Freesurfer Vs. Manual Tracing: Detecting Future Cognitive Decline In Healthy Older Adults At-Risk For Alzheimer's Disease

Alissa Butts
Marquette University

FREESURFER VS MANUAL TRACING: DETECTING FUTURE COGNITIVE DECLINE IN HEALTHY OLDER ADULTS AT-RISK FOR ALZHEIMER'S DISEASE

by

Alissa M. Butts, M.S.

A Dissertation submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy

Milwaukee, Wisconsin

August 2013

ABSTRACT
FREESURFER VS MANUAL TRACING: DETECTING FUTURE COGNITIVE DECLINE IN HEALTHY OLDER ADULTS AT-RISK FOR ALZHEIMER'S DISEASE

Alissa M. Butts, M.S.

Marquette University, 2013

Alzheimer's disease (AD) is a neurodegenerative pathological process that is thought to begin years prior to observable symptom onset. The hippocampus appears to be particularly vulnerable to the underlying brain pathology of AD. Hippocampal volume is a sensitive measure in predicting conversion from mild cognitive impairment to AD, but less is known regarding the use of hippocampal volume in asymptomatic individuals at risk for AD who eventually decline. The inconsistent findings may, in part, be due to the chosen method of hippocampal segmentation. FreeSurfer (FS) and manual tracings (MT) are two common segmentation techniques that have unique costs and benefits. The present study directly compared hippocampal volumes generated by FS and MT in a longitudinal design assessing cognitively healthy elders, with varying degree of risk for AD, over a 4.5-year period. After 4.5 years, 15 participants demonstrated cognitive decline, while 45 remained stable. The results suggest FS consistently produced larger hippocampal volumes than MT, but neither method distinguished between groups at baseline. Longitudinally, individuals who declined experienced a more progressive pattern of atrophy compared to those who remained stable. These data suggest that hippocampal volume over time may be a useful variable in determining cognitive change over time, with the addition of other known risk factors, such as genetic risk. This study also suggests that in presymptomatic individuals, MT may not provide added benefit over the use of the more cost-effective FS.

ACKNOWLEDGMENTS

Alissa M. Butts, M.S.

I could not have completed this project without the support of many, and for that, I am very grateful. In particular, the support from friends and family gave me a constant sense of perspective and encouragement that was necessary to maintain perseverance throughout graduate school. My fellow graduate students, who became dear friends throughout this process, have not only helped me to complete graduate school, but have also helped me in my personal growth throughout the years. Similarly, thank you to all of the support staff, faculty, and supervisors at Marquette University, as well as at the Medical College of Wisconsin, Milwaukee VA, and the West Virginia University Medical School. I am indebted to you all for the time and effort that was generously given to foster my professional and personal growth. Finally, thank you to my advisor, Dr. Nielson, and additional faculty within the PIN research group for the intellectual contributions throughout this project and over the years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FreeSurfer vs Manual Tracing: Detecting Future Cognitive Decline in Healthy Older Adults At-Risk for Alzheimer's Disease

Alzheimer's disease (AD) is a growing public health concern that alters the quality of life of the affected individual and the individual's family. There is a need to identify individuals who are at risk for disease development before they demonstrate clinical symptoms because there is no current cure or treatment for AD that substantively slows disease progression. Indeed, redefined diagnostic criteria, set forth by the National Institute on Aging and the Alzheimer's Association, encourage the exploration toward identifying the earliest markers of AD development, with much promise being demonstrated by certain neuroimaging techniques, such as structural MRI (McKhann et al., 2011). However, limitations such as ambiguous clinical cutoffs and the lack of agreed upon routine standards for measurements prevent acceptance of these tools in the clinical setting (Jack, Albert, et al., 2011). Demonstration of the predictive clinical utility and agreement regarding standard methods in research settings may provide the scientific foundations necessary for the application of biomarkers in the clinical setting. Ideally, in vivo markers will be established that are sensitive enough to clinically predict future cognitive decline in seemingly healthy individuals.

Reliable, early distinction between those who will eventually develop AD from those who will only experience the milder cognitive changes associated with aging, will require identification of subtle differences in key brain regions affected at the earliest stages of the underlying brain disease. At the earliest stages, this degeneration is most apparent in medial temporal lobe (MTL) structures and in particular, the hippocampus.

Longitudinal, noninvasive neuroimaging techniques have shown that relative to healthy aging, hippocampal volume declines at an increased rate in AD (Costafreda et al., 2011). Reductions in hippocampal volume have even been shown in mild cognitive impairment (MCI) (Woodard et al., 2009), which is often a prodromal stage of AD (Petersen, 2004). Fewer studies involving hippocampal volume and magnitude of change have investigated conversion from cognitively healthy to cognitive decline, and fewer still have directly compared the methods used to characterize hippocampal volumes. Automated methods are cost-effective, but manually derived volumes are the 'gold standard' due to the afforded precision (Jack, Barkhof, et al., 2011) and the limitations of early automated methods. The volume differences and magnitude of atrophy change over time may be subtle, if indeed present, in individuals who eventually decline relative to those who do not decline. Thus, subtle and accurate *in vivo* hippocampal measures may be a key biomarker used to identify individuals who will develop AD, and thus, who would likely benefit most from early intervention.

Prevalence, Incidence, and Public Health Implications of AD

The high prevalence and incidence rates of Alzheimer's disease is a growing public health concern and is creating a large financial burden for caregivers and society. An estimated 200,000 people under the age of 65 ("2011 Alzheimer's disease facts and figures," 2011) and 5.4 million people over the age of 65 (Hebert, Scherr, Bienias, Bennett, & Evans, 2003) are living with AD. By 2050, the number of AD cases is expected to triple to between 11-16 million (Hebert et al., 2003). AD is the fifth leading cause of death in individuals over the age of 65 (Minino, Arias, Kochanek, Murphy, &

Smith, 2002) and is the only disease within the top 10 leadings causes of death without a current treatment to cure, reverse, or reliably slow the disease progression ("2011 Alzheimer's disease facts and figures," 2011). The estimated cost of caring for individuals with dementia was estimated at $183 billion in 2011 and is estimated to rise to $1.1 trillion by 2050 ("2011 Alzheimer's disease facts and figures," 2011). However, if the onset of this disease could be delayed by 5 years, the prevalence rate may decrease by 50%, and a 10-year delay in symptom onset may virtually eliminate AD (DeKosky & Marek, 2003). Delaying AD symptom onset is possible if clinicians can provide an early intervention to individuals with underlying AD pathology before symptoms emerge.

During the "asymptomatic, preclinical phase" of AD, the afflicted have pathophysiological disease markers, but still appear cognitively healthy (Sperling et al., 2011). It is likely that an AD intervention would be more effective in this earliest, presymptomatic stage of the disease. Yet, the ability to differentiate cognitively healthy elders who are able to withstand the disease burden from those who are not, and eventually become symptomatic remains elusive. An advanced understanding of the additive implications of various risk factors for AD development, and their impact on the individual over time, might lead to the identification of specific characteristics, or biomarkers, that will reliably predict who will eventually develop AD.

Predominant Risk Factors for Alzheimer's Disease

During the asymptomatic, preclinical phase of Alzheimer's disease, the individual is presumed to have risk factors that increase the likelihood of eventual decline. It is possible that understanding and exploring the individual and additive contributions of risk factors such as, demographic characteristics, including age and sex, genetic components, and a pre-dementia diagnosis, will lead to the development of a profile of characteristic that identify who will later decline into the clinical phases of AD. There are many other implicated ‘risk factors’ for AD that affect the individual at the community or societal level, such as education and socioeconomic status (Sattler, Toro, Schonknecht, & Schroder, 2012); however, for the purposes of this specific study the discussion will focus more on biological factors. Although there is no single characteristic that determines the absolute development of AD, there is increasing evidence to suggest that certain factors dramatically increase one’s risk.

Age.

The neurodegenerative nature of AD results in the emergence of symptoms typically after the fifth decade, and as one grows older the likelihood of developing AD increases dramatically. Of the cohort of Americans over the age of 65, one in eight (13%) will have AD, whereas nearly one in two (43%) Americans over the age of 85 will have AD (Hebert et al., 2003). Of the total number of individuals with AD, approximately 4% are under the age of 65 years (early onset), 6% are between the ages of 65-74 years, 45% are between the ages of 75-84 years, and 45% are 85 years or older ("2011 Alzheimer's disease facts and figures," 2011). Roughly, the prevalence rate of AD doubles with every five years of age over the age of 65 (Musicco, 2009). Insult or injury from a vascular event, diabetes, head trauma, accumulated distress from stressful life events, or risk due to genetic factors may each cause the brain to be more vulnerable to the changes that occur at the microbiological level (Herrup, 2010). Therefore, the longer the individual lives, the more likely it is that he or she will experience enough insult or injury to deplete their individual cognitive reserve (Herrup, 2010) and begin to manifest the clinical symptoms of AD.

Sex.

The uneven distribution of AD prevalence rates across sex suggests that the AD process differentially affects men and women through a currently unknown mechanism. Of the 5.2 million individuals living with AD, 3.4 million are women, whereas 1.8 million are men ("2011 Alzheimer's disease facts and figures," 2011). European studies consistently find higher incidence rates of AD in women compared to men, with women being 1.6 times as likely to develop AD than men (Gao, Hendrie, Hall, & Hui, 1998). This is consistent with the higher incidence of AD in women in American based samples (Vina & Lloret, 2010). A woman's lifetime risk of developing AD is nearly one in five (17.2%), though it is approximately one in ten (9.1%) for men (Seshadri, 2006).

The leading hypothesis as to why more women tend to develop AD than men is because the life expectancy of females is longer than their male counterparts (Vina & Lloret, 2010). Alternatively, a study conducted with Italian women found more pregnancies in women with AD compared to healthy elders, suggesting that a high level of estrogen is a potential contributor to AD development (Colucci et al., 2006). One study demonstrated greater metabolism reduction in AD targeted brain regions in women relative to men, suggesting women may have reduced cognitive reserve relative to men (Perneczky, Drzezga, Diehl-Schmid, Li, & Kurz, 2007). These differences remained even after controlling for disease severity, age, and education, placing them at a greater risk of cognitive decline (Perneczky et al., 2007). To date, there is not an agreed upon rationale for why there is a discrepancy in prevalence rates in AD across sex. Given that increased age is a clear risk factor for AD and women tend to live longer than men, it is possible that certain characteristics of women lead to their increased prevalence rate of AD. For example, while education may increase cognitive reserve, women historically have had fewer educational opportunities than men, and high estrogen levels due to multiple pregnancies are unique to women, both of which may contribute to the higher prevalence rates of AD in women relative to men.

Genetics.

Although no singular factor has been found to definitively lead to a diagnosis of AD, it is believed to be a disease with strong genetic contributions. Individuals with a positive family history, as defined as report of AD in a parent or sibling, are more likely to develop AD than those without a family history of AD (Green et al., 2002). Additionally, the risk of developing AD by 85 years old is about 43.7% for African American individuals and about 27% for Caucasian individuals (Green et al., 2002). Caucasian females with a first-degree relative with AD are more likely (31.2%) to develop AD than their male counterparts (20.4%). African American females with a first-degree relative with AD are no more likely (46.7%) to develop AD compared to their male counterparts (40.1%), though an increased likelihood is trending toward significance (Green et al., 2002). These statistics suggest at least a partial biological component to AD, in that sex and ethnicity are uniquely informative and interact with a positive family history to increase an individual's risk of developing AD.

In addition to a positive family history of AD, the impact of a number of genes and proteins on an individual's likelihood of developing AD has been explored. In particular, four genetic markers have received specific investigation in studies of the risk for late-onset AD (i.e., the most common form of AD): a) the Translocase of Outer Mitochondrial Membrane-40 homolog (TOMM40) gene, via the length of its polymorphic poly-T variant, rs10524523, connecting TOMM40 to the APOE gene (Roses et al., 2009); b) the β–amyloid precursor protein (β-APP), associated with mitochondrial dysfunction (Corder et al., 1993); c) mutations of the presenilin-1 (PSEN1) or presenilin-2 (PSEN2) proteins, which may contribute to Aβ accumulation in the extracellular space (Strittmatter et al., 1993); and d) apolipoprotein-E (APOE) ε4, which has to date shown by far the strongest association with AD development. By virtue of its predictive ability, APOE will be the only gene further discussed in this paper.

APOE is polymorphic biochemically and has three different allele forms: ε2, ε3, and ε4. Every individual carries two alleles in some combination of the three forms. While all combinations are known to occur, ε3ε3 occurs most commonly in the general population (68%) (Corrada, Paganini-Hill, Berlau, & Kawas, 2012), while ε2ε2 is the most rare (<1%) (Hyman et al., 1996). The ε4 allele has been shown to confer the greatest risk of developing AD, while ε2 may confer some protection (Martins, Oulhaj, de Jager, & Williams, 2005). Although carrying the ε4 allele does not mean that an individual *will* develop AD, the risk of AD development increases from 20%, for individuals who carry no ε4 allele, to 46.6% for individuals with one ε4 allele (heterozygous), and to 90% for individuals with both ε4 alleles (homozygous) (Corder et al., 1993). The ε2/ε4 combination is found in 2% of the general population, ε3ε4 in 15%, and ε4ε4 is found in <1% of the population (Corrada et al., 2012).

The effects and importance of APOE ε4 can be readily seen when examining the age of onset of AD and the general odds ratios, or all-cause probability of developing AD. The age of AD onset, based on examination, was shown to be reduced to 68 years on average when carrying two ε4 alleles compared to an average age of onset of 84 years in non-carriers (Corder et al., 1993). Moreover, it has been estimated that nearly all individuals homozygous for the ε4 allele will develop AD by the age of 80 (Corder et al., 1993), and will experience a faster rate of decline relative to non-carriers (Martins et al., 2005). In addition, the risk of AD with ε4 may be greater for women: Bretsky et al. (1999) found that women, but not men, with at least one copy of the ε4 allele were 7.8 times more likely to develop AD than women without the ε4 allele. The odds ratio (OR) of developing AD for non-carriers of ε4 (i.e., ε2/ε2, ε2/ε3, or ε3ε3 genotypes) is 0.6, yet it is 2.6 for the ε2/ε4 genotype, 3.2 for the ε3/ε4 genotype (Saunders et al., 1993), and is 14.9 if the individual is homozygous (ε4/ε4) for the APOE ε4 allele (Farrer et al., 1997).

Although the ε4 allele appears to be a strong indicator of AD development, APOE status alone cannot predict who will definitively develop AD, or even who will manifest clinical symptoms. As such, because genetics provide a strong predictive foundation, genetic risk factors in addition to other known risk factors, may provide a more exact indicator of cognitive decline.

Mild Cognitive Impairment.

One of the greatest risk factors for AD is a clinical diagnosis of mild cognitive impairment (MCI). The general term of MCI refers to a state of functioning in which an individual has some degree of cognitive impairment relative to healthy aging, but is not severe enough to meet diagnostic criteria for dementia (Petersen et al., 2001). Thus, MCI is often considered to be a transitional stage between healthy aging and AD because this diagnosis marks the initiation of symptom onset. The following discussion explains what is expected of memory change over time in healthy aging, how that differs in MCI and AD, and how that can be applied to detecting the earliest risk for AD.

Hippocampus and Memory in Aging, MCI, and AD

Alzheimer's disease is a progressive neurodegenerative disorder with a distinct clinical presentation of cognitive deficits and corresponding neuropathological characteristics that create a profile of changes far exceeding those that occur as a part of healthy aging. This decline is gradual over time but it reflects a marked reduction from the individual's previous level of functioning (APA, 2000). In particular, episodic memory in AD appears to represent the most dramatic and earliest change in memory, and in cognition in general (Dubois et al., 2007), which correlates to the earliest neural changes occurring in the MTL (Braak & Braak, 1991). The hippocampus, which is essential for memory, within the MTL appears to be asymmetrical as a part of normal aging, which may contribute its vulnerability in disease states. Better understanding of the pattern of memory change over time in normal aging and AD, and the integrity of associated supporting anatomical regions, may help to inform which feature(s) might be most indicative of insidious cognitive decline in cognitively intact individuals.

Volume differences between the left and right hippocampi across many age groups have been well documented, wherein there is a right greater than left tendency (Giedd et al., 1996; Honeycutt & Smith, 1995; Wolf et al., 2001). The etiology of this asymmetry is unknown, yet it is possible that it has evolutionary meaning. That is, the right hippocampus has been linked to spatial navigation and memory, and the left hippocampus has been linked to more verbally mediated memory processes (Kilpatrick et al., 1997). It may be that humans relied more on spatial navigation skills throughout development, giving rise to larger baseline right hippocampi that have continued to be observed in more recent neuroimaging studies. Asymmetrical differences have been documented even in children and adolescents. One study examining the hippocampus in individuals between the ages of 4- to 18- years old found that the hippocampus, bilaterally, increased with age and that the right hippocampus was larger than the left hippocampus (Giedd et al., 1996). This pattern of asymmetry, wherein the right hippocampus is shown to be larger than the left hippocampus, has been consistently documented in healthy young adults (Honeycutt & Smith, 1995), as well as older adults (Wolf et al., 2001). Further volume reductions, or atrophy, of the left hippocampus as a result of a disease process has been shown to be strongly correlated with poorer memory performances, particularly for verbal memory tests (Kilpatrick et al., 1997). Some studies have documented even smaller baseline left hippocampal volumes in individuals who eventually convert to dementia compared those who remain stable (Fox et al., 1996). As such, the asymmetry of the hippocampus may lead some individuals to be more vulnerable to disease insults, and as such, may be a useful indicator of detecting future cognitive decline.

In normal aging, various cognitive skills change over time to different degrees and at different rates. That is, certain skills demonstrate a more significant change, while others may change more subtly, and some may decline earlier while others may be more stable. Memory in particular, is a cognitive domain that is variably affected by the aging process, and is drastically affected in AD. Semantic memory, also known as ‘knowledge’ or memory for factual information, appears to be relatively preserved in healthy aging (Hoff, 2009). In contrast, episodic memory, or memory for information with temporal and/or personal relevance, tends to decline more with age than semantic memory (Grady & Craik, 2000). Functional neuroimaging studies reveal an increase in neuronal activity in memory-facilitating structures, such as the hippocampus, in healthy older adults relative to younger individuals (Nielson et al., 2006). This pattern of increased neuronal activity, which has been shown in numerous studies using various tasks (e.g., (Cabeza & Nyberg, 1997; Grady & Craik, 2000; Grady, McIntosh, & Craik, 2005; Langenecker & Nielson, 2003; Nielson et al., 2006; Nielson, Langenecker, & Garavan, 2002; Woodard et al., 2010), has most often been explained as neural compensation or ‘recruitment.’ That is, data from these studies suggest that as the brain ages, it ‘recruits’ additional neural activation to support performance during cognitive demand (Nielson et al., 2006).

In contrast to the typical cognitive changes associated with aging, individuals with MCI show a change in cognition beyond what is expected to occur with normal aging.

MCI has been recently re-categorized as within the “symptomatic, pre-dementia phase” (Albert et al., 2011). It is thought to be a transitional state between healthy aging and dementia. Indeed, 12% of patients diagnosed with MCI convert to dementia each year, with 31% to 44% of the converters progressing eventually to AD (Busse, Hensel, Guhne, Angermeyer, & Riedel-Heller, 2006). The individuals who decline from MCI to AD tend to be those with prominent memory disturbance, relative to other cognitive skills (Petersen, 2004). In contrast, only 1-2% of the general population is thought to convert from cognitively intact to dementia in a one-year interval (Petersen & Morris, 2003). Part of the high MCI conversion rate may be due to underlying neural changes that are occurring as part of the AD trajectory. Specifically, the functional recruitment, or hyperactivation, in the hippocampus seen in healthy aging is even more pronounced during memory tasks in those with MCI, suggesting that the brain is attempting to compensate for the disease process (Woodard et al., 2009). Yet, as MCI progresses to AD, as evidenced by increased deterioration of episodic memory skills and additional cognitive deficits (Dubois et al., 2007), the brain cannot compensate for the disease process, and the functional hippocampal activation in AD then drops off to less than that of healthy controls (Celone et al., 2006; Dickerson et al., 2005). Therefore, the ability to accurately diagnose MCI and understand its conversion to AD, while important, *is likely too late in the disease process to intervene* to an impactful degree because at this point of symptom onset, neuropathological change has already occurred and the current medications only potentially provide brief management of symptoms (Hong-Qi, Zhi-Kun, & Sheng-Di, 2012).

It may be beneficial to use the functional and cognitive characteristics of healthy aging and diseased aging (i.e. MCI and AD) as a gauge to help determine what factors may be suggestive of an individual who is at risk for cognitive decline. Importantly, although individuals who are at risk for AD perform similarly on cognitive tasks to those not at risk, they exhibit hippocampal hyperactivation that is comparable to those with MCI (Bondi, Houston, Eyler, & Brown, 2005; Seidenberg et al., 2009). The increased recruitment seen in those at risk for AD, suggests that not all "cognitively intact" individuals are on the trajectory of normal aging. Instead, these data indicate the presence of early degeneration affecting the function, and likely the structure, of the hippocampus. Thus, interventions have the most potential for efficacy during this pre-symptomatic stage, because of the early stage of disease progression. Examining asymptomatic individuals at risk for developing AD and following them over time will allow for an observation of change over time as many develop MCI, and some develop AD. More, studying these asymptomatic individuals before symptoms emerge might allow for the discovery of predictors for the development of MCI and AD.

Structural Effects on Hippocampus in the AD Course

Following the trajectory of an individual's memory ability via neuropsychological testing is certainly useful; it can clearly delineate change over time and may eventually be used to indicate symptom onset. Yet, it is well understood that the neuropathological changes underlying AD begin long before symptoms are measurable with neuropsychological testing (Morris, 2005). Therefore, accurate early identification of neuropathological changes in at-risk elders is essential to advancing the treatment or prevention of AD. Unfortunately, at the earliest stage of disease development individuals at risk for developing AD may perform cognitively within their expected performance range, when underlying brain changes suggest a disease trajectory. That is, neuropsychological tests effectively evaluate overt cognitive functioning, but there may be times in which information regarding the covert integrity of the brain will provide additional evidence that an individual is likely to decline.

At the molecular level, the pathological process of AD includes many histopathological changes in the brain prior to and following symptom onset. In fact, the underlying neurodegenerative process of AD is thought to begin decades before the onset of clinical symptoms (Morris, 2005). Histologically, AD is differentiated from other dementias by the presence of intracellular neurofibrillary tangles and extracellular beta-amyloid plaques (Jellinger, Danielczyk, Fischer, & Gabriel, 1990). It is the presence, location, and the effects of these signature proteins that likely contribute to the cognitive deficits associated with AD.

Beta-amyloid is a normally occurring protein in the brain; however, in AD over-productivity or insufficient disposal of this protein results in aggregation and progressive deposition, which leads to the conglomeration of plaques (Petrella, Coleman, & Doraiswamy, 2003). The beta-amyloid accumulation is thought to begin early in the pathogenesis of AD and leads to additional insults to the brain, such as local inflammatory changes, neurofibrillary degeneration, and neurotransmitter deficits in memory facilitating structures (Walsh & Selkoe, 2004). Plaques and tangles associated with AD pathology have been found to first arise in the entorhinal cortex of the MTL, then to spread to the adjacent hippocampus of the MTL, and finally to proliferate into the neocortex (Braak & Braak, 1991). This pattern of plaque and tangle emergence in the MTL is consistent with the earliest clinical symptom of marked and progressive impairment in the acquisition of episodic memories, for which, MTL structures play an important role.

In addition to the plaques and tangles, synaptic loss is thought to be a contributing factor to the clinical deficits observed in individuals with AD pathology (Arendt, 2009), with the hippocampus appearing to be particularly vulnerable to synaptic loss (Scheff, Price, Schmitt, DeKosky, & Mufson, 2007). The accumulated loss of individual synapses over time is thought to give rise to eventual loss of grey matter (Kassem et al., 2012). Indeed, the overall volume loss that occurs in AD is greater than that which is observed as a part of healthy aging (Juottonen, Laakso, Partanen, & Soininen, 1999), with approximately 12.2% lower total gray matter volume in individuals with AD relative to cognitively healthy elders (Karas et al., 2004). Fjell et al. (2009) examined the brain volumes of healthy older individuals over multiple time points and found that atrophy occurred at the rate of approximately 0.5% per year in the temporal and prefrontal cortices (PFC), two regions involved in memory performance. The negative correlation between age and volume was strongest for the entorhinal cortex and bilateral hippocampus, such that the older an individual, the smaller the volume of the MTL (Fjell et al., 2009). The hippocampus in particular has been shown to be 12% smaller in individuals with AD compared to non-demented elders, with similar reductions in the left and right hippocampus (Scher et al., 2007). Structural changes (e.g. medial temporal lobe atrophy) presumably due to AD pathology are notable in the pre-dementia stage of MCI. Relative to other cortical regions, entorhinal cortex thickness and the volume of the bilateral hippocampus have been shown to be the most sensitive structural differences in differentiating MCI and healthy controls (Desikan et al., 2010). Indeed, individuals with MCI have been documented as having smaller hippocampal volumes than healthy controls, with a particularly smaller left hippocampus and a trending toward significantly smaller right hippocampus (Woodard et al., 2010). These cross-sectional studies highlight the change in hippocampal volume in individuals with MCI and AD as a result of the AD process.

In addition to the cross-sectional volume differences that have been observed throughout the course of AD pathology, the rate of atrophy in AD-targeted regions over time in longitudinal paradigms has been documented and used to explore potential prediction models of cognitive decline. In one study examining the unique contribution of regions within the MTL, found that individuals with smaller entorhinal cortex, but not hippocampus after controlling for the entorhinal cortex, were more likely to convert to AD (Stoub et al., 2005). In this sample of 58 nondemented individuals, 14 converted to AD and 11 of those converters met criteria for MCI at baseline. This finding may imply that entorhinal cortex atrophy is more indicative of future decline than the hippocampus. However, remarkably, the rate of hippocampal volume loss bilaterally has been shown to be more rapid (1.18% per year) than entorhinal cortex atrophy (0.53% per year) in healthy older adults over time (Raz, Rodrigue, Head, Kennedy, & Acker, 2004). In AD, the rate of bilateral hippocampal atrophy has been shown to be approximately twice the rate observed in healthy aging (Jack et al., 1998). Furthermore, in a study comparing MTL regions over time, both entorhinal cortex and hippocampal atrophy had large effect sizes in comparing converters and nonconverters, but the hippocampal effect size was larger than that of the entorhinal cortex (Desikan et al., 2008). Additionally, other conversion studies have identified the hippocampus as a greater indicator of future decline. For example, one conversion study examining parahippocampal gyrus volume, including the entorhinal cortex, and hippocampus found hippocampal volume to be a better predictor of MCI to AD conversion than non-hippocampal structures, such as the entorhinal cortex (Visser, Verhey, Hofman, Scheltens, & Jolles, 2002). A final study found baseline hippocampal volume and rate of atrophy bilaterally provided greater predictability in MCI to AD conversion than other regions (Henneman et al., 2009). Together, these studies highlight the importance of medial temporal lobe atrophy in predicting cognitive decline, with somewhat inconclusive results as to whether a single MTL structure is uniquely reliable in predicting decline. Still, hippocampal atrophy has been shown to be one of the most important structures in identifying individuals who decline over time from MCI to AD. Needed, are more studies that assess the utility of cross-sectional and longitudinal hippocampal measurements in predicting decline in cognitively intact elders at risk for AD who eventually convert.

Consistent with the notion that the pathological processes of AD begin years before cognitive deficits are notable (Morris, 2005), it is possible that the AD-related morphological changes may extend to the asymptomatic, preclinical phase, when the individual is still cognitively intact yet is at risk for AD development. There appear to be inconsistent findings in the literature regarding the presence of hippocampal differences in those at risk compared to not at risk. For example, one study observed smaller volumes bilaterally in asymptomatic women at risk for developing AD relative to those not at risk (Cohen, Small, Lalonde, Friz, & Sunderland, 2001). In contrast, Burggren et al. (2008) found comparable hippocampi in cognitively healthy individuals at risk compared to those not at risk. Part of the inconsistencies between these studies may be due to the methods that were used to measure hippocampal volume and change over time. For example, the method used to measure hippocampal volume in Burggren et al. was different from the method used in Cohen et al. and therefore may account for the different findings. Instead of using complete manual derivations of the hippocampus, as was done in the Cohen et al. study, Burggren et al. began by manually separating the white matter of the MTL from the cerebral spinal fluid (CSF) and then used an automated cortical unfolding method to further define MTL subregions, including those of the hippocampus (Burggren et al., 2008). Using this approach, essentially utilizing automated techniques to label hippocampal tissue, they did not find group differences in hippocampal volume between groups (Burggren et al., 2008). In contrast, Cohen et al. used fully manualized methods to define hippocampal volume, and did find differences between groups (Cohen et al., 2001). The discrepant findings in these two studies suggest that, cross-sectionally, hippocampal volume differences between those at high risk compared to low risk, if present, may be subtle. Observing the pattern of hippocampal atrophy over time may also be useful in identifying indicators of future cognitive decline.

Few studies have been conducted that follow *cognitively intact* individuals longitudinally and explore the role of the hippocampus in conversion over time. An older study using serial MRI measurements, with manual tracings, found that healthy participants who later decline had smaller baseline hippocampi relative to healthy participants who remained stable (Kaye et al., 1997). Yet, the participants in the Kaye et al. study were at least 84 years old at study entry, and those who declined were older and had poorer cognitive performance to begin the study than those who remained stable. Therefore, it may be possible that a different disease process was detected in that study. In a more recent longitudinal study, larger manually derived total hippocampal volumes in cognitively intact elders at risk for AD were shown to be protective against cognitive decline after an 18-month interval (Woodard et al., 2010). These studies suggest that there may be subtle quantifiable differences in brain structure between cognitively healthy individuals at risk for AD who eventually decline and cognitively healthy individuals who remain stable over time. Yet, Jak et al. (2007) did not find a difference in baseline total hippocampal volume between risk groups; however, they did find a greater rate of atrophy measured longitudinally in those at risk relative to those not at risk, using manual methods (Jak, Houston, Nagel, Corey-Bloom, & Bondi, 2007). This study also did not find hippocampal volume to be significantly related to relative cognitive decline over time, although declining and stable participants were mixed amongst the risk groups (Jak et al., 2007), which may have contributed to the lack of significant correlation with the hippocampal volumes. Notably, the goal of their study was to examine the impact of risk status on hippocampal volume over time. Given that no single risk factor has been shown to be 100% predictive of AD development, it would be valuable to observe hippocampal volume change over time in cognitively healthy individuals who eventually convert.

To summarize, a number of studies have shown subtle hippocampal volume differences at various time points in cognitively intact individuals who are at risk for AD. In addition, some studies have suggested that the rate of change over time in hippocampal volume differs in those at higher risk compared with those at lower risk, whereas others have not supported these findings. Some of the inconsistencies may be due to the variability in the methods used across studies with this population. As such, clear group definition and structure inclusion may help to clarify potential group differences. That is, because these differences may be subtle, structural MRI studies that seek to be sensitive to the slight hippocampal volume differences in individuals who eventually decline may require careful awareness of the boundaries of surrounding structures and account for the possibility of 'partial voluming', or the inclusion of tissue from adjacent brain regions that are not targeted for inclusion. As such, a method with exceptional refinement capabilities may be warranted to assess for possible hippocampal volume differences at the earliest stage of AD. Therefore, a direct comparison in longitudinal data of two commonly used methods of isolating hippocampal tissue would be valuable toward helping to identify the most sensitive approach to discern early, subtle, hippocampal differences between AD risk groups.

Rationale for Present Study: FreeSurfer vs. Manual Hippocampal Volume Measurement.

Clinical classification studies that involve determining volume differences in specific brain structures employ either automated methods, such as FreeSurfer, or manually defined regions of interest, both of which are accompanied by specific costs and benefits. FreeSurfer is the predominant automated approach that is used to label cortical and subcortical brain structures, and it is compatible with multiple computer operating systems as well as functional and structural brain imaging software programs (Fischl, 2012). Notably, 20 versions of FS have been released between March 2006 (Version 3.0.0) and May 2011 (Version 5.1.0). FreeSurfer, which will be described in further detail in the methods section, is available for free download (http://surfer.nmr.mgh.harvard.edu) and labels regions of interest (ROI) through a series of tissue-isolating steps. As such, the use of FreeSurfer requires neuroimaging software processing knowledge and training. That is, FS requires running various commands and scripts to transform raw scanning data into appropriately processed numerical data representing each individual scan. Additional knowledge is then required to prepare the data for group comparisons and apply these numerical values to neuroimaging software programs such as Analysis of Functional NeuroImaging (AFNI) and Statistical Parametric Mapping (SPM). For volumetric studies, once the segmentation scripts are completed, the volumes may then need to be transformed into visually represented ROIs through the use of programs, such as AFNI or SPM. Although multiple neuroimaging processing steps are required, and many steps take multiple hours to execute, once the command begins, it will run independently without the constant attention of the researcher.

In contrast, manual tracings require a greater amount of labor, necessitating the constant attention of the researcher to complete the tracing, and are thus more time consuming than FreeSurfer (Jack, Barkhof, et al., 2011). Furthermore, manual tracings also require that the tracer have neuroanatomical knowledge and training (Jack, Barkhof, et al., 2011). Despite these potential costs of resources, tracings may allow for greater specificity in isolating voxels that include tissue belonging to relevant structures, and as such are considered to be the 'gold standard' in deriving hippocampal volume (Jack, Barkhof, et al., 2011). Thus, both approaches require neuroimaging and neuroanatomy knowledge, but the particular skills, as well as the costs and benefits, vary between method. Because FreeSurfer is one of the most commonly used automated methods and manual tracings are considered to be the 'gold standard' in defining hippocampal volume, these two approaches will be the focus of the present paper.

There are only a small number of studies that directly compare volumes generated by FreeSurfer and manual tracings, particularly as they relate to predicting cognitive decline. Both FreeSurfer and manual tracings have been shown to accurately differentiate clinical populations, such as AD and semantic dementia, from healthy controls (Lehmann et al., 2010). One study compared manual tracings to FreeSurfer (Version 4, released in August 2007) in participants with MCI, early AD, elders with cognitive complaints but normal neuropsychological testing, and healthy elderly controls (Shen et al., 2010). This study found manually traced and FreeSurfer volumes were highly correlated to one another and both were able to detect smaller hippocampal volumes in the clinical groups (early AD and MCI) relative to the nonclinical groups (healthy controls and elders with cognitive complaints but normal neuropsychological testing, CC) (Shen et al., 2010). The authors also report that both techniques showed the same pattern of atrophy AD > aMCI > CC (Shen et al., 2010). They suggest that the cognitive complaints group may mimic a preMCI group, which may lend to the theory that those with impending cognitive decline may not show hippocampal volume differences prior to measurable cognitive change. However, this study also showed that FreeSurfer yielded larger hippocampal volumes relative to the manual tracings, particularly in the hippocampal tail, and included a number of small areas that were presumably not hippocampal tissue but were instead adjacent MTL regions (Shen et al., 2010), which may have led to an inability to detect group differences in the CC group. Additionally, an older version of FS was used, and they did not follow the cognitive complaints group over time to determine if they eventually demonstrated decline on formal neuropsychological testing.

In another study with a sample of MCI, AD, and healthy controls, FreeSurfer (Version 4.0.2, released in December 2007) hippocampal volumes were compared to manually traced hippocampal volumes by evaluating percent overlap, percent volume difference, Pearson correlations, and Bland-Altman plots between FS and MT across groups (Sanchez-Benavides et al., 2010). This study demonstrated that, similarly to Shen et al. (2010), both FreeSurfer and manual tracings revealed a pattern of increasingly smaller volumes with an increase in disease severity (healthy controls > MCI > AD) (Sanchez-Benavides et al., 2010). Based on the similar pattern of hippocampal volume across groups, the strong correlation between the two measures (Pearson correlation = 0.84), and the Bland-Altman plots, they concluded that the two methods have “acceptable” interchangeability, but they acknowledge that FreeSurfer volumes were 10% larger on average than manual tracings (Sanchez-Benavides et al., 2010). In this study, a direct comparison between the manually traced volumes and FreeSurfer volumes revealed a 78% percent overlap between the two methods, but consistently found FreeSurfer to yield larger hippocampal volumes than the manual tracings in all participant groups (Sanchez-Benavides et al., 2010). The authors caution, that while FS may be used in general studies comparing populations, using FS as a primary volumetric method may increase the risk of Type II error, particularly in populations with atrophic brains, such as those with MCI and AD, or other clinical processes (Sanchez-Benavides et al., 2010).

The occurrence of FreeSurfer yielding larger volumes has been documented even in healthy populations (Cherbuin, Anstey, Reglade-Meslin, & Sachdev, 2009; Morey et al., 2009), particularly in the hippocampus (Lehmann et al., 2010). That is, a direct comparison between FreeSurfer and manual tracings yielded 23% larger left and 29% larger right hippocampal volumes generated by FreeSurfer (Cherbuin et al., 2009). This considerable difference in volumes derived from FreeSurfer relative to manual tracings was found in a large sample of healthy individuals with well-defined structural boundaries. They further demonstrated an even greater volume difference for individuals with indications of abnormally structured hippocampi (Cherbuin et al., 2009). This difference in methodology suggests that the potential for costly over-inclusion of tissue with FreeSurfer is even more likely when using this method in individuals with less well defined structural boundaries, as is likely the case for diseased populations and potentially cognitively healthy individuals that are on a disease trajectory.

For example, some studies have directly compared manual tracings to two automated methods, one of which was FreeSurfer, and have found a pattern of over-inclusion in FreeSurfer volumes, similar to that reported in earlier studies, in individuals with Major Depressive Disorder (Morey et al., 2009; Tae, Kim, Lee, Nam, & Kim, 2008). More specifically, in one study, FreeSurfer hippocampal volumes were 35% larger than manually traced hippocampal volumes in participants with Major Depression, leading the authors to conclude that manual tracings are preferable to FreeSurfer (Tae et al., 2008). The difference between the methods found in this population (35% larger FreeSurfer volumes) was even greater than the difference reported by Cherbuin et al. (2009) (23-29% larger FreeSurfer volumes) for a healthy population.

In another study, Morey et al. (2009) compared FS and another automated method to manually traced hippocampal volumes in a sample of individuals with depression compared to controls. They acknowledge manual tracings to be the standard because it is considered to be the closest representation of true hippocampal anatomy (Morey et al., 2009). The variance associated with FreeSurfer was greater than the manual tracings, but the correlation between the manual tracings and FreeSurfer was strong (R = .82), with a strong coefficient of determination (r^2 = .67), suggesting that 67% of the variance was shared by the two methods. Based on the greater variance associated with FreeSurfer, they suggest that a slightly larger sample size would be required if FreeSurfer is a utilized method to find the same effect size as manual tracings (Morey et al., 2009). Further, they found FreeSurfer to overlap with manual tracings by 82%. The between group comparison was also revealing in that they found FreeSurfer hippocampal volumes of individuals with depression were 9% smaller compared to controls, with a moderate effect size ($\eta^2 = .08$). They did not report on group differences found by manual tracings (Morey et al., 2009). These studies highlight the potential need for precise hippocampal measurement particularly in special populations, such as those with underlying disease and neurodegenerative pathology.

In summary, these studies suggest that both manual tracings and FreeSurfer can be used to distinguish between clinical groups and healthy controls, although they also reveal a pattern of FreeSurfer having a tendency toward producing larger absolute volumes. This greater absolute volume difference may not affect group comparisons when attempting to distinguish between groups with large volume differences, as may be the case when comparing clinical and nonclinical groups. However, attempting to distinguish between groups with subtle volume differences may require a greater attention to detail and more refined volume measurement capabilities. That is, the ability to detect the subtle, if present, differences that may exist between cognitively healthy individuals who later decline and those who remain stable may require a method with exceptional refinement ability.

Importantly, a newer version of FreeSurfer (Version 5.1.0, released May 2011) incorporates a number of software adjustments to presumably account for the problem of voxel over-inclusion. More specifically, this newer version includes GEnerative Model-based Segmentation (GEMS) that allows for improved hippocampal definition, as well as hippocampal subfield derivations (Van Leemput et al., 2009). While this upgrade may improve hippocampal segmentation, it is currently unknown if it has been improved enough to distinguish between individuals who eventually decline from those who remain stable. As such, a direct comparison of this new FreeSurfer version to manual tracings may help inform if one technique has a greater indication for use in detecting individuals who eventually decline.

Although time-intensive in comparison to FreeSurfer, it is possible that the additional refinement afforded by the manual editing of the hippocampal volumes may still be important in the ability to distinguish between hippocampal volumes of cognitively intact individuals who will later demonstrate decline from those who will remain stable. If manual tracings reliably provide this necessary refinement, then the cost of training and resources may be worth the knowledge gained for treatment planning and prognosis. Therefore, if hippocampal volume is an important indicator of future cognitive decline in lengthy longitudinal studies, it is important to assess whether the method of determining volume influences the sensitivity of such studies. Thus, the present study aims to explore the currently unaddressed possibility that FreeSurfer (Version 5.1.0) and manual tracings differ in their ability to distinguish between individuals who eventually decline from those who remain stable in a sample of cognitively intact elders.

Our Previous Research Toward Early Detection of AD.

To better understand the AD process with the goal of identifying a biomarker for AD development, we have been exploring the impact of various biological markers and risk factors for AD. Our population has included older adults who were cognitively intact based on neuropsychological testing at study entry 4.5 years ago. Although all of these individuals were determined to be cognitively intact, some subsets had specific risks for AD, including a positive AD family history and at least one APOE ε4 allele. Each testing visit consisted of a neuropsychological battery and an MRI with functional and structural acquisitions. Participants were retested at an 18-month and a 4.5-year follow-up session.

The majority of our work thus far has focused on cross-sectional differences in spatial extent and magnitude of activation, during our functional MRI task, in individuals at risk relative to other groups (e.g., Seidenberg et al., 2009). We have also begun conducting longitudinal analyses in which we can explore factors relevant to cognitive change over time. At the 18-month retest, we showed that age, education, sex, and a positive family history of dementia failed to adequately predict cognitive decline (Woodard et al., 2010). However, several factors did provide significant and independent predictive power. Specifically, smaller baseline hippocampal volume was a valuable and independent additive predictor of individuals who declined over the 18-month interval (p = .016) (Woodard et al., 2010). Data from the retest 4.5 years beyond baseline have not yet been published. The 18-month retest findings highlight the importance of precise hippocampal volume measurement to the prediction of cognitive decline (Woodard et al., 2010), which is somewhat complex and varies by the method used.

There have been a limited number of preliminary analyses conducted on approximately 60 participants who have already undergone the 4.5-year follow-up testing. In our previous prediction publication (Woodard et al., 2010), ”decline” was defined as a reduction from baseline performance to 18-month follow-up performance of at least one standard deviation on at least one of three principal outcome cognitive indices (DRS-2 total score, RAVLT Sum of Trials 1-5, RAVLT Delayed Recall). These measures are described in further detail in the methods section. Based on this definition, 45 of the 60 individuals remained cognitively stable from baseline to the 4.5-year follow-up, while the remaining 15 individuals exhibited decline. Of the 45 "stable" participants, 14 were +ε4 (31.1%) and 10 (66.7%) of the "declining" participants were +ε4. In a separate 18-month follow-up study not yet published, we showed that individuals who decline over 18 months have *smaller baseline hippocampal volumes* than individuals who remain stable over the 18-month interval. We do not yet know, however if this volume difference is a harbinger of continued cognitive decline over time at a rate steeper than those with larger baseline volumes. Examining cognitive change over a 4.5-year interval in comparison with baseline hippocampal volume is an important next step toward understanding the importance of hippocampal volume to predicting cognitive decline and its role in risk for AD.

Goals of the Present Study.

As described, few studies examining hippocampal volume have followed cognitively healthy individuals over time, and fewer still have attempted to directly compare the ability of manual versus FreeSurfer (Version 5.1.0) hippocampal segmentation in differentiating future cognitive decline. Therefore, this study will specifically compare the ability of automated FreeSurfer hippocampal volumes to the ability of manually traced hippocampal volumes to identify individuals who eventually decline in a sample of cognitively intact elders followed over 4.5 years. The following hypotheses were tested in the current study:

Hypothesis 1: Manually traced (MT) baseline hippocampal volumes will more optimally differentiate between Stable and Declining participants compared to FreeSurfer (FS). This was anticipated because manual tracings allow for more refined hippocampal volume measurement than FreeSurfer, which is likely necessary to differentiate between the subtle hippocampal volume differences in this asymptomatic – preclinical phase. That is, if there are baseline volume differences between cognitively intact individuals who eventually decline and stable participants, then it would be expected that MT volumes would differ significantly between groups, while FS volumes may not. If the FS volumes distinguish significantly between groups, the effect size for that difference was expected to be smaller than the effect size for the difference obtained from MT.

Hypothesis 2: The rate of hippocampal atrophy from baseline to the 4.5-year follow-up would differ by method. Manually traced volumes were expected to show a greater amount of atrophy between groups compared to the FreeSurfer volumes, such that there was an anticipated interaction between methods across the time points between groups. That is, at the follow-up assessment the groups were clinically different, as determined by group status, and this difference between the groups was expected to be large enough such that the greater variance associated with FS would have less of an impact on its ability to detect differences. That is, the degree of shared variance and overlap between FS and MT reported in previous studies with different clinical groups and similar methodology (e.g. Morey et al. 2009), suggests at the follow-up, both methods may accurately differentiate the cognitively different groups. However, at the baseline assessment, where the groups are equal on cognitive measures, the difference between groups, if present, are likely to be subtle enough that the greater variance associated with FS, as reported in previous studies (e.g. Morey et al., 2009), was expected to impact the ability of FS to detect group differences, whereas the MT was expected to be refined enough to do so.

Objectives Summary. Evaluating measurement of hippocampal atrophy over time is important because greater atrophy has been linked to poorer clinical outcomes. Therefore, it is important to compare manually traced hippocampal volumes to the automated FreeSurfer volumes since they are two of the most commonly used techniques in these volumetric studies. If manually traced volumes were to generate a statistically significant distinction in asymptomatic individuals over time relative to the FreeSurfer volumes, then that could be evidence for the importance of using manual tracings over FreeSurfer in a clinical setting. Alternatively, if the two methods are not shown produce significant volume differences between groups, then that could be evidence for using the automated FreeSurfer technique over manual tracings.

Method

Participants

This study made use of archival data. Participants were a subset of 60 healthy older adults. The participants were included from a larger sample of 459 adults from the community who were recruited via newspaper advertisement and met inclusion/exclusion criteria (see below). Following a telephone screen, 81 individuals agreed to undergo a neuropsychological evaluation, APOE genotyping from blood samples, and a MRI scan. FreeSurfer volumes were not available for five participants and of the remaining participants, 60 received both neuropsychological testing and a MRI scan at the third follow-up assessment. At study entry, of the 60 participants 75% were female, the mean age was about 72 years (M_{age} = 71.6 years, *SD* = 4.5 years), and the mean education was about 15 years ($M_{education}$ = 14.9 years, *SD* = 2.6 years). The interval time from the first scan to the second scan was about 554 days (*M* = 553.9 days, *SD* = 33.9 days) and from the second scan to the third scan was about 1191 days (*M* = 1191.4 days, *SD* = 105.0 days).

For APOE genotyping, DNA was isolated with Gentra Systems Autopure 1.5 for Large Sample Nucleic Acid Purification and was amplified using a PCR method (Saunders et al., 1996). Family history was defined as a report of a clear clinical diagnosis of AD or a reported history of gradual decline in memory and other cognitive functions, confusion, or judgment problems without a formal diagnosis of AD prior to death, in a first-degree relative. Of the total 60 participants, 24 (40%) had at least one copy of the ε4 allele. Thirty-two participants (53.3%) had a positive family history, while 28 (46.7%) did not have a family history.

Inclusion into the study required a negative history of cognitive impairment and/or dementia, neurological disease or medical illnesses, major psychiatric disturbance (e.g. Geriatric Depression Scale score greater than 15), or substance use meeting DSM-IV Axis I criteria. In addition, all participants were right-handed, based on the Edinburgh

Handedness Inventory (Oldfield, 1971). Informed consent was obtained consistent with the Declaration of Helsinki and institutional guidelines established by the Medical College of Wisconsin Human Subjects Review Committee; all participants received financial compensation for their participation.

Neuropsychological Assessment

All participants underwent baseline and follow-up neuropsychological testing and were determined to be cognitively intact, based on local norms determined at study entry. To generate the normative data, 91 local individuals with similar age, education, and ethnicity were tested. Participants were considered ‘cognitively intact’ if they fell within at least one standard deviation of the population mean. Separate means were created for males and females. The neuropsychological battery included the Mini-Mental State Examination (MMSE) (Folstein, Folstein, & McHugh, 1975), Mattis Dementia Rating Scale-2 (DRS-2) (Jurica, Leitten, & Mattis, 2001; Mattis, 1988), Rey Auditory Verbal Learning Test (RAVLT) (Rey, 1958), Geriatric Depression Scale (Yesavage et al., 1983), and Lawton Instrumental Activities of Daily Living Scale (ADL) (Lawton & Brody, 1969). The RAVLT and the DRS-2 was used to define decline. As such, these two tests are described in more detail below. All participants underwent a follow-up neuropsychological assessment approximately 18-months and 4.5 years after study entry. Alternate forms of the DRS-2 and the RAVLT were used during the initial follow-up examination for the majority of participants (see below). Because there is only one alternate form of the DRS-2, the DRS-2 Original version was used at study entry and at the 4.5-year follow-up.

The Dementia Rating Scale-2 (DRS-2) is intended for individuals over the age of 55-years-old (Strauss, 2006). This cognitive screening measure incorporates five subscales, including Attention, Initiation/Perseveration, Construction, Conceptualization, and Memory (Strauss, 2006). The individual subscale scores are then added together to create a Total score. Individual subscale alpha coefficients range from .75 to .95 (Vitaliano et al., 1984). Internal consistency of the Total score has been shown to be greater than .70 (Smith et al., 1994). Alternate form reliability for the Total score has been shown to be .82 (Schmidt, Mattis, Adams, & Nestor, 2005). Overall, the DRS-2 has been shown to have adequate reliability and vailidity coefficients. Good construct validity, with strong correlations (*r* = .78) with the Mini-Mental State Exam, and as such has often been used to determine cognitive decline (Bobholz & Brandt, 1993).

The RAVLT is a verbally mediated list learning test that includes 15 items (Rey, 1958). The individual is asked to listen carefully as the list of 15 words is read aloud, and is then asked to recall as many words from that list as possible in any order. There are five total learning trials in which the participant listens to the list and recalls as many words as possible. Following the fifth trial, a second list of 15 novel words (i.e., a distractor list) is read and the individual is asked to recall this new list of words in any order. Immediately after this distracter list, the individual is asked to again recall as many words from the original list as possible. Following a 20-minute delay, the individual is again asked to recall the original list. Finally, the individual is asked to correctly identify words from the original list during a verbal discrimination task. Thus, this measure evaluates an individual’s ability to learn new information, retain this new information in the presence of interference, and recognize this information among distracters (Knight, McMahon, Skeaff, & Green, 2007). The RAVLT has shown good reliability, particularly for the total score (Trials 1-5), delay recall score and Trial 5 (Strauss, 2006), with the internal reliability of the total score demonstrating a coefficient alpha of .90 (van den Burg & Kingma, 1999). Strong convergent validity ($r > .75$) has also been demonstrated, particularly between the delay recall and total score (van den Burg & Kingma, 1999), as well as other verbal memory measures, including verbal memory tests in the Wechsler Memory Scales (e.g., see Salthouse, Hancock, Meinz, & Hambrick, 1996). The stability of the test has been established in populations over 65-years-old (Estevez-Gonzalez, Kulisevsky, Boltes, Otermin, & Garcia-Sanchez, 2003; Knight et al., 2007) and it has been used frequently in the literature to assess cognitive decline over time (Estevez-Gonzalez et al., 2003; Woodard et al., 2010). As many as six alternate versions have been created with similar difficulty (Hawkins, Dean, & Pearlson, 2004), with alternate form reliability typically greater than .60 (Strauss, 2006). In the present study, three different forms were used for the majority of participants.

During the data collection process, test administration errors were made in which 10 participants were given the same version of the RAVLT at the 18-month and 4.5-year follow-up. Data from these 10 participants were reviewed and two of the 10 participants met criteria for Decline, whereas the remaining eight participants met criteria for Stable. All 10 participants were retained in analyses to maintain sufficient power. See limitations section for further discussion.

Image Acquisition

Anatomical images were acquired as part of a larger MRI protocol conducted on a General Electric (Waukesha, WI) Signa Excite 3.0 Tesla short bore scanner. High resolution, three-dimensional spoiled gradient recalled at steady-state (SPGR) anatomic images were acquired (TE = 3.9 ms; TR = 9.5 ms; inversion recovery (IR) preparation time = 450 ms; flip angle = 12 degrees; number of excitations (NEX) = 2; slice thickness = 1.0 mm; FOV = 24 cm; resolution = 256 x 224). Foam padding was used to reduce head movement within the coil.

Measurement of Hippocampal Volumes

FreeSurfer Hippocampal Volumes. FreeSurfer is a free software application for neuroimaging analyses and is available by download (http://surfer.nmr.mgh.harvard.edu). In the present study, FreeSurfer version 5.1.0 was used. Among other available tools, FreeSurfer can be used to label subcortical structures, such as the hippocampus. The details of FreeSurfer procedures can be found in previous publications (e.g., see Fischl & Dale, 2000). Briefly, brain matter is isolated by removing any non-brain matter after correcting for motion effects and applying the average of the volumetric T1 images (Segonne et al., 2004). A Talairach transformation, to adequately orient the brain into a standard reference space, is applied to the remaining images before deep gray and white matter is segmented (Fischl et al., 2002). An automatic deformation of the surface identifies the greatest shift in signal intensity from the cerebrospinal fluid (CSF) to the gray matter (Segonne, Pacheco, & Fischl, 2007), which is then used to create a boundary between structures, such as the hippocampus, that are adjacent to CSF filled ventricles by applying an intensity normalization (Sled, Zijdenbos, & Evans, 1998) and tessellation of the boundary (Fischl, Liu, & Dale, 2001). Finally, Freesurfer performs a topology correction of the identified brain matter (Fischl et al., 2001). The standard protocol for longitudinal processing with FS Version 5.1.0 was used for the present study (Reuter, Rosas, & Fischl, 2010). Briefly, the longitudinal data are first segmented cross-sectionally. Next, a template, or a base, is created from the cross-sectional segmentations to create an unbiased comparison point. Then, the cross-sectional scans are re-sampled to this template to allow for the detection of change over time. The creation and use of an unbiased template within this longitudinal processing stream works to reduce random variation inherent across multiple scans (Reuter et al., 2010).

Manually Traced Hippocampal Volumes. Hippocampal volumes from all three time points were manually traced in AFNI (Cox, 1996) on a T_1-weighted SPGR image by three raters (one primary rater and two reliability raters) blinded to participant group membership, time point and genetic risk status. The segmented hippocampal mask generated from FreeSurfer was overlaid on the individual's anatomical image. Left and right hippocampal volumes were traced independently. In the sagittal plane, raters edited the overlaid FreeSurfer mask to exclude any non-hippocampal voxels included by FreeSurfer and to add any hippocampal voxels excluded by FreeSurfer. In the coronal plane, the mask was further refined to exclude the fimbria while retaining hippocampal structures (alveus, uncal apex, cornu ammonis, subiculum, gyrus of retzius and fasciola cineria) as outlined in (Duvernoy, 2005).

Using the atlas, the primary rater (AMB) met with the two alternate raters to discuss the goal and method of manual tracing. Included, was training the alternate raters on the proper use of Linux and AFNI as they function as a platform for completing manual tracings. The primary rater demonstrated a manual tracing step-by-step, and then assigned practice tracings for the alternate raters. Once the practice tracings were complete, the primary rater reviewed the overlaps and met again with the alternate raters to discuss areas of agreement and disagreement, as they relate to the training atlas. This process began with face-to-face meetings, then continued with internet mediated meetings as needed until all raters achieved sufficient reliability (ICC > .80).

The primary rater completed all manual tracings, and the two additional raters traced a total of 10% of different randomly selected scans (5% each). A script was created to compare the alternate tracing to the primary tracing, by generating a quantity of voxles uniquely identified by the alternate rater and primary rater, respectively, as well as the overall agreement between the two tracings. The unique voxels were then summed to the overlap voxels for each tracer, and that total value was then entered into the intraclass correlation (ICC) statistic. Within SPSS, the ICC Statistic was chosen, with the Two-way Mixed Model and Absolute options selected. The resulting ICC values for the two raters with the primary tracer were 0.93 and 0.94, respectively, see Figure 1. Additionally, intra-rater drift (i.e., test-retest reliability) was also assessed on the primary rater. A total of 5% of tracings were retraced blind to participant group or characteristics and blind to the original tracing. The intraclass correlation was 0.95 (see Figure 1). Although the ICC values for the alternate raters and the primary rater were high, there was a notable absolute volume difference. The average volume for the first alternate rater was 2997 mm^3, and was 2528 mm^3 for the respective tracings for the primary rater. The average volume for the second alternate rater was 2038 mm^3, and was 2446 mm^3 for the respective tracings for the primary rater. This discrepancy suggests that the primary tracer was slightly more conservative compared to the first alternate rater, yet slightly more liberal compared to the second alternate rater. Together, this may suggest that had either one of the alternate rater been the primary rater, the overall set of manual tracings my have been systematically more conservative or liberal than the tracings in the present data set. Despite these mean volume differences, the ICC values were sufficient and strongly correlated with one another. Additionally, this method of manual hippocampal measurement has been used and accepted in previous publications (Hantke et al., 2013; Woodard et al., 2009; Woodard et al., 2010).

Hippocampal volumes were normalized by using a covariance approach (Buckner et al., 2004; Raz et al., 2005) to account for individual variation in brain size using the following formula:

$$\text{Adjusted volume} = \text{raw volume} - b * (\text{ICV} - \text{mean ICV})$$

Where the adjusted volume is the intracranial volume (ICV) corrected volume that was entered into analyses. Raw volume is the volume of the target region of interest (ROI) (i.e. hippocampus). The ‘b’ is the slope of a regression of a region of interest volume on the ICV, such that one slope value was generated per ROI to represent the slope for the sample. Finally, mean ICV is the average ICV, which is generated by FreeSurfer, for the whole sample. This approach to normalization accounts not just for ICV, but also the relative effects of ICV.

Definition of Cognitive Decline. Cognitive decline was defined as a reduction from baseline performance to 4.5-year follow-up performance of at least one standard deviation (SD) on at least one of the three principal outcome indices (DRS-2 total score, RAVLT Sum of Trials 1-5 [T1-5], RAVLT Delayed Recall [DR]). A residualized change score was generated for each individual and each cognitive measure by predicting follow-up scores using baseline scores; this procedure adjusts for baseline performance, practice effects, and regression to the mean (McSweeny, 1993; Temkin, Heaton, Grant, & Dikmen, 1999). Participants with standardized residuals of -1.0 or lower were assigned to the cognitively declining group. The remaining participants were classified as cognitively stable.

Power Analysis. A power analysis for each hypothesis was conducted using GPower. In previously discussed studies with similar methodology, medium effect sizes have been found when comparing hippocampal generating methods between clinical groups (e.g. Morey et al., 2009). Therefore, the medium effect sizes that were anticipated, along with repeated measures of Method (FS and MT) and Time (Baseline, 18-month follow-up, 4.5-year follow-up) for the second hypothesis were entered into a power analysis for the first and second hypothesis separately. Additionally, a high correlation among repeated measures was expected because the scanner was consistent throughout data acquisition of this longitudinal study, and FreeSurfer is an automated method that demonstrates high reproducibility (*R* = .82) of hippocampal volumes across multiple scans (Morey et al., 2009). Furthermore, the manually tracings, which may be more susceptible to variation over time, were mixed across sessions and the raters completing the manual edits were blinded to session. Thus, the risk of poor correlation over time was reduced.

Therefore, the power analysis for the first hypothesis indicated that for a mixed ANOVA, with partial eta square set at .06 for a medium effect, alpha at 0.05, power set at 0.8, two groups (Declining and Stable), four measurements (MT, FS, Left, Right), 0.8 correlation among repeated measures, 22 participants would be required. Given that there are 60 participants in the present study, the proposed design should have sufficient power to test the first hypothesis. The power analysis for the second hypothesis indicated that for a mixed ANOVA, with partial eta square set at .06 for a medium effect, as discussed above, alpha at 0.05, power set at 0.8, two groups (Declining and Stable), six measurements (MT and FS volumes measured at three time points), 0.8 correlation among repeated measures, 24 participants would be required. These power analyses suggest that the present study should have sufficient power.

Statistical Analyses

Hypothesis 1 aimed at assessing for group differences at baseline, while Hypothesis 2 assessed group differences over time in a longitudinal study. A 2 Method (FS, MT) x 2 Side (Left, Right) x 2 Group (Declining, Stable) mixed ANOVA was conducted to test Hypothesis 1. A 3 Time (baseline, 18 months, 4.5 years) x 2 Method (FS, MT) x 2 Group (Declining, Stable) mixed ANOVA was conducted to test

Hypothesis 2. Statistics were analyzed using SPSS version 18.0 for Windows. The level of significance was set at alpha = .05 for initial analyses.

Results

Identification of Cognitive Decline

Out of the 60 total participants, 15 (25%) declined by at least one standard deviation on at least one of the principal neuropsychological outcomes measures (DRS-2, RAVLT Trials 1-5, RAVLT Delay). These 15 participants comprised the "Declining" group, whereas the remaining 45 participants comprised the "Stable" group.

Ten of the 15 Declining participants (68%) compared with only 14/45 of the Stable participants (31%) carried an ε4 allele (Fisher's Exact Test $p = .031$) (see Table 1). The groups did not differ on other demographic variables, including age, education, family history or sex (see Table 1). Baseline performance on the DRS-Total, RAVLT Trials 1-5 and RAVLT Delay also did not differ between groups (see Table 1). Yet, as expected, Declining participants had poorer scores than Stable participants on all three neuropsychological measures at the 4.5 year follow up [DRS-2 Total, $F\,(1,14.99) = 6.9$, $p = .019$, $\eta^2 = .06$; RAVLT Trials 1-5, $F\,(1,18.27) = 10.41$, $p = .005$, $\eta^2 = .01$; RAVLT Delay, $F\,(1,17.00) = 15.82$, $p = .001$, $\eta^2 = .02$. Values are adjusted for significant tests of homogeneity of variances, see Table 2].

Hypothesis 1

The ability of FreeSurfer and manual tracings to distinguish baseline hippocampal volumes in future Declining participants from Stable participants was assessed using a 2

Method (FS, MT) x 2 Side (Left, Right) x 2 Group (Declining, Stable) mixed ANOVA. There were significant main effects for Method (FS volumes > MT volumes), *F* (1, 57) = 625.87, *p* < .001, η^2 = .915, see Figure 2), and Side (right > left; *F* (1, 57) = 23.64, *p* < .001, η^2 = .290, see Figure 3). There was no main effect of Group, *F* (1, 57) = .514 *p* < .476, η^2 < .009, see Figure 2.

Further clarifying the main effects of Method and Side, there was a significant interaction between Method and Side, *F* (1, 57) = 12.326 *p* = .001, η^2 = .175. Post hoc analyses revealed that within each method, right hippocampal volumes were larger than left, and that FS volumes were greater than MT volumes for both the left and right hippocampus (left hippocampus: *p* < .001, η^2 = .897; right hippocampus: *p* < .001, η^2 = .910; see Figure 4). The 2- and 3-way interactions involving Group were not significant: a) Method x Group, *F* (1, 57) = .131, *p* = .718, η^2 = .002, see Figure 5; b) Side by Group, *F* (1, 57) = 1.344, *p* = .251, η^2 = .023, see Figure 6; c) Method by Side by Group, *F* (1, 57) = 1.713, *p* = .196, η^2 = .029, see Figure 7.

Additional comparison between the hippocampal measurements was conducted with Bland-Altman plots (Bland & Altman, 1999). This approach, also used in similar studies (Sanchez-Benavides et al., 2010), depicts the consistency between FS and MT by plotting the mean difference between the two measures against the mean for each participant. As can be seen in Figures 8, 9 and 10 the vast majority of hippocampal volumes fell within the expected range and neither methodology exhibited evidence of systematic error. Thus, other than overall volume differences between methods, FS and MT appear to measure hippocampal volume in consistent and comparable ways.

One possibility for not finding significant baseline group effects in the mixed ANOVA may be the occurrence of low power to detect such differences. Although the power analysis and previously discussed literature indicated that this sample size would be sufficient power to test for differences, it is possible that there is a degree of variability in the present data that is leading to insufficient power. It may also be the case that no group differences were detected because there were no group differences to detect in these data, or the effect is very small. As originally proposed, to further assess these two contrasting possibilities, the results were reconsidered adjusting the alpha level from .05 to .10 to better protect against possible Type II error. Still, this adjustment made no difference with respect to significance of effects. Another possibility for not finding group differences may be that the data are influenced by another variable.

Because APOE ε4 status has been shown to be a meaningful factor in cognitive decline, a follow-up mixed ANOVA including APOE status was conducted to determine if APOE inheritance was greatly influencing the ability to detect differences between Stable and Declining participants. Supporting the value of APOE, a Method x Side x APOE mixed ANOVA revealed a significant interaction F (1, 58) = 7.732, p = .007, η^2 = .118, see Figure 11. Post hoc analyses reveal that within each method, APOE status did not differ within left and right volumes. Additionally, ε4-negative participants showed a pattern of right greater than left hippocampal volumes as measured by both FS and MT. Notably, ε4-positive participants showed the pattern of right greater than left hippocampal volumes as measured by FS, but not MT, see Figure 11. Therefore, it may be that APOE status is also an important variable in distinguishing between groups, and that using a combination of variables, such as hippocampal volume and APOE status may be useful.

To further assess the clinical meaningfulness of the interaction, Decline status was added to the mixed Method x Side x APOE ANOVA. This 4-way mixed ANOVA revealed an interaction trend, F (1, 56) = 3.315, p = .074, η^2 = .056, see Figure 12. The same FS >MT effects remained, (p's < .001), but post hoc analyses revealed specific differences between groups. In particular, ε4-negative Stable participants had greater right than left hippocampal volumes, as measured by both FS (p < .001) and MT (p = .004), but ε4-negative Declining participants (n = 5) had larger right than left volumes, as measured by FS (p =.003), but not MT (p = .639). Furthermore, ε4-positive Stable participants had larger right than left volumes as measured by FS (p =.021), but not by MT (p = .298), whereas ε4-positive Declining participants did not differ, see Figure 12. It may be that with additional subjects to improve power, this interaction may become statistically significant.

Further clarifying this marginally significant 4-way interaction, a Method x Decline x APOE mixed ANOVA was separately conducted for left and right hippocampi. The results of these analyses yielded a significant main effect of Method for both left (p < .001, η^2 = .888) and right hippocampi (p < .001, η^2 = .899), see Figure 13, with FS producing larger volumes than MT. Left, but not right, hippocampal volumes showed a trend for a Method by APOE interaction F (1, 56) = 3.936, p = .052, η^2 = .066. Post hocs show right volumes were greater than left in both ε4-negative and ε4-positive participants (*p's* < .001), but that within method, ε4-positive participants did not differ from ε4-negative participants (FS: *p* =.405; MT: *p* = .341), see Figure 13.

The relationship between hippocampal volume measurement, APOE, and cognition was further assessed in a way that may be less limited by low power and possibly suboptimal groups distinction. To assess these relationships, hierarchical multiple regression analyses were conducted to predict the change in cognitive performance from baseline to the 4.5 year assessment. The residualized change scores for RAVLT Trials 1-5, RAVLT Delay, and DRS-2 Total were used as the predicted variables for each regression because this technique adjusts for baseline performance, practice effects, and regression to the mean. Initial regressions with age, sex, and education entered in at Step 1, and with APOE and hippocampal volume entered in at Step 2, showed that demographic factors did not significantly contribute to the models predicting cognitive change, but APOE and hippocampal volume provided some predictive utility. Therefore, because the residualized change score inherently adjusts for some variance that may be associated with demographic factors, follow-up regressions were conducted replacing the demographics for hippocampal volume in Step 1 and APOE in Step 2. Given the relatively small sample size and the high inter-correlations of MT and FS, no more than three predictor variables were included and each method was entered into separate regressions.

Models 1-4 attempted to predict RAVLT 1-5 with hippocampal volume entered at Step 1 and APOE entered at Step 2, see Table 3. Model 1, which included total FS volume, and Model 2, which included left and right FS volumes, were not significant (*p* =

.302 and *p* =.425, respectively). Similarly, Models 3 and 4, which included total MT volume and left and right MT volumes, respectively, were not significant (*p's* < .280). Next, Models 1-4 were used to predict RAVLT Delay score, see Table 4. Model 1 showed that FS total volume alone was not significant (*p* = .087), but the model became significant after adding APOE (*p* = .002). Left and right FS volumes alone were marginally significant (*p* = .053), but again adding APOE significantly improved the overall predictive ability of Model 2 (*p* =. 004). In Model 3, total MT volume marginally predicted RAVLT Delay alone (*p* = .062), but the overall model improved when adding APOE (*p* .002). Separate left and right MT alone were not predictive (*p* = .151), but together with APOE, Model 4 was significant (*p* =.007). Finally, Models 1-4 were used to predict DRS-2 Total scores, see Table 5. As with RAVLT 1-5, Model 1 with FS total volume and Model 2 with left and right FS volumes were not significant (*p* = .093 and *p* = .108, respectively). A similar pattern of not significantly predicting DRS-2 Total score was found for Model 3, using MT total volume, and Model 4, using left and right MT volumes (*p* = .311 and *p* = .509, respectively). These regressions may suggest that baseline hippocampal volume may provide limited contribution in predicting change in cognitive performance, but is not necessarily sufficient without the inclusion of APOE.

Another useful method less restricted by small cell sizes and power is the receiver operating characteristic (ROC) curve, which has been used in similar studies (Cherbuin et al., 2009; Lehmann et al., 2010; Tae et al., 2008). This approach provides an indication of sensitivity and specificity of a measure, in this case the ability of FS and MT to correctly classify decline. The results of the ROC curve with baseline volumes suggest that neither baseline MT nor FS volumes correctly classified decline above chance levels, (all area under the curve (AUC) < .505, *p's* > .08) see Table 6, Figure 14.

Finally, sagittal and coronal overlap images of FS and MT baseline volumes for one Declining and one Stable participant were created, see Figures 15 and 16. A qualitative review reveals a degree of tissue included by FS, but not MT around the boarder of the hippocampus. There were indications of possible over-inclusion anteriorly adjacent to the amygdala by FS. This single participant depiction supports the quantitative finding of FS generating overall larger volumes than MT.

Hypothesis 2

To test the hypothesis that FreeSurfer and manually traced hippocampal volumes would vary by group at different time points was assessed by conducting a 3 Time (baseline, 18 months, 4.5 years) x 2 Method (FS, MT) x 2 Group (Declining, Stable) mixed ANOVA, with alpha = .05. There were significant main effects for both Time, F (2, 57) = 27.85, $p < .001$, $\eta^2 = .324$, see Figure 17, and Method, F (1, 57) = 635.57 $p < .001$, $\eta^2 = .916$, see Figure 18. Specifically, FS volumes were larger than MT volumes. Post hoc comparisons showed that the Time main effect reflected smaller volumes at each successive time point. There was no significant main effect for Group, F (1, 57) = 1.535, $p = .220$, $\eta^2 = .026$, see Figure 19.

Importantly, there was a significant interaction of Group by Time, F (1, 57) = 3.966, $p = .022$, $\eta^2 = .064$, see Figure 20. Post hoc analyses revealed that in Stable participants, 18-month volume did not differ significantly from baseline ($p = .168$), but volume at 4.5 years was significantly smaller than at baseline ($p < .001$) and at 18 months (*p* = .001) compared to baseline. However, the Declining participants had a more progressive pattern of atrophy, where hippocampal volume was smaller each time successive time point (*p*'s < .012). Yet, no contrasts differed significantly between groups at any time point, see Figure 20. None of the other interactions reached significance: a) Group by Method, *F* (1, 57) < .001, *p* = .996, η^2 < .001, see Figure 21; b) Method by Time, *F* (1, 57) = 2.387, *p* = .096, η^2 = .040, see Figure 22; c) Group by Method by Time, *F* (1, 57) = .552, *p* = .577, η^2 = .009, see Figure 23. Based on the same potential issues with power as were discussed for Hypothesis 1, the analysis was reconsidered using an adjusted alpha = .10, but no additional effects reached significance using this cutoff. Again, this may suggest that the absence of group differences may not be due to low power, or it may suggest that if group differences are present, the effects are very small.

To further explore longitudinal changes in hippocampal volume, APOE inheritance was included in the analysis. A separate mixed Time x APOE ANOVA for MT and FS was conducted. These analyses revealed a significant Time x APOE interaction as measured by FS [*F* (2, 57) = 4.192, *p* = .017, η^2 = .067], but not MT [*F* (2, 57) = .389, *p* = .678, η^2 = .007], see Figure 24. Post hoc analyses revealed that when measured by FS, the volumes of ε4-negative participants changed little from baseline to the 18-month follow-up (*p* =.164). Yet significant reduction was apparent at the 4.5-year follow-up, compared to the 18-month follow-up and baseline (*p* = .019, *p* =.006, respectively). In contrast, ε4-positive participants had somewhat smaller hippocampal volumes at the 18-month follow-up compared to baseline (*p* = .063), and significantly smaller 4.5-year volumes than baseline and the 18-month follow-up (*p's* < .001). At each time point, hippocampal volumes did not differ between ε4-positive and ε4-negative participants. Yet, there was no time by APOE interaction as measured by MT, such that ε4-positive and ε4-negative participants did not differ at each time point or over each time point, see Figure 24.

As with the first hypothesis, additional analyses were done to assess the relationship between hippocampal volume, APOE, and cognitive performance, see Table 7. Hippocampal volume at the 4.5-year follow-up appears to be more strongly related to cognitive outcome than APOE inheritance. In particular, a Pearson correlation demonstrated many strong correlations between cognitive performance and MT and FS volumes at the 4.5-year follow-up. Notably, although APOE is significantly correlated with decline status, *t*-test comparisons between APOE and hippocampal volumes at the 4.5-year follow-up were not significant. Similarly, a *t*-test comparing cognitive measures to APOE status revealed a significant association with RAVLT Delay score at the 4.5-year follow-up, wherein ε4-negative participants had higher scores than ε4-positive participants (*p* = .003), yet all other comparisons were not significant, see Table 7.

To determine if cognitive performance could be predicted by hippocampal volume and APOE at the 4.5-year follow-up, hierarchical multiple regressions were conducted. The Step and Model structures for this regression analysis on 4.5-year volumes were identical to those used for the first hypothesis on baseline volumes. Residualized change scores were used as the predicted outcomes, with 4.5-year hippocampal volumes and APOE as the predictive variables, because demographics were not significantly contributing. Models 1-4 attempted to predict RAVLT 1-5 performance, see Table 8. Model 1, using FS total volume alone was significant (*p* = .039), but adding APOE reduced the overall significance of the model (*p* = .091). In Model 2, using left and right FS volumes alone or with APOE was not significant (*p* =.142). In Model 3, MT total volume alone predicted RAVLT 1-5 (*p* =.014), and remained significant after adding APOE (*p* = .042). Left and right MT volumes alone were significant (*p* = .037), but adding APOE to Model 4 marginalized the significance of the overall model (*p* = .068). Models 1-4 were then used to predict RAVLT Delay score, see Table 9. In Model 1, FS total volumes alone (*p* =.027) and with APOE (*p* = .001) significantly predicted cognitive performance. Left and right FS volumes alone were marginally significant (*p* =.058), but the overall model became significant when adding APOE (*p* = .004), Model 2. A similar pattern was found with MT, such that total volumes (*p* =.052), as well as left and right volumes (*p* =.083) alone were marginally significant, whereas adding APOE significantly enhanced the overall models (Model 3: *p* = .003; Model 4: *p* = .007). Finally, Models 1-4 were used to predict DRS-2 Total score, see Table 10. Model 1, with FS total volume alone was significant (*p* =.008), and remained significant after adding APOE (*p* = .019). This same pattern was found for Model 2, using left and right FS volumes (*p* = .031). With MT, total volume alone was significant (*p* = .044), but adding APOE reduced the overall significance of the model (*p* = .089), Model 3. Left and right MT volume alone was significant (*p* =.026), and the overall model remained significant when adding APOE (*p* =.030), Model 4.

To supplement the previous comparison of hippocampal volumes at each time point across groups, a separate analysis was conducted to assess the degree of atrophy from one time point to another. This evaluation allowed for another observation of rate of atrophy over a period of time between groups across methods. Therefore, three change scores were computed for each method: 1) change from baseline to 18-month follow-up; 2) change from 18-month follow-up to 4.5-year follow-up; and 3) change from baseline to 4.5-year follow-up. These change scores were entered into three separate 2 Method (FS, MT) x 2 Group (Declining, Stable) ANOVA tests, with alpha = .05. The results revealed no significant differences between Declining and Stable participants in the amount of change between interval assessments: a) change from baseline to 18-month follow-up, FS: *p* =.110; MT: *p* =.500; b) change from 18-months to 4.5 years: FS: *p* =.094; MT: *p* =.489; c) change from baseline to 4.5 years FS: *p* = .067 (after accounting for violation of homogeneity of variance); MT: *p* = .163, see Figure 25.

Although there were no significant interactions when comparing the degree of change from one time point to another between Group and Method, there was a notable finding when assessing the degree of change between time points with hippocampal volume and APOE. A one-way ANOVA revealed the change in FS volume to the 4.5-year follow-up from baseline and the 18-month assessment differed by APOE (*p* = .023 and .029, respectively). For both change intervals, ε4-positive participants demonstrated greater change than ε4-negative participants. These were the only two significant findings with the ANOVA comparing hippocampal volumes and APOE, see Table 7.

Discussion

Recently revised diagnostic criteria for Alzheimer's disease (AD) (McKhann et al., 2011) urge for further research of biological factors associated with AD, such as hippocampal volume, in an effort to identify presymptomatic individuals who will eventually convert to AD (Sperling et al., 2011). To date, studies have been inconsistent as to whether individuals at risk for AD and those who eventually experience cognitive decline have smaller baseline hippocampal volumes than those who remain healthy (Burggren et al., 2008; Cohen et al., 2001; Jak et al., 2007; Kaye et al., 1997; Woodard et al., 2010). Furthermore, the rate of atrophy in presymptomatic individuals has as of yet seen only very limited exploration, and has produced somewhat inconsistent results. The disparate findings may be in part related to methodological differences in hippocampal volume measurement (i.e. FreeSurfer versus manual tracings). Therefore, the current study aimed to investigate the utility of hippocampal volume in a sample of cognitively intact elders at risk for AD measured over three time points. At the 4.5-year follow-up, a subsample of individuals demonstrated cognitive decline. Hippocampal volumes as measured by FreeSurfer (FS) Version 5.1.0 were directly compared to manually traced (MT) hippocampal volumes at all three time points to determine if one method was superior to another at detecting meaningful group differences. More specifically, the first hypothesis was that MT, but not FS, would detect group differences at baseline between eventual Declining and Stable participants. The second hypothesis was that the rate of hippocampal atrophy from the baseline to the 4.5-year follow-up assessment would differ between FS and MT, such that MT were expected to show a greater amount of atrophy between groups compared to FS.

Aim 1: Baseline Hippocampal Comparisons Between Method, Side and Group

General baseline hippocampal volume differences.

The first aim of the study was to compare baseline hippocampal volumes, as measured by FS to those measured by MT, and to assess for possible group differences in the hippocampal volume of cognitively healthy elders who later declined compared to those who remained stable. As expected, FS yielded overall larger hippocampal volumes relative to MT. This finding is very consistent with the literature demonstrating that FS produces larger hippocampal volumes than MT (Burggren, Small, Sabb, & Bookheimer, 2002; Cherbuin et al., 2009; Jak et al., 2007; Lehmann et al., 2010; Morey et al., 2009; Sanchez-Benavides et al., 2010). FreeSurfer is a segmentation technique initially developed to delineate cortical structures (Fischl, 2012; Fischl & Dale, 2000). Although a relative strength of FS is its automated nature that allows for rapid segmentation of high quantities of data, the ability of FS to segment subcortical structures has been less precise, resulting in consistently larger reported hippocampal volumes (Burggren et al., 2002; Cherbuin et al., 2009; Jak et al., 2007; Lehmann et al., 2010; Morey et al., 2009; Sanchez-Benavides et al., 2010). While this version of FS, Version 5.1.0, was thought to have made improvements to subcortical segmentation (Van Leemput et al., 2009), the present data suggest that compared to the 'gold standard' manual tracings, it continues to yield larger overall hippocampal volumes. This discrepancy between the methods is notable because when the method chosen differs between research studies, it may bias the generalized findings and comparison across studies, which may be in part related to the inconsistencies within the literature.

In the present study, baseline hippocampal volumes as measured by FS, were 52% larger than MT. This percentage is larger than the 26% larger volumes that have previously been reported in healthy individuals (Cherbuin et al., 2009), and is larger even than the 35% that has been documented in clinical samples (Tae et al., 2008). By side, the present study found 50% larger left and 54% larger right hippocampal volumes as measured by FS compared to MT. Again, these values are larger than the 23% larger left and 29% larger right hippocampal volumes that have previously been reported (Cherbuin et al., 2009). Furthermore, in the present study, the standard error of baseline FS volumes was 35% greater than MT. This variance is greater than the comparable variance between the two methods found in the previous study with healthy individuals (Cherbuin et al., 2009). Yet, a greater variance associated with FS has been documented in other studies with clinical populations (Morey et al., 2009). Therefore, the present study provides further demonstration that even this newer version of FreeSurfer produces overall larger hippocampal volumes with greater variance compared to MT.

Hippocampal volumes were 3624.4 mm^3 on average for Declining and 3715.3 mm^3 for Stable participants in the present study. When examined by method separately, the manually traced volumes were 2917 mm^3 and FS volumes were 4422.6 mm^3, on average. Earlier studies have examined the histological and MRI generated volume of the hippocampus post-mortem. In a study that compared the post-mortem volume of the hippocampus in young individuals and elders who died of non-neurological causes to patients with AD, the average volume was shown to decrease with age, with a dramatic reduction in AD (Simic, Kostovic, Winblad, & Bogdanovic, 1997). Specifically, the authors showed the volume of the hippocampus in the younger sample, with a mean age of 30-years-old, was 3,731 mm^3, hippocampal volume was 3,484 mm^3 in the control group (mean age of 80 years), and was 2,409 mm^3 in those with AD (mean age 84.4 years) (Simic et al., 1997). Additionally, another study comparing post-mortem histological volumes to MRI volumes found very strong correlations between the MRI and histological measurements (Bobinski et al., 2000). The volumes for the MRI measurements were 3083 mm^3 for the control and 2169 mm^3 for the AD group, and the histology volumes were 3203 mm^3 and 2257 mm^3 for the two groups, respectively (Bobinski et al., 2000). Together, these studies suggest that the manually traced measurements obtained in the present study may be a slight under-quantification of expected volumes for healthy elders and those that might be on a mild disease trajectory, whereas the FS volumes may be producing greater over-estimation of true hippocampal volume.

One explanation for the observed volume differences between methods in the present study is that FS is yielding larger volumes than expected. Yet, another possible explanation is that manual tracing is providing an underestimation of true hippocampal volume. This latter explanation may also provide rationale as to why the FS volumes were 52% larger in the present study, when earlier studies with healthy and clinical samples have found 26% and 10-35% greater FS volumes, respectively (Cherbuin et al., 2009; Tae et al., 2008). The histological studies on post-mortem hippocampi within this population may suggests that slight biases in both methods may be contributing to the greater difference between FS and MT observed in the present study. That is, the 52% difference may be mediated by a slight under-estimation of manually traced volumes and an over-estimation by FS. Together, the contrasting biases between the two approaches may have led to the overall greater amount of difference between the two methods compared to previous studies.

Laterality baseline hippocampal volume differences.

In addition to FS yielding larger volumes than MT, there also was a laterality difference detected in the present study where baseline right hippocampal volumes were about 5% greater than left volumes. Although many similar studies collapse across hemisphere to explore a “total hippocampal” volume (Henneman et al., 2009) and many studies have not found laterality differences (Kaye et al., 1997; Raz et al., 2004; Sanchez-Benavides et al., 2010), differences between left and right hippocampi in various samples have been well documented (Jack et al., 2003; Shi, Liu, Zhou, Yu, & Jiang, 2009; Woodard et al., 2009). Furthermore, the present study demonstrated a marginally significant interaction between side and APOE inheritance. Compared to the left, the right hippocampus was 3% larger in ε4-positive and 7% larger in ε4-negative participants. Additionally, compared to ε4-negative participants, the left hippocampus was 1% larger, but the right hippocampus was 4% smaller in ε4-positive participants. These marginally significant findings may build upon previous studies that have not found early APOE risk differences (Burggren et al., 2002; Jak et al., 2007), by suggesting that APOE may influence the right greater than left pattern that has been documented, even in asymptomatic individuals.

In addition to cognitively healthy samples, as was the case in the baseline comparison of the present study, the pattern of right greater than left hippocampal volume has been found in later stages of AD, with indications of smaller differences between the hemispheres as the disease progresses (Shi et al., 2009). We previously demonstrated that individuals with MCI showed smaller left, and trending right, hippocampal volumes compared to those at risk and healthy controls, supporting the finding that the right hippocampus was larger than the left hippocampus (Woodard et al., 2009). Also consistent with the present study, a large-scale clinical trial of individuals with probable AD demonstrated right hippocampal volumes were approximately 5% larger than left hippocampal volumes (Jack et al., 2003). Taken together, data from the present study lend further support to the right greater than left hippocampal volume pattern reported in the literature, and further extend the possibility of laterality differences according to APOE inheritance in cognitively intact individuals. This may have important implications for understanding the role of APOE and its potential impact on hippocampal volume and structure, which may be contributing to the imposed risk for AD development.

Laterality by method with baseline hippocampal volumes.

In addition to showing overall right greater than left baseline hippocampal volume differences, the present study demonstrated that this pattern is discernable with both FS and MT. This finding is consistent with a recent paper directly comparing MT and FS in a sample of individuals with AD, MCI, cognitive complaints with normal neuropsychological data, and healthy controls. In this previous study, both MT and FS detected larger right hippocampal volumes relative to left hippocampal volumes across all groups (Shen et al., 2010). Collectively, these data may suggest that regardless of which method is used to segment the hippocampus, right hippocampal volumes will likely be larger than left hippocampal volumes. This may have research and clinical utility in that, if lateral volume differences are expected, FS may be just as reliable as MT in detecting the relative difference to the left hippocampus. Additionally, the laterality differences that are detected by both FS and MT may also be important as disease progresses. That is, if the left hippocampus is smaller than the right, then it may have a focally reduced amount of cognitive reserve, or a reduced ability to withstand disease insults relative to the right hippocampus. Given that memory impairment is a hallmark feature of AD (APA, 2000), and verbal memory is strongly represented by the left hippocampus (Reminger et al., 2004), it lends further support to the theory that the left hippocampus may be more susceptible to the early effects of dementia pathology (Thompson et al., 2007).

Notably, the way in which cognitive change was measured in the present study was dependent upon verbal memory performance, as well as a general cognitive screen that included verbal memory in addition to other cognitive domains. That the declining individuals did not show a relative laterality difference compared to stable individuals may suggest that the left hippocampus has not necessarily experienced atrophy, but rather the laterality differences that were detected reflect the general asymmetry observed in the hippocampus (Honeycutt & Smith, 1995).

Stable and Declining comparisons with baseline hippocampal volumes.

The hypothesis that Declining participants would have smaller baseline hippocampi than Stable participants, as measured by MT but not FS, was not supported in the present study. That is, the interaction between method and side did not differ between Stable and Declining participants, as expected, and neither FS nor MT baseline measurements differentiated Stable and Declining participants. This null finding is in contrast to an earlier study where we showed that individuals with smaller baseline hippocampal volumes were more likely to experience cognitive decline than individuals with larger baseline hippocampal volumes (Woodard et al., 2010). The present data are also in contrast to what we found in a previous analysis of unpublished data that showed MT, but not FS Version 5.0.0, detected smaller volumes in Declining participants compared to Stable participants.

There are a few key distinctions between the present study and our previous studies involving hippocampal volume differences in eventually Declining individuals. In both of these previous studies, cognitive decline was determined based on cognitive change from baseline to the 18-month follow-up compared to the 4.5-year follow-up in the present study. Furthermore, while some participants in the present analyses were in included in the previous studies, the overall sample in the present study was 33% smaller than the previous prediction study (Woodard et al., 2010) and 20% smaller than the previous study comparing MT and FS over the 18-month interval (unpublished data). Therefore, although the power analysis for the present study indicated sufficient power with the 60 participants in the present study, it is possible that this study suffered from low power, and as such was not able to detect group differences at baseline.

Additionally, the earlier studies used a different version of FS (V 5.0.0, released August 2010) and a different set of manual tracings. That is, because the present study is a separate experiment comparing a different version of FS to MT, all included tracings were unique to this study. As such, another explanation for the lack of reproducibility is a susceptibility of human variance associated with the manual tracings, which has been a critique of this segmentation method (Cherbuin et al., 2009). Although inter-rater reliabilities were strong with the primary tracer (ICC = 0.93 and 0.94), as was the intra-rater reliability (ICC = 0.95) to assess rater drift in the present study, it is possible that this set of hippocampal tracings varied somewhat from the tracings in the earlier study. Given that the primary tracer in the present study was also involved in the tracing of hippocampal volumes for the earlier study, and that the tracing method was identical, it is also likely that the differing results are due to a combination of the newer version of FS and a differently defined Decline group, with low power.

Yet another explanation is the potential for cognitive inconsistency across time points that may have resulted in different cognitive groupings. That is, there may be natural variability in cognitive assessments across individuals that may influence group distinction when adding in an additional follow-up assessment. Natural inconsistency in addition to a relatively narrow cognitive assessment, which included an administration error in one measure for a subset of participants, may also influence cognitive consistency and introduce variance into the participant groupings. Given that the sample size was relatively small, even a small amount of cumulative variance may have undermined the ability to detect group differences.

Although in the present study MT were not shown to identify presymptomatic individuals, this method has been shown to detect volume difference in predementia individuals (Kaye et al., 1997). That is, an earlier study found smaller baseline hippocampal volumes in individuals who later decline compared to those who remain stable (Kaye et al., 1997), which suggest that the present study should have detected group differences. Yet, there are several notable differences between the Kaye et al. study and the present one that may account for the inconsistencies. First, although the resolution between scans were similar, the original Kaye et al. study used a 1.5T with 4 mm slices, while the present study used a 3T with 1 mm slices. It is possible that with 4mm slices, Kaye et al. was more vulnerable to partial voluming effects, or the occurrence of including tissue belonging to adjacent structures that were partially, but not fully representative of hippocampal volume. Partial voluming may have been more likely to occur in the group with more poorly differentiated anatomy due to disease than the control group. Additionally, the predementia group differed in age and cognition at the baseline assessment in Kaye et al., whereas the Stable and Declining groups in the present study were not different across these variables at baseline. Furthermore, the participants in the Kaye et al. study were older at study entry (at least 84 years old, with a mean age of 92), whereas those in the present study were at least 65 years old at study entry (mean age of 72 years). This may suggest that the present study is asking a similar theoretical question about two decades prior to the sample in the Kaye et al. study.

While the present study did not find group differences in presymptomatic individuals, studies have shown that FS and MT are able to distinguish between symptomatic and asymptomatic individuals. The present study is consistent with Lehmann et al. (2010) who documented larger FS volumes relative to MT in AD, Semantic dementia, and controls, but Lehmann et al. also found both methods to correctly classify the different groups. Similarly, a study comparing FS and MT in MCI, AD, and controls found 10% larger FS than MT volumes, with acceptable interchangeability between the two methods as demonstrated by Bland-Altman plots (Sanchez-Benavides et al., 2010). Further still, 35% larger FS volumes, with greater associated variance, have been documented in other clinical populations that were distinguishable from control groups by both methods (Morey et al., 2009; Tae et al., 2008). These studies are consistent with the present study in showing acceptable agreement between the two methods as demonstrated by Bland-Altman plots, yet the present study found 52% larger FS compared to MT volumes, with 35% more variance associated with FS than MT. Therefore, while the prior studies with clinical populations demonstrated that both MT and FS distinguished between symptomatic and asymptomatic individuals, the present study showed that neither method distinguished presymptomatic individuals. The present study, then, extends the findings of these earlier studies with symptomatic individuals by showing that while absolute volumes may differ between FS and MT, neither method may be able to reliably distinguish between eventual declining from stable individuals while they are in the presymptomatic stage, at least given the parameters of the present study.

This study was one of the few to compare FS and MT in presymptomatic individuals, and the inability to distinguish between eventual declining and stable participants here may suggest that the effect, if present is very small and difficult to reproduce. The current results are consistent with a recent study that found both FS and MT were able to correctly differentiate aMCI and AD participants from cognitively intact individuals, but found no differences between intact individuals who had cognitive complaints and intact individuals without complaints (Shen et al., 2010). The authors suggest that the cognitive complaints group may mimic a “preMCI” group, which may lend to the theory that those with impending cognitive decline may not show hippocampal volume differences compared to other cognitively intact individuals prior to measurable cognitive change. The Shen et al. study though is entirely cross-sectional; it is not known whether the participants with cognitive complaints eventually exhibited more cognitive decline over time than those without complaints. Thus, the present study builds upon their study by incorporating longitudinal follow-up. The lack of hippocampal volume difference in eventual declining and stable individuals suggests that such volume differences may not appear in the presymptomatic stage. It may be that measurable symptoms, whether cognitive and or behavioral, must be present to have associative detectable volume changes.

The influence of APOE inheritance.

Another explanation as to why group differences were not found in the present data may be related to the influence of the APOE allele on hippocampal volume. We have previously shown that the ability to predict cognitive decline at an 18-month follow-up assessment with a combination of structural and functional neuroimaging variables is enhanced when including APOE inheritance as a predictive variable (Hantke et al., 2013; Woodard et al., 2010). Unlike these previous prediction studies that aimed to predict decline with a number of variables, the goal of the present study was to directly compare two different methods of hippocampal volume measurement within the context of differentiating eventual stable and declining participants. Therefore, assessing the contribution of APOE in the present study was not the primary aim, but APOE was examined as it might influence the variables of interest. As such, a number of analyses incorporating APOE as an additional between subjects factor were conducted to explore the contribution of APOE in the present study. Indeed, a larger number of individuals in the Declining group had at least one copy of the ε4 allele (68%) compared to the Stable group (31%). This disproportional distribution of the ε4 allele is consistent with the reported risk of cognitive decline procured by the presence of even just one ε4 allele (Saunders et al., 1993).

When entered into statistical analyses as an additional between subjects variable, a marginally significant interaction was found between Method, Side, Decline status and APOE with baseline hippocampal volumes. In particular, this marginally significant interaction suggested that ε4-negative Stable participants showed the laterality effect (right > left) when measured by both methods, and ε4-negative Declining participants (*n* =5) had larger left hippocampal volumes as measured by FS only. Additionally, ε4-positive Stable participants (*n* = 14) had larger right compared to left volumes as measured by FS, but not MT, whereas ε4-positive Declining participants (*n* = 10) did not differ. Given the resulting small cell sizes in some subgroups, there was likely not sufficient power to fully detect differences, particularly with the relatively large amount variance of the smaller Declining group (SE = 88.62; as compared to the Stable group, SE = 52.10). Despite this, these data suggest that there likely is a valuable interaction with APOE, such that it may interact with the left and right hippocampi differently in presymptomatic individuals, and MT and FS may uniquely detect this interaction. To comprehensively explore this interaction, additional participants are needed to represent the differing subgroups, and as such, the present data would serve as a strong set of pilot data to infer future hypotheses.

Because the present study was likely limited by small cell sizes, a number of hierarchical regressions were conducted to examine factors associated with cognition without the restrictions imposed by grouping variables. The present study found that baseline hippocampal volume and APOE could predict the change in RAVLT Delay performance from baseline to the 4.5-year follow-up, but not RAVLT 1-5 or DRS-2 Total performance. That is, FS left and right volumes and MT total volume both were marginally significant predictors of RAVLT Delay, but both models significantly benefitted from the addition of APOE. Therefore, it may be that at baseline, at least when participants are asymptomatic, APOE inheritance has a stronger influence on future cognitive outcome than hippocampal size alone. Therefore, to increase the ability to predict cognitive decline, hippocampal volume alone may not be sufficient, and other risk factors, such as APOE and perhaps other brain regions, may provide additive predictive ability.

The relationship between APOE inheritance and hippocampal volume has been well documented and explored given the connection with AD pathology, such as the beta amyloid accumulation. Molecular studies have shown that the ε4 allele has a stronger association with increased amounts of amyloid plaques relative to the ε2 and ε3 alleles (Mahley & Huang, 2012). The ε4 allele has also been implicated in affecting proper and sufficient clearance of beta amyloid, perhaps contributing to large accumulation in those individuals with AD (Bien-Ly, Gillespie, Walker, Yoon, & Huang, 2012). Additionally, the ε4 allele is thought to be involved in alterations of synaptic activity, mitochondrial functioning, as well as the cytoskeletal structure and integrity of the cell (Mahley & Huang, 2012). Together, the strong relationship between APOE and hippocampal structure is expected. Indeed, ε4 carriers have been shown in some studies to have smaller hippocampal volumes and an excelled rate of atrophy relative to non-carriers (Jak et al., 2007; Schuff et al., 2009), perhaps due to the additive impact of these alterations in cell characteristics. Overall, APOE and in particular the ε4 allele, has been consistently shown to be a valuable indicator of change to the function and structure of the hippocampus.

Together, these data suggest that APOE may have a strong and unique involvement in detecting future cognitive decline and may interact with the method used. This may suggest that hippocampal volume measurement, along with a combination of other factors, such as APOE inheritance, may provide the strongest predictive ability. Furthermore, although there were absolute volume differences between FS and MT, in the present study, group differences were not detectable with either method. Therefore, within this context, MT was not found to afford any advantage over FS.

Aim 2: Longitudinal Change in Hippocampal Volume by Group and Side

To supplement the investigation of differences in baseline hippocampal volumes between future declining and stable individuals, the second aim was to investigate atrophy over time. More specifically, it was expected that the rate of hippocampal atrophy from the baseline to 4.5-year follow-up assessment, and at each time point, would differ between FS and MT, such that MT were expected to show a greater amount of atrophy between groups compared to FS.

Atrophy measured longitudinally between groups.

As expected, hippocampal volumes were smaller at the 18-month and 4.5-year follow-up assessments compared to the baseline assessment, demonstrating an overall trend of atrophy over time. Furthermore, there was a greater amount of atrophy occurring between the 18-month to 4.5-year follow-up assessment compared to the baseline to 18-month follow-up. Additionally, the present study demonstrated an interaction between time and decline status, such that Declining and Stable participants showed a different rate of atrophy over time. Compared to Stable participants, Declining participants showed a more progressive pattern of atrophy over time, differing at each follow-up assessment, with a more dramatic reduction in hippocampal volume at the 4.5 year follow-up and an overall greater amount of atrophy than the Stable group. This more drastic change in hippocampal volume reflects the decline in cognitive status at the 4.5-year assessment, suggesting that hippocampal volume is a useful indicator of cognitive decline.

These data are consistent with the literature showing hippocampal atrophy occurs over time at a rate of about 1.18% per year as a result of healthy aging (Raz et al., 2004), and about twice that rate in AD (Jack et al., 1998). The rate of hippocampal atrophy has previously been shown to reliably predict conversion from MCI to AD (Henneman et al., 2009; Visser et al., 2002). The results of the present study suggests that during an earlier stage of the disease process, when the individual is beginning to show mild cognitive symptoms, measurable hippocampal volume differences exist between the those who are declining and those who remain cognitively stable.

The finding that declining participants show a more progressive rate of atrophy is consistent with an earlier study that conducted serial MRI measurements in controls, MCI, and AD over time (Jack et al., 2000). At the 3-year follow-up assessment, they found the declining groups to demonstrate a greater rate of hippocampal atrophy relative to those who remained stable (Jack et al., 2000). Notably, the present study defined decline by a memory measure (RAVLT) and general cognitive status (DRS-Total Score), whereas Jack et al. used the MMSE and CDR to determined cognitive status. The similar findings occurring in the context of different measures used to define decline, suggest that the RAVLT and DRS-Total score are adequate measures of cognitive decline, and that a pattern of more progressive atrophy occurring in declining individuals should be reproducible with reliable cognitive measures. Additionally, the present study further extends the findings in Jack et al. by including a third time point. That is, the present study shows that not only do hippocampal volumes decrease over time, but that the amount of atrophy from the 18-month to 4.5 year follow-up is greater than the atrophy between baseline and the 18-month follow-up. This may suggest that the rate of atrophy becomes more progressive over time and is not necessarily linear. Yet, in the present study, the interval between the second and third scan was greater than the interval between the first and second scan, which may lead to a biased temporal gradient of hippocampal change over time. To account for the uneven intervals, an annual percent change of hippocampal volume was computed following the method of Jack et al. 2000. With this correction, the annual percent change of hippocampal volume of the Declining group was greater (1.60% per year) than the Stable group (.63% per year), but there was no annual difference between the two different intervals. That is, although the present data do not support a non-linear pattern of annual change, the data do support a greater rate of annual hippocampal atrophy in the Declining relative to the Stable participants.

A differing pattern of atrophy between cognitively healthy individuals who eventually decline and those who remain stable has not always been found. Kaye et al. (1997) followed cognitively healthy individuals over serial MRI measurements and showed that those who declined had a similar rate of atrophy than those who remained stable. One explanation for the inconsistent findings may be that the declining group had smaller volumes and were older to start out the study than the stable group, whereas in the present study groups were similar in age and hippocampal volume at baseline. That is, it may be that when volume differences are first emerging, the pattern of atrophy may be non-linear, and may stabilize over time as found in Kaye et al. Another possible explanation for the differences between the two findings may be related to the differences in age of the samples. Participants were an average of 72 years old in the present study and 92 years in Kaye et al. Therefore, it is possible that “declining” group captured and followed in Kaye et al. was different than the “declining” group in the present study. That is, a disease process that is structurally but not yet cognitively measureable at 92 years old, is likely different from a neurodegenerative disease process that is structurally detectable at 72 years-old. Therefore, a different trajectory of atrophy in two potentially differently disease processes and cognitive groups is not surprising.

FreeSurfer vs. manual tracing effects over time.

In addition to finding that cognitively declining individuals demonstrate a more progressive rate of atrophy over time, the present study showed that FS measurements detected an overall greater amount of atrophy than MT over the 4.5-year study. Although not initially expected, one possibility is that this finding reflects the greater amount of variance associated with FS relative to MT. Indeed, over the 4.5-year study, the variance associated with FS was 53% larger than the variance associated with MT, which is greater than the 35% greater variance in FS found at baseline. A larger amount of variance associated with FS has been shown in previous cross-sectional studies (Morey et al., 2009). Together, these data may suggest that the error associated with FS increases over time. One explanation for this finding may be that the anatomical boundaries become more poorly differentiated during this atrophic process, causing FS to have increasing difficulty in discerning proper boundaries. Manual tracings may have less associated variance because this technique may benefit from a human perspective and rational knowledge of adjacent anatomical structures in defining proper boundaries. Yet, another possibility is that MT may be too conservative in its definition of hippocampal tissue. Indeed, a qualitative review of one Declining and one Stable participant suggests that the two methods do not agree on boarder definition and perhaps in the more anterior portion of the hippocampus. If FS is quantifying atrophy of hippocampal tissue, in addition to adjacent regions that also may be losing tissue, then the overall amount of atrophy detected by FS would be greater than that detected by the more conservative MT. In that sense, the larger volumes produced by FS, may be leading to the greater amount of atrophy observed over time.

Of course, it is also possible that FS may actually be a more sensitive technique in detecting atrophy over time and that MT may be too conservative of an approach. That is, the voxels included in FS and not MT quantification may be just as sensitive, if not more sensitive, in differentiating future cognitive decline. Thus, it is also possible that the larger volumes associated with FS is beneficial rather than detrimental. Indeed, in the present data, MT was not able to distinguish between future declining individuals. Therefore, it may be beneficial to trace the regions FS includes to determine if MT is also detecting change, in order to better understand why FS is showing greater atrophy over time than MT. This analysis may be beneficial for the continuing refinement of FS, and may lend to the further support of FS in large-scale research and potentially clinical settings.

Differences between groups over time.

Despite greater associated error and consistently larger volumes, FS did not differ from MT in the ability to differentiate stable and declining individuals at each time point, or across intervals, which is inconsistent with what was anticipated. Individuals who declined were hypothesized to differ from those who were stable as measured by MT because of additional refinement, but not by FS because of the tendency to yield larger volumes. Furthermore, in contrast to expectation, Declining and Stable groups did not differ from one another at any time point. It is possible that group differences at each time point were not detected because the differences were not yet large enough to statistically detect. An interaction with method may not have been found because FS and MT similarly measured the pattern of hippocampal size in each group. That is, Declining participants may have had non-significantly smaller hippocampal volume than Stable participants, and FS volumes were larger than MT across groups, thus potentially washing out a significant interaction effect. An alternative explanation could be that structural differences did not exist between groups. Because the variance in the smaller Declining group was 73% greater than the variance in the Stable group, it was likely difficult to detect group differences with either method, even if they did exist. An examination for outliers revealed one person who was slightly under the 2 SD below the mean cut off score for manually traced volumes at the 4.5-year follow-up, but no other potential outliers were discovered. This suggests that there was not a strong influence of outliers contributing to the large amount of observed variance in the Declining group inhibiting the ability to detect group differences.

As with the first aim, because there were concerns related to power due to a small Declining group, hierarchical regressions were conducted to determine if hippocampal volume and APOE could predict cognitive scores. For both RAVLT 1-5 and DRS-2 Total, adding APOE did not necessarily enhance the relationship, but APOE was beneficial when added to RAVLT Delay performance. In comparison to baseline hippocampal measurements, the 4.5-year volumes were more strongly related to cognitive performance, often even in the absence of APOE inheritance. Furthermore, even though there was a strong relationship at 4.5 years, baseline hippocampal volumes were more heavily dependent upon APOE and did not adequately predict this future pattern of cognitive performance. This may suggest that structural changes may occur in parallel to cognitive changes.

Although group differences by method were not found, there were some notable strengths of the present study. In particular, the longitudinal study design within an aging population using a combination of cognitive and advanced neuroimaging procedures is relatively rare. Even within a longitudinal paradigm, the capability of assessing hippocampal volume at more than two time points is rare and is called for when possible (Henneman et al., 2009), because serial MRI and cognitive assessments allow for variations in how the data may be explored and understood. Additionally, the cognitive assessment in the present study is believed to capture a good representative sample of cognitive decline over time. Associating structural changes to cognitive change is of particular value, given that cognitive status is necessary to interpret the meaningfulness of structural changes (Jak et al., 2007). Relatedly, longitudinal approaches are of statistical value in that each individual serves as his or her own control. The results of this study suggests that if detecting subtle group differences is the aim in lengthy longitudinal studies, MT may not provide reliable benefit beyond the more cost-effective FS.

Limitations

There also are several notable limitations to the present study that may have influenced the results and the subsequent conclusions. First, clearly power may have been a limitation of this study. Despite the power analysis and previous studies supporting the use of this sample size, the variance observed in the Declining group in particular, suggests that this study may have suffered from low power. Insufficient power increases the chances of failing to reject the null hypothesis when it is false. Specifically for this study, group differences were not observed between stable and declining participants across methods, but it is possible that differences were not observed because the sample size was too small and the variance was too great.

Additionally, as mentioned in the methods section, test administration errors were made in which 10 participants were given the same version of the RAVLT at the 18-month and the 4.5-year follow-up. This is a limitation due to the potential for practice effects and contamination of the groups. Data from these 10 participants were reviewed and two of the 10 participants met criteria for "Declining," whereas the remaining eight participants met criteria for "Stable." As proposed, all 10 participants were retained in analyses to maintain power. Some data have shown that elderly individuals given the same form yearly for three years demonstrate little practice effect (Mitrushina & Satz, 1991). Yet, a previous analysis of this administration error indicated that those who received the same form at the 18-month and 4.5- year follow-up within our larger sample of research participants benefited from practice effects for both RAVLT 1-5 and RAVLT Delay (unpublished data). As such, it is very possible that the eight individuals who were classified as “Stable” in this study may have benefited enough from practice effects to be classified as Stable, when they otherwise would have performed more consistent with our definition of “Decline.” This likelihood of group contamination could limit the ability to detect group effects.

It may also be important to track which measures are detecting a decline in cognition. A qualitative review of the Declining group revealed two out of the 15 participants showed a clear decline, by declining on all three outcome measures. Four out of 15 declined on two outcome measures, and the remaining nine participants declined on one measure. Additionally, nine participants declined on the RAVLT Delay, eight on the RAVLT 1-5, and six on the DRS-2. Together, these data suggest that the current Declining group may be comprised of individuals who experienced only a subtle reduction in cognitive performance, and that the RAVLT may be a slightly more sensitive measure to this subtle decline. It may be that more group differences would be reliably detected if there were more individuals in the Declining group that experienced a more drastic decline.

Although the change in cognitive performance is a measurable and meaningful amount of decline, it is important to note that the reduction, for the majority in the group, is subtle and may not represent a large change in functional status. That is, in a clinical setting, the subtle change detected in about 10 of the participants in the Declining group may result in a caution to follow-up to monitor for further cognitive change, but perhaps no clinical diagnosis. However, the more drastic decline detected in the two participants who declined on all three measures, may warrant a closer assessment of possible accompanying functional decline and consideration of clinical diagnosis. Although there is no reliable treatment for a neurodegenerative process, there are some medications that have shown some benefit of slowing progression of cognitive decline for some individuals (Dantoine et al., 2006; Kozauer & Katz, 2013; Tariot et al., 2001). If biomarkers were able to reliably detect future cognitive decline, it may be useful to consider a trial of the available medications to assess individual response to treatment. As more targets for intervention become available, perhaps more reliable and effective interventions may be generated to offer to those at risk for or at the early stages of cognitive decline.

Another measurement consideration is that way in which cognition was assessed in the present study to define cognitive status. In the present study, a verbal learning task (RAVLT) and screen of several cognitive domains (DRS-2) were used to assess cognition. It is possible that only using these measures resulted in an incomplete cognitive assessment, such that some individuals may have declined in a manner or cognitive domain that was not thoroughly assessed. It is possible, then, that some individuals who experienced a decline in cognition over the 4.5-year interval remained in the Stable instead of Declining group. While a more comprehensive neuropsychological assessment may have been beneficial in some respects, using these measures and this definition of decline, we have previously shown group differences in cognitively intact individuals who eventually decline compared to those who remain stable (Hantke et al., 2013; Woodard et al., 2010).

Another potential limitation is the use of FS as a mask on which the manual tracings were edited. That is, it is possible that beginning with an automated FS mask may have introduced the potential for inherent bias in manual tracings. However, this concern was addressed in the following ways. First, the mask offered a starting point, wherein voxels were both subtracted *and* added as needed to adequately cover hippocampal tissue. Some form of automated method to help isolate the hippocampus is typical as a starting point for any manual tracing approach. Second, beginning with a mask generated by FS likely introduced the same amount of bias as any other automated method would have introduced. More, if an alternative method were used for this purpose, analyses would have been needed to determine its effects on the results. Because the goal of this study was to determine whether FS or manual tracing was better able to distinguish future and gradual cognitive decline, comparing unedited FS measurements with manual tracings that were edits of FS seemed the most parsimonious approach. Importantly, all raters were blinded to session number, participant number, genetic risk status, and neuropsychological testing performance to prevent biases in tracings introduced by the raters. This method of manual hippocampal measurement has been used and accepted in previous publications (Hantke et al., 2013; Woodard et al., 2010).

Future Directions

The results of the present study provide a framework from which additional research may build upon in further investigation of useful biomarkers of AD. The present study used archival data to test out the presented hypotheses. Since this study, additional participants have undergone the 4.5-year follow-up scan and neuropsychological assessment. A necessary and logical follow-up study would be to use the additional participants that have been collected to comprehensively explore this potential interaction between FS and MT hippocampal volume measurement and cognitive decline, along with APOE ε4 inheritance, to determine if this type of structural analysis may be a useful indicator of future cognitive decline. In particular, there were indications that a valuable 4-way interaction between Method, Side, Decline, and APOE at baseline may emerge as significant if additional power was procured to test out this interaction. Additionally, the individuals that were administered the repeat RAVLT form should be re-evaluated to obtain a more accurate representation of their current cognitive functioning. Indeed, the participants that were administered the incorrect RAVLT form at the 4.5-year follow-up are projected to return for a repeat assessment of memory function that will ideally clarify the Stable and Declining groups.

The present study suggests that detecting future cognitive decline, and particularly attempting to capture the neuropathology of AD at an early stage, is complex and may require cross-sectional and longitudinal paradigms, as well as a combination of indicators. That is, the present data suggest that APOE may be useful in increasing the ability to detect future cognitive decline, and it may be important to explore the additive benefit of incorporating additional memory facilitating brain regions in future volumetric studies. In particular, the function and structure of the posterior cingulate (Protas et al., 2013), as well as the caudate and thalamus (Ryan et al., 2013), have been recently identified as regions that may be involved in the early stages of AD. Future studies should consider combining the impact of changes in these regions, which may help to improve the ability to distinguish future decline and perhaps eventually predict AD.

General Conclusion

The present study sought to compare two commonly used hippocampal segmentation techniques in a structural analysis of distinguishing between future declining from stable individuals in a sample of cognitively healthy elders. The scope of the present study is in line with the recent call for further exploration of predictive factors in individuals who later develop Alzheimer's disease (Albert et al., 2011; Sperling et al., 2011). Hippocampal volume has been raised as a potential factor in predicting decline, even in presymptomatic elders (Woodard et al., 2010).

Taken together, the present study did not find detectable baseline hippocampal volume differences in asymptomatic individuals who later exhibit cognitive decline. However, over time, Declining participants did show a more progressive amount of atrophy, and this pattern was detectable by FS and MT alike. These data speak to the usefulness of combining static measurements with the longitudinal measurements, because the overall pattern may be more meaningful. That is, although Declining participants did not have smaller hippocampal volumes at any time point, as was hypothesized, they did show an overall more progressive rate of atrophy over time. Furthermore, although FS was consistently shown to produce larger volumes than MT, there were no statistical differences between MT and FS in the ability to distinguish decline status. Therefore, because MT is vastly more time consuming and less practical, and although FS produced larger volumes than MT, MT may not be more beneficial than FS in detecting hippocampal atrophy over time. Future studies with a larger sample are needed to more comprehensively explore the impact of APOE status in hippocampal measurement and cognitive decline over time. A tool that allows for rapid and reliable hippocampal measurement would serve to benefit ongoing research efforts in the quest to predict Alzheimer's disease, and potentially confer a degree of clinical utility.

References

2011 Alzheimer's disease facts and figures. (2011). *West Virginia Medical Journal, 107*(3), 82-83.

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., . . . Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia, 7*(3), 270-279. doi: 10.1016/j.jalz.2011.03.008 [doi]

APA. (2000). *Diagnostic and Statistical Manual of Mental Disorders (IV-TR)* (4 ed.). Washington, D.C.: American Psychiatric Association.

Arendt, T. (2009). Synaptic degeneration in Alzheimer's disease. *Acta Neuropathologica, 118*(1), 167-179. doi: 10.1007/s00401-009-0536-x

Bien-Ly, N., Gillespie, A. K., Walker, D., Yoon, S. Y., & Huang, Y. (2012). Reducing human apolipoprotein E levels attenuates age-dependent Abeta accumulation in mutant human amyloid precursor protein transgenic mice. *Journal of Neuroscience, 32*(14), 4803-4811. doi: 10.1523/jneurosci.0033-12.2012

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135-160.

Bobholz, J. H., & Brandt, J. (1993). Assessment of cognitive impairment: relationship of the Dementia Rating Scale to the Mini-Mental State Examination. *Journal of Geriatric Psychiatry and Neurology, 6*(4), 210-213.

Bobinski, M., de Leon, M. J., Wegiel, J., Desanti, S., Convit, A., Saint Louis, L. A., . . . Wisniewski, H. M. (2000). The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in Alzheimer's disease. *Neuroscience, 95*(3), 721-725.

Bondi, M. W., Houston, W. S., Eyler, L. T., & Brown, G. G. (2005). fMRI evidence of compensatory mechanisms in older adults at genetic risk for Alzheimer disease. *Neurology, 64*(3), 501-508. doi: 10.1212/01.WNL.0000150885.00929.7E [doi]

Braak, H., & Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica, 82*(4), 239-259.

Buckner, R. L., Head, D., Parker, J., Fotenos, A. F., Marcus, D., Morris, J. C., & Snyder, A. Z. (2004). A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. *NeuroImage, 23*(2), 724-738. doi: 10.1016/j.neuroimage.2004.06.018

Burggren, A. C., Small, G. W., Sabb, F. W., & Bookheimer, S. Y. (2002). Specificity of brain activation patterns in people at genetic risk for Alzheimer disease. *American Journal of Geriatric Psychiatry, 10*(1), 44-51.

Burggren, A. C., Zeineh, M. M., Ekstrom, A. D., Braskie, M. N., Thompson, P. M., Small, G. W., & Bookheimer, S. Y. (2008). Reduced cortical thickness in hippocampal subregions among cognitively normal apolipoprotein E e4 carriers. *NeuroImage, 41*(4), 1177-1183. doi: 10.1016/j.neuroimage.2008.03.039 [doi]

Busse, A., Hensel, A., Guhne, U., Angermeyer, M. C., & Riedel-Heller, S. G. (2006). Mild cognitive impairment: long-term course of four clinical subtypes. *Neurology, 67*(12), 2176-2185. doi: 10.1212/01.wnl.0000249117.23318.e1 [doi]

Cabeza, R., & Nyberg, L. (1997). Imaging cognition: an empirical review of PET studies with normal subjects. *Journal of Cognitive Neuroscience, 9*(1), 1-26.

Celone, K. A., Calhoun, V. D., Dickerson, B. C., Atri, A., Chua, E. F., Miller, S. L., . . . Sperling, R. A. (2006). Alterations in memory networks in mild cognitive impairment and Alzheimer's disease: an independent component analysis. *Journal of Neuroscience, 26*(40), 10222-10231. doi: 10.1523/JNEUROSCI.2250-06.2006

Cherbuin, N., Anstey, K. J., Reglade-Meslin, C., & Sachdev, P. S. (2009). In vivo hippocampal measurement and memory: a comparison of manual tracing and automated segmentation in a large community-based sample. *Public Library of Science, 4*(4), e5265. doi: 10.1371/journal.pone.0005265 [doi]

Cohen, R. M., Small, C., Lalonde, F., Friz, J., & Sunderland, T. (2001). Effect of apolipoprotein E genotype on hippocampal volume loss in aging healthy women. *Neurology, 57*(12), 2223-2228.

Colucci, M., Cammarata, S., Assini, A., Croce, R., Clerici, F., Novello, C., . . . Tanganelli, P. (2006). The number of pregnancies is a risk factor for Alzheimer's disease. *European Journal of Neurology, 13*(12), 1374-1377. doi: 10.1111/j.1468-1331.2006.01520.x [doi]

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., . . . Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science, 261*(5123), 921-923.

Corrada, M. M., Paganini-Hill, A., Berlau, D. J., & Kawas, C. H. (2012). Apolipoprotein E genotype, dementia, and mortality in the oldest old: The 90+ Study. *Alzheimer's and Dementia*. doi: 10.1016/j.jalz.2011.12.004 [doi]

Costafreda, S. G., Dinov, I. D., Tu, Z., Shi, Y., Liu, C. Y., Kloszewska, I., . . . Simmons, A. (2011). Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. *NeuroImage, 56*(1), 212-219. doi: 10.1016/j.neuroimage.2011.01.050 [doi]

Cox, R.W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29*, 162-173.

Dantoine, T., Auriacombe, S., Sarazin, M., Becker, H., Pere, J. J., & Bourdeix, I. (2006). Rivastigmine monotherapy and combination therapy with memantine in patients with moderately severe Alzheimer's disease who failed to benefit from previous cholinesterase inhibitor treatment. *International Journal of Clinical Practice, 60*(1), 110-118. doi: 10.1111/j.1368-5031.2005.00769.x

DeKosky, S. T., & Marek, K. (2003). Looking backward to move forward: early detection of neurodegenerative disorders. *Science, 302*(5646), 830-834. doi: 10.1126/science.1090349 [doi]

Desikan, R. S., Cabral, H. J., Settecase, F., Hess, C. P., Dillon, W. P., Glastonbury, C. M., . . . Fischl, B. (2010). Automated MRI measures predict progression to Alzheimer's disease. *Neurobiology of Aging, 31*(8), 1364-1374. doi: 10.1016/j.neurobiolaging.2010.04.023 [doi]

Desikan, R. S., Fischl, B., Cabral, H. J., Kemper, T. L., Guttmann, C. R., Blacker, D., . . . Killiany, R. J. (2008). MRI measures of temporoparietal regions show differential rates of atrophy during prodromal AD. *Neurology, 71*(11), 819-825. doi: 10.1212/01.wnl.0000320055.57329.34

Dickerson, B. C., Salat, D. H., Greve, D. N., Chua, E. F., Rand-Giovannetti, E., Rentz, D. M., . . . Sperling, R. A. (2005). Increased hippocampal activation in mild cognitive impairment compared to normal aging and AD. *Neurology, 65*(3), 404-411. doi: 10.1212/01.wnl.0000171450.97464.49

Dubois, B., Feldman, H. H., Jacova, C., Dekosky, S. T., Barberger-Gateau, P., Cummings, J., . . . Scheltens, P. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *The Lancet Neurology, 6*(8), 734-746. doi: 10.1016/S1474-4422(07)70178-3 [doi]

Duvernoy, Henri, M. (2005). *The Human Hippocampus: Functional Anatomy, Vascularization and Serial Sections with MRI.* (Third ed.). New York: Springer.

Estevez-Gonzalez, A., Kulisevsky, J., Boltes, A., Otermin, P., & Garcia-Sanchez, C. (2003). Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer's disease: comparison with mild cognitive impairment and normal aging. *International Journal of Geriatric Psychiatry, 18*(11), 1021-1028. doi: 10.1002/gps.1010

Farrer, L. A., Cupples, L. A., Haines, J. L., Hyman, B., Kukull, W. A., Mayeux, R., . . . van Duijn, C. M. (1997). Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *Journal of the American Medical Association, 278*(16), 1349-1356.

Fischl, B. (2012). FreeSurfer. *NeuroImage, 62*(2), 774-781. doi: 10.1016/j.neuroimage.2012.01.021

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

Fischl, B., Liu, A., & Dale, A. M. (2001). Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Transactions of Medical Imaging, 20*(1), 70-80. doi: 10.1109/42.906426 [doi]

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., . . . Dale, A. M. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron, 33*(3), 341-355.

Fjell, A. M., Walhovd, K. B., Fennema-Notestine, C., McEvoy, L. K., Hagler, D. J., Holland, D., . . . Dale, A. M. (2009). One-year brain atrophy evident in healthy aging. *Journal of Neuroscience, 29*(48), 15223-15231. doi: 10.1523/JNEUROSCI.3252-09.2009 [doi]

Folstein, M F, Folstein, S E, & McHugh, P R. (1975). "Mini-Mental State": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*, 189-198.

Fox, N. C., Warrington, E. K., Freeborough, P. A., Hartikainen, P., Kennedy, A. M., Stevens, J. M., & Rossor, M. N. (1996). Presymptomatic hippocampal atrophy in Alzheimer's disease. A longitudinal MRI study. *Brain, 119 (Pt 6)*, 2001-2007.

Gao, S., Hendrie, H. C., Hall, K. S., & Hui, S. (1998). The relationships between age, sex, and the incidence of dementia and Alzheimer disease: a meta-analysis. *Archives of General Psychiatry, 55*(9), 809-815.

Giedd, J. N., Vaituzis, A. C., Hamburger, S. D., Lange, N., Rajapakse, J. C., Kaysen, D., . . . Rapoport, J. L. (1996). Quantitative MRI of the temporal lobe, amygdala, and hippocampus in normal human development: ages 4-18 years. *Journal of Comparative Neurology, 366*(2), 223-230. doi: 10.1002/(SICI)1096-9861(19960304)366:2&lt;223::AID-CNE3&gt;3.0.CO;2-7

Grady, C. L., & Craik, F. I. (2000). Changes in memory processing with age. *Current Opinion in Neurobiology, 10*(2), 224-231.

Grady, C. L., McIntosh, A. R., & Craik, F. I. (2005). Task-related activity in prefrontal cortex and its relation to recognition memory performance in young and old adults. *Neuropsychologia, 43*(10), 1466-1481. doi: 10.1016/j.neuropsychologia.2004.12.016 [doi]

Green, R. C., Cupples, L. A., Go, R., Benke, K. S., Edeki, T., Griffith, P. A., . . . Farrer, L. A. (2002). Risk of dementia among white and African American relatives of patients with Alzheimer disease. *Journal of American Medical Academy, 287*(3), 329-336.

Hantke, N., Nielson, K. A., Woodard, J. L., Breting, L. M., Butts, A., Seidenberg, M., . . . Rao, S. M. (2013). Comparison of semantic and episodic memory BOLD fMRI activation in predicting cognitive decline in older adults. *Journal of the International Neuropsychological Society, 19*(1), 11-21. doi: 10.1017/s1355617712000951

Hawkins, K. A., Dean, D., & Pearlson, G. D. (2004). Alternative forms of the Rey Auditory Verbal Learning Test: a review. *Behav Neurol, 15*(3-4), 99-107.

Hebert, L. E., Scherr, P. A., Bienias, J. L., Bennett, D. A., & Evans, D. A. (2003). Alzheimer disease in the US population: prevalence estimates using the 2000 census. *Archives of Neurology, 60*(8), 1119-1122. doi: 10.1001/archneur.60.8.1119 [doi]

Henneman, W. J., Sluimer, J. D., Barnes, J., van der Flier, W. M., Sluimer, I. C., Fox, N. C., . . . Barkhof, F. (2009). Hippocampal atrophy rates in Alzheimer disease: added value over whole brain volume measures. *Neurology, 72*(11), 999-1007. doi: 10.1212/01.wnl.0000344568.09360.31 [doi]

Herrup, K. (2010). Reimagining Alzheimer's disease--an age-based hypothesis. *Journal of Neuroscience, 30*(50), 16755-16762. doi: 10.1523/JNEUROSCI.4521-10.2010 [doi]

Hoff, P. R. & Mobbs, C.V. . (2009). *Handbook of the Neuroscience of Aging*. Burlington, MA: Academic Press.

Honeycutt, N. A., & Smith, C. D. (1995). Hippocampal volume measurements using magnetic resonance imaging in normal young adults. *Journal of Neuroimaging, 5*(2), 95-100.

Hong-Qi, Y., Zhi-Kun, S., & Sheng-Di, C. (2012). Current advances in the treatment of Alzheimer's disease: focused on considerations targeting Abeta and tau. *Transl Neurodegener, 1*(1), 21. doi: 10.1186/2047-9158-1-21

Hyman, B. T., Gomez-Isla, T., West, H., Briggs, M., Chung, H., Growdon, J. H., & Rebeck, G. W. (1996). Clinical and neuropathological correlates of apolipoprotein E genotype in Alzheimer's disease. Window on molecular epidemiology. *Annals of the New York Academy of Sciences, 777*, 158-165.

Jack, C. R., Jr., Albert, M. S., Knopman, D. S., McKhann, G. M., Sperling, R. A., Carrillo, M. C., . . . Phelps, C. H. (2011). Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia, 7*(3), 257-262. doi: 10.1016/j.jalz.2011.03.004 [doi]

Jack, C. R., Jr., Barkhof, F., Bernstein, M. A., Cantillon, M., Cole, P. E., Decarli, C., . . . Foster, N. L. (2011). Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. *Alzheimers Dement, 7*(4), 474-485 e474. 10.1016/j.jalz.2011.04.007 [doi]

Jack, C. R., Jr., Petersen, R. C., Xu, Y., O'Brien, P. C., Smith, G. E., Ivnik, R. J., . . . Kokmen, E. (2000). Rates of hippocampal atrophy correlate with change in clinical status in aging and AD. *Neurology, 55*(4), 484-489.

Jack, C. R., Jr., Petersen, R. C., Xu, Y., O'Brien, P. C., Smith, G. E., Ivnik, R. J., . . . Kokmen, E. (1998). Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. *Neurology, 51*(4), 993-999.

Jack, C. R., Jr., Slomkowski, M., Gracon, S., Hoover, T. M., Felmlee, J. P., Stewart, K., . . . Petersen, R. C. (2003). MRI as a biomarker of disease progression in a therapeutic trial of milameline for AD. *Neurology, 60*(2), 253-260.

Jak, A. J., Houston, W. S., Nagel, B. J., Corey-Bloom, J., & Bondi, M. W. (2007). Differential cross-sectional and longitudinal impact of APOE genotype on hippocampal volumes in nondemented older adults. *Dementia and Geriatric Cognitive Disorders, 23*(6), 382-389. doi: 10.1159/000101340 [doi]

Jellinger, K., Danielczyk, W., Fischer, P., & Gabriel, E. (1990). Clinicopathological analysis of dementia disorders in the elderly. *Journal of Neuroscience, 95*(3), 239-258.

Juottonen, K., Laakso, M. P., Partanen, K., & Soininen, H. (1999). Comparative MR analysis of the entorhinal cortex and hippocampus in diagnosing Alzheimer disease. *American Journal of Neuroradiology, 20*(1), 139-144.

Jurica, P.J., Leitten, C.L., & Mattis, S. (2001). *Dementia Rating Scale-2 professional manual*. Lutz, FL: Psychological Assessment Resources.

Karas, G. B., Scheltens, P., Rombouts, S. A., Visser, P. J., van Schijndel, R. A., Fox, N. C., & Barkhof, F. (2004). Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. *NeuroImage, 23*(2), 708-716. doi: 10.1016/j.neuroimage.2004.07.006 [doi]

Kassem, M. S., Lagopoulos, J., Stait-Gardner, T., Price, W. S., Chohan, T. W., Arnold, J. C., . . . Bennett, M. R. (2012). Stress-Induced Grey Matter Loss Determined by MRI Is Primarily Due to Loss of Dendrites and Their Synapses. *Molecular Neurobiology*. doi: 10.1007/s12035-012-8365-7

Kaye, J. A., Swihart, T., Howieson, D., Dame, A., Moore, M. M., Karnos, T., . . . Sexton, G. (1997). Volume loss of the hippocampus and temporal lobe in healthy elderly persons destined to develop dementia. *Neurology, 48*(5), 1297-1304.

Kilpatrick, C., Murrie, V., Cook, M., Andrewes, D., Desmond, P., & Hopper, J. (1997). Degree of left hippocampal atrophy correlates with severity of neuropsychological deficits. *Seizure, 6*(3), 213-218.

Knight, R. G., McMahon, J., Skeaff, C. M., & Green, T. J. (2007). Reliable Change Index scores for persons over the age of 65 tested on alternate forms of the Rey AVLT. *Archives of Clinical Neuropsychology, 22*(4), 513-518. doi: 10.1016/j.acn.2007.03.005

Kozauer, N., & Katz, R. (2013). Regulatory innovation and drug development for early-stage Alzheimer's disease. *New England Journal of Medicine, 368*(13), 1169-1171. doi: 10.1056/NEJMp1302513

Langenecker, S. A., & Nielson, K. A. (2003). Frontal recruitment during response inhibition in older adults replicated with fMRI. *NeuroImage, 20*(2), 1384-1392. doi: 10.1016/S1053-8119(03)00372-0 [doi]

Lawton, M. P., & Brody, E. M. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. *Gerontologist, 9*(3), 179-186.

Lehmann, M., Douiri, A., Kim, L. G., Modat, M., Chan, D., Ourselin, S., . . . Fox, N. C. (2010). Atrophy patterns in Alzheimer's disease and semantic dementia: a comparison of FreeSurfer and manual volumetric measurements. *NeuroImage, 49*(3), 2264-2274. doi: 10.1016/j.neuroimage.2009.10.056 [doi]

Mahley, R. W., & Huang, Y. (2012). Apolipoprotein e sets the stage: response to injury triggers neuropathology. *Neuron, 76*(5), 871-885. doi: 10.1016/j.neuron.2012.11.020

Martins, C. A., Oulhaj, A., de Jager, C. A., & Williams, J. H. (2005). APOE alleles predict the rate of cognitive decline in Alzheimer disease: a nonlinear model. *Neurology, 65*(12), 1888-1893. doi: 10.1212/01.wnl.0000188871.74093.12 [doi]

Mattis, S. (1988). *Dementia Rating Scale professional manual*. Odessa, Florida: Psychological Assessment Resources.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Jr., Kawas, C. H., . . . Phelps, C. H. (2011). 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. *Alzheimer's and Dementia, 7*(3), 263-269. doi: 10.1016/j.jalz.2011.03.005 [doi]

McSweeny, J. A., Naugle, R. I., Chelune, G. J., Luders, H. . (1993). "T Scores for change": An illustration of a regression approach to depicting change in clinical neuropsychology. *Clinical Neuropsychologist, 7*(3), 300-312. doi: 10.1080/13854049308401901

Minino, A. M., Arias, E., Kochanek, K. D., Murphy, S. L., & Smith, B. L. (2002). Deaths: final data for 2000. *National Vital Statistics Reports, 50*(15), 1-119.

Mitrushina, M., & Satz, P. (1991). Effect of repeated administration of a neuropsychological battery in the elderly. *Journal of Clinical Psychology, 47*(6), 790-801.

Morey, R. A., Petty, C. M., Xu, Y., Hayes, J. P., Wagner, H. R., 2nd, Lewis, D. V., . . . McCarthy, G. (2009). A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. *NeuroImage, 45*(3), 855-866. doi: 10.1016/j.neuroimage.2008.12.033

Morris, J. C. (2005). Early-stage and preclinical Alzheimer disease. *Alzheimer Disease and Associated Disorders, 19*(3), 163-165.

Musicco, M. (2009). Gender differences in the occurrence of Alzheimer's disease. *Functional Neurology, 24*(2), 89-92.

Nielson, K. A., Douville, K. L., Seidenberg, M., Woodard, J. L., Miller, S. K., Franczak, M., . . . Rao, S. M. (2006). Age-related functional recruitment for famous name recognition: an event-related fMRI study. *Neurobiology of Aging, 27*(10), 1494-1504.

Nielson, K. A., Langenecker, S. A., & Garavan, H. (2002). Differences in the functional neuroanatomy of inhibitory control across the adult life span. *Psychology of Aging, 17*(1), 56-71.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*(1), 97-113.

Perneczky, R., Drzezga, A., Diehl-Schmid, J., Li, Y., & Kurz, A. (2007). Gender differences in brain reserve : an (18)F-FDG PET study in Alzheimer's disease. *Journal of Neurology, 254*(10), 1395-1400. doi: 10.1007/s00415-007-0558-z

Petersen. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine, 256*(3), 183-194. doi: 10.1111/j.1365-2796.2004.01388.x [doi]

Petersen, Doody, R., Kurz, A., Mohs, R. C., Morris, J. C., Rabins, P. V., . . . Winblad, B. (2001). Current concepts in mild cognitive impairment. *Archives of Neurology, 58*(12), 1985-1992.

Petersen, & Morris, J. C. (2003). Clinical Features. In R. C. Petersen (Ed.), *Mild Cognitive Impairment: Aging to Alzheimer's Disease*. New York: Oxford University Press, Inc.

Petrella, J. R., Coleman, R. E., & Doraiswamy, P. M. (2003). Neuroimaging and early diagnosis of Alzheimer disease: a look to the future. *Radiology, 226*(2), 315-336.

Protas, H. D., Chen, K., Langbaum, J. B., Fleisher, A. S., Alexander, G. E., Lee, W., . . . Reiman, E. M. (2013). Posterior cingulate glucose metabolism, hippocampal glucose metabolism, and hippocampal volume in cognitively normal, late-middle-aged persons at 3 levels of genetic risk for Alzheimer disease. *Journal of the American Medical Association Neurology, 70*(3), 320-325. doi: 10.1001/2013.jamaneurol.286

Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., Williamson, A., . . . Acker, J. D. (2005). Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. *Cerebral Cortex, 15*(11), 1676-1689. doi: 10.1093/cercor/bhi044

Raz, N., Rodrigue, K. M., Head, D., Kennedy, K. M., & Acker, J. D. (2004). Differential aging of the medial temporal lobe: a study of a five-year change. *Neurology, 62*(3), 433-438.

Reminger, S. L., Kaszniak, A. W., Labiner, D. M., Littrell, L. D., David, B. T., Ryan, L., . . . Kaemingk, K. L. (2004). Bilateral hippocampal volume predicts verbal memory function in temporal lobe epilepsy. *Epilepsy and Behavior, 5*(5), 687-695. doi: 10.1016/j.yebeh.2004.06.006

Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: a robust approach. *NeuroImage, 53*(4), 1181-1196. doi: 10.1016/j.neuroimage.2010.07.020

Rey, A. (1958). *L'examen clinique en psychologie*. Paris: Presses Universitaires de France.

Roses, A. D., Lutz, M. W., Amrine-Madsen, H., Saunders, A. M., Crenshaw, D. G., Sundseth, S. S., . . . Reiman, E. M. (2009). A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. *Pharmacogenomics Journal, 10*(5), 375-384. doi: 10.1038/tpj.2009.69 [doi]

Ryan, N. S., Keihaninejad, S., Shakespeare, T. J., Lehmann, M., Crutch, S. J., Malone, I. B., . . . Fox, N. C. (2013). Magnetic resonance imaging evidence for presymptomatic change in thalamus and caudate in familial Alzheimer's disease. *Brain, 136*(Pt 5), 1399-1414. doi: 10.1093/brain/awt065

Salthouse, T. A., Hancock, H. E., Meinz, E. J., & Hambrick, D. Z. (1996). Interrelations of age, visual acuity, and cognitive functioning. *Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 51*(6), P317-330.

Sanchez-Benavides, G., Gomez-Anson, B., Sainz, A., Vives, Y., Delfino, M., & Pena-Casanova, J. (2010). Manual validation of FreeSurfer's automated hippocampal segmentation in normal aging, mild cognitive impairment, and Alzheimer Disease subjects. *Psychiatry Research, 181*(3), 219-225. doi: 10.1016/j.pscychresns.2009.10.011

Sattler, C., Toro, P., Schonknecht, P., & Schroder, J. (2012). Cognitive activity, education and socioeconomic status as preventive factors for mild cognitive impairment and Alzheimer's disease. *Psychiatry Research, 196*(1), 90-95. doi: 10.1016/j.psychres.2011.11.012

Saunders, Hulette, O., Welsh-Bohmer, K.A., Schmechel, D.E., Crain, B., Burke, J.R., . . . C. (1996). Specificity, sensitivity, and predictive value of apolipoprotein-E genotyping for sporadic Alzheimer's disease. *Lancet, 348*(9020), 90-93.

Saunders, Schmader, K., Breitner, J. C., Benson, M. D., Brown, W. T., Goldfarb, L., . . . et al. (1993). Apolipoprotein E epsilon 4 allele distributions in late-onset Alzheimer's disease and in other amyloid-forming diseases. *Lancet, 342*(8873), 710-711.

Scheff, S. W., Price, D. A., Schmitt, F. A., DeKosky, S. T., & Mufson, E. J. (2007). Synaptic alterations in CA1 in mild Alzheimer disease and mild cognitive impairment. *Neurology, 68*(18), 1501-1508. doi: 10.1212/01.wnl.0000260698.46517.8f

Scher, A. I., Xu, Y., Korf, E. S., White, L. R., Scheltens, P., Toga, A. W., . . . Launer, L. J. (2007). Hippocampal shape analysis in Alzheimer's disease: a population-based study. *NeuroImage, 36*(1), 8-18. doi: 10.1016/j.neuroimage.2006.12.036 [doi]

Schmidt, K. S., Mattis, P. J., Adams, J., & Nestor, P. (2005). Alternate-form reliability of the Dementia Rating Scale-2. *Archives of Clinical Neuropsychology, 20*(4), 435-441. doi: 10.1016/j.acn.2004.09.011

Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L. M., Trojanowski, J. Q., . . . Weiner, M. W. (2009). MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain, 132*(Pt 4), 1067-1077. doi: 10.1093/brain/awp007

Segonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., & Fischl, B. (2004). A hybrid approach to the skull stripping problem in MRI. *NeuroImage, 22*(3), 1060-1075. doi: 10.1016/j.neuroimage.2004.03.032

Segonne, F., Pacheco, J., & Fischl, B. (2007). Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Translational Medical Imaging, 26*(4), 518-529. doi: 10.1109/TMI.2006.887364 [doi]

Seidenberg, M., Guidotti, L., Nielson, K. A., Woodard, J. L., Durgerian, S., Zhang, Q., . . . Rao, S. M. (2009). Semantic knowledge for famous names in mild cognitive impairment. *Journal of the International Neuropsychological Society, 15*(1), 9-18. doi: 10.1017/S1355617708090103 [doi]

Seshadri, S. (2006). Elevated plasma homocysteine levels: risk factor or risk marker for the development of dementia and Alzheimer's disease? *Journal of Alzheimer's Disease, 9*(4), 393-398.

Shen, L., Saykin, A. J., Kim, S., Firpi, H. A., West, J. D., Risacher, S. L., . . . Flashman, L. A. (2010). Comparison of manual and automated determination of hippocampal volumes in MCI and early AD. *Brain Imaging and Behavior, 4*(1), 86-95. doi: 10.1007/s11682-010-9088-x [doi]

Shi, F., Liu, B., Zhou, Y., Yu, C., & Jiang, T. (2009). Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: Meta-analyses of MRI studies. *Hippocampus, 19*(11), 1055-1064. doi: 10.1002/hipo.20573

Simic, G., Kostovic, I., Winblad, B., & Bogdanovic, N. (1997). Volume and number of neurons of the human hippocampal formation in normal aging and Alzheimer's disease. *Journal of Comparative Neurology, 379*(4), 482-494.

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

Smith, G. E., Ivnik, R. J., Malec, J. F., Kokmen, E., Tangalos, E., & Petersen, R. C. (1994). Psychometric Properties of the Mattis Dementia Rating Scale. *Assessment, 1*(2), 123-132.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., . . . Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia, 7*(3), 280-292. doi: 10.1016/j.jalz.2011.03.003 [doi]

Stoub, T. R., Bulgakova, M., Leurgans, S., Bennett, D. A., Fleischman, D., Turner, D. A., & deToledo-Morrell, L. (2005). MRI predictors of risk of incident Alzheimer disease: a longitudinal study. *Neurology, 64*(9), 1520-1524. doi: 10.1212/01.WNL.0000160089.43264.1A [doi]

Strauss, E., Sherman, Elisabeth M.S., & Spreen, O. (2006). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary* (3rd ed.). New York, NY: Oxford University Press.

Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., & Roses, A. D. (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences U S A, 90*(5), 1977-1981.

Tae, W. S., Kim, S. S., Lee, K. U., Nam, E. C., & Kim, K. W. (2008). Validation of hippocampal volumes measured using a manual method and two automated methods (FreeSurfer and IBASPM) in chronic major depressive disorder. *Neuroradiology, 50*(7), 569-581. doi: 10.1007/s00234-008-0383-9

Tariot, P. N., Cummings, J. L., Katz, I. R., Mintzer, J., Perdomo, C. A., Schwam, E. M., & Whalen, E. (2001). A randomized, double-blind, placebo-controlled study of the efficacy and safety of donepezil in patients with Alzheimer's disease in the nursing home setting. *Journal of the American Geriatrics Society, 49*(12), 1590-1599.

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: a comparison of four models. *Journal of the International Neuropsychological Society, 5*(4), 357-369.

Thompson, P. M., Hayashi, K. M., Dutton, R. A., Chiang, M. C., Leow, A. D., Sowell, E. R., . . . Toga, A. W. (2007). Tracking Alzheimer's disease. *Annals of the New York Academy of Sciences, 1097*, 183-214. doi: 10.1196/annals.1379.017

van den Burg, W., & Kingma, A. (1999). Performance of 225 Dutch school children on Rey's Auditory Verbal Learning Test (AVLT): parallel test-retest reliabilities with an interval of 3 months and normative data. *Archives of Clinical Neuropsychology, 14*(6), 545-559.

Van Leemput, K., Bakkour, A., Benner, T., Wiggins, G., Wald, L. L., Augustinack, J., . . . Fischl, B. (2009). Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. *Hippocampus, 19*(6), 549-557. doi: 10.1002/hipo.20615

Vina, J., & Lloret, A. (2010). Why women have more Alzheimer's disease than men: gender and mitochondrial toxicity of amyloid-beta peptide. *Journal of Alzheimer's Disease, 20 Suppl 2*, S527-533. doi: 10.3233/JAD-2010-100501 [doi]

Visser, P. J., Verhey, F. R., Hofman, P. A., Scheltens, P., & Jolles, J. (2002). Medial temporal lobe atrophy predicts Alzheimer's disease in patients with minor cognitive impairment. *Journal of Neurology, Neurosurgery, and Psychiatry, 72*(4), 491-497.

Vitaliano, P. P., Breen, A. R., Russo, J., Albert, M., Vitiello, M. V., & Prinz, P. N. (1984). The clinical utility of the dementia rating scale for assessing Alzheimer patients. *Journal of Chronic Diseases, 37*(9-10), 743-753.

Wolf, H., Grunwald, M., Kruggel, F., Riedel-Heller, S. G., Angerhofer, S., Hojjatoleslami, A., . . . Gertz, H. (2001). Hippocampal volume discriminates between normal cognition; questionable and mild dementia in the elderly. *Neurobiology of Aging, 22*(2), 177-186.

Woodard, J. L., Seidenberg, M., Nielson, K. A., Antuono, P., Guidotti, L., Durgerian, S., . . . Rao, S. M. (2009). Semantic memory activation in amnestic mild cognitive impairment. *Brain, 132*(Pt 8), 2068-2078. doi: 10.1093/brain/awp157 [doi]

Woodard, J. L., Seidenberg, M., Nielson, K. A., Smith, J. C., Antuono, P., Durgerian, S., . . . Rao, S. M. (2010). Prediction of cognitive decline in healthy older adults using fMRI. *Journal of Alzheimer's Disease, 21*(3), 871-885. doi: 10.3233/JAD-2010-091693 [doi]

Yesavage, J A, Brink, T L, Rose, T L, Lum, O, Huang, V, Adey, M, & Leirer, V O. (1983). Development and validation of a geriatric depression screening scale: A preliminary report. *Journal of Psychiatric Research, 17*, 37-49.

Table 1

Participant Characteristics and Neuropsychological Performance at Baseline.

	Stable (*n* = 45)		Declining (*n* = 15)			
Variables	Mean	SD	Mean	SD	*p*	η^2
Demographics						
Age (yrs)	71.29	4.50	72.53	4.41	.36	.01
Education (yrs)	15.11	2.71	14.20	2.18	.24	.02
						r_φ
Sex	12M, 33F	--	3M, 12F	--	.74	.07
Family History	22FH+, 23FH-	--	10FH+, 5FH-	--	.37	.23
APOE inheritance	14ε4+, 31ε4-	--	10ε4+, 5ε4-	--	***.03***	***.31***
Neuropsychological Testing						η^2
DRS-2 Total	141.13	2.17	139.47	4.44	.18	.06
RAVLT Trials 1-5	50.31	8.10	48.13	8.72	.38	.01
RAVLT DR	10.16	2.57	9.33	2.79	.30	.02

Note. All indices represent raw scores. M = male; F = female; FH = Family History; FH- = no family history, FH+ = positive family history; APOE = Apolipoprotein-E ε4 inheritance; ε4+ = at least one ε4 allele; ε4 - = no ε4 allele; DRS-2= Mattis Dementia Rating Scale-2; RAVLT = Rey Auditory Verbal Learning Test; DR = Delayed Recall.

Table 2

Neuropsychological Performance at the 4.5-year Follow-up Assessment.

	Stable (*n* = 45)		Declining (*n* = 15)			
Neuropsychological Testing	Mean	SD	Mean	SD	*p*	η^2
DRS-2 Total	141.49	2.27	136.67	6.00	***.019***	***.22***
RAVLT Trials 1-5	51.00	7.53	40.87	11.36	***.005***	***.21***
RAVLT DR	10.58	2.37	6.00	4.24	***.001***	***.32***

Note. All indices represent raw scores. DRS-2= Mattis Dementia Rating Scale, Second Ed; RAVLT = Rey Auditory Verbal Learning Test; DR = Delayed Recall.

Table 3

Hierarchical Multiple Regression with Baseline Hippocampal Volumes and APOE Predicting Residualized Change Scores for RAVLT 1-5.

	R	R^2	R^2 Change	*B*	*SE*	*β*	*p*
Model 1							
Step 1	.157	.025					
FS Total				< .001	< .001	.157	.231
Step 2	.203	.041	.017				
FS Total				< .001	< .001	.156	.233
APOE				-.258	.260	-.129	.326
Model 2							
Step 1	.160	.026					
FS Left				< .001	< .001	.129	.578
FS Right				< .001	.001	.036	.878
Step 2	.220	.048	.023				
FS Left				.001	.001	.230	.354
FS Right				< .001	.001	-.065	.793
APOE				-.325	.282	-.162	.253
Model 3							
Step 1	.177	.031					
MT Total				< .001	< .001	.177	.175
Step 2	.209	.044	.012				
MT Total				< .001	< .001	.165	.210
APOE				-.233	.262	-.111	.396
Model 4							
Step 1	.248	.061					
MT Left				.001	.001	.428	.105
MT Right				- .001	.001	-.243	.355
Step 2	.276	.076	.015				
MT Left				.001	.001	.435	.101
MT Right				- .001	.001	-.263	.318
APOE				-.244	.260	-.122	.351

Note. Statistical significance: *$p < .05$; ** $p < .01$; FS Total = FreeSurfer total hippocampal volume collapsed across side; FS Left = FreeSurfer left hippocampus; FS Right = FreeSurfer right hippocampus; MT Total = manual tracing total hippocampal volume collapsed across side; MT Left = Manual tracings left hippocampus; MT Right = Manual tracing right hippocampus; R^2 = amount of variance explained by IVs; R^2 Change = additional variance in DV; *B* = Unstandardized coefficient; *β* = Standardized coefficient; *SE* = Standard Error.

Table 4

Hierarchical Multiple Regression with Baseline Hippocampal Volumes and APOE Predicting Residualized Change Scores for RAVLT Delay.

	R	R^2	R^2 Change	*B*	*SE*	*β*	*p*
Model 1							
Step 1	.223	.050					
FS Total				< .001	< .001	.223	.087
Step 2	.448	.200**	.151**				
FS Total				< .001	< .001	.221	.067
APOE				-.779	.238	-.388	.002
Model 2							
Step 1	.313	.098					
FS Left				- .001	.001	-.263	.241
FS Right				.001	< .001	.492	.031
Step 2	.455	.207**	.109**				
FS Left				< .001	.001	-.041	.854
FS Right				.001	.001	.271	.232
APOE				-.713	.257	-.355	.007
Model 3							
Step 1	.242	.059					
MT Total				< .001	< .001	.242	.062
Step 2	.438	.192**	.133**				
MT Total				< .001	< .001	.202	.097
APOE				-.737	.240	-.367	.003
Model 4							
Step 1	.253	.064					
MT Left				< .001	.001	-.023	.930
MT Right				.001	.001	.273	.298
Step 2	.441	.195**	.131**				
MT Left				< .001	.001	-.002	.994
MT Right				.001	.001	.211	.392
APOE				-.731	.243	-.364	.004

Note. Statistical significance: *$p < .05$; ** $p < .01$; FS Total = FreeSurfer total hippocampal volume collapsed across side; FS Left = FreeSurfer left hippocampus; FS Right = FreeSurfer right hippocampus; MT Total = manual tracing total hippocampal volume collapsed across side; MT Left = Manual tracings left hippocampus; MT Right = Manual tracing right hippocampus; R^2 = amount of variance explained by IVs; R^2 Change = additional variance in DV; *B* = Unstandardized coefficient; *β* = Standardized coefficient; *SE* = Standard Error.

Table 5

Hierarchical Multiple Regression with Baseline Hippocampal Volumes and APOE Predicting Residualized Change Scores for DRS-2 Total.

	R	R^2	R^2 Change	*B*	*SE*	*β*	*p*
Model 1							
Step 1	.235	.055					
FS Total				< .001	< .001	.235	.071
Step 2	.282	.080	.025				
FS Total				< .001	< .001	.234	.071
APOE				-.315	.255	-.157	.222
Model 2							
Step 1	.248	.062					
FS Left				.001	.001	.253	.268
FS Right				< .001	.001	-.006	.980
Step 2	.320	.102	.040				
FS Left				.001	.001	.388	.110
FS Right				< .001	.001	-.140	.561
APOE				-.434	.273	-.216	.118
Model 3							
Step 1	.140	.020					
MT Total				< .001	< .001	.140	.287
Step 2	.200	.040	.021				
MT Total				< .001	< .001	.124	.346
APOE				-.290	.262	-.145	.273
Model 4							
Step 1	.140	.020					
MT Left				< .001	.001	.058	.827
MT Right				< .001	.001	.086	.747
Step 2	.201	.040	.021				
MT Left				< .001	.001	.067	.803
MT Right				< .001	.001	.062	.818
APOE				-.291	.265	-.145	.277

Note. Statistical significance: *p < .05; ** p < .01; FS Total = FreeSurfer total hippocampal volume collapsed across side; FS Left = FreeSurfer left hippocampus; FS Right = FreeSurfer right hippocampus; MT Total = manual tracing total hippocampal volume collapsed across side; MT Left = Manual tracings left hippocampus; MT Right = Manual tracing right hippocampus; R^2 = amount of variance explained by IVs; R^2 Change = additional variance in DV; *B* = Unstandardized coefficient; *β* = Standardized coefficient; *SE* = Standard Error.

Table 6

Areas Under the ROC Curve for Baseline and 4.5-year Hippocampal Volumes

		FreeSurfer AUC	*p*	Manual Tracing AUC	*p*
Baseline					
	Left	.505	.952	.446	.533
	Right	.437	.468	.455	.603
	Total	.476	.778	.430	.417
4.5 Years					
	Left	.439	.479	.372	.140
	Right	.348	.080	.366	.122
	Total	.396	.084	.354	.077

Note. AUC values indicate areas under the ROC (receiver operating characteristic) curve. Statistical significance: **p* < .05.

Table 7

Correlations of 4.5-year Follow-up Hippocampal Volumes and Cognitive Measures

	RAVLT 1-5	RAVLT Delay	DRS-2 Total	APOE
FreeSurfer				
Left Hipp	*r* =.271 *p* = .036	*r* =.229 *p* = .078	*r* =.430 *p* = .001	*p* = .631
Right Hipp	*r* =.305 *p* = .018	*r* =.326 *p* = .011	*r* =.425 *p* = .001	*p* = .168
FS Total Hipp	*r* =.302 *p* = .019	*r* =.292 *p* = .024	*r* =.448 *p* < .001	*p* = .328
Manual Tracings				
Left Hipp	*r* =.318 *p* = .013	*r* =.262 *p* = .043	*r* =.426 *p* = .001	*p* = .401
Right Hipp	*r* =.323 *p* = .012	*r* =.335 *p* = .009	*r* =.330 *p* = .010	*p* = .105
MT Total Hipp	*r* =.332 *p* = .010	*r* =.310 *p* = .016	*r* =.390 *p* = .002	*p* = .200
APOE Status				
ε4 Negative	50.0 (6.8)	10.5 (2.5)	141.0 (2.8)	
ε4 Positive	46.2 (12.5) *p* = .188	7.8 (4.2) *p* = .009	139.3 (6.1) *p* = .204	

Note. Left Hipp = Left Hippocampus; Right Hipp = Right Hippocampus; Total Hipp = right and left hippocampal volumes combined; RAVLT 1-5 = Rey Auditory Verbal Learning Test Trials 1-5 raw score; RAVLT Delay = Rey Auditory Verbal Learning Test Delay Recall raw score; DRS-2 Total = Dementia Rating Scale -2 Total Score; APOE Status = ApolipoProtein E 4 positive of negative; *r* = Pearson Correlation; *p* = alpha level set at 0.05 for 2-tailed. APOE Status values indicate Mean (Standard Deviation).

Table 8

Hierarchical Multiple Regression with 4.5-Year Hippocampal Volumes and APOE Predicting Residualized Change Scores for RAVLT 1-5

	R	R^2	R^2 Change	*B*	*SE*	*β*	*p*
Model 1							
Step 1	.267	.071*					
FS Total				< .001	< .001	.267	.039
Step 2	.284	.081	.009				
FS Total				< .001	< .001	.255	.051
APOE				-.194	.257	0.097	.453
Model 2							
Step 1	.280	.079					
FS Left				.001	< .001	.280	.213
FS Right				< .001	< .001	< .001	.998
Step 2	.303	.092	.043				
FS Left				.001	< .001	.311	.173
FS Right				< .001	< .001	-.046	.841
APOE				-.237	.263	-.118	.371
Model 3							
Step 1	.315	.099*					
MT Total				< .001	< .001	.315	.014
Step 2	.324	.105*	.006				
MT Total				< .001	< .001	.301	.021
APOE				-.158	.255	-.079	.538
Model 4							
Step 1	.330	.109*					
MT Left				.001	.001	.351	.166
MT Right				< .001	.001	-.025	.921
Step 2	.345	.119	.010				
MT Left				.001	.001	.381	.139
MT Right				< .001	.001	-.072	.780
APOE				-.206	.260	-.103	.432

Note. Statistical significance: **p* < .05; ** *p* < .01; FS Total = FreeSurfer total hippocampal volume collapsed across side; FS Left = FreeSurfer left hippocampus; FS Right = FreeSurfer right hippocampus; MT Total = manual tracing total hippocampal volume collapsed across side; MT Left = Manual tracings left hippocampus; MT Right = Manual tracing right hippocampus; R^2 = amount of variance explained by IVs; R^2 Change = additional variance in DV; *B* = Unstandardized coefficient; *β* = Standardized coefficient; *SE* = Standard Error.

Table 9

Hierarchical Multiple Regression with 4.5-Year Hippocampal Volumes and APOE Predicting Residualized Change Scores for RAVLT Delay

	R	R^2	R^2 Change	*B*	*SE*	*β*	*p*
Model 1							
Step 1	.285	.081*					
FS Total				< .001	< .001	.285	.027
Step 2	.456	.208**	.127**				
FS Total				< .001	< .001	.239	.050
APOE				-.720	.239	-.359	.004
Model 2							
Step 1	.308	.095					
FS Left				< .001	< .001	-.050	.820
FS Right				.001	< .001	.348	.119
Step 2	.458	.210**	.115**				
FS Left				< .001	< .001	.040	.850
FS Right				< .001	< .001	.211	.326
APOE				-.700	.245	-.349	.006
Model 3							
Step 1	.252	.063					
MT Total				< .001	< .001	.252	.052
Step 2	.433	.187**	.124**				
MT Total				< .001	< .001	.192	.119
APOE				-.717	.243	-.357	.005
Model 4							
Step 1	.289	.084					
MT Left				< .001	.001	-.148	.561
MT Right				.001	.001	.408	.113
Step 2	.439	.193**	.109**				
MT Left				< .001	.001	-.049	.841
MT Right				.001	.001	.250	.316
APOE				-.687	.249	-.342	.008

Note. Statistical significance: *$p < .05$; ** $p < .01$; FS Total = FreeSurfer total hippocampal volume collapsed across side; FS Left = FreeSurfer left hippocampus; FS Right = FreeSurfer right hippocampus; MT Total = manual tracing total hippocampal volume collapsed across side; MT Left = Manual tracings left hippocampus; MT Right = Manual tracing right hippocampus; R^2 = amount of variance explained by IVs; R^2 Change = additional variance in DV; *B* = Unstandardized coefficient; *β* = Standardized coefficient; *SE* = Standard Error.

Table 10

Hierarchical Multiple Regression with 4.5-Year Hippocampal Volumes and APOE Predicting Residualized Change Scores for DRS-2 Total.

	R	R^2	R^2	*B*	*SE*	*β*	*p*
			Change				
Model 1							
Step 1	.341	.116**					
FS Total				< .001	< .001	.341	.008
Step 2	.260	.130**	.013				
FS Total				< .001	< .001	.326	.011
APOE				-.233	.250	-.116	.355
Model 2							
Step 1	.356	.127*					
FS Left				.001	< .001	.346	.116
FS Right				< .001	< .001	.012	.954
Step 2	.382	.146*	.019				
FS Left				.001	< .001	.382	.085
FS Right				< .001	< .001	-.043	.846
APOE				-.285	.255	-.142	.269
Model 3							
Step 1	.261	.068*					
MT Total				< .001	< .001	.261	.044
Step 2	.285	.081	.013				
MT Total				< .001	< .001	.241	.067
APOE				-.236	.258	-.118	.364
Model 4							
Step 1	.346	.120*					
MT Left				.002	.001	.573	.025
MT Right				-.001	.001	-.302	.230
Step 2	.383	.147*	.027				
MT Left				.002	.001	.622	.016
MT Right				-.001	.001	-.380	.140
APOE				-.341	.256	-.170	.189

Note. Statistical significance: *p < .05; ** p < .01; FS Total = FreeSurfer total hippocampal volume collapsed across side; FS Left = FreeSurfer left hippocampus; FS Right = FreeSurfer right hippocampus; MT Total = manual tracing total hippocampal volume collapsed across side; MT Left = Manual tracings left hippocampus; MT Right = Manual tracing right hippocampus; R^2 = amount of variance explained by IVs; R^2 Change = additional variance in DV; *B* = Unstandardized coefficient; *β* = Standardized coefficient; *SE* = Standard Error.

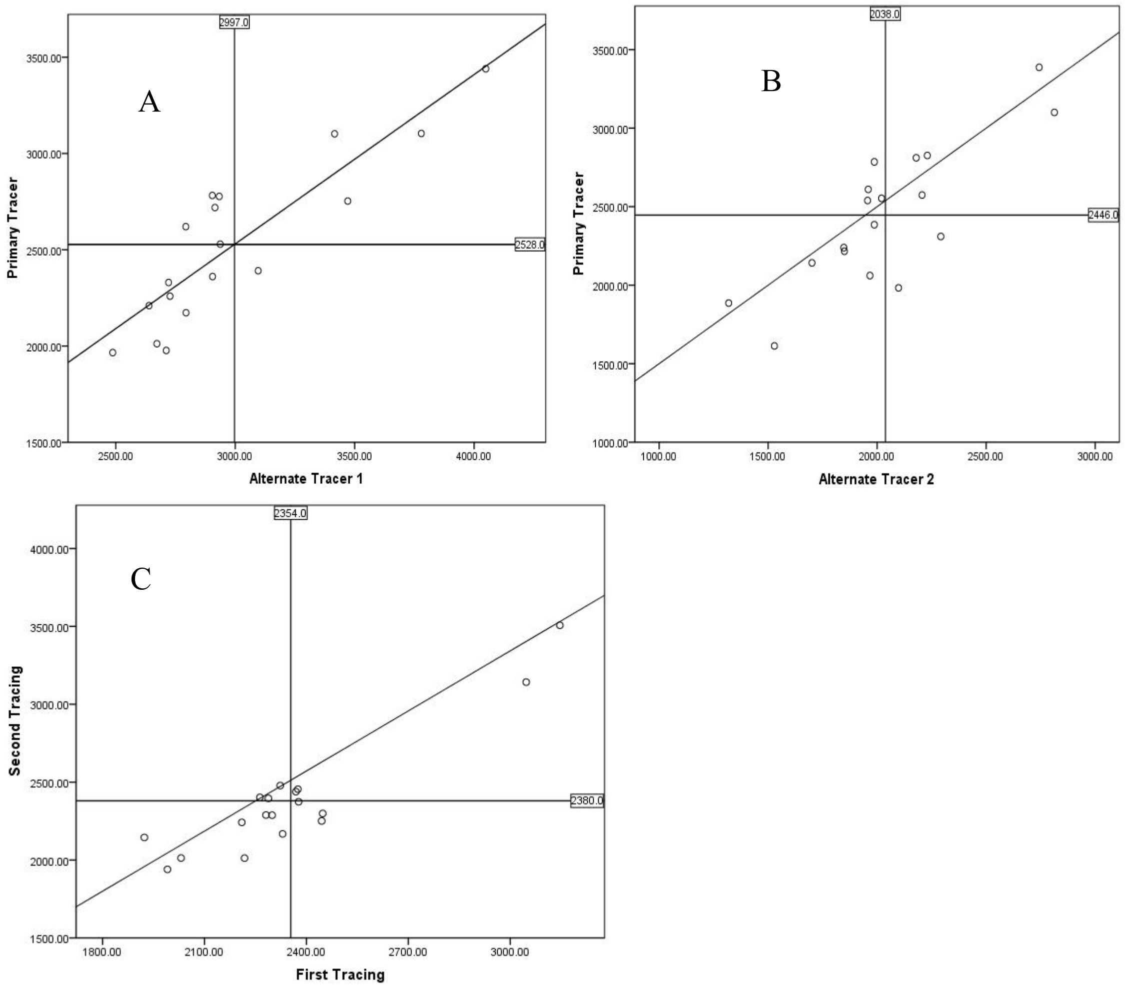

Figure 1. Intraclass correlation (ICC) plots for manual hippocampal volume measurements. Intraclass correlations were strong for Alternate Rater 1 (A) and Alternate Rater 2 (B), as was intra-rater reliability (C). The average volume of the alternate rater is shown as the reference line for the X-axis, and the average volume of the primary rater is shown as the reference line for the y-axis.

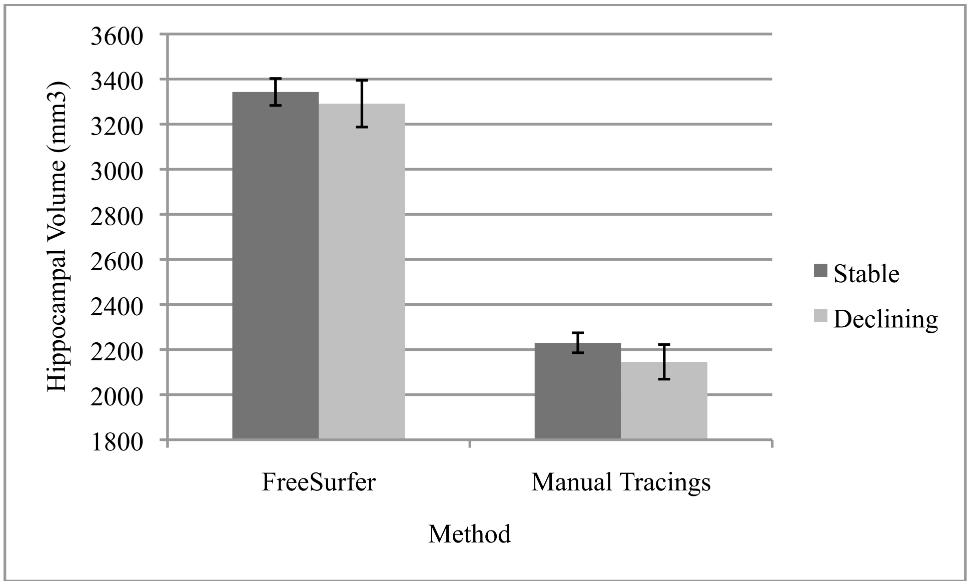

Figure 2. Separate hippocampal volume comparisons by method and group. FreeSurfer produced significantly larger estimates of baseline hippocampal volume than did manual tracing. Collapsed across methods, there was no significant baseline volume difference between Stable and Declining Participants. Error bars represent SEM.

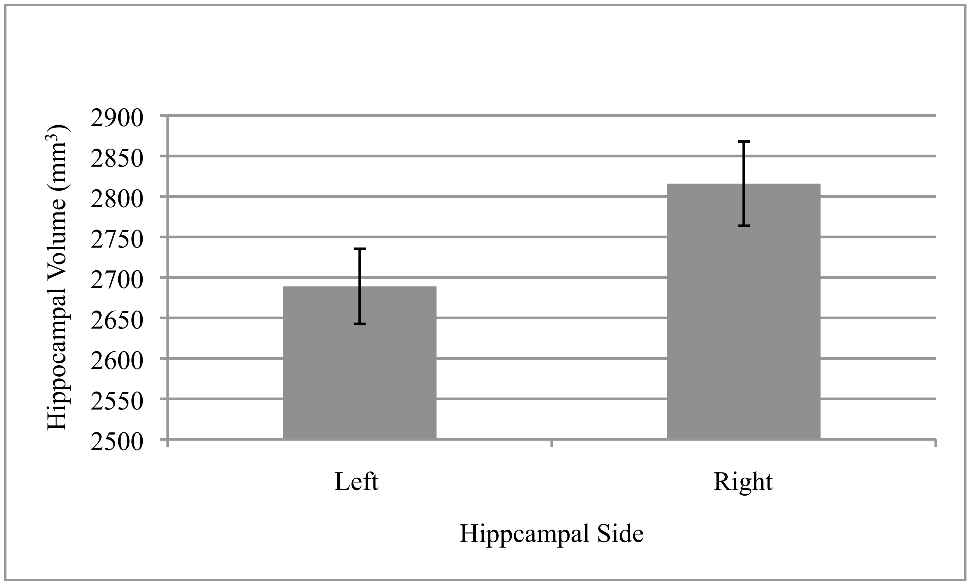

Figure 3. Comparison of left and right baseline hippocampal volume. Collapsed across methods and group, right baseline hippocampal volumes were significantly larger than left baseline hippocampal volumes. Error bars represent SEM.

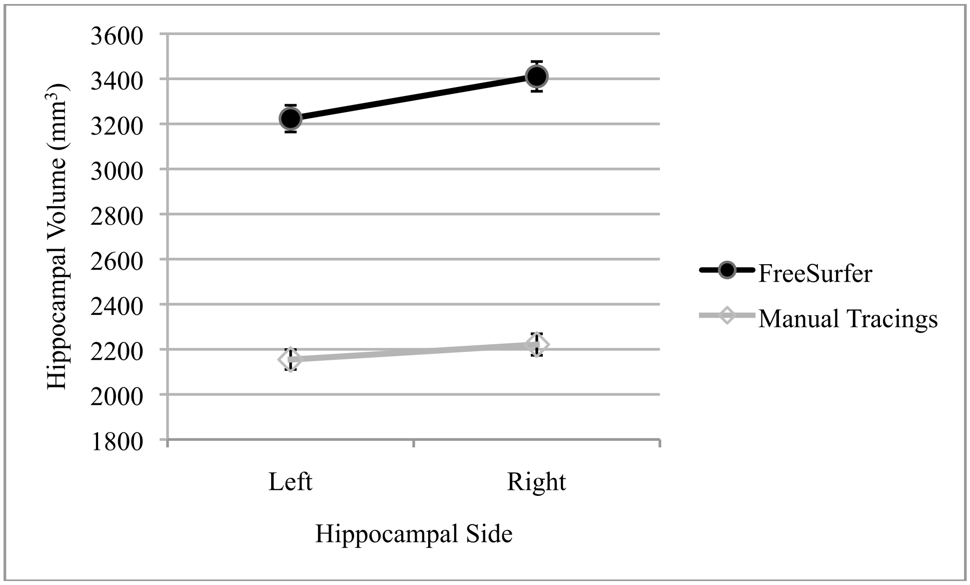

Figure 4. Baseline hippocampal volume comparison of method by side. FreeSurfer produced significantly larger left and right baseline hippocampal volumes compared to manual tracings. For each method, right hippocampal volumes were larger than left volumes. Error bars represent SEM.

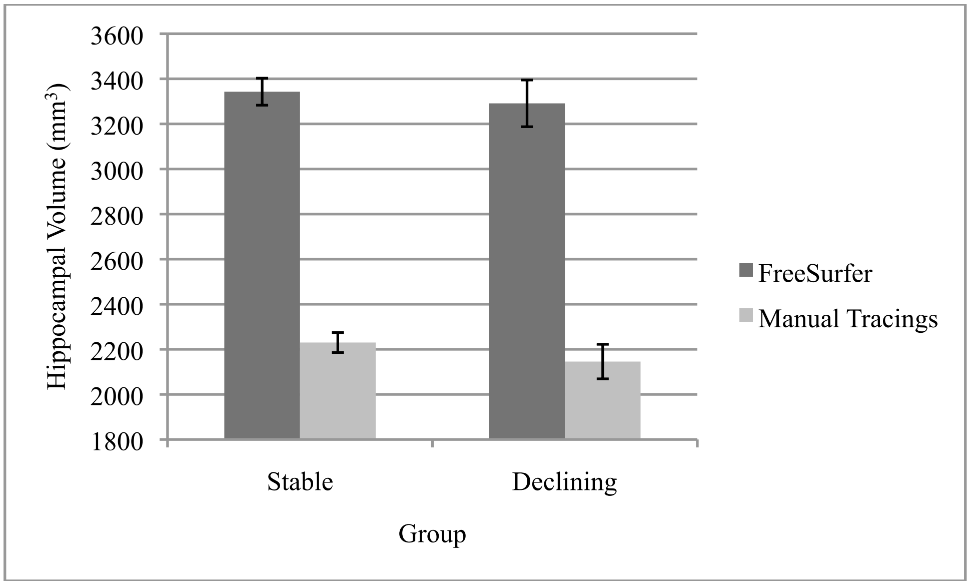

Figure 5. Baseline hippocampal volume comparison of method by group. Stable and Declining participants did not have significantly different baseline hippocampal volumes as measured by FreeSurfer or manual tracing. Error bars represent SEM.

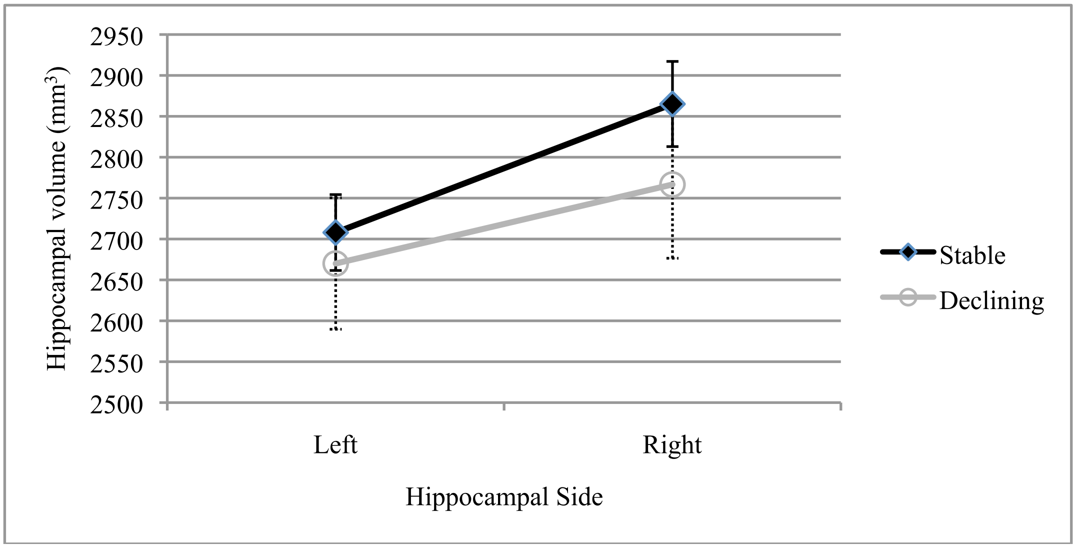

Figure 6. Baseline hippocampal volume comparison of group by side. Left and right hippocampal volumes did not differ significantly between Stable and Declining participants. Error bars represent SEM.

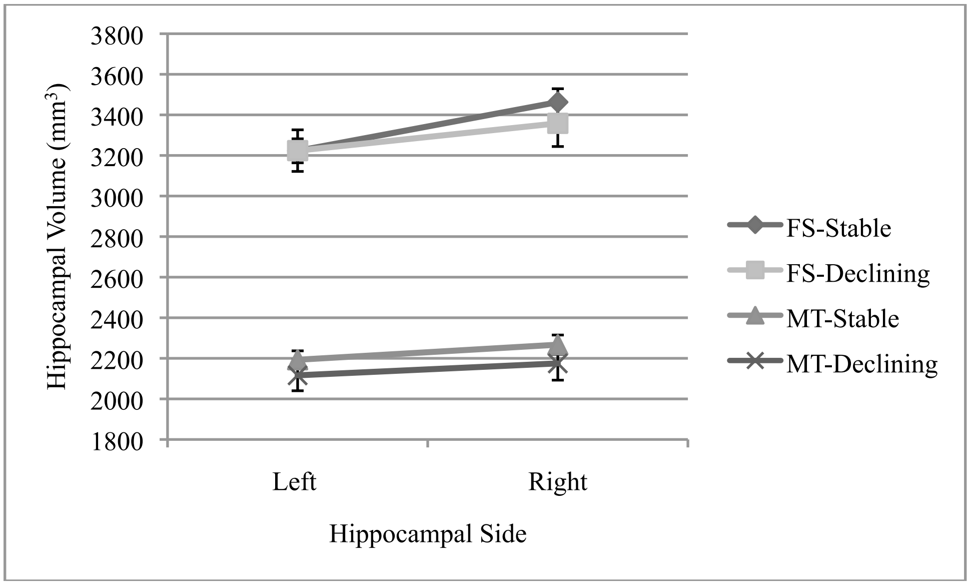

Figure 7. Baseline hippocampal volume comparison of method by side by group. Bilaterally, baseline hippocampal volumes measured by FreeSurfer and manual tracing not differ significantly between Stable and Declining participants. Error bars represent SEM.

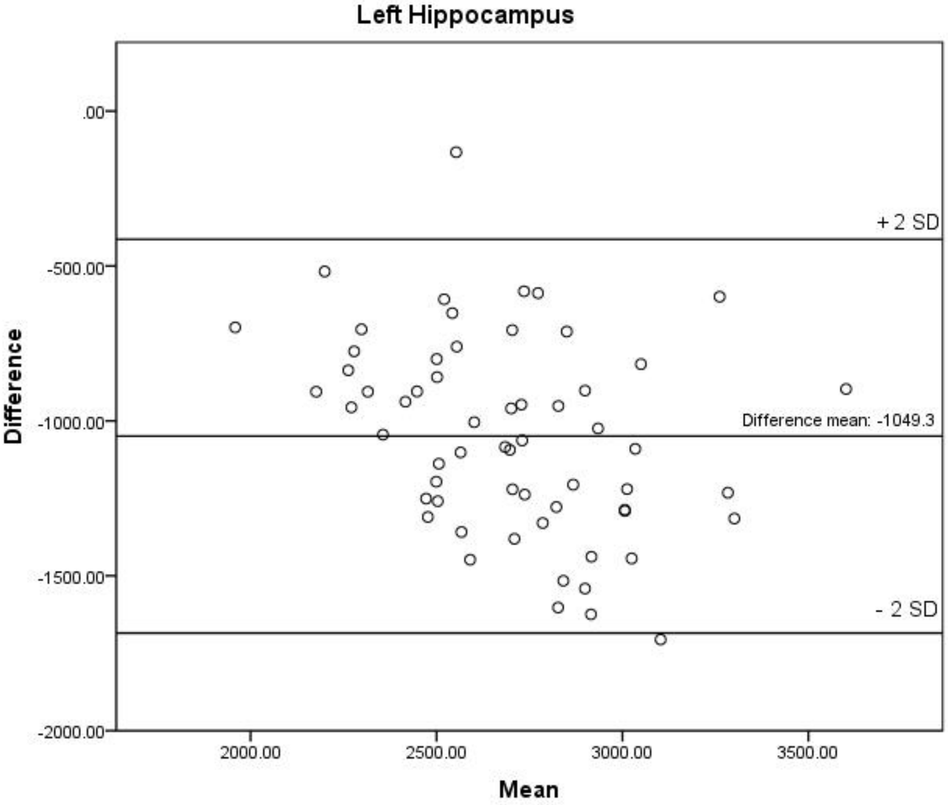

Figure 8. Bland-Altman plot for baseline left hippocampal volumes. The vast majority of left hippocampal measurements between FreeSurfer and manual tracings fall within two standard deviations of the difference mean, or within the expected range.

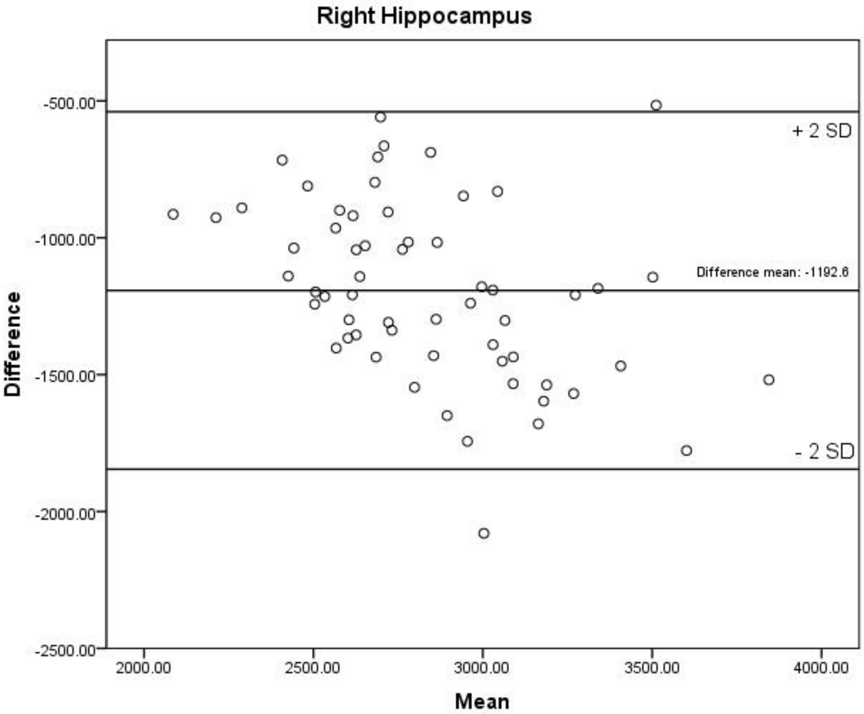

Figure 9. Bland-Altman plot for baseline right hippocampal volumes. The vast majority of right hippocampal measurements between FreeSurfer and manual tracings fall within two standard deviations of the difference mean, or within the expected range.

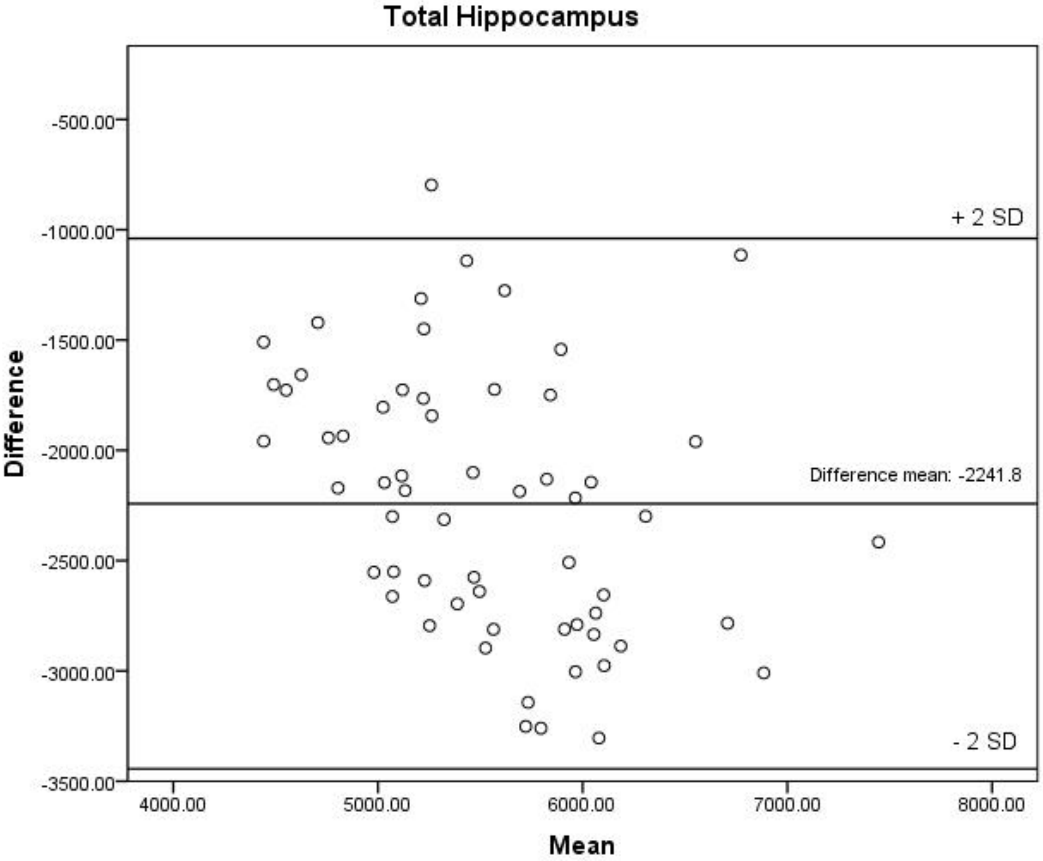

Figure 10. Bland-Altman plot for baseline total hippocampal volumes. The vast majority of left hippocampal measurements between FreeSurfer and manual tracings fall within two standard deviations of the difference mean, or within the expected range.

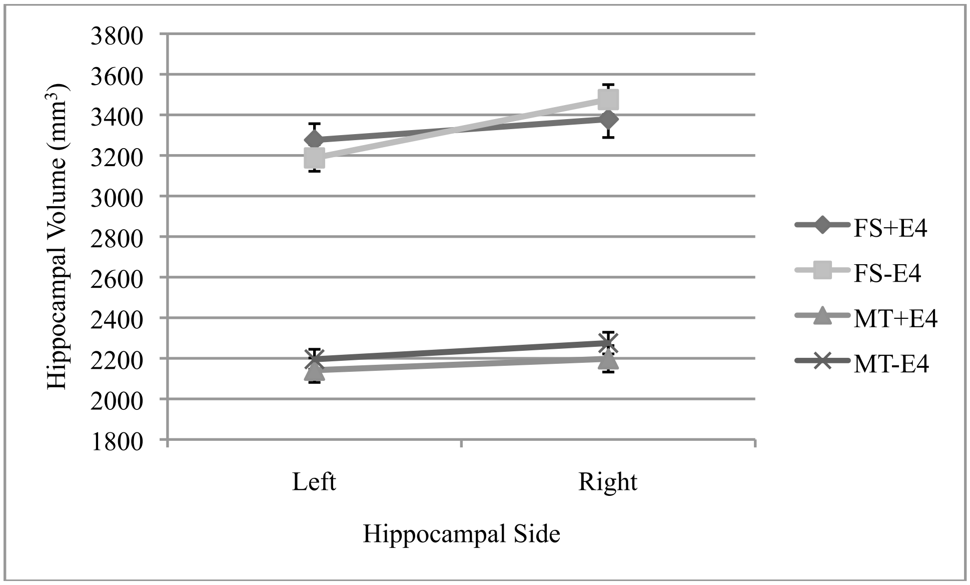

Figure 11. Baseline hippocampal volumes comparisons of method by side by APOE. ε4-negative participants had larger right than left hippocampal volumes as measured by both FreeSurfer (FS) and manual tracings (MT). Yet, ε4-positive participants had larger right than left hippocampal volumes measured by FS, but not by MT. Error bars represent SEM.

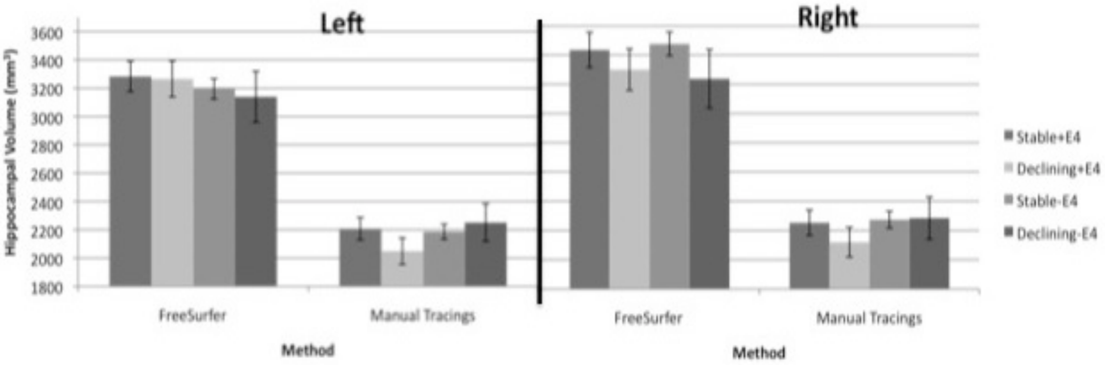

Figure 12. Comparison of baseline hippocampal volume of method by side by decline status by APOE. Stable ε4-negative and ε4 –positive participants had greater left compared to right hippocampal volumes, as measured by both FreeSurfer (FS) and manual tracings (MT). This pattern was found in ε4-negative Declining participants with FS, but not MT. Declining ε4-positive participants did not differ between side for either method. Error bars represent SEM.

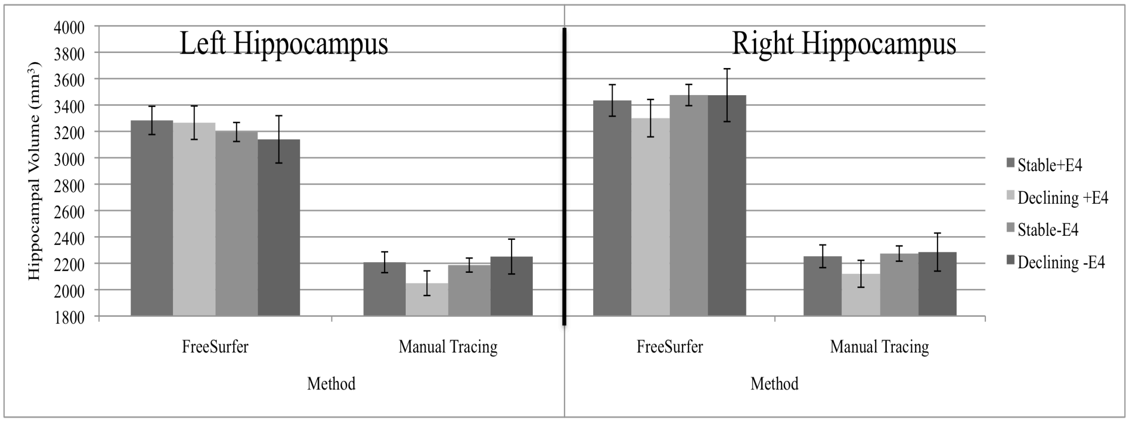

Figure 13. Baseline hippocampal volume comparison of method by decline status by APOE for the left and right hippocampus separately. FreeSurfer volumes were greater than manually traced volumes. ε4-positive participants did not differ from ε4-negative participants. There was not a significant between declining and stable participants. Error bars represent SEM.

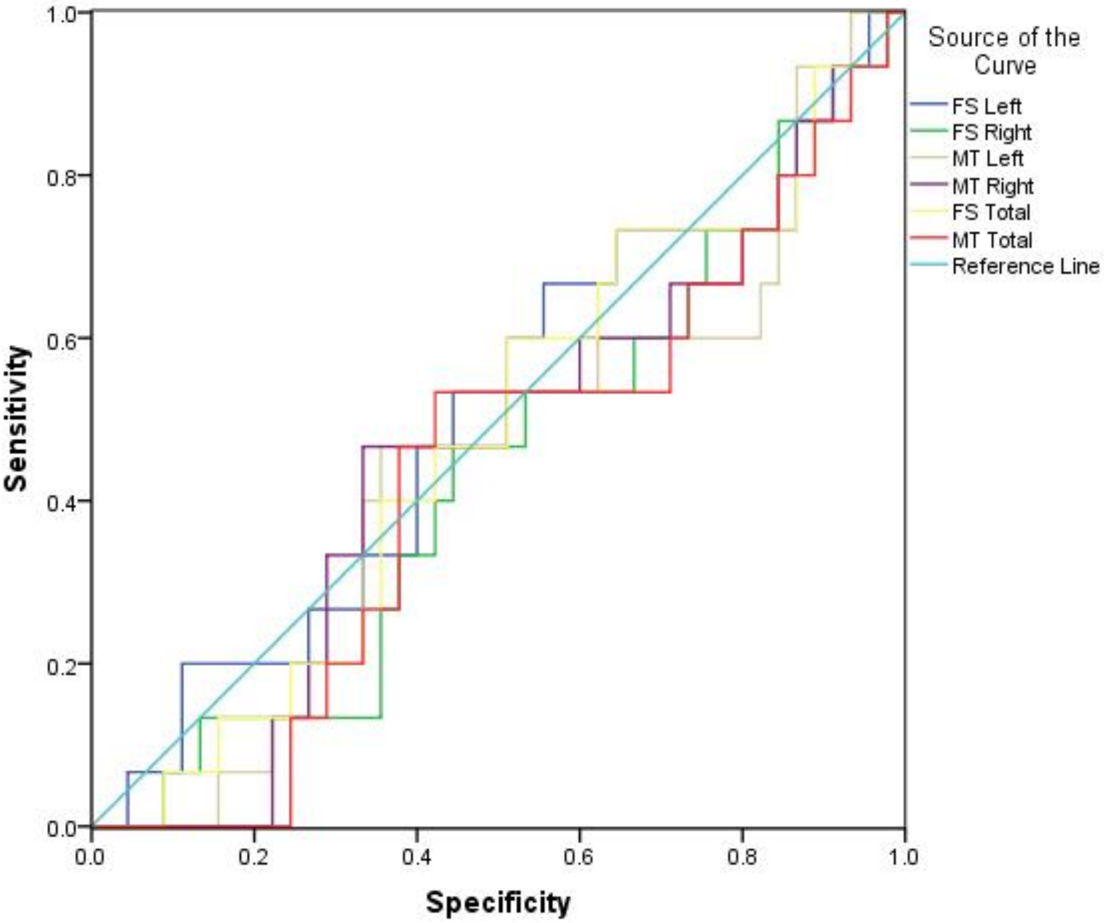

Figure 14. Receiver operating characteristic (ROC) curve for baseline hippocampal volumes. Baseline hippocampal volumes measured by FreeSurfer (FS) and manual tracings (MT) were not able to significantly classify cognitive decline.

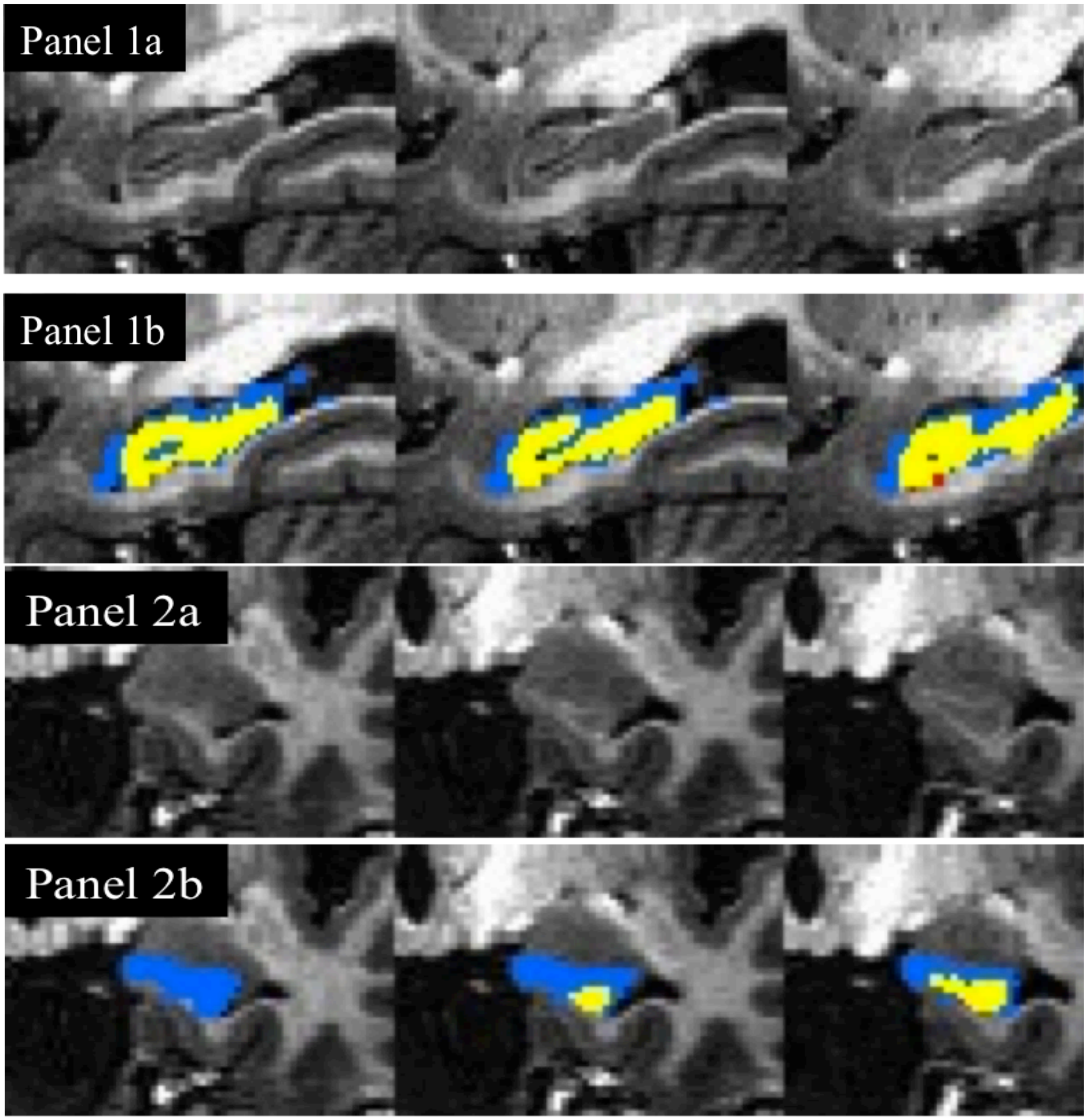

Figure 15. Sagittal and coronal FreeSurfer and manual tracing in one Stable participant. Panel 1a = Sagittal SPGR of hippocampus; Panel 1b = Sagittal overlap of FreeSurfer (FS) and manual tracing (MT). Panel 2a = Coronal SPGR of hippocampus; Panel 2b = Sagittal overlap of FS and MT. Blue = tissue uniquely identified by FS; red = tissue uniquely identified by MT; yellow = overlap between FS and MT. FreeSurfer yielded overall larger volumes than MT.

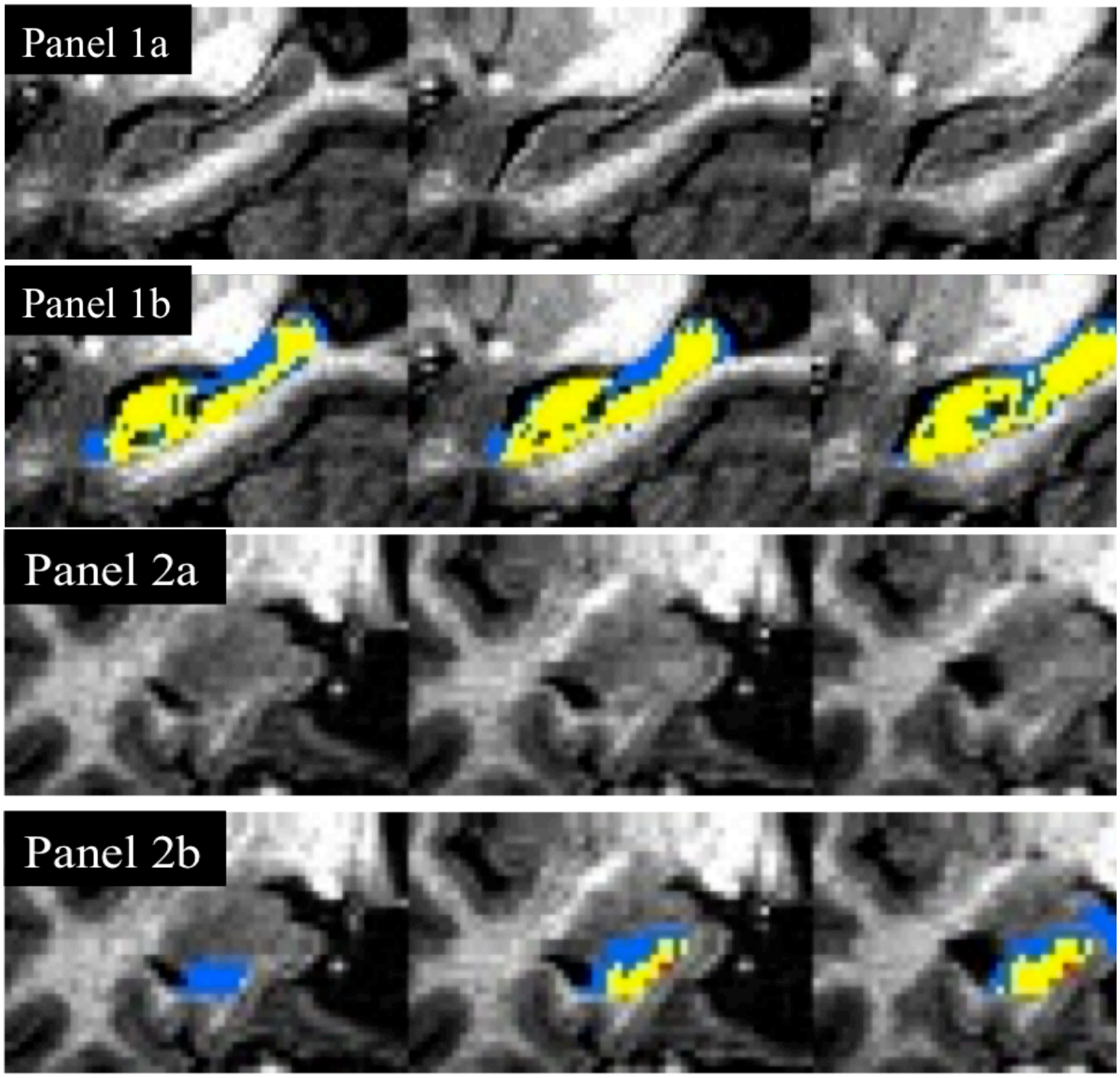

Figure 16. Sagittal and coronal FreeSurfer and manual tracing in one Declining participant. Panel 1a = Sagittal SPGR of hippocampus; Panel 1b = Sagittal overlap of FreeSurfer (FS) and manual tracing (MT). Panel 2a = Coronal SPGR of hippocampus; Panel 2b = Sagittal overlap of FS and MT. Blue = tissue uniquely identified by FS; red = tissue uniquely identified by MT; yellow = overlap between FS and MT. FreeSurfer yielded overall larger volumes than MT.

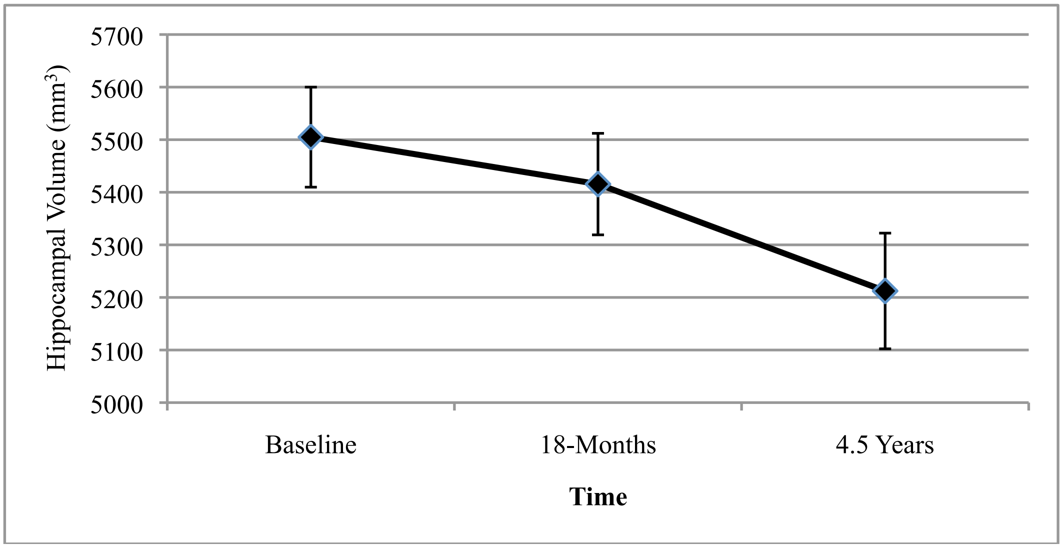

Figure 17. Comparison of hippocampal volume at each time point. Hippocampal volumes significantly reduced over time. Error bars represent SEM.

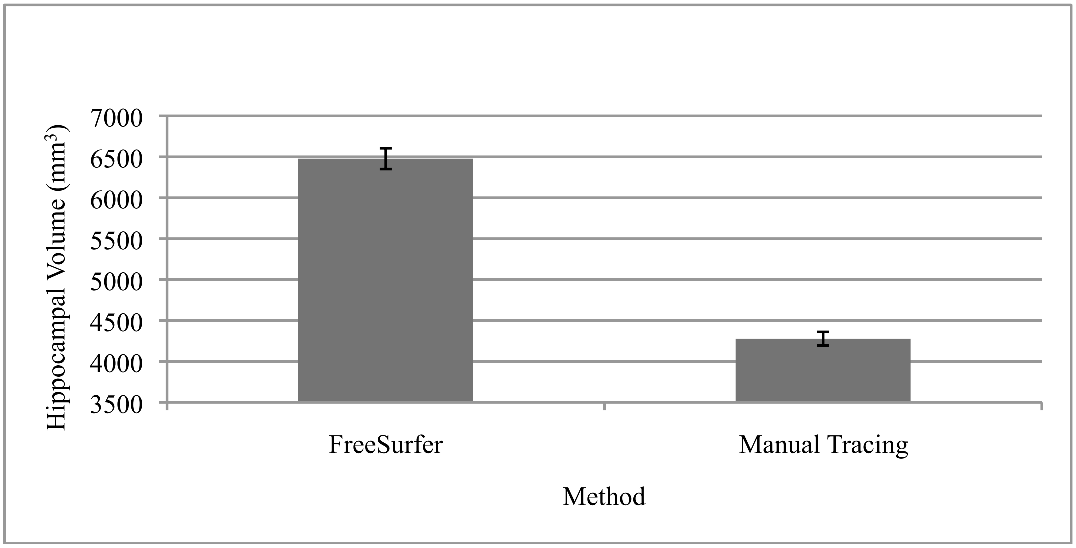

Figure 18. Comparison of total hippocampal volume by method. Collapsed across all time points, FreeSurfer produced significantly larger hippocampal volumes than manual tracings. Error bars represent SEM.

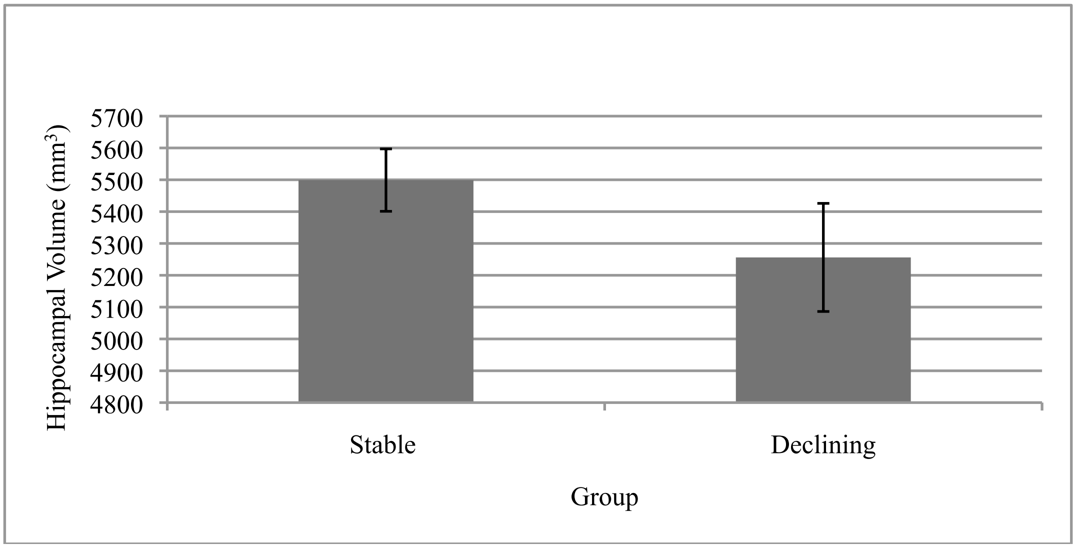

Figure 19. Comparison of total hippocampal volume by group. Collapsed across all three time points, hippocampal volumes did not differ between Stable and Declining participants. Error bars represent SEM.

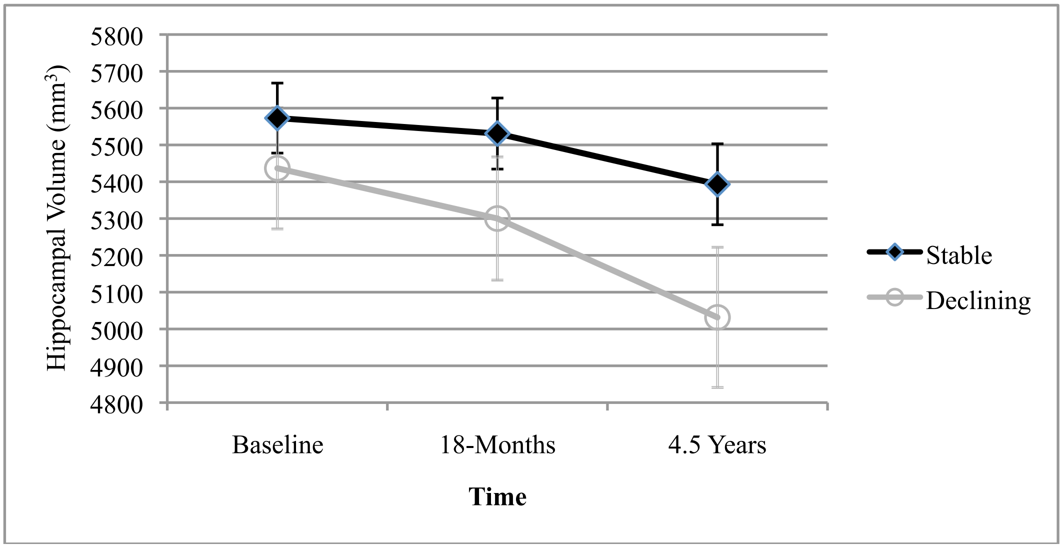

Figure 20. Comparison of hippocampal volumes across time by group. The hippocampal volume of Stable participants did not differ from baseline to the 18-month follow-up, but were significantly smaller at the 4.5-year follow up compared to baseline. Declining participants had significantly smaller hippocampal volumes at each successive time point. Error bars represent SEM.

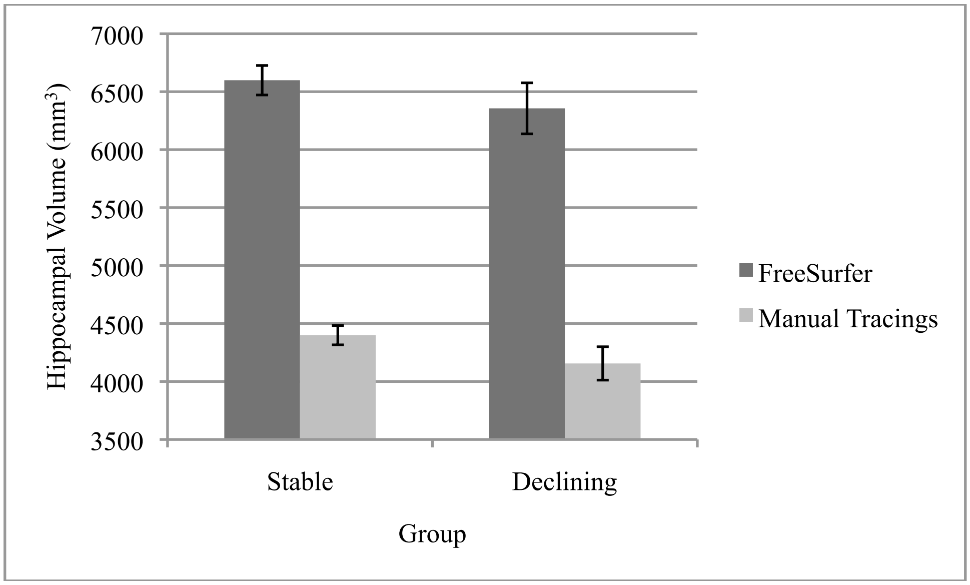

Figure 21. Comparison of hippocampal volumes over time by group and method. Collapsed across time, FreeSurfer and manually traced hippocampal volumes did not differ between Stable and Declining participants. Error bars represent SEM.

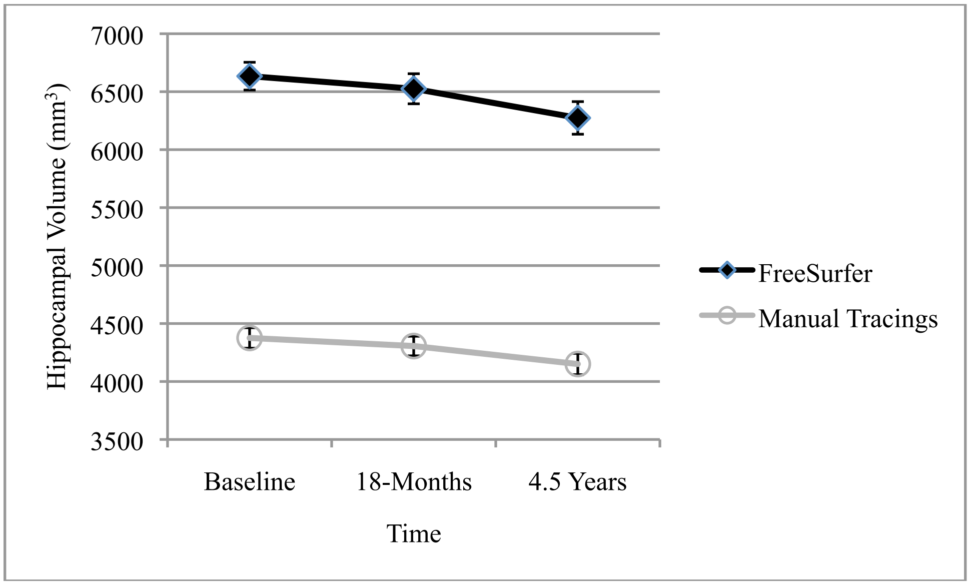

Figure 22. Comparison of hippocampal volumes across time by method. FreeSurfer and manually traced hippocampal volumes did not differ significantly across time points. Error bars represent SEM.

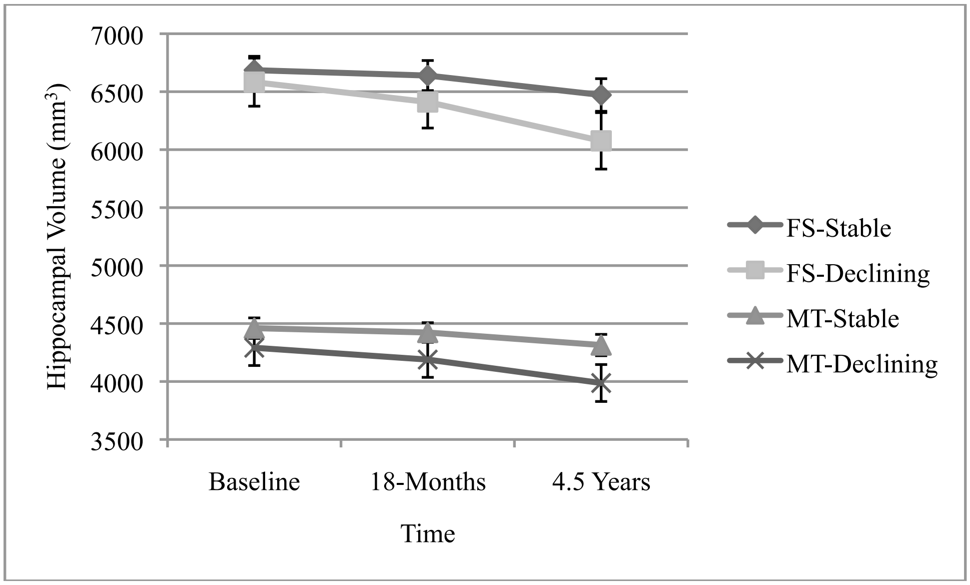

Figure 23. Hippocampal volume comparison across time by method and group. Manually traced and FreeSurfer hippocampal volumes did not differ significantly between Stable and Declining participants at each time point. Error bars represent SEM.

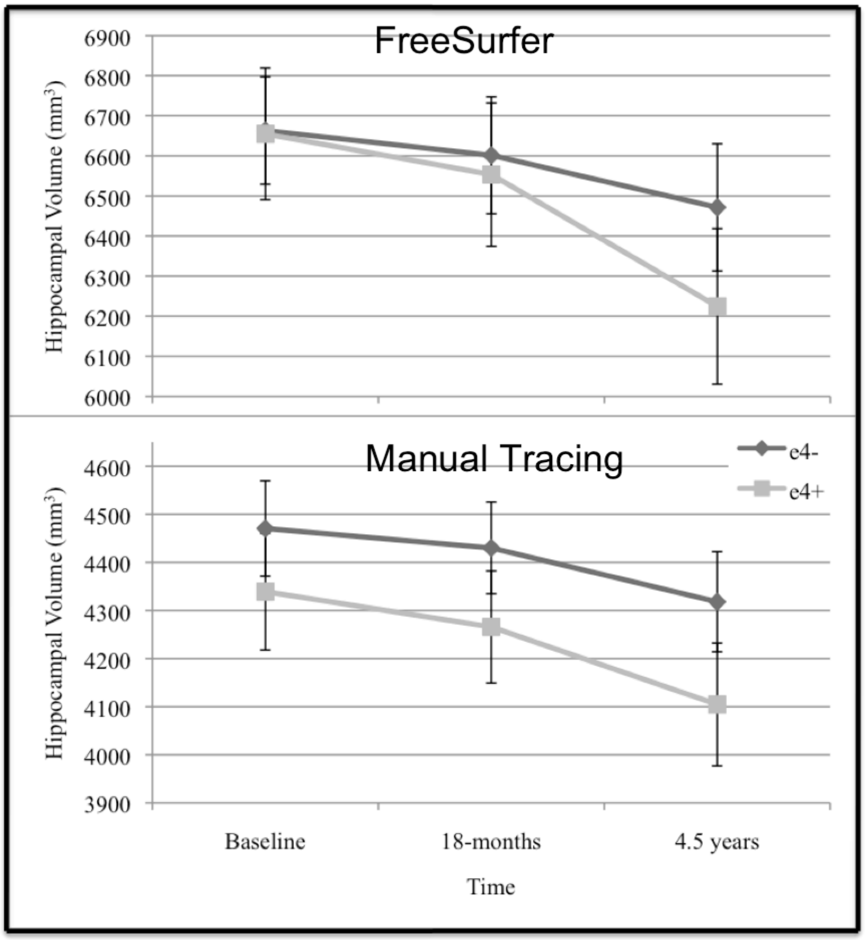

Figure 24. Comparison of hippocampal volume over time by APOE for FreeSurfer and manual tracings separately. When measured with FreeSurfer (FS), ε4-positive participants had smaller volumes at the 4.5-year compared to the 18-month assessment, but did not differ from the 18-month compared to baseline assessment. ε4-positive participants showed marginally smaller 18-month compared to baseline volumes, and significantly smaller 4.5-year compared to 18-month volumes. This pattern was not found in manual tracings, which yielded no significant differences. Error bars represent SEM.

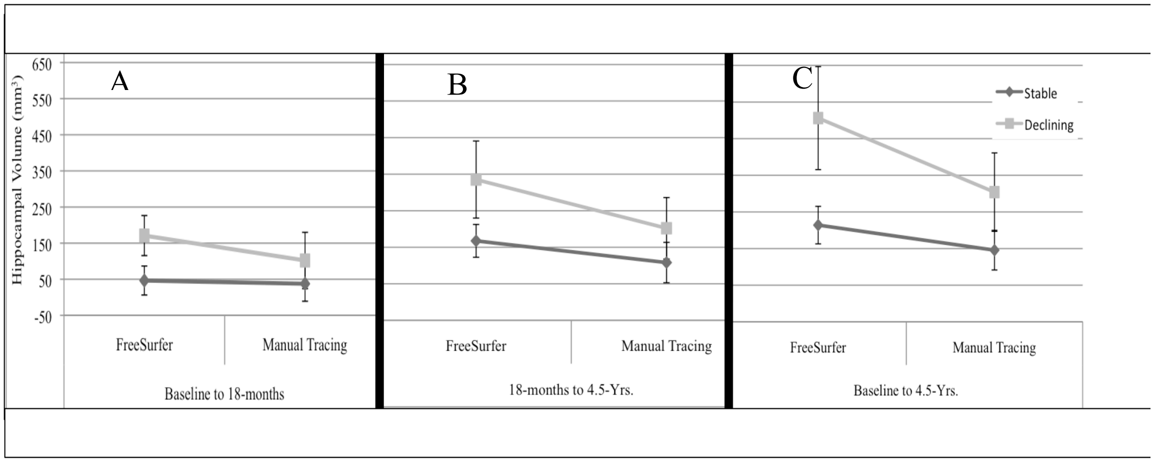

Figure 25. Three separate comparisons of hippocampal volume of method by group for each change interval. There were no significant differences between the amount of change in hippocampal volume between Stable and Declining individuals at any interval. The amount of change from baseline to 4.5 year volumes as measured by FreeSurfer was marginally greater in Declining compared to Stable participants. Data are presented as absolute values, such that higher values indicate greater atrophy. Error bars represent SEM.