AD pathologies.

"'Exceptional Aging' as well as protection against AD dementia will come from 'net sum' protection against all the components of the AD biomarker cascade."[92]

APOE4 carriers as compared to non-carriers do appear to demonstrate differences in recruitment of the default mode network as measured by global efficiency. Given that a network can be controlled in multiple ways by varying types of nodes serving in multiple positions within the neural system, the input of *APOE4* status may serve different roles and affect the manner in which the system runs. We worked to understand how *APOE4* affects AD pathologies related to functioning in order to encourage, provide insight, and propel the future direction of study on the effect of *APOE4* on the neural basis of cognitive reserve, and to propose areas of intervention. Our study implicates an interactive relationship between global efficiency and Amyloid-β, which is objectively modeled in the FCI graph. The algorithm employed by FCI functions through the use of a constraint-based algorithm to determine whether the relationship between two variables is causal, due to a latent variable, or undetermined. The FCI graph displays the causal structure of variables given the presence and absence of Amyloid-β. *APOE4* carriership contributes to variability in Amyloid-β, and Amyloid-β contributes to effects on global efficiency. Further, the presence of Amyloid-β directly affects functioning and is also directly impacted by age. In the absence of Amyloid-β, *APOE4* carriership contributes to global efficiency; the FCI graph places *APOE4* as a contributor to global efficiency. Notably, in multiple regression models within this work, *APOE4* carriers as compared to non-carriers demonstrate differences in recruitment of the default mode network as measured by global efficiency.

Interactions of multiple contributing biomarkers were identified within this work, reflecting the confirmed multifaceted nature of the disease. The Amyloid-β-global efficiency interactions surfaced as a common theme throughout this dissertation. The dynamics of Amyloid-β appear to be influenced by the following factors, which in some cases appear to be *APOE4* carriership specific: a) in a large group sample, Amyloid-β and global efficiency interact to affect functioning (as demonstrated in regression models and as displayed in FCI graphs), however, upon stratification, in tests of regression and in FCI, the contributions of Amyloid-β to global functioning are only sustained within the *APOE4* carriers group; b) Amyloid-β is influenced by age, which remains constant throughout all models; c) Amyloid-β is influenced by *APOE4* carriership; and d) Amyloid-β directly contributes to functioning. As previously mentioned, it is possible that upon initiation of The Toxic Cycle (Figure 2.3), *APOE4* carriers lose “reserves” built by education at a faster rate or are unable to rebuild effectively. The carriership of a single *APOE4* allele in healthy younger and older adults is associated with changes in neural activation, throughout the working memory encoding phase. Investigators of failing compensatory mechanism,[53] concluded that *APOE4* may be associated with “early” compensatory mechanisms compared to non-carriers and that these compensatory mechanisms may fail comparatively earlier in older *APOE4* carriers than in non-carriers. The measure of global efficiency built in this work using structural covariance matrices could be further evaluated as a potential biomarker for predicting AD in *APOE4* carriers.

5.3 Limitations

The main limitation of this demonstrated work is that it showcases constrained disease variables, a single network, and gene. To fully understand the biological underpinnings, the models ought to be applied to *multiple* targeted disease variables, and *additional* networks and genes. The pipeline is constructed to accommodate the addition of domain knowledge and was built to allow the input of different variables, networks, and genes. Therefore, more work can be done, and more questions answered. Specifically, the pipeline’s graph network model enhanced by genomic data should be vetted and considered for application to clinical settings in the early identification of neurodegeneration. The algorithm’s network measures may allow for the automated identification of relative and gradual regional decreases that are invisible to the naked eye. Such application would contribute to the exploration of artificial intelligence in radiology. Pertaining to the statistical models, multiple regression models are built to understand the association between multiple regressor variables on a single outcome variable, however regression does not provide insight into the cause of the predictor or outcome variable, nor does it provide information pertaining to the directionality of the relationship. To overcome this and to understand the causal relationship between variables we sought to complement our regression findings by putting our effort into using causal inference. For this we used algorithms that are tailored to take into account latent relationships to discover true causal associations. The fast causal inference algorithm that we applied to our model has the ability to detect bias, and unlike regression models, does not assume that there are no latent confounders. This means the algorithm can determine whether the statistical dependence of a regressor of an outcome variable may be biased by an unmeasured common cause of other regressors (latent confounder), which could be variables that do not actually influence the outcome. We applied causal inference to complement our regression models by inferring directionality and independence of our variables. We also sought to remove conditional independence from our model and to instead summarize our findings in a causative graph.

5.4 Conclusion

This work satisfied the aims of study: (1) the development of a neuroinformatics pipeline for the replicable collection, manipulation, and analysis of data (2) the employment of the neuroinformatics pipeline to evaluate the potential impact of specific allele carriership on what is recognized as a resilience mechanism in the context of AD. It appears that global efficiency, Amyloid-β, and the interaction between *APOE4* carrier status and CR has a significant effect on functioning. The summary of findings are provided in the table below (Table 5.1).

Table 5.1

Summary of Neuroscience Findings

Objective	Finding
1. Do education and *APOE* genotype differentially impact GMD?	Education and *APOE* genotype differentially impact the middle temporal region and the whole brain measure.
2. Does *APOE genotype* impact clinical functioning, and if so, is the effect mediated by global efficiency?	*APOE4* genotype impacts clinical functioning and the effect is mediated by global efficiency.
3. Does *APOE* genotype demonstrate differences in network recruitment (specifically, global efficiency of the default mode network)?	*APOE4* carriers and non-carriers demonstrate differences in Default Mode Network recruitment.

While there is a growing body of evidence describing the heterogeneity of AD, as well as literature detailing failed therapeutic attempts, it is worth considering whether, in addition to the heterogeneity of the disease presentation where *APOE4* carriers become sicker faster, *APOE4* responds differently to therapies or resilience mechanisms that are promoted in the field to equally benefit all carrier types.

In our preliminary work, we found that when we match *APOE4* carriers with non-carriers of the same level of education, sex, age, and Amyloid-β level, the effect of education on the density of gray matter does not appear in *APOE4* carriers. Whether *APOE4* non-carriers are more likely to benefit from education, or whether *APOE4* carriers lose the benefits of education more quickly, remains unresolved. If we consider this problem in the framework of the Amyloid-β cascade, it is possible that upon initiation of the toxic cycle, *APOE4* carriers lose reserve built by education at a faster rate and have reduced capacity to protect and repair neurons after injury. In fact, some investigators have speculated in their results that *APOE4* may be associated with "early" compensatory mechanisms compared to non-carriers, and that these compensatory mechanisms may then fail earlier in older *APOE4* carriers than in non-carriers.

The studies presented in this work drew on the methodological strengths of health informatics, biostatistics, and neuroscience presented in two-folds; the development of a neuroinformatics pipeline, and the employment of the neuroinformatics pipeline to address impactful complexities in neuroscience. The neuroinformatics pipeline was structured as an automated, replicable pathway, and optimized to run via a single, open-source computing tool (R). We used R as an environment for statistical computing and the functionalities of R were leveraged to construct a SQL database, process samples, build functions, and compute our composite measure for neurodegeneration. The primary strength of this work is that the framework can be extended to other networks, including additional disease mechanisms and neuroimaging measures.

Our data set was created using Structured Query Language (SQL) commands, which were annotated and stored as reproducible scripts written in R and subsequently published to the open-source code repository GitHub (https://github.com). Scripts and details of the study sample were included in the manuscript. The work followed appropriate exclusions and manipulations of the data set, testing for linearity, distribution trends, and evaluation of error values. Although there were significant differences in several of the factors comparing *APOE4* carriers and non-carriers, our primary features of analysis were not statistically different. For this work, education was used as a surrogate for cognitive reserve, and there our samples of *APOE4* carriers and non-carriers were not statistically different, nor were they different in sex, age, or measures of whole brain GMD.

This work uniquely contributes to health informatics through the construction of a neuroinformatics pipeline which combines multimodal biomedical data (neuroimaging, genomics, cognition, and clinical), employs database management, automated computing, graph theory, and biostatistics to answer complex clinical questions. This work contributes to science by proposing a method to measure and monitor brain health, providing additional insight into the mechanistic underpinnings of *APOE4* allele carriership underlying AD pathology. In the age of artificial intelligence, the algorithm's network measures may allow for the automated identification and measurement of minute gradual changes in the brain. The whole of this work will continue to be vetted and adapted for clinical application to the early identification of neurodegenerative disease.

Bibliography

1. Knopman DS. Lowering of amyloid-beta by β-secretase inhibitors-some informative failures. *N Engl J Med*. 2019;380(15):1476.

2. Pievani M, de Haan W, Wu T, Seeley WW, Frisoni GB. Functional network disruption in the degenerative dementias. *The Lancet Neurology*. 2011;10(9):829-843.

3. Marcus D, Harwell J, Olsen T, et al. Informatics and data mining tools and strategies for the human connectome project. *Frontiers in neuroinformatics*. 2011;5:4.

4. Marcus DS, Harms MP, Snyder AZ, et al. Human connectome project informatics: Quality control, database services, and data visualization. *Neuroimage*. 2013;80:202-219.

5. Hao X, Yao X, Yan J, et al. Identifying multimodal intermediate phenotypes between genetic risk factors and disease status in alzheimer’s disease. *Neuroinformatics*. 2016;14(4):439-452.

6. Akil H, Martone ME, Van Essen DC. Challenges and opportunities in mining neuroscience data. *Science*. 2011;331(6018):708-712.

7. Medaglia JD, Lynall M, Bassett DS. Cognitive network neuroscience. *J Cogn Neurosci*. 2015;27(8):1471-1491.

8. Toga AW, Crawford KL, Alzheimer's Disease Neuroimaging Initiative. The informatics core of the alzheimer's disease neuroimaging initiative. *Alzheimer's & Dementia*. 2010;6(3):247-256.

9. May A. Experience-dependent structural plasticity in the adult human brain. *Trends Cogn Sci (Regul Ed)*. 2011;15(10):475-482.

10. Mahley RW, Huang Y. Apolipoprotein e sets the stage: Response to injury triggers neuropathology. *Neuron*. 2012;76(5):871-885.

11. Vemuri P, Gunter JL, Senjem ML, et al. Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies. *Neuroimage*. 2008;39(3):1186-1197.

12. G.A. Ascoli, M. Halavi, Neuroinformatics. *Encyclopedia of Neuroscience*. 2009; 477-484. doi.org/10.1016/B978-008045046-9.00872-X.

13. Stern Y. What is cognitive reserve? theory and research application of the reserve concept. *Journal of the International Neuropsychological Society*. 2002;8(3):448-460.

14. Katzman R, Aronson M, Fuld P, et al. Development of dementing illnesses in an 80-year-old volunteer cohort. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*. 1989;25(4):317-324.

15. Alladi S, Bak TH, Mekala S, et al. Impact of bilingualism on cognitive outcome after stroke. *Stroke*. 2016;47(1):258-261.

16. Mortimer JA, Snowdon DA, Markesbery WR. Head circumference, education and risk of dementia: Findings from the nun study. *Journal of clinical and experimental neuropsychology*. 2003;25(5):671-679.

17. Vemuri P, Weigand SD, Przybelski SA, et al. Cognitive reserve and alzheimer's disease biomarkers are independent determinants of cognition. *Brain*. 2011;134(5):1479-1492.

18. Vemuri P, Lesnick TG, Przybelski SA, et al. Association of lifetime intellectual enrichment with cognitive decline in the older population. *JAMA neurology*. 2014;71(8):1017-1024.

19. Vemuri P, Lesnick TG, Przybelski SA, et al. Effect of lifestyle activities on alzheimer disease biomarkers and cognition. *Ann Neurol*. 2012;72(5):730-738.

20. Vemuri P, Lesnick TG, Przybelski SA, et al. Effect of intellectual enrichment on AD biomarker trajectories: Longitudinal imaging study. *Neurology*. 2016;86(12):1128-1135.

21. Vemuri P, Fields J, Peter J, Klöppel S. Cognitive interventions in alzheimer's and parkinson's diseases: Emerging mechanisms and role of imaging. *Curr Opin Neurol*. 2016;29(4):405.

22. Tsai MS, Tangalos EG, Petersen RC, et al. Apolipoprotein E: Risk factor for alzheimer disease. *Am J Hum Genet*. 1994;54(4):643.

23. Mahley RW, Huang Y. Apolipoprotein e sets the stage: Response to injury triggers neuropathology. *Neuron*. 2012;76(5):871-885.

24. Cohen S, Janicki-Deverts D, Doyle WJ, et al. Chronic stress, glucocorticoid receptor resistance, inflammation, and disease risk. *Proceedings of the National Academy of Sciences*. 2012;109(16):5995-5999.

25. Sheffler J, Moxley J, Sachs-Ericsson N. Stress, race, and APOE: Understanding the interplay of risk factors for changes in cognitive functioning. *Aging & mental health*. 2014;18(6):784-791.

26. McEwen BS, Gray JD, Nasca C. Recognizing resilience: Learning from the effects of stress on the brain. *Neurobiology of stress*. 2015;1:1-11.

27. McEwen BS, Bowles NP, Gray JD, et al. Mechanisms of stress in the brain. *Nat Neurosci*. 2015;18(10):1353.

28. Lee BK, Glass TA, James BD, Bandeen-Roche K, Schwartz BS. Neighborhood psychosocial environment, apolipoprotein E genotype, and cognitive function in older adults. *Arch Gen Psychiatry*. 2011;68(3):314-321.

29. Kimbrel N, Hauser M, Garrett M, et al. Effect of the APOE epsilon 4 allele and combat exposure on PTSD and psychiatric comorbidity. 2014;44(6):664-665.

30. Lee BK, Glass TA, James BD, Bandeen-Roche K, Schwartz BS. Neighborhood psychosocial environment, apolipoprotein E genotype, and cognitive function in older adults. *Arch Gen Psychiatry*. 2011;68(3):314-321.

31. Vyas S, Rodrigues AJ, Silva JM, et al. Chronic stress and glucocorticoids: From neuronal plasticity to neurodegeneration. *Neural Plast*. 2016;2016.

32. Sterling P, Eyer J. Allostasis: A new paradigm to explain arousal pathology. In: *Handbook of life stress, cognition and health.* Oxford, England: John Wiley & Sons; 1988:629-649.

33. McEwen BS. Physiology and neurobiology of stress and adaptation: Central role of the brain. *Physiol Rev*. 2007;87(3):873-904. doi: 10.1152/physrev.00041.2006

34. Peavy GM, Lange KL, Salmon DP, et al. The effects of prolonged stress and APOE genotype on memory and cortisol in older adults. *Biol Psychiatry*. 2007;62(5):472-478. doi: 10.1016/j.biopsych.2007.03.013

35. Boyce WT, Ellis BJ. Biological sensitivity to context: I. an evolutionary-developmental theory of the origins and functions of stress reactivity. *Dev Psychopathol*. 2005;17(2):271-301.

36. McEwen BS, Magarinos AM. Stress and hippocampal plasticity: Implications for the pathophysiology of affective disorders. *Hum Psychopharmacol*. 2001;16(S1):S19. doi: 10.1002/hup.266

37. McEwen BS. Stress and the aging hippocampus. *Front Neuroendocrinol*. 1999;20(1):49-70. doi: 10.1006/frne.1998.0173

38. Cairney J, Krause N. Negative life events and age-related decline in mastery: Are older adults more vulnerable to the control-eroding effect of stress? *J Gerontol B Psychol Sci Soc Sci*. 2008;63(3):162. doi: 10.1093/geronb/63.3.s162

39. Vyas S, Rodrigues AJ, Silva JM, et al. Chronic stress and glucocorticoids: From neuronal plasticity to neurodegeneration. *Neural Plast*. 2016;2016:6391686. doi: 10.1155/2016/6391686

40. Lim YY, Hassenstab J, Goate A, et al. Effect of BDNFVal66Met on disease markers in dominantly inherited alzheimer's disease. *Ann Neurol*. 2018;84(3):424-435. doi: 10.1002/ana.25299

41. Lim YY, Villemagne VL, Laws SM, et al. APOE and BDNF polymorphisms moderate amyloid β-related cognitive decline in preclinical alzheimer's disease. *Mol Psychiatry*. 2015;20(11):1322-1328. doi: 10.1038/mp.2014.123

42. Sheffler J, Moxley J, Sachs-Ericsson N. Stress, race, and APOE: Understanding the interplay of risk factors for changes in cognitive functioning. *Aging Ment Health*. 2014;18(6):784-791. doi: 10.1080/13607863.2014.880403

43. Sachs-Ericsson N, Corsentino E, Collins N, Sawyer K, Blazer DG. Problems meeting basic needs moderate the association between the APOE epsilon4 allele and cognitive decline. *Aging Ment Health*. 2010;14(2):138-144. doi: 10.1080/13607860903421060

44. Glahn DC, Kent JW, Sprooten E, et al. Genetic basis of neurocognitive decline and reduced white-matter integrity in normal human brain aging. *Proc Natl Acad Sci U S A*. 2013;110(47):19006-19011. doi: 10.1073/pnas.1313735110

45. Mahley RW. Apolipoprotein E: From cardiovascular disease to neurodegenerative disorders. *J Mol Med*. 2016;94(7):739-746. doi: 10.1007/s00109-016-1427-y

46. Misonou H, Morishima-Kawashima M, Ihara Y. Oxidative stress induces intracellular accumulation of amyloid beta-protein (abeta) in human neuroblastoma cells. *Biochemistry*. 2000;39(23):6951-6959.

47. Paola D, Domenicotti C, Nitti M, et al. Oxidative stress induces increase in intracellular amyloid beta-protein production and selective activation of betaI and betaII PKCs in NT2 cells. *Biochem Biophys Res Commun*. 2000;268(2):642-646. doi: 10.1006/bbrc.2000.2164

48. Tamagno E, Bardini P, Obbili A, et al. Oxidative stress increases expression and activity of BACE in NT2 neurons. *Neurobiol Dis*. 2002;10(3):279-288. 49. Bien-Ly N, Andrews-Zwilling Y, Xu Q, Bernardo A, Wang C, Huang Y. C-terminal-truncated apolipoprotein (apo) E4 inefficiently clears amyloid-beta (abeta) and acts in concert with abeta to elicit neuronal and behavioral deficits in mice. *Proc Natl Acad Sci U S A*. 2011;108(10):4236-4241. doi: 10.1073/pnas.1018381108

50. Castellano JM, Kim J, Stewart FR, et al. Human apoE isoforms differentially regulate brain amyloid-β peptide clearance. *Sci Transl Med*. 2011;3(89):89ra57. doi: 10.1126/scitranslmed.3002156

51. Kim J, Jiang H, Park S, et al. Haploinsufficiency of human APOE reduces amyloid deposition in a mouse model of amyloid-β amyloidosis. *J Neurosci*. 2011;31(49):18007-18012. doi: 10.1523/JNEUROSCI.3773-11.2011

52. Chen H, Ji Z, Dodson SE, et al. Apolipoprotein E4 domain interaction mediates detrimental effects on mitochondria and is a potential therapeutic target for alzheimer disease. *J Biol Chem*. 2011;286(7):5215-5221. doi: 10.1074/jbc.M110.151084

53. Filbey FM, Chen G, Sunderland T, Cohen RM. Failing compensatory mechanisms during working memory in older apolipoprotein E-epsilon4 healthy adults. *Brain Imaging Behav*. 2010;4(2):177-188. doi: 10.1007/s11682-010-9097-9

54. Overgaard, S., Przybelski, S., Machulda, M., Mielke, M., Knopman, D., & Lowe, V. et al. (2015). O2-09-03: Evidence of reduced neuronal plasticity to education in ApoE-ε4 carriers. *Alzheimer's & Dementia*, *11*(7S_Part_4), P195-P195. doi: 10.1016/j.jalz.2015.07.179

55. Kimbrel NA, Hauser MA, Garrett M, et al. Effect of the apoe e4 allele and combat exposure on ptsd among iraq/afghanistan-era veterans. *Depress Anxiety*. 2015;32(5):307-315. doi: 10.1002/da.22348

56. Aschbacher K, O'Donovan A, Wolkowitz OM, Dhabhar FS, Su Y, Epel E. Good stress, bad stress and oxidative stress: Insights from anticipatory cortisol reactivity. *Psychoneuroendocrinology*. 2013;38(9):1698-1708. doi: 10.1016/j.psyneuen.2013.02.004

57. Theendakara V, Patent A, Peters Libeu CA, et al. Neuroprotective sirtuin ratio reversed by ApoE4. *Proc Natl Acad Sci U S A*. 2013;110(45):18303-18308. doi: 10.1073/pnas.1314145110

58. Shinohara M, Sato N. The Roles of Apolipoprotein E, Lipids, and Glucose in the Pathogenesis of Alzheimer's Disease. Adv Exp Med Biol. 2019;1128:85-101. doi: 10.1007/978-981-13-3540-2_5

59. Hua, X., Ching, C., Mezher, A., Gutman, B. A., Hibar, D. P., Bhatt, P., Leow, A. D., Jack, C. R., Jr, Bernstein, M. A., Weiner, M. W., Thompson, P. M., & Alzheimer's Disease Neuroimaging Initiative (2016). MRI-based brain atrophy rates in ADNI phase 2: acceleration and enrichment considerations for clinical trials. *Neurobiology of aging*. 2016 Jan;37:26-37. doi: 10.1016/j.neurobiolaging.2015.09.018

60. Liu CC, Liu CC, Kanekiyo T, Xu H, Bu G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat Rev Neurol. 2013 Feb;9(2):106-18. doi: 10.1038/nrneurol.2012.263 .

61. Brown BM, Peiffer JJ, Martins RN. Multiple effects of physical activity on molecular and cognitive signs of brain aging: Can exercise slow neurodegeneration and delay alzheimer's disease? *Mol Psychiatry*. 2013;18(8):864-874. doi: 10.1038/mp.2012.162

62. Sato N, Morishita R. The roles of lipid and glucose metabolism in modulation of β-amyloid, tau, and neurodegeneration in the pathogenesis of alzheimer disease. *Front Aging Neurosci*. 2015;7:199. doi: 10.3389/fnagi.2015.00199

63. Soto I, Graham LC, Richter HJ, et al. APOE stabilization by exercise prevents aging neurovascular dysfunction and complement induction. *PLoS Biol*. 2015;13(10):e1002279. doi: 10.1371/journal.pbio.1002279

64. Grefkes C, Fink GR. Reorganization of cerebral networks after stroke: New insights from neuroimaging with connectivity approaches. *Brain*. 2011;134(Pt 5):1264-1276. doi: 10.1093/brain/awr033

65. Grefkes C, Fink GR. Reorganization of cerebral networks after stroke: New insights from neuroimaging with connectivity approaches. *Brain*. 2011;134(Pt 5):1264-1276. doi: 10.1093/brain/awr033

66. van Praag H, Kempermann G, Gage FH. Running increases cell proliferation and neurogenesis in the adult mouse dentate gyrus. *Nat Neurosci*. 1999;2(3):266-270. doi: 10.1038/6368

67. Brown J, Cooper-Kuhn CM, Kempermann G, et al. Enriched environment and physical activity stimulate hippocampal but not olfactory bulb neurogenesis. *Eur J Neurosci*. 2003;17(10):2042-2046. doi: 10.1046/j.1460-9568.2003.02647.x

68. van Praag H, Kempermann G, Gage FH. Neural consequences of environmental enrichment. *Nat Rev Neurosci*. 2000;1(3):191-198. doi: 10.1038/35044558

69. Vemuri P, Lesnick TG, Przybelski SA, et al. Association of lifetime intellectual enrichment with cognitive decline in the older population. *JAMA Neurol*. 2014;71(8):1017-1024. doi: 10.1001/jamaneurol.2014.96370. Alwis DS, Rajan R. Environmental enrichment and the sensory brain: The role of enrichment in remediating brain injury. *Front Syst Neurosci*. 2014;8:156. doi: 10.3389/fnsys.2014.00156

71. Johansen-Berg H, Scholz J, Stagg CJ. Relevance of structural brain connectivity to learning and recovery from stroke. *Frontiers in systems neuroscience*. 2010;4:146.doi: 10.3389/fnsys.2010.00146

72. Pfeffer RI, Kurosaki TT, Harrah Jr CH, Chance JM, Filos S. Measurement of functional activities in older adults in the community. *J Gerontol*. 1982;37(3):323-329. doi: 10.1093/geronj/37.3.323. PMID: 7069156

73. Ritter K, Schumacher J, Weygandt M, Buchert R, Allefeld C, Haynes JD. Multimodal prediction of conversion to Alzheimer's disease based on incomplete biomarkers. Alzheimers Dement (Amst). 2015 Apr 30;1(2):206-15. doi: 10.1016/j.dadm.2015.01.006.

74. Weiner MW, Veitch DP, Aisen PS, et al. The alzheimer's disease neuroimaging initiative: A review of papers published since its inception. *Alzheimers Dement*. 2012;8(1 Suppl):1. doi: 10.1016/j.jalz.2011.09.172

75. Stage E, Duran T, Risacher SL, Goukasian N, Do TM, West JD, Wilhalme H, Nho K, Phillips M, Elashoff D, Saykin AJ, Apostolova LG. The effect of the top 20 Alzheimer disease risk genes on gray-matter density and FDG PET brain metabolism. Alzheimers Dement (Amst). 2016 Dec 19;5:53-66. doi: 10.1016/j.dadm.2016.12.003

76. Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, Mass.: MIT Press.

77. Ogarrio JM, Spirtes P, Ramsey J. A Hybrid Causal Search Algorithm for Latent Variable Models. JMLR Workshop Conf Proc. 2016 Aug;52:368-379.

78. Petersen RC, Aisen PS, Beckett LA, et al. Alzheimer's disease neuroimaging initiative (ADNI): Clinical characterization. *Neurology*. 2010;74(3):201-209. doi: 10.1212/WNL.0b013e3181cb3e25.

79. McKhann GM, Knopman DS, Chertkow H, et al. The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Dementia*. 2011;7(3):263. doi: 10.1016/j.jalz.2011.03.005.

80. Fischl B, Dale AM. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc Natl Acad Sci U S A*. 2000;97(20):11050-11055.. doi: 10.1073/pnas.200033797.

81. Kabbara A, El Falou W, Khalil M, Wendling F, Hassan M. The dynamic functional core network of the human brain at rest. *Sci Rep*. 2017;7(1):2936. doi: 10.1038/s41598-017-03420-6.

82. Desikan RS, Ségonne F, Fischl B, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*. 2006;31(3):968-980. doi: 10.1016/j.neuroimage.2006.01.021

83. Kong X, Liu Z, Huang L, et al. Mapping individual brain networks using statistical similarity in regional morphology from MRI. *PLoS ONE*. 2015;10(11):e0141840. doi: 10.1371/journal.pone.0141840.

84. Kim H, Shin J, Han CE, et al. Using individualized brain network for analyzing structural covariance of the cerebral cortex in alzheimer's patients. *Front Neurosci*. 2016;10:394. doi: 10.3389/fnins.2016.0039

85. Rubinov M, Sporns O. Complex network measures of brain connectivity: Uses and interpretations. *Neuroimage*. 2010;52(3):1059-1069. doi: 10.1016/j.neuroimage.2009.10.003

86. Volpe, G., & Volpe, G. (2020). BRAPH - Brain Analysis using Graph Theory. Retrieved 5 November 2020, from http://braph.org/

87. brainGraph: Graph theory analysis of brain MRI data version 2.2.0 from CRAN. Retrieved 5 November 2020, from https://rdrr.io/cran/brainGraph/

88. Latora V, Marchiori M. Efficient behavior of small-world networks. *Phys Rev Lett*. 2001;87(19):198701. doi: 10.1103/PhysRevLett.87.198701

89. Landau SM, Horng A, Jagust WJ. Memory decline accompanies subthreshold amyloid accumulation. *Neurology*. 2018;90(17):e1460. doi: 10.1212/WNL.0000000000005354

90. Kim YJ, Seo SW, Park SB, et al. Protective effects of APOE e2 against disease progression in subcortical vascular mild cognitive impairment patients: A three-year longitudinal study. *Sci Rep*. 2017;7(1):1910. doi: 10.1038/s41598-017-02046-y

91. Wu L, Zhao L. ApoE2 and Alzheimer's disease: Time to take a closer look. *Neural Regen Res*. 2016;11(3):412-413. doi: 10.4103/1673-5374.179044

92. Vemuri, P. “Exceptional brain aging” without Alzheimer’s disease: triggers, accelerators, and the net sum game. *Alz Res Therapy* 10, 53 (2018). doi: 10.1186/s13195-018-0373-z

93. Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik, 1*, 269-271. doi: 10.1007/BF01386390

94. Ewers M, Insel PS, Stern Y, Weiner MW; Alzheimer's Disease Neuroimaging Initiative (ADNI). Cognitive reserve associated with FDG-PET in preclinical Alzheimer disease. *Neurology*. 2013;80(13):1194-201. doi: 10.1212/WNL.0b013e31828970c2

95. Ow, S. Y., & Dunstan, D. E. (2014). A brief overview of amyloids and Alzheimer's disease. *Protein science : a publication of the Protein Society*, *23*(10), 1315–1331.

96. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Appendix: R Code

A.1. Overview

This appendix serves as a guide to the code developed for the Neuroinformatics Pipeline and its Application to Gene-environment Interaction in Neurodegenerative Disease. In 2017, the emerging neuroinformatics compact to produce open source and replicable methods prompted the initial development and documentation of this work which was achieved through: (1) Intentional limitation of the number of platforms used (2) employment of open-source computing software (R-Project[96]) (3) brevity in scripting yet thorough documentation of code (4) community storage allowing for open use of, and comment on, the application and supporting materials (5) proof of concept established using publicly available data.

The data used in preparation of this manuscript were approved by and obtained from ADNI. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is

coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The R code that has been created for and used in this original work is presented in the following sections. This R code represents four main sections (2-6) as described in Figure 3.1 of the thesis and also printed below. In each of the following sections, the purpose of the pipeline section is described, and the R code is provided.

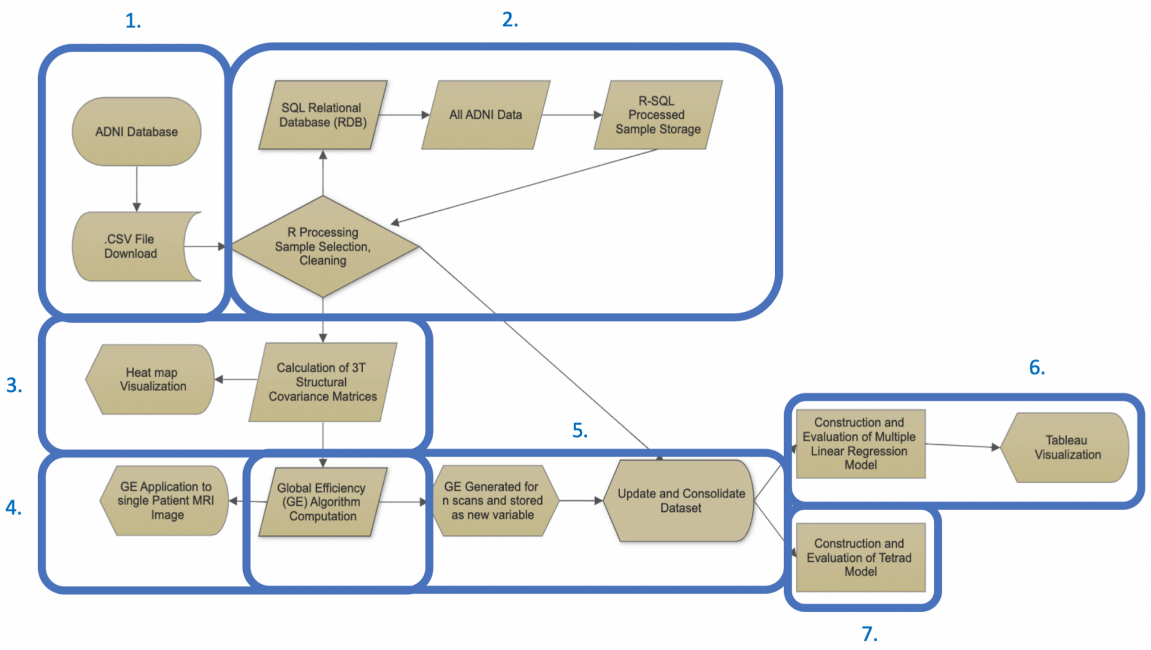
A.1. ADNI Database R-SQL Upload (Figure 3.1 Section 2.)

A.2. ADNI Data Processing (Figure 3.1 Section 3.)

A.3. Local and Global Efficiency Calculation (Figure 3.1 Section 4.)

A.4. Adding newly generated measure to database (Figure 3.1 Section 5.)

Since this analysis was performed, Tetrad and Tableau have begun working on R wrappers to allow for seamless integration. Thus eventually, Steps F and G of this pipeline ought to be included as R code. The code for sections 2-6 of the thesis research "Development of a Neuroinformatics Pipeline and its Application to Gene-environment Interaction in Neurodegenerative Disease" is stored in the repository: https://github.com/shaunaovergaard/neuroinformatics.git


1.
2.
ADNI Database
.CSV File Download
SQL Relational Database (RDB)
All ADNI Data
R-SQL Processed Sample Storage
R Processing Sample Selection, Cleaning
3.
Heat map Visualization
Calculation of 3T Structural Covariance Matrices
4.
GE Application to single Patient MRI Image
Global Efficiency (GE) Algorithm Computation
5.
GE Generated for n scans and stored as new variable
Update and Consolidate Dataset
6.
Construction and Evaluation of Multiple Linear Regression Model
Tableau Visualization
Construction and Evaluation of Tetrad Model
7.

Neuroinformatics pipeline in sectioned format (Image refers to Figure 3.1 of the thesis). The visualization follows the construction of a neuroinformatics pipeline and the consideration of tools and analysis techniques that are both generalizable and replicable. The method is described through a sectioned process map: (1) Data acquisition (2) Sample processing and storage (3) Computation and visualization of brain structural covariance (4) New variable generation through the statistical computation of graph theoretical metrics (5) Merging of datasets (6) Variable manipulation and construction of regression model (7) Objective validation of the model performed using the FCI search algorithm. The Tetrad software output provided visualizations of graphs. Additionally, Tableau was used to create a visualization of interaction findings. Processes corresponding to sections 2-6 are automated (using a single R script, provided in the appendix).

A.2. ADNI Database R-SQL Upload

The following section of code refers to Section 2 of the pipeline (Figure 3.1 of the thesis):

Section 2. Sample Processing and Storage	● Disparate datasets are prepared and loaded into a relational database for sample construction. ● Local Structured Query Language (SQL) phpMyAdmin Database is created. ● R is used to load and prepare the data, which are subsequently uploaded to the SQL database.

```
       ##  The following script installs RMySQL, which is a Database Interface
(DBI) and 'MySQL' Driver for R: https://cran.r-
project.org/web/packages/RMySQL/index.html
       ##  Notes: Password and user may need to be configured -
/Library/WebServer/Documents/phpmyadmin/config.inc.php

       ##  Diagnosis File contains: ADSXLIST.csv       DXSUM_PDXCONV_ADNIALL.csv
BLCHANGE.csv

       ##  Neuropsychological File contains: ADASSCORES.csv
       GDSCALE.csv
       ##  ADAS_ADNI1.csv                      ITEM.csv  ADAS_ADNIGO23.csv
              ITEM_DICT.csv  ADNI_CBBRESULTS.csv                   MMSE.csv
       ##  **ADNI_Methods_UWNPSYCHSUM_20160112.pdf**  MOCA.csv    CCI.csv
                     MODHACH.csv  CDR.csv             NEUROBAT.csv   ECOGPT.csv
                     NPI.csv
       ##  ECOGSP.csv                          NPIQ.csv    FAQ.csv
              UWNPSYCHSUM_02_22_17.csv     FCI.csv
       UWNPSYCHSUM_DICT_02_23_17.csv

       ##  Biospecimen File contains: APOERES.csv           APOERES_DICT.csv

       ##  MR_Image_Analysis File contains: UCSFFSX51_08_01_16.csv
       UCSFFSX51_11_02_15_V2.csv  MRINCLUSIO.csv

       install.packages('RMySQL') ##Installs RMySQL package
       require(RMySQL) ##Ensures that commands used reference the RMySQL package


       library(RMySQL)

       mydb = dbConnect(MySQL(), user='root',
password='yourpasswordusedatSQLinstall', host='localhost')
       dbSendQuery(mydb, "CREATE DATABASE Diagnosis")
       dbSendQuery(mydb, "USE Diagnosis")

       conn<-dbConnect(RMySQL::MySQL(), host="localhost", dbname="Diagnosis",
user="root", password="yourpasswordusedatSQLinstall") ##Establishes connection
to local instance of MySQL DB
       setwd("pathtoDiagnosistablesfolder")
       ##Build "Diagnosis" tables
       df<-read.csv("BLCHANGE.csv")
       dbWriteTable(conn,"BLCHANGE", df)
       df<-read.csv("ADSXLIST.csv")
       dbWriteTable(conn,"ADXLIST", df)
       df<-read.csv("DXSUM_PDXCONV_ADNIALL.csv")
       dbWriteTable(conn,"DXSUM_PDXCONV_ADNIALL", df)


       ##Build "Neuropsychological" tables in DB
```

```
        ##create "Neuropsychological" database using RMySQL in R
        mydb = dbConnect(MySQL(), user='root',
password='yourpasswordusedatSQLinstall', host='localhost')
        dbSendQuery(mydb, "CREATE DATABASE Neuropsychological")
        dbSendQuery(mydb, "USE Neuropsychological")

        conn<-dbConnect(RMySQL::MySQL(), host="localhost",
dbname="Neuropsychological", user="root",
password="yourpasswordusedatSQLinstall") ##Establishes connection to local
instance of MySQL DB
        setwd("pathtoNeuropsychologicaltablesfolder")

        ## create "ADASSCORES" table
        df<-read.csv("ADASSCORES.csv")
        dbWriteTable(conn,"ADASSCORES", df)

        ## create "GDSCALE" table
        df<-read.csv("GDSCALE.csv")
        dbWriteTable(conn,"GDSCALE", df)

        ## create "ADAS_ADNI1" table
        df<-read.csv("ADAS_ADNI1.csv")
        dbWriteTable(conn,"ADAS_ADNI1", df)

        ## create "ADAS_ADNIGO23" table
        df<-read.csv("ADAS_ADNIGO23.csv")
        dbWriteTable(conn,"ADAS_ADNIGO23", df)

        ## create "ITEM_DICT" table
        df<-read.csv("ITEM_DICT.csv")
        dbWriteTable(conn,"ITEM_DICT", df)

        ## create "ADNI_CBBRESULTS" table
        df<-read.csv("ADNI_CBBRESULTS.csv")
        dbWriteTable(conn,"ADNI_CBBRESULTS", df)

        ## create "MMSE" table
        df<-read.csv("MMSE.csv")
        dbWriteTable(conn,"MMSE", df)

        ## create "MOCA" table
        df<-read.csv("MOCA.csv")
        dbWriteTable(conn,"MOCA", df)

        ## create "CCI" table
        df<-read.csv("CCI.csv")
        dbWriteTable(conn,"CCI", df)

        ## create "MODHACH" table
        df<-read.csv("MODHACH.csv")
        dbWriteTable(conn,"MODHACH", df)

        ## create "CDR" table
        df<-read.csv("CDR.csv")
        dbWriteTable(conn,"CDR", df)

        ## create "NEUROBAT" table
        df<-read.csv("NEUROBAT.csv")
        dbWriteTable(conn,"NEUROBAT", df)

        ## create "ECOGPT" table
        df<-read.csv("ECOGPT.csv")
        dbWriteTable(conn,"ECOGPT", df)
```

```
	## create "NPI" table
	df<-read.csv("NPI.csv")
	dbWriteTable(conn,"NPI", df)

	## create "ECOGSP" table
	df<-read.csv("ECOGSP.csv")
	dbWriteTable(conn,"ECOGSP", df)

	## create "ECOGSP" table
	df<-read.csv("ECOGSP.csv")
	dbWriteTable(conn,"ECOGSP", df)

	## create "NPIQ" table
	df<-read.csv("NPIQ.csv")
	dbWriteTable(conn,"NPIQ", df)

	## create "FAQ" table
	df<-read.csv("FAQ.csv")
	dbWriteTable(conn,"FAQ", df)

	## create "UWNPSYCHSUM_02_22_17" table
	df<-read.csv("UWNPSYCHSUM_02_22_17.csv")
	dbWriteTable(conn,"UWNPSYCHSUM_02_22_17", df)

	## create "FCI" table
	df<-read.csv("FCI.csv")
	dbWriteTable(conn,"FCI", df)

	## create "UWNPSYCHSUM_DICT_02_23_17" table
	df<-read.csv("UWNPSYCHSUM_DICT_02_23_17.csv")
	dbWriteTable(conn,"UWNPSYCHSUM_DICT_02_23_17", df)


	###########################

	##Build "Biospecimen_Results" tables in DB
	##create "Biospecimen_Results" database using RMySQL in R
	mydb = dbConnect(MySQL(), user='root', 
password='yourpasswordusedatSQLinstall', host='localhost')
	dbSendQuery(mydb, "CREATE DATABASE Biospecimen_Results")
	dbSendQuery(mydb, "USE Biospecimen_Results")
	conn<-dbConnect(RMySQL::MySQL(), host="localhost", 
dbname="Biospecimen_Results", user="root", 
password="yourpasswordusedatSQLinstall") ##Establishes connection to local 
instance of MySQL DB

	## change working directory
	setwd("pathtoBiospecimen_Resultsfolder")

	## create "APOERES" table
	df<-read.csv("APOERES.csv")
	dbWriteTable(conn,"APOERES", df)

	## create "APOERES_DICT" table
	df<-read.csv("APOERES_DICT.csv")
	dbWriteTable(conn,"APOERES_DICT", df)

	#########################################

	##Build "MR_Image_Analysis" database using RMySQL in R
	mydb = dbConnect(MySQL(), user='root', 
password='yourpasswordusedatSQLinstall', host='localhost')
```

```
       dbSendQuery(mydb, "CREATE DATABASE MR_Image_Analysis")
       dbSendQuery(mydb, "USE MR_Image_Analysis")
       conn<-dbConnect(RMySQL::MySQL(), host="localhost",
dbname="MR_Image_Analysis", user="root",
password="yourpasswordusedatSQLinstall") ##Establishes connection to local
instance of MySQL DB

       ## change working directory
       setwd("pathtoMR_Image_Analysisfolder")

       ## create "UCSFFSX_11_02_15" table
       df<-read.csv("UCSFFSX51_08_01_16.csv")
       dbWriteTable(conn,"UCSFFSX51_08_01_16", df)

       ## create "UCSFFSX_DICT_08_01_14" table
       df<-read.csv("UCSFFSX51_DICT_08_01_14.csv")
       dbWriteTable(conn,"UCSFFSX51_DICT_08_01_14", df)

       ## create "MRINCLUSIO" table
       df<-read.csv("MRINCLUSIO.csv")
       dbWriteTable(conn,"MRINCLUSIO", df)

       ##  create "MAYOADIRL_MRI_FMRI_11_07_17" table
       df<-read.csv("MAYOADIRL_MRI_FMRI_11_07_17.csv")
       dbWriteTable(conn, "MAYOADIRL_MRI_FMRI_11_07_17", df)

       ##  create "UCSFFSL51ALL_08_01_16" table
       df<-read.csv("UCSFFSL51ALL_08_01_16.csv")
       dbWriteTable(conn, "UCSFFSL51ALL_08_01_16", df)

       ##  create "UCSFFSX_11_02_15" table
       df<-read.csv("UCSFFSX_11_02_15.csv")
       dbWriteTable(conn, "UCSFFSX_11_02_15", df)

       ##  create "UCSFFSX51_08_01_16" table
       df<-read.csv("UCSFFSX51_ADNI1_3T_02_01_16.csv")
       dbWriteTable(conn, "UCSFFSX51_ADNI1_3T_02_01_16", df)


       #####################################

       ##Build "Subject_Characteristics" database using RMySQL in R
       mydb = dbConnect(MySQL(), user='root',
password='yourpasswordatsqlinstall', host='localhost')
       dbSendQuery(mydb, "CREATE DATABASE Subject_Characteristics")
       dbSendQuery(mydb, "USE Subject_Characteristics")
       conn<-dbConnect(RMySQL::MySQL(), host="localhost",
dbname="Subject_Characteristics", user="root",
password="yourpasswordatsqlinstall") ##Establishes connection to local instance
of MySQL DB

       ## change working directory
       setwd("pathtoSubject_Characteristicsfolder")

       ## create "PTDEMOG" table
       df<-read.csv("PTDEMOG.csv")
       dbWriteTable(conn,"PTDEMOG", df)

       ######################################
       ##Build "Detached" database using RMySQL in R
       dbSendQuery(mydb, "CREATE DATABASE Detached")
       dbSendQuery(mydb, "USE Detached")
       conn<-dbConnect(RMySQL::MySQL(), host="localhost", dbname="Detached",
user="root", password="yourpasswordatsqlinstall") ##Establishes connection to
local instance of MySQL DB

       ## change working directory
       setwd("yourpathtoADNI_dirfolder")
       ##  create "adnimerge" table
       df<-read.csv("adnimerge.csv")
       dbWriteTable(conn,"adnimerge", df)

       ## change working directory
       setwd("yourpathtoADNI_dirfolder")
       ##  create "ADNIMERGE_DICT" table
       df<-read.csv("ADNIMERGE_DICT.csv")
       dbWriteTable(conn,"ADNIMERGE_DICT", df)

       ## change working directory
       setwd("/yourpreferredlocation")
       ##  create "DMN_atlas" table
       df<-read.csv("DMN.csv")
       dbWriteTable(conn,"DMN_atlas", df)
```

A.3. ADNI Data Processing

The following code refers to Section 3 of the pipeline (Figure 3.1 of the thesis):

Section 3. Computation and Visualization of Brain Structural Covariance	● Algorithm is applied to individual 3T MRI data and covariance matrices are produced. ● Heat maps employing a diverging color scheme are generated to describe the intensity of covariance between nodes.

```
        ########
        ##  In the below we are trying to get as many processed scans as possible
        ##  MR_Image_Analysis.UCSFFSX51_08_01_16 is a table that has thickness
data for the DMN ROIs.
        ##  The aim of this code will be to retain as many subjects as possible
from this set, while acquiring APOE E4 data, education, cognition, pathology
variables. TBD
        ##  In order to view all databases and navigate through tables, one may
choose to open up the local instance of the database -
http://localhost/phpmyadmin/index.php
        ##  MySQL Server Instance on the machine must be running to allow for
client connections to DB.
        ##  The specific SQL DB tables that are being used are
MR_Image_Analysis.UCSFFSX51_08_01_16.csv, Detached.adnimerge,
Biospecimen_Results.APOERES
        ##  A view was created, 'datanoacc1047allcols' within the
MR_Image_Analysis DB and is later used as a base from which to draw distinct
RIDs of cases with non-accelerated T1 scans.
        ##  ...within the already processed dataset "UCSFFSX51_08_01_16"

        dbDisconnect(conn) # close connections prior to establishing new
connection
        conn<-dbConnect(RMySQL::MySQL(), host="localhost",
dbname="MR_Image_Analysis", user="root", password="yourpasswordatsqlinstall")
##Establishes connection to local instance of MySQL DB

        dbListTables(conn) # lists tables within the called database.

        ##  Here we are creating a view of distinct RIDs of subjects who have
processed thickness data and a Non-Accelerated T1 scan. This list will be used
as our reference set.
        dbSendQuery(conn,"
        CREATE
        ALGORITHM = UNDEFINED
        VIEW `nonacc1047`
        AS SELECT DISTINCT RID FROM `UCSFFSX51_08_01_16`
        WHERE IMAGETYPE='Non-Accelerated T1'")

        ##  Using subquery as a check, here we create a table of all columns for
the 1047 cases from the table MR_Image_Analysis.UCSFFSX51_08_01_16.
        dbSendQuery(conn,"
        CREATE
        ALGORITHM = UNDEFINED
        VIEW ‘datanoacc1047allcols’
        SELECT * FROM MR_Image_Analysis.UCSFFSX51_08_01_16 as a
        WHERE a.IMAGETYPE = 'Non-Accelerated T1'
        AND a.RID IN
        (SELECT RID FROM MR_Image_Analysis.nonacc1047)")

        ##########################################################################
        ##  The following packages are required:
```

```
    #install.packages(“ggplot2”)
    library(ggplot2)


    ########################################################################

    dbDisconnect(conn) # close connections prior to establishing new
connection
    conn<-dbConnect(RMySQL::MySQL(), host="localhost",
dbname="MR_Image_Analysis", user="root", password="yourpasswordatsqlinstall")
##Establishes connection to local instance of MySQL DB

    dbListTables(conn) # lists tables within the called db.


    ##Run the Select statement (SELECT * FROM
MR_Image_Analysis.datanoacc1047allcols as a, Detached.adnimerge as b,
Biospecimen_Results.APOERES as c
    #WHERE a.RID=b.RID AND a.RID=c.RID) in phpmyadmin, exporting data, and
saving to your preferred folder location as a csv file (removes rownames
column)
    #d3<-dbFetch(rs1, n=500) ##MySQL(max.con = 16, fetch.default.rec = 500)
    #dbHasCompleted(rs1)
    #dim(d3) #checks dimensions of d3


    ###
    d3<-read.csv(file.choose(""))
    colnames(d3)
    dim(d3)
    d4<-d3 #just renaming so that we don't overwrite original file

    d4<-d4[ , -which(names(d4) %in% c("row_names"))] #removing row_names
columns so that the table can be smoothly written in the SQL DB
    ##This is done as it generates rownames of its own and would throw an
error if asked to duplicate a column name)
    dim(d4)


    dbWriteTable(conn,"ADNI_973_all", d4)

    ###
    ##  Create a dataset that is ordered by RID
    d4 <- d4[order(d4$RID),]
    ##  Unlist and create a new column that ranks EXAMDATE chronologially
within each RID group
    d4$Order.by.group <- unlist(with(d4, tapply(EXAMDATE.1, RID, function(x)
rank(x,ties.method= "first"))))

    colnames(d4)
    table(d4$RID,d4$Order.by.group)
    #####
    dbWriteTable(conn, "ADNI_973_ALL_ranked_examdate", d4)
    dbListTables(conn)

    #####
    ##  Select only the cases that = number 1 within the new
d3$Order.by.group - these will be the first/baseline scans.

    d5 <- d4[which(d4$Order.by.group=='1'),]
    dim(d5) #973 x 483
```

```
       dbDisconnect(conn) # close connections prior to establishing new
connection
       conn<-dbConnect(RMySQL::MySQL(), host="localhost",
dbname="MR_Image_Analysis", user="root", password="yourpasswordatsqlinstall")
       ##Establishes connection to local instance of MySQL DB

       dbListTables(conn) # lists tables within the called db.


       dbWriteTable(conn, "ADNI_973_ALL_ordergrp1", d5)
       dbListTables(conn)
       write.csv(d5, 'yourpreferredlocation/ADNI_973_ALL_ranked_examdate.csv')
```

A.4. Local and Global Efficiency Calculation

The following code refers to Section 4 of the pipeline (Figure 3.1 of the thesis):

Section 4. New variable generation through graph theory	● Global efficiency (E_{global}) calculated based on the covariance matrix. ● E_{global} output is generated for a single subject and corresponds to a heat map. ● E_{global} is generated for the entire sample.

```
       #########################################################################
################################
       ##### Call the variables in the order that you want them, consider order
of regions in brain map ########
       #########################################################################
################################

       ##  The "mainvars" are the areas corresponding to the default mode
network (DMN), mainvars can be changed to whichever regions are of interest.
       ##  The suffix "TA" refers to Thickness Average, "TS" refers to Thickness
Standard Deviation
       ##  The names of these variables are preserved as the original ADNI Image
Analysis file, ' '  column names

       mainvars<-c("RID", "ST113TA", ##Begins thickness average
                   "ST111TA",
                   "ST103TA",
                   "ST109TA",
                   "ST98TA",
                   "ST95TA",
                   "ST93TA",
                   "ST73TA",
                   "ST54TA",
                   "ST52TA",
                   "ST44TA",
                   "ST50TA",
                   "ST39TA",
                   "ST36TA",
                   "ST34TA",
                   "ST14TA",
                   "ST113TS", ##Begins thickness standard deviation
                   "ST111TS",
                   "ST103TS",
                   "ST109TS",
                   "ST98TS",
                   "ST95TS",
                   "ST93TS",
                   "ST73TS",
                   "ST54TS",
                   "ST52TS",
                   "ST44TS",
                   "ST50TS",
                   "ST39TS",
                   "ST36TS",
                   "ST34TS",
                   "ST14TS")
       #########################################################################
##########################################################
       ##################  Now that "mainvars" and order have been established,
here we are changing names of variables to the ROI naming convention of the
Desikan-Killiany atlas  ##########
       #########################################################################
###############################################################

       d5mainvars <- d5[mainvars] #selecting only the variables corresponding to
the "mainvars" list within the sample
       dim(d5mainvars) #establishing dimensions, should appear as number of
cases x number of regions plus RID (973 x 33)
       colnames(d5mainvars)

       ##################  This is where subjects can be selected
###########################
       dim(d5mainvars)
       globaleff<-NA
       ne_rRAC<-NA
       ne_rPREC<-NA
       ne_rPHIP<-NA
       ne_rPCG<-NA
       ne_rMOF<-NA
       ne_rLOF<-NA
       ne_rIST<-NA
       ne_rCAC<-NA
       ne_lRAC<-NA
       ne_lPREC<-NA
       ne_lPHIP<-NA
       ne_lPCG<-NA
       ne_lMOF<-NA
       ne_lLOF<-NA
       ne_lIST<-NA
       ne_lCAC<-NA


       nodaleff<-NA
       d=NULL
       it<-NA
       sit99<-NA
       localsit<-NA
       for (thing in 1:973) {
         #TATS<-cbind(t(d5[1, 2:17]), t(d5[1,18:33]) )
         TATS<-cbind(t(d5mainvars[thing, 2:17]), t(d5mainvars[thing,18:33]) )

         #print(head(TATS))
         #}

         TATSdf<-data.frame(TATS)
         nodej<- c("rRAC",
                   "rPREC",
                   "rPHIP",
                   "rPCG",
                   "rMOF",
                   "rLOF",
                   "rIST",
                   "rCAC",
                   "lRAC",
                   "lPREC",
                   "lPHIP",
                   "lPCG",
                   "lMOF",
                   "lLOF",
                   "lIST",
                   "lCAC")
         TATSdf$nodej<-nodej

         #now add new column that is nodei

         TATSdf$nodei="rRAC" #make the whole column repeat "rRAC"
         View(TATSdf)
         #print(head(TATSdf))
         #}

         TATSdf$imu<-TATSdf[1,1] #make the whole column be item [1,1] which is
 first row, first column
         TATSdf$isd<-TATSdf[1,2] #make the whole column be item [1,2]

         ##  Rename columns, then reorder
         colnames(TATSdf)
         head(TATSdf)

         colnames(TATSdf) <- c("jmu","jsd",
                               "nodej",
                               "nodei",
                               "imu",
                               "isd")

         TATS_rRAC<-TATSdf
         dim(TATS_rRAC)

         ## rPREC
         TATSdf$nodei="rPREC"
         TATSdf$imu<-TATSdf[2,1]
         TATSdf$isd<-TATSdf[2,2]
         head(TATSdf)
         #dim(TATSdf)
         TATS_rPREC<-TATSdf
         dim(TATS_rPREC)
         head(TATS_rPREC)

         ##  rPHIP
         TATSdf$nodei="rPHIP"
         TATSdf$imu<-TATSdf[3,1]
         TATSdf$isd<-TATSdf[3,2]
         head(TATSdf)
         #dim(TATSdf)
         TATS_rPHIP<-TATSdf
         dim(TATS_rPHIP)
         head(TATS_rPHIP)

         ## rPCG
         TATSdf$nodei="rPCG"
         TATSdf$imu<-TATSdf[4,1]
         TATSdf$isd<-TATSdf[4,2]
         head(TATSdf)
         #dim(TATSdf)
         TATS_rPCG<-TATSdf
         dim(TATS_rPCG)
         head(TATS_rPCG)


         ## rMOF
         TATSdf$nodei="rMOF"
         TATSdf$imu<-TATSdf[5,1]
```

```
TATSdf$isd<-TATSdf[5,2]
head(TATSdf)
#dim(TATSdf)
TATS_rMOF<-TATSdf
dim(TATS_rMOF)
head(TATS_rMOF)


## rLOF
TATSdf$nodei="rLOF"
TATSdf$imu<-TATSdf[6,1]
TATSdf$isd<-TATSdf[6,2]
#head(TATSdf)
#dim(TATSdf)
TATS_rLOF<-TATSdf
#dim(TATS_rLOF)
#head(TATS_rLOF)


## rIST
TATSdf$nodei="rIST"
TATSdf$imu<-TATSdf[7,1]
TATSdf$isd<-TATSdf[7,2]
#head(TATSdf)
#dim(TATSdf)
TATS_rIST<-TATSdf
#dim(TATS_rIST)
#head(TATS_rIST)

## rCAC
TATSdf$nodei="rCAC"
TATSdf$imu<-TATSdf[8,1]
TATSdf$isd<-TATSdf[8,2]
head(TATSdf)
#dim(TATSdf)
TATS_rCAC<-TATSdf
dim(TATS_rCAC)
head(TATS_rCAC)

## lLOF
TATSdf$nodei="lRAC"
TATSdf$imu<-TATSdf[9,1]
TATSdf$isd<-TATSdf[9,2]
head(TATSdf)
#dim(TATSdf)
TATS_lRAC<-TATSdf
##checks
dim(TATS_lRAC)
head(TATS_lRAC)
TATS_lRAC[9,3]


## lPREC
TATSdf$nodei="lPREC"
TATSdf$imu<-TATSdf[10,1]
TATSdf$isd<-TATSdf[10,2]
head(TATSdf)
#dim(TATSdf)
TATS_lPREC<-TATSdf
##checks
dim(TATS_lPREC)
head(TATS_lPREC)
TATS_lPREC[10,3]
```

```
## lPHIP
TATSdf$nodei="lPHIP"
TATSdf$imu<-TATSdf[11,1]
TATSdf$isd<-TATSdf[11,2]
head(TATSdf)
#dim(TATSdf)
TATS_lPHIP<-TATSdf
##checks
dim(TATS_lPHIP)
head(TATS_lPHIP)
TATS_lPHIP[11,3]


## lPCG
TATSdf$nodei="lPCG"
TATSdf$imu<-TATSdf[12,1]
TATSdf$isd<-TATSdf[12,2]
#head(TATSdf)
#dim(TATSdf)
TATS_lPCG<-TATSdf
##checks
dim(TATS_lPCG)
head(TATS_lPCG)
TATS_lPCG[12,3]

## lMOF
TATSdf$nodei="lMOF"
TATSdf$imu<-TATSdf[13,1]
TATSdf$isd<-TATSdf[13,2]
#head(TATSdf)
#dim(TATSdf)
TATS_lMOF<-TATSdf
##checks
dim(TATS_lMOF)
head(TATS_lMOF)
TATS_lMOF[13,3]

## lLOF
TATSdf$nodei="lLOF"
TATSdf$imu<-TATSdf[14,1]
TATSdf$isd<-TATSdf[14,2]
#head(TATSdf)
#dim(TATSdf)
TATS_lLOF<-TATSdf
##checks
dim(TATS_lLOF)
head(TATS_lLOF)
TATS_lLOF[14,3]


## lIST
TATSdf$nodei="lLIST"
TATSdf$imu<-TATSdf[15,1]
TATSdf$isd<-TATSdf[15,2]
#head(TATSdf)
#dim(TATSdf)
TATS_lIST<-TATSdf
##checks
dim(TATS_lIST)
head(TATS_lIST)
```

```
TATS_lIST[15,3]

## lCAC
TATSdf$nodei="lCAC"
TATSdf$imu<-TATSdf[16,1]
TATSdf$isd<-TATSdf[16,2]
#head(TATSdf)
#dim(TATSdf)
TATS_lCAC<-TATSdf
##checks
dim(TATS_lCAC)
head(TATS_lCAC)
TATS_lCAC[16,3]


#####

TATSdf$nodej

#print(head(TATSdf))

#}

newdf<-rbind(TATS_rRAC,
             TATS_rPREC,
             TATS_rPHIP,
             TATS_rPCG,
             TATS_rMOF,
             TATS_rLOF,
             TATS_rIST,
             TATS_rCAC,
             TATS_lRAC,
             TATS_lPREC,
             TATS_lPHIP,
             TATS_lPCG,
             TATS_lMOF,
             TATS_lLOF,
             TATS_lIST,
             TATS_lCAC)
dim(newdf)
View(newdf)
table(newdf$nodei)
newdf$nodepair<-c(1:256)
subj1sc1<-newdf
head(subj1sc1)

#print(head(subj1sc1))
#}


#write.csv(subj1sc1, 'yourpreferredlocation/subj1sc1.csv', row.names=T)
#dim(subj1sc1)
#colnames(subj1sc1)
test<-as.data.frame(subj1sc1, rownames=TRUE)

#View(testing2) #review data layout
##this is based on the final file created using 'adding_sd.R'


adjacency_matrix<-function(csvfile){ #can input a dataframe or csv file
  imu=csvfile[,5] # fifth column will be stored as imu (mean of node i)
  jmu=csvfile[,1] # jmu (mean of node j)
  isd=csvfile[,6] # isd (standard deviation of node i)
```

```
           jsd=csvfile[,2] # jsd (standard deviation of node j)
           Zij <- ((imu - jmu) / jsd) #computes z score of ij
           Zji <- ((jmu - imu) / isd) #computes z score of ji
           Cij <- ((abs(Zij))+(abs(Zji)))/2 #computes covariance value for the z
scores of ij and ji
           return(Cij)}

         #print(Cij)
         #################

         ## run function on dataset
         dim(test)
         mat<-adjacency_matrix(test)

         M_test <- matrix(data=mat, nrow=16)
         M_test
         dim(M_test)

         #print(head(M_test))
         #}


         rownames(M_test)<-c("rRAC",
                             "rPREC",
                             "rPHIP",
                             "rPCG",
                             "rMOF",
                             "rLOF",
                             "rIST",
                             "rCAC",
                             "lRAC",
                             "lPREC",
                             "lPHIP",
                             "lPCG",
                             "lMOF",
                             "lLOF",
                             "lIST",
                             "lCAC")

         colnames(M_test)<-c("rRAC",
                             "rPREC",
                             "rPHIP",
                             "rPCG",
                             "rMOF",
                             "rLOF",
                             "rIST",
                             "rCAC",
                             "lRAC",
                             "lPREC",
                             "lPHIP",
                             "lPCG",
                             "lMOF",
                             "lLOF",
                             "lIST",
                             "lCAC")

         ########
         # Basics: heatmap visualization using default (guidance:
http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-
software-and-data-visualization)

         library(reshape2)
         conn_matrix<-M_test
```

```
        melted_conn_matrix <- melt(conn_matrix)
        head(melted_conn_matrix)
        melted_conn_matrix
        library(ggplot2)
        ggplot(data = melted_conn_matrix, aes(x=Var1, y=Var2, fill=value)) +
          geom_tile()

        ####### Colormap using ggplot2 ##########

        library(reshape2)
        #conn_matrix<-read.csv(file.choose())
        conn_matrix<-M_test
        melted_conn_matrix <- melt(conn_matrix, na.rm = TRUE)
        #melted_conn_matrix<-melt(conn_matrix_bin)
        ######
        # Create ggheatmap
        ggheatmap <- ggplot(melted_conn_matrix, aes(Var1, Var2, fill = value))+
          # ggtitle("subj1sc1")+
          geom_tile(color = "blue")+
          #scale_fill_gradient2(low = "white", high = "blue", mid = "white",
          #midpoint = 0.5, limit = c(0,1),
          scale_fill_gradient2(low = "white", high = "black", mid = "white",
                               #midpoint = 0.5, limit = c(0,1),
                               space = "Lab",
                               name="Structural Covariance") +
          theme_minimal()+ # minimal theme
          theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                           size = 12, hjust = 1),
                axis.text.y = element_text(vjust = 1, size = 12, hjust = 1))+
          coord_fixed()
        print(ggheatmap)

        range(melted_conn_matrix$value)

        ggheatmap +
          #geom_text(aes(Var1, Var2, label = value), color = "black", size = 1)
+

          theme(
            axis.title.x = element_blank(),
            axis.title.y = element_blank(),
            panel.grid.major = element_blank(),
            panel.border = element_blank(),
            axis.ticks = element_blank()

          )


        #######
        dim(M_test)
        g<-graph_from_adjacency_matrix(M_test, weighted=TRUE)
        dist_bw_nodes<-shortest.paths(g)
        row_length <- nrow(dist_bw_nodes) #16
        inverse_dist <- 1 / dist_bw_nodes #shortest path length
        #inverse_dist <- 1 / D #shortest path length
        nodal_efficiency <-
colSums((inverse_dist*is.finite(inverse_dist))/(row_length-1), na.rm=T)
        # is.finite(inverse_dist) returns a corresponding true/false matrix as
to whether or not a number is finite/infinite (e.g. 0)
        # here we are trying to detect the diagonal line, removing it and in
the row_length-1 computation removing the i=j node (16-1) in the calculation of
mean.
        global_efficiency<- sum(nodal_efficiency)/row_length
```

```
  print(nodal_efficiency)
  print(global_efficiency)

  #d=rbind(d,data.frame(nodal_efficiency))
  test<-data.frame(nodal_efficiency)
  NE_rRAC=test[1,]
  NE_rPREC=test[2,]
  NE_rPHIP=test[3,]
  NE_rPCG=test[4,]
  NE_rMOF=test[5,]
  NE_rLOF=test[6,]
  NE_rIST=test[7,]
  NE_rCAC=test[8,]
  NE_lRAC=test[9,]
  NE_lPREC=test[10,]
  NE_lPHIP=test[11,]
  NE_lPCG=test[12,]
  NE_lMOF=test[13,]
  NE_lLOF=test[14,]
  NE_lIST=test[15,]
  NE_lCAC=test[16,]

  ne_rRAC[thing]<-c(NE_rRAC)
  ne_rPREC[thing]<-c(NE_rPREC)
  ne_rPHIP[thing]<-c(NE_rPHIP)
  ne_rPCG[thing]<-c(NE_rPCG)
  ne_rMOF[thing]<-c(NE_rMOF)
  ne_rLOF[thing]<-c(NE_rLOF)
  ne_rIST[thing]<-c(NE_rIST)
  ne_rCAC[thing]<-c(NE_rCAC)
  ne_lRAC[thing]<-c(NE_lRAC)
  ne_lPREC[thing]<-c(NE_lPREC)
  ne_lPHIP[thing]<-c(NE_lPHIP)
  ne_lPCG[thing]<-c(NE_lPCG)
  ne_lMOF[thing]<-c(NE_lMOF)
  ne_lLOF[thing]<-c(NE_lLOF)
  ne_lIST[thing]<-c(NE_lIST)
  ne_lCAC[thing]<-c(NE_lCAC)


  globaleff[thing]<-c(global_efficiency)
  #nodaleff[thing]<-c(nodal_efficiency)

}


Eglob<-as.data.frame(globaleff)
#head(Eglob)
Eglob$RID<-d5$RID

NE_DMN<-as.data.frame(cbind(ne_rRAC,
ne_rPREC,
ne_rPHIP,
ne_rPCG,
ne_rMOF,
ne_rLOF,
ne_rIST,
ne_rCAC,
ne_lRAC,
ne_lPREC,
ne_lPHIP,
ne_lPCG,
ne_lMOF,
ne_lLOF,
ne_lIST,
ne_lCAC))

d5$ne_rRAC<-NE_DMN$ne_rRAC
d5$ne_rPREC<-NE_DMN$ne_rPREC
d5$ne_rPHIP<-NE_DMN$ne_rPHIP
d5$ne_rPCG<-NE_DMN$ne_rPCG
d5$ne_rMOF<-NE_DMN$ne_rMOF
d5$ne_rLOF<-NE_DMN$ne_rLOF
d5$ne_rIST<-NE_DMN$ne_rIST
d5$ne_rCAC<-NE_DMN$ne_rCAC
d5$ne_lRAC<-NE_DMN$ne_lRAC
d5$ne_lPREC<-NE_DMN$ne_lPREC
d5$ne_lPHIP<-NE_DMN$ne_lPHIP
d5$ne_lPCG<-NE_DMN$ne_lPCG
d5$ne_lMOF<-NE_DMN$ne_lMOF
d5$ne_lLOF<-NE_DMN$ne_lLOF
d5$ne_lIST<-NE_DMN$ne_lIST
d5$ne_lCAC<-NE_DMN$ne_lCAC

RID<-d5$RID
Efficiency_DMN<-as.data.frame(cbind(RID,globaleff,ne_rRAC,
                              ne_rPREC,
                              ne_rPHIP,
                              ne_rPCG,
                              ne_rMOF,
                              ne_rLOF,
                              ne_rIST,
                              ne_rCAC,
                              ne_lRAC,
                              ne_lPREC,
                              ne_lPHIP,
                              ne_lPCG,
                              ne_lMOF,
                              ne_lLOF,
                              ne_lIST,
                              ne_lCAC))


colnames(Efficiency_DMN)
head(Efficiency_DMN)
```

A.5. Adding newly generated measure to database

The following refers to Section 5 of the pipeline (Figure 3.1 of the thesis):

Section 5. Merging of Datasets	● Newly generated data are merged with a foundational data set based on subject ID, creating a new variable in the dataset.

```
	#conn<-dbConnect(RMySQL::MySQL(), host="localhost", dbname="Detached", 
user="root", password="yourpasswordatsqlinstall")
	#dbWriteTable(conn, "Efficiency_DMN", Efficiency_DMN)
	head(d5)
	d5$globeff<-Eglob$globaleff #add new column to primary dataframe
	colnames(d5)
	d5$globeff
	#write.csv(Eglob, 'yourpreferredlocation/globeff973.csv', row.names=T)
```