ICER

INSTITUTE FOR CLINICAL AND ECONOMIC REVIEW

Diagnostic Tests for Alzheimer's Disease:
Generating and Evaluating Evidence to Inform Insurance Coverage Policy

Steven D. Pearson, MD, MSc, FRCP
Daniel A. Ollendorf, MPH, ARM
Jennifer A. Colby, PharmD

in conjunction with the
ICER Alzheimer's Disease Diagnostics Policy Development Group

December 10, 2012

Table of Contents

Policy Development Group

The Policy Development Group (PDG) is an independent group brought together by the Institute for Clinical and Economic Review (ICER) and composed of academic experts, patient advocates, clinicians, epidemiologists, ethicists, and medical policy representatives of stakeholder groups including health plans and manufacturers. PDG members and their affiliations are listed below; in addition, declared conflicts of interest and other potential influences on judgment are described in Appendix A to this document.

PDG Participant Name and Affiliation
Mark Bondi, PhD, ABPP/CN Professor of Psychiatry, University of California at San Diego (UCSD) School of Medicine Staff Neuropsychologist, VA San Diego Healthcare System
Robin Cisneros National Director, Medical Technology Assessment and Products, Kaiser Permanente
Joshua T. Cohen, PhD Deputy Director, Center for the Evaluation of Value and Risk in Health Research Associate Professor of Medicine, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center
Michele DiPalo Director, Health Services Evaluation, Blue Cross & Blue Shield of Massachusetts
Howard Feldman, MD, FRCP Executive Associate Dean, Research Professor of Neurology, Department of Medicine, University of British Columbia
G. Scott Gazelle, MD, MPH, PhD Director, Institute for Technology Assessment Professor in Radiology, Harvard Medical School Professor, Department of Health Policy and Management, Harvard School of Public Health
David Holtzman, MD Andrew B. and Gretchen P. Jones Professor and Chairman of Neurology Associate Director, Alzheimer's Disease Research Center, Washington University
Bradley T. Hyman, MD, PhD Director, Massachusetts Alzheimer's Disease Research Center Director, Alzheimer's Unit, MGH Institute for Degenerative Disease John B. Penney Jr. Professor of Neurology, Harvard Medical School
Kejal Kantarci, MD, MSc Associate Consultant, Division of Neuroradiology Associate Professor of Radiology, Mayo Clinic
Jason Karlawish, MD Associate Professor of Medicine, University of Pennsylvania School of Medicine Associate Director, Penn Memory Center Director, Alzheimer's Disease Center's Education and Information Transfer Core
Robert McDonough, MD Senior Medical Director, Clinical Research and Policy Development, Aetna, Inc.
Mark Mintun, MD Chief Medical Officer, Avid Radiopharmaceuticals, Inc. (a division of Eli Lilly)

Jeffrey R. Petrella, MD Associate Professor of Radiology, Duke University Medical Center
Creighton H. Phelps, PhD Chief, Dementias of Aging Branch (Acting) & Director, Alzheimer's Disease Centers Program Division of Neuroscience, National Institute on Aging (NIA)/ National Institutes of Health (NIH)
Murray Raskind, MD Director, Mental Health Services, VA Puget Sound Health Care System Professor and Vice-Chair, Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine
James Rollins, MD Director, Division of Items and Devices, Coverage & Analysis, Office of Clinical Standards and Quality (OCSQ), Centers for Medicare and Medicaid Services (CMS)
Alan Rosenberg, MD Vice President of Medical and Clinical Pharmacy Policy, Wellpoint, Inc.
Reisa A. Sperling, MD, MMSc Director, Center for Alzheimer's Research and Treatment Director, Education and Information Transfer Core, Massachusetts Alzheimer's Disease Research Center Co-Leader, Neuroimaging Program, Massachusetts Alzheimer's Disease Research Center Associate Professor of Neurology, Harvard Medical School
William H. Thies, PhD Chief Medical and Scientific Officer, Alzheimer's Association
Kathleen Welsh-Bohmer, PhD, ABCN Director, Bryan Alzheimer's Disease Research Center Professor of Psychiatry and Behavioral Sciences and Assistant Professor in Medicine, Duke University School of Medicine
Kristine Yaffe, MD Chief of Geriatric Psychiatry and Director of the Memory Disorders Clinic, San Francisco VA Medical Center Roy and Marie Scola Endowed Chair in Psychiatry; Associate Chair of Clinical and Translational Research; Professor, Department of Psychiatry, Neurology, and Epidemiology and Biostatistics, University of California at San Francisco
Eric Yuen, MD Vice President and Head of Clinical Development Johnson & Johnson – Janssen Alzheimer Immunotherapy

Purpose

In April, 2011, for the first time in 27 years, diagnostic criteria for Alzheimer's disease (AD) were revised by clinical and policy experts under the leadership of the National Institutes of Health and the Alzheimer's Association. These new guidelines were the first to include consideration of the findings of multiple forms of biomarker tests and highlighted the major changes that had occurred in how experts think about AD and design studies of potential treatments. This evolution in diagnostic criteria has come at a time when research on different diagnostic techniques is expanding rapidly, many new treatments intended to delay the progression of AD are undergoing evaluation, and public interest in obtaining access to promising tests and treatments is growing.

Outside of their uses in drug development and clinical research trials, the current clinical value of performing any type of formal diagnostic or biomarker testing for AD is controversial. The possible benefits of testing have been posited to include: helping identify patients, possibly even individuals with no symptoms, for initiation of treatment; helping clinicians decide whether additional diagnostic evaluation is necessary to look for disorders other than AD that impair cognitive function; reassuring patients who receive a negative test; and allowing patients with a positive test to plan their future more effectively. But at a time when available treatments for AD are unable to improve long-term outcomes, there is no consensus regarding the clinical benefits of diagnostic testing, and most imaging and biomarker tests for AD are not covered by public or private insurers. As the search for more effective treatments for AD continues, many questions remain about how to design clinical trials so that it is possible to evaluate different tests for AD in a way that will generate "adequate" evidence not only for patients and clinicians, but for insurers as well. The goal of this project is to seize the opportunity to address this policy need in a collaborative and pro-active manner. It is collaborative because this white paper is the product of a process involving input from patient advocates, clinicians, clinical researchers, manufacturers, and insurers. It is pro-active because the specific aim of the project is to define the standards by which evidence will be evaluated for coverage, both at the current time and after the potential advent of more effective treatments. This aim will be accomplished by providing specific research recommendations to help clinical researchers and manufacturers generate the level of evidence required to meet these standards.

The sections that follow include background on AD and the evolving paradigm of its diagnosis. Among the diagnostic tests under consideration most are considered "biomarkers," which can conceptually be used in several different ways to guide drug development, diagnose patients, and evaluate the outcomes of treatment. The focus of this white paper is on the evidence needed to validate tests as tools in the diagnosis of AD, but the broader potential use of different tests will be discussed. The white paper presents a summary of key diagnostic tests in development along with a framework for how technology assessment groups and insurers will evaluate the evidence on these tests as part of the coverage determination process. To establish a context for this framework, a summary of current clinical guidelines as well as major public and private payer coverage policies is provided, as is an overview of ongoing and planned clinical studies. Finally, major themes from each of these sections are incorporated into a set of recommendations to guide future research in diagnostic testing for AD. The ultimate goal is to frame a research agenda that can address not only the perceived needs of researchers, regulators, clinicians, and patients, but

can also incorporate the perspective of insurers responsible for evaluating the evidence to develop coverage and reimbursement policies for diagnostic testing for AD.

This white paper was written by the staff of ICER at the Massachusetts General Hospital's Institute for Technology Assessment (www.icer-review.org). ICER reached out to leading US-based clinical researchers and representatives of patient advocacy groups, life science manufacturers, and public and private insurers to form an Alzheimer's Disease Diagnostics Policy Development Group, whose goal was to discuss the state of the research and develop recommendations to guide future AD diagnostics research. The recommendations of this white paper are framed to serve as a guide for Alzheimer's research globally, but the input of the Policy Development Group was focused on the insurer-research environment in the United States. Although strongly informed by this input, the content of this white paper represents an attempt by ICER to bridge many different perspectives and therefore should not be interpreted as reflecting the opinions of individual Policy Development Group members or the organizations with which they are affiliated.

Executive Summary of Recommendations

Insurers and technology assessment groups look for evidence that can persuade them that diagnostic tests for AD improve patient outcomes. All the other potential uses for diagnostic tests, and in particular biomarkers, in drug development and clinical trial design are viewed as important by payers, but coverage determinations will be driven largely by whether insurers believe that there is adequate evidence to demonstrate that the use of a test will improve patient outcomes.

Based on this perspective, the following targeted recommendations are intended to guide the development of further research that will help create a body of evidence adequate to meet all relevant evidentiary standards. These recommendations include those intended to frame the broader research agenda to establish on more solid ground our understanding of the relationship between diagnostic tests and the course of AD, as well as recommendations focused on clinical trial design of studies designed to measure the effectiveness of test-and-treat strategies for AD.

Broad Research Agenda Recommendations

1. *In the current era of AD treatments of limited effectiveness, randomized controlled trials should be performed to evaluate diagnostic tests with potential overall net health benefits.*

2. *Develop a framework for assessing the social and economic impact of diagnosis of pre-clinical AD.*

3. *Looking to a future when there are more effective treatments for AD, continue to conduct biomarker studies in selected populations as well as in large population-based cohorts to evaluate the natural history of AD as well as the prognostic value of multiple combinations of neuropsychological testing and biomarkers.*

4. *Develop consensus standards for biomarker test deployment and interpretation.*

5. *As certain biomarkers gain validation for use as predictive of progression of disease, it will be important to study their predictive accuracy across the full spectrum of AD.*

6. *In studies that have used positive biomarker tests as inclusion criteria (enrichment design studies), include in baseline tests other potential biomarkers that can also be evaluated (nested marker-by-treatment-interaction studies). Ideally, always include additional test options that would be simpler, more accessible, and less expensive than the "gold standard" set of biomarkers used to qualify for inclusion.*

7. *Given that diagnostic testing for AD may involve expensive tests such as imaging, radionuclide tests, and CSF biomarkers, and that future therapies for AD may themselves be quite expensive, certainly on the cumulative, population-based level, evidence on comparative value should be included as a goal of the research agenda for AD diagnostics.*

8. *Given that many important clinical and economic outcomes occur years after diagnostic testing, a broad research agenda will benefit from the use of simulation modeling (decision analysis).*

Trial Design

1. *Design clinical trial protocols to enhance the generalizability of results to typical clinical practice.*

2. *Use a common set of consensus-based diagnostic test measurement thresholds and patient outcome measures.*

3. *Complement unusual enriched populations in early studies with studies that enroll representative patient populations in order to enhance the generalizability of results to real-world clinical practice.*

4. *Broaden the potential treatment population in trials.*

5. *For effective therapeutic agents developed through enrichment designs, consider further analyses to evaluate whether the original enrichment criteria were so narrow that less stringent enrichment criteria might identify many other patients who would benefit from treatment.*

6. *Retrospective assessment of a prognostic biomarker can only be done using data from well-conducted randomized controlled trials and with prospectively stated hypotheses, analysis techniques, and patient populations, with a pre-defined and standardized assay and scoring system for "positive" results. In other words: data mining should not be done to search retrospectively for combinations of clinical characteristics and biomarker results that are correlated with positive treatment outcomes.*

7. *Consider clinically-equivalent but lower-cost diagnostic strategies in translating trial results to clinical practice.*

Introduction

Alzheimer's disease (AD) is the most common underlying contributor to dementia, affecting over 35 million persons worldwide, a figure that is expected to nearly double by 2030 given the aging of the world's population (Alzheimer's Disease International, 2012). It has been estimated that AD and other dementias contribute a greater percentage of years lived with disability among those over age 60 than any other single category of disease (Mathers, 2003), account for approximately half of individuals aged 75 or older who are functionally dependent (Agüero-Torres, 1998), and are associated with a two- to five-fold increased risk of death (Helzner, 2008; Mehta, 2008). The total cost attributed to AD was estimated to be greater than $600 billion worldwide in 2010, including the costs of medical care, residential and community care, and uncompensated support provided by family and other caregivers (Alzheimer's Disease International, 2012).

AD is characterized by the accumulation of plaques comprised of the amyloid-β protein, and of neurofibrillary tangles of hyperphosphorylated forms of the protein tau, both of which are thought to contribute to neuron degradation and death and the subsequent cognitive, memory, and behavioral symptoms of AD (McKhann, 1984). Unfortunately, the only definitive method for diagnosing AD has been a combination of clinical diagnosis and findings of the pathognomonic plaques and tangles on examination of the brain at autopsy (McKhann, 1984). Various batteries of clinical assessment questions have been in use for nearly three decades and, while able to detect mild cognitive disorders, their accuracy in distinguishing between AD and other dementias or even non-dementia syndromes is variable (Jobst, 1998; Mayeux, 1998). The clinical value of brain imaging studies suggesting AD pathology is also still uncertain, as there are consistent reports of "positive" amyloid findings in the brain in approximately one-third of the cognitively normal older adult population (Jack, 2009). This has created a conundrum for the AD research community. As leading experts have asked, "How do older individuals spend years with a 'head full of amyloid' and remain apparently healthy?" (Sperling, 2011b)

AD therefore has neither a gold standard diagnostic test (prior to autopsy), nor a single test with which clinical researchers can confidently track the potential impact of treatment intended to prevent or slow the progression of disease. Meanwhile, the general view of the time course and potential therapeutic window for AD has changed, and clinical trials have begun to shift focus toward treatment of earlier, milder cases of AD. As a consequence, interest in identifying diagnostic biomarkers for AD has been growing, and new methods are rapidly emerging, including new imaging-based tests, markers detectable in blood, plasma, and cerebrospinal fluid (CSF), and even retinal scans. And, as always, these new diagnostic approaches are being developed in the context of intense patient, clinician, and public interest. Responding to the current and future impact of AD, the U.S. Department of Health and Human Services (HHS), at the behest of the White House, recently developed a "National Plan to Address Alzheimer's Disease" that describes five far-reaching and ambitious goals: (1) prevent and effectively treat AD by 2025; (2) enhance care quality and efficiency; (3) expand supports for persons with AD and their families; (4) enhance public awareness and engagement; and (5) enhance the data and monitoring capability to track progress against the clinical goals (U.S. Department of Health and Human Services, 2012). This white paper seeks to contribute to the ability of the AD research community to achieve these goals by developing a clearer picture of how to design studies that will produce high-quality evidence on various diagnostic approaches for AD.

The Evolving Paradigm of Alzheimer's Disease and the Role of Biomarkers

The mainstay of AD diagnosis in the clinical and research community until recently was a set of criteria based on a history of insidious onset and gradual progression of memory and other cognitive impairments, with confirmation through neuropsychological testing. These "clinical criteria" were first promulgated in 1984 under the auspices of a workgroup convened by the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's Disease and Related Disorders Association (ADRDA) (McKhann, 1984). Behind this effort lay the premise that clinical symptoms and pathology would always be tightly linked in AD (Jack(a), 2011), and therefore the criteria focused on the presence of symptoms indicative of deficits in memory and/or other cognitive functions, as well as absence of pathological evidence of other disorders that could account for the symptoms.

While these clinical criteria have proven useful in identifying patients with dementia and those with milder forms of cognitive impairment that are likely to progress to dementia, they are not as helpful in distinguishing between patients whose cognitive symptoms are due to AD pathology and those whose symptoms arise from other causes.   Even when imaging or CSF tests confirm likely AD pathology, research conducted since the introduction of the clinical criteria has established that the correspondence between symptoms and AD pathology is not always consistent, and the progression of cognitive symptoms and AD pathology does not track concurrently on an identical pathway. For example, as noted earlier, extensive AD pathology can be present in the absence of any obvious cognitive symptoms (Davis, 1999; Knopman, 2003; Price, 1999). Studies have also shown that changes in cerebral pathology may occur 10 or more years before the emergence of cognitive decline or dementia (Sperling(b), 2011; Jack, 2009). In a recent cohort study of patients with the autosomal dominant form of AD, changes in amyloid-β and other biomarkers began to occur 15-25 years before the expected onset of dementia (Bateman(b), 2012).

These observations have been interpreted by many in the AD clinical and research communities to imply an ordered, but staggered sequence to the development of AD pathology and its clinical consequences. According to this new view, rather than developing simultaneously, amyloid and neurofibrillary pathology occur over different time scales. Amyloid pathology is thought to develop first during a 10-15 year long preclinical phase, whereas the development of neurofibrillary pathology (tangled tau proteins) begins later in this phase and accelerates before the emergence of the symptomatic phase of AD (Jack(a), 2011).

This hypothesis of a sequential relationship between AD pathology and cognitive symptoms lies at the foundation of the new criteria for the diagnosis of AD developed in 2011 by a multi-stakeholder workgroup convened by the National Institute on Aging (NIA) and the Alzheimer's Association (AA) (Jack(a), 2011; Sperling(a), 2011; Albert, 2011; McKhann, 2011). The key difference between the new criteria and the earlier clinical criteria from 1984 is the incorporation of biomarkers and their role in defining three different stages of AD: 1) preclinical; 2) mild cognitive impairment (MCI); and 3) AD dementia. The goal of including preclinical AD in the new diagnostic guidelines was to facilitate the possibility of future presymptomatic/preclinical treatment of AD, but the workgroup members argued that the criteria were only appropriate for research purposes at this time since the

extent to which biomarkers of AD pathology predict a cognitively normal individual's subsequent clinical course remains to be clarified. The framework for preclinical AD includes three stages, all of which depend upon biomarkers: 1) asymptomatic amyloidosis; 2) amyloidosis plus neurodegeneration (e.g., high CSF tau); and 3) amyloidosis plus neurodegeneration plus subtle cognitive decline.

As subtle cognitive decline worsens, the next diagnostic category after the preclinical period is mild cognitive impairment (MCI). For MCI the workgroup presented two sets of diagnostic criteria: one set of "core clinical criteria" that could be used by clinicians without access to advanced imaging techniques or CSF analysis; and another set of criteria for use in clinical research settings that incorporate imaging and CSF biomarkers. Although the workgroup emphasized that further work is needed to validate the criteria that depend upon biomarkers, the pattern of biomarker results is presented as a tool to judge the level of certainty in a diagnosis of MCI due to AD as opposed to other possible causes. Thus, given a patient who meets the core clinical criteria for MCI, positive amyloid and neuronal injury test results produce "high likelihood" that the MCI is due to AD, whereas positive tests on only one or the other test lead to "intermediate likelihood," and negative amyloid and neuronal tests lead to the "lowest" likelihood of MCI due to AD.

The distinction between MCI and AD dementia in the new diagnostic criteria depends on findings from a set of core clinical criteria, and not on the levels or progression of biomarker findings. As with MCI, however, biomarker results are incorporated as a tool for judging the certainty that dementia is due to AD pathology and not some other cause. These categories of relative certainty are called "probable AD dementia," "possible AD dementia," and "dementia unlikely to be due to AD." The workgroup did not propose the use of AD biomarker tests for routine diagnostic purposes among patients with dementia at the present time for four reasons: 1) the core clinical criteria provide reasonable diagnostic accuracy and utility in most patients; 2) more research needs to be done to ensure that criteria that include the use of biomarkers have been appropriately designed; 3) there is limited standardization of biomarkers from one locale to another; and 4) access to biomarkers is limited to varying degrees in community settings. Despite these concerns, however, the workgroup deemed biomarker tests useful "as optional clinical tools for use where available and when deemed appropriate by the clinician" (McKhann, 2011). The 5th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) is also in the process of changing diagnostic criteria for MCI and dementia to reflect the use of biomarker testing.

The new diagnostic criteria for pre-clinical AD, MCI, and AD dementia represent an important sea change in the view of the clinical research community on the utility of amyloid and neurodegenerative biomarkers for identifying patients with different phases of AD. The utility of biomarkers is framed cautiously as applicable largely in research efforts, but the explicit algorithms linking biomarker test results to the probability that cognitive symptoms are due to AD expresses a clear conviction in the new paradigm of AD as a disease that begins with a decades-long course of pre-symptomatic pathophysiological changes in amyloid accumulation and resulting neurodegeneration that predates cognitive changes. This broad acceptance of the potential role of biomarkers, particularly in their ability to identify individuals with minimal or no symptoms who are at high risk of progressing to full AD dementia, has led to biomarker incorporation in many trials as means to "enrich" the population of eligible patients with those most likely to have AD pathology and to progress rapidly without successful treatment. And as the community of clinical experts who lead these trials have gained increasing evidence of the ability of biomarker tests to discriminate

between patients with and without AD, and even to help identify asymptomatic patients at high risk for AD, the diffusion of biomarker testing from the research into the clinical domain seems ever more likely. The new paradigm in the course of AD has therefore shifted research efforts toward earlier intervention, and highlighted the importance of distinguishing the evidence that will be needed to move biomarkers from research tools into broader use as diagnostic and prognostic tests in the clinical arena.

The Different Roles of Biomarkers in AD Research and Practice

A biomarker is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (Colburn, 2000). In clinical practice biomarkers can be potentially used in a variety of ways: 1) to predict the course of illness if not treated (prognostic test); 2) to identify patients with an underlying pathology that is the target of a specific therapeutic agent (predictive, or "theragnostic" test, alternatively spelled "theranostic"); 3) to evaluate the results of treatment (surrogate outcome test); and 4) to diagnose patients with a disease (diagnostic test) (Hampel, 2010; Blennow, 2010). In the practice of oncology, biomarkers have been established for various tumor types, with many biomarkers now serving as predictive tests that can prospectively identify individuals who are likely to have a favorable clinical outcome, such as improved survival, slower progression of illness, or decreased toxicity, from a specific treatment (Mandrekar, 2009). Validated predictive oncologic biomarkers now widely used in clinical practice include *KRAS* as a predictor of efficacy of panitumumab and cetuximab in advanced colorectal cancer; HER-2 as a predictor of the efficacy of trastuzumab in breast cancer; and epidermal growth factor receptor (EGFR) as a predictive marker of response to tyrosine kinase inhibitor treatment for non-small cell lung cancer (Mandrekar, 2009).

The current role of biomarkers in AD is quite different. There are no biomarkers that have been validated as a prognostic test, a surrogate outcome test, or a diagnostic test for use in routine clinical practice. Nevertheless, given the current paradigm for the pathogenesis and evolution of AD there are several important roles that biomarkers currently play in drug development. First, biomarkers such as amyloid and tau tests can be used in animal studies and in early Phase I and II human trials to establish whether a drug hits its intended target and alters the intended biochemical mechanism. Small, short-term trials that provide biochemical evidence of an effect of the drug on the central pathogenic processes are of great value to make go/no-go decisions before embarking on larger Phase III clinical trials. Second, when designing Phase III trials biomarkers can be used as tools to "enrich" the sample of patients enrolled to include only those – from among all those meeting clinical criteria – who also have the underlying pathology that is the intended target of the drug.

Potentially, biomarkers could also be used as surrogate outcomes in clinical trials of AD therapeutics, but before a biomarker will be accepted by regulators as a surrogate outcome in Phase III studies, a link between the treatment-induced change in the biomarker and the desired clinical outcome has to be firmly established (Hampel, 2010). A biomarker can serve as a surrogate outcome only when it can be considered as a substitution for a clinically relevant end point that is a direct measure of how a patient feels, functions or survives. This requires evidence from randomized clinical trials that improvements in the surrogate end point lead consistently to improvement in the relevant patient outcome (Katz, 2004; Temple, 1999). While some trials of

therapeutic agents for AD have shown a dose-dependent reduction in the amount of CSF amyloid biomarkers, there has as yet been no firm evidence linking change in a biomarker to the course of clinically significant end points (Blennow, 2010).

So what of the potential role of biomarkers in the diagnosis of AD? Much of the rest of this white paper will be concerned with the evidence framework for validating biomarkers as diagnostic tests, along with corresponding research design features that will be needed to generate the necessary data.   The question of how to evaluate the possible clinical value of biomarker tests that can rule-out the possibility of AD pathology will also be addressed, even when a "positive" result on the same test cannot confirm a positive diagnosis. Throughout, the discussion on biomarker tests will be framed by the recent shift in the focus of clinical research toward therapeutic trials involving patients with minimal symptoms of dementia or even asymptomatic individuals. The result of this shift, driven by the change in the paradigm of the course of AD pathology and illness, has been the increasing use of biomarkers to identify and enroll individuals who have evidence of the underlying pathological mechanism that is the specific target of an experimental drug. Whereas in oncology the general pattern has been for biomarker validation to occur in trials involving patients with an established pathology-based diagnosis, the evolution of biomarker evidence in AD will proceed largely from therapeutic research studies among individuals who do not yet meet the existing diagnostic criteria for AD. Therefore, if in some future clinical trial researchers find that a particular targeted treatment benefits a biomarker-defined subset of individuals, it will be said that the biomarker profile for treatment-responsive patients represents a diagnostic test for a new condition: "treatment-responsive" (early) AD. This linkage between biomarker, therapeutic agent, and improved outcome is the hallmark of what has been called the "theragnostic" model of drug development (Blennow, 2010). As is discussed in later sections of this paper, if this occurs in AD, the biomarker will be viewed as a validated diagnostic test. However, there will be many questions about its ability to identify patients outside the initial trial population who may benefit from treatment; there will also be questions about whether other diagnostic approaches might be more accurate, equally accurate but less invasive or less expensive, or even less accurate but more widely accessible. These are some of the questions that researchers, clinicians, patients, and insurers will be asking, even after we cross the threshold into a new landscape of "treatment-responsive" AD.

Methods of Diagnostic Testing for Alzheimer's Disease

As noted earlier, the mainstay of the diagnosis of AD for research and clinical purposes until very recently has been a set of clinical criteria in place since 1984. Clinical criteria are relatively reliable in diagnosing AD dementia and MCI but are not as helpful in distinguishing between patients whose cognitive symptoms are due to AD pathology and those whose symptoms arise from other causes. Thus it has been estimated that approximately 15-20% of patients currently enrolled via clinical criteria in clinical trials for AD do not have true AD pathology (Rinne, 2010). In the clinical setting, up to 25% of patients who meet clinical criteria for AD dementia are believed not to have the underlying pathology (Klatka, 1996; Pearl, 1997; Rasmusson, 1996; Schneider, 2010). The recent changes in the understanding of the natural history of AD and the progression of cognitive decline have resulted in an unprecedented level of interest in the development of biomarkers that can help researchers and clinicians identify preclinical AD and MCI due to AD.

In addition to clinical assessment in which the history, time course, degree of functional cognitive decline, and types of signs and symptoms of a disease process are ascertained both from informants and patients, other tests are often utilized to augment the clinical assessment such as neuropsychological tests and additional biomarkers. In the current standardized evaluation of patients who may be developing dementia, the additional tests that are recommended currently include a brain imaging study (MRI or CT scan of the brain to rule out structural causes of dementia) and plasma levels of B12 as well as thyroid function tests. Additional blood and other tests may be warranted depending on the clinical presentation. Neuropsychological tests and biomarkers currently used for the diagnosis of different phases of AD as well as for monitoring patient outcomes are described in the sections below. It is important to note that each test may serve multiple purposes, however (i.e., diagnostic, prognostic, and/or theragnostic).

Neuropsychological tests

There is a wide assortment of neuropsychological tests used in the clinic and across research studies of MCI and dementia. Recent consensus workgroups on the diagnosis of MCI (Albert et al. 2011) and vascular contributions to cognitive impairment (Gorelick et al. 2011) both stress the necessity for a comprehensive cognitive battery and its importance in determining whether there is objective evidence of cognitive decline and, if present, the severity of that decline. Broadly speaking, Gorelick et al. (2011) note that the diagnosis of dementia or MCI "must be based on cognitive testing, and a minimum of 4 cognitive domains should be assessed: executive [functions]/attention, memory, language, and visuospatial functions." Albert et al. (2011) also emphasize that neuropsychological assessment is optimal for objectively assessing the degree of cognitive impairment for an individual. Both workgroups explain that scores on tests for individuals with MCI are typically 1 to 1.5 standard deviations below the mean for their age and education matched peers on culturally appropriate normative data and are preferred over qualitative descriptions of cognitive symptoms.

There are a variety of neuropsychological tests that are useful for identifying those patients with MCI who have a high likelihood of progressing to dementia. Among the most valuable are those tests assessing episodic memory (e.g., learning and memory for word-lists, stories, geometric

figures). A number of longitudinal studies have shown that episodic memory performance in cognitively normal individuals can predict subsequent dementia many years prior to diagnosis (for review and meta-analysis, see Bäckman, 2005 and Twamley, 2006). Word-list tests of episodic memory have been among the most sensitive neuropsychological markers of the preclinical period of Alzheimer's disease, often outperforming a variety of CSF and imaging biomarkers (see Devanand, 2008; Gomar, 2011; Heister, 2011; Jedynak, 2012; Landau, 2010). Because other domains of cognition can be impaired among individuals with MCI and preclinical AD (Brandt, 2009; Clark, 2012; Mickes, 2007), it is also important to examine tests of executive functions (e.g. reasoning, problem-solving, and planning), language (e.g. naming, fluency), visuospatial skills, and attentional control (e.g. simple and divided attention).

Examples of available and commonly utilized neuropsychological tests are listed below.

Episodic Memory
• Rey Auditory Verbal Learning Test
• California Verbal Learning Test
• Logical Memory (story paragraphs)
• Rey-Osterrieth Complex Figure

Executive Functions/Attention
• Trail Making Test
• Wisconsin Card Sorting Test
• Digit Span

Language
• Boston Naming Test
• Verbal Fluency

Visuospatial Skills
• Block Design
• Clock Drawing
• Complex Figure Copy

The most common scales used in clinical trials are the Alzheimer's Disease Assessment Scale – Cognitive Behavior (ADAS-Cog), the Mini-Mental State Examination (MMSE), and the Clinical Dementia Rating (CDR) scale, although these tests lack sensitivity to the mildest forms of cognitive impairment (Salmon, 2002; Chang, 2011; Jedynak, 2012).

Examples of available and commonly utilized cognitive and functional ability tests are listed below.

• Mini-Mental State Examination (MMSE): used to briefly assess domains such as memory, attention and orientation through the administration of 11 different tasks (Feldman, 2008).

• Alzheimer's Disease Assessment Scale – Cognitive Behavior (ADAS-Cog): used to evaluate cognitive behavior in areas such as language, orientation and memory-related realms such as word-recall and word-recognition (Peña-Casanova, 1997).

- Clinical Dementia Rating (CDR): used to evaluate 6 domains of cognitive and functional abilities, including personal care, judgment and problem solving through a semi-structured interview process (Morris, 2011).

Finally, while not a cognitive screening test, the Neuropsychiatric Inventory (NPI) is also commonly used to assess behavioral changes, and related caregiver distress, in 12 distinct areas such as anxiety, apathy/indifference and night-time behavior disturbances through an interview process (Cummings, 1997).

Imaging

Multiple imaging modalities have been developed for use in AD, either to document changes in brain volume and structure, metabolism, or blood perfusion indicative of AD, or to measure amyloid-related biomarkers associated with cognitive decline. Four imaging modalities have been used as secondary end points in clinical trials on AD: structural magnetic resonance imaging (sMRI), functional MRI (fMRI), magnetic resonance spectroscopy (MRS), and positron emission tomography (PET). These techniques provide information on the regional distribution of changes on a macroscopic (fMRI, PET, MRS) or a mesoscopic (i.e., intermediate measurement between microscopic and macroscopic, as with MRI) scale. These and other key imaging tests are described below.

- *PET amyloid-β imaging*: involves use of radioactive compounds such as florbetapir (Amyvid™), along with PET imaging, to identify areas of accumulation of amyloid-β plaques. Amyvid is the first compound in its class to receive FDA approval (in April 2012) for use in AD diagnosis, and is labeled as an adjunct diagnostic tool. The manufacturer's label indicates that a negative scan result indicates a reduced likelihood that any cognitive impairment is due to AD, but that a positive scan does not establish a diagnosis of AD or any other neurodegenerative disease, as excess amyloid-β is also found in adults with normal cognition. Amyvid is also not intended for use in predicting the development of dementia or in monitoring the effectiveness of any current or investigational therapy (Amyvid™ package insert, 2012).
- *Fluorodeoxyglucose positron emission tomography (FDG-PET):* used to evaluate cerebral glucose metabolism using a radioactive tracer, flurodeoxyglucose (FDG). Hypoactive glucose metabolism has been associated with both MCI and AD (Small, 2008).
- *Structural magnetic resonance imaging* (*sMRI*): used to evaluate atrophy of different brain structures such as the medial temporal lobe and the parietal cortex, as well as loss of hippocampal volume, which has been associated with progression from MCI to AD (Jack(b), 2011).
- *Magnetic resonance spectroscopy (MRS):* provides quantitative biochemical measures of compounds in brain tissue. The best established MRS marker is an amino acid (NAA) which reflects the functional status of neuronal mitochondria. A reduction of NAA levels independent of brain atrophy is a consistent finding in AD (Kantarci, 2007).
- *Structural computed tomography (sCT):* also used for evaluation of brain atrophy, and may be utilized for patients unable to undergo MRI due to the presence of metal (e.g. a pacemaker) (Wattjes, 2009).
- *Functional MRI (fMRI):* used in combination with cognitive assessment, such as evaluation of a memory encoding task, to map resulting neural activity in the brain (Machulda, 2003).

- *Single-photon emission computed tomography (SPECT):* used with a variety of radioactive tracers to assess areas of hypoperfusion, or decreased regional cerebral blood flow, which have been associated with the early stages of AD (Mitsumoto, 2009; Matsuda, 2007).

Cerebrospinal Fluid (CSF)

Four CSF biomarkers have been evaluated in a large number of independent studies: amyloid-β_{40}, amyloid-β_{42}, total tau, and phosphorylated tau. Amyloid and phosphorylated tau reflect the core elements of the disease process in AD as it is currently known: levels of brain amyloid and neurofibrillary tangles. Conversely, total tau is a nonspecific marker for axonal damage that mirrors the activity of the neurodegenerative process. The combination of elevated levels of total tau together with reduced levels of amyloid-β_{42} or reduced amyloid-β_{42}/ amyloid-β_{40} ratio is a consistent finding in biomarker studies of patients with different stages of AD, including MCI (Hampel, 2010).

Emerging Testing Methods

In addition to the tests described above, all of which have been evaluated in multiple clinical studies, a variety of newer tests are being explored for their diagnostic and prognostic capabilities in AD. These include newer imaging- and CSF-based techniques as well as novel approaches to biomarker measurement intended to further increase convenience and potentially reduce cost. These emerging tests include:

- *Plasma Amyloid- β_{40} and Amyloid- β_{42}*: studies performed to date have yielded inconclusive results regarding the correlation between plasma concentrations and accumulation in the brain (Humpel, 2011).
- *Other plasma-based biomarkers*: those currently under investigation include markers of inflammation such as C-reactive protein, interleukins, and tumor necrosis factor, as well as homocysteine and a variety of lipoproteins (Cedazo-Minguez, 2010). In addition to individual biomarkers, research is also ongoing on multiple protein arrays and their correlation with neuropsychological testing results (O'Bryant, 2011).
- *Vision-based biomarkers:* a variety of techniques are being explored to identify patients with AD, including:
  - measurement of blood vessel width in the eye through the use of retinal photography
  - evaluation of specific eye movements as a means to discriminate between neurodegenerative diseases
  - changes in the retinal nerve fiber layer as a marker of cerebral axonal degeneration
- *New imaging approaches,* including:
  - SPECT for assessment of dopaminergic dysfunction
  - PET for evaluation of glycine and histamine receptor levels
  - Use of a novel fMRI index for brain activity
- *New CSF biomarkers*, including:
  - Alternate forms of amyloid-β
  - Visinin-like protein-1, a marker of neuronal injury
  - Inflammatory markers such as interleukins

A Conceptual Approach to Evaluating Evidence on Alzheimer's Disease Diagnostic Tests

Insurers, as well as patients and clinicians, have one overriding question that they expect to have answered about potential diagnostic tests for Alzheimer's disease: is there adequate evidence to demonstrate that a diagnostic test improves patient outcomes? New tests may not help improve outcomes for many reasons: they may have poor accuracy, leading to many false positive and/or false negative results; they may be accurate but may be no better than existing clinical methods of diagnosis; they may be better than existing clinical methods but not change clinical decisions about treatment; they may even change clinical decisions about treatment but the treatment choices available may not provide significant clinical benefits. Thus the central question asked by insurers and other stakeholders about diagnostic tests for AD reflects the reality that there are many reasons why tests may not necessarily lead to improved patient outcomes. It also highlights that a judgment must be made about whether evidence is strong enough or persuasive enough to be "adequate" to merit insurance coverage. In this section we present an analysis of conceptual frameworks for the evaluation of evidence on AD diagnostic tests. Our goal is to describe how evidence on AD diagnostic tests will be evaluated by insurers and by technology assessment groups that often evaluate evidence to help guide insurers in coverage decision-making. This view is intended to help clarify where the evidence "gaps" are today, what kinds of outcomes in diagnostic studies should be sought in the future, and how the "strength" of evidence will be judged when decisions are made about whether the evidence is adequate to demonstrate that a diagnostic test improves patient outcomes. This information is critical in framing recommendations for the design of the diagnostic test research agenda of the future.

Analytic frameworks
An analytic framework is one useful way to visualize the many links in an evidentiary chain that must be considered when judging whether tests ultimately improve patient outcomes. A simplified analytic framework for AD diagnostics is shown on the following page.

Simplified Analytic Framework: Diagnostic Testing for Alzheimer's Disease
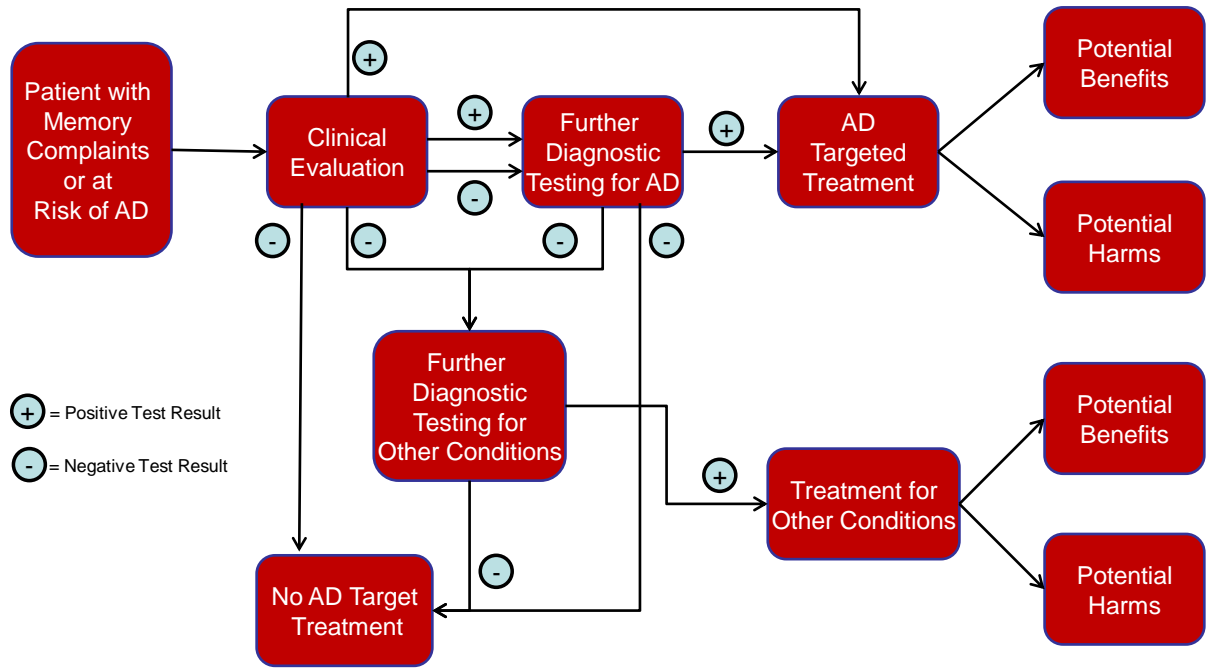
In this simplified framework, patients at risk for AD or with memory or other cognitive complaints are first evaluated clinically, often including the administration of some type of neuropsychological testing. For patients who are found not to have cognitive impairment, three consequent actions are possible: 1) no further investigation or treatment; 2) further diagnostic testing for other conditions; or 3) further diagnostic testing specifically for AD pathology to confirm a "negative" diagnosis.

If the clinical evaluation is positive or equivocal for MCI, dementia, or an AD-associated disease such as stroke due to amyloid angiopathy, the clinician could move directly either to targeted AD treatment or the clinician could choose to conduct further diagnostic testing to confirm or rule-out AD. This step in the analytic framework highlights the issue that judgments regarding the clinical value of AD diagnostic testing will depend upon whether there is evidence that further diagnostic testing adds information that makes the diagnosis more accurate than clinical evaluation alone. An evaluation of "accuracy" requires two elements: 1) judgment of the pre-test probability of illness based on the findings of the clinical evaluation; and 2) evidence on the sensitivity and specificity of *further* testing – not simply the sensitivity and specificity of the tests vs. some gold standard, but the sensitivity and specificity of the tests as they would be used in the sequence of evaluation in real-world clinical practice.

Not only does the diagnostic testing need to contribute to a more accurate diagnosis, but the analytic framework also clarifies that evidence or a strong assumption is needed that testing changes the decisions regarding subsequent testing for other conditions, treat/no-treat decisions, or the selection of treatment that would have been made without the tests. The need for evidence on these linkages between test results and test actions is true whether the focus is on the clinical value of a positive test result or on the clinical value of a negative test result. For example, as can be seen in the analytic framework, there is the potential for a negative diagnostic test for AD to affect the conduct of subsequent testing looking for other possible conditions. This is part of the potential clinical impact of a test that seeks to "rule-out" AD, but any potential impact requires that diagnosis of an alternative condition will lead to a course of treatment that produces overall improvement in health outcomes.

Given that no test is perfect, the evidence on a diagnostic test must be available to evaluate the consequences of false positive and false negative test results in changing treatment decisions and outcomes. The rates of false positive and false negative results of any test are dependent not only on their sensitivity and specificity, but on the underlying prevalence of the disease in the population to be tested. Thus evidence to evaluate this step in the analytic framework will require evidence allowing judgment of the performance and outcomes of testing in patient populations with different underlying rates of illness.

The final step in the analytic framework links the treatment or treatments available to the harms and benefits of treatment. Which harms and which benefits that are relevant for consideration of insurance coverage will be considered in subsequent sections of this paper. But the framework makes clear that the balance of all harms and benefits must be considered in order to evaluate the net health benefits of treatment. Ultimately, it is the ability of a diagnostic test to produce a higher net health benefit for patients that will guide insurance coverage decisions. The evidentiary links between all the steps in this analytic framework highlight the many areas in which evidence will be sought and evaluated by technology assessment groups and insurers seeking to make a judgment about the potential impact on net health benefit caused by the introduction of a new diagnostic test for AD.

Evidence Hierarchies, Analytic Validity, Clinical Validity, and Clinical Utility
An analytic framework can be helpful in clarifying the many steps in a logic chain for which evidence or a strong assumption must exist in order to demonstrate that the introduction of a new diagnostic test will improve patient outcomes. There are two other commonly used conceptual approaches to identifying the type of evidence needed for the evaluation of diagnostics. The first of these is an evidence hierarchy developed in the 1990s for imaging tests composed of six levels of evidence, and the second a conceptual model arising out of efforts to evaluate genetic tests that has three levels of evidence: "analytic validity," "clinical validity," and "clinical utility." We will describe both models because although the underlying concepts are similar the language and specifics are different enough that researchers seeking to understand how technology assessment groups and insurers will judge evidence on diagnostics should be familiar with both.

The evidence hierarchy approach, which assigns higher importance to different kinds of evidence, has gained wide use in evaluations of the literature on therapeutic interventions, with randomized, double-blind clinical trials (RCTs) and systematic reviews at the top of the hierarchy due to their perceived lower risks of bias, and observational studies and case series at lower levels.

But a different evidence hierarchy is often used for diagnostic tests, as RCTs are often not feasible and key patient outcomes of interest may lie many years in the future following the use of a test. In the early 1990s, Fryback and Thornbury developed an influential hierarchy of evidence specifically for imaging tests, but the hierarchy has been widely applied to all forms of diagnostic testing (Fryback, 1991). The hierarchy is presented in Table 1 below. Each level of evidence is shown with corresponding examples of the relevant outcome measures for studies at that level. We also include the analogous level of evidence from the genetic testing model that is discussed in parallel.

Table 1. Fryback and Thornbury hierarchy of evidence for diagnostic testing and genetic testing evidence categories promulgated by the Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group.

Diagnostic Imaging Evidence Hierarchy Level	Genetic Testing Evidence Category	Example of Outcome Measures
1. Technical Efficacy	1. Analytic validity	Interpretable scan resolution, accuracy and reliability of tests of CSF proteins to measure CSF protein levels, inter-reader and inter-laboratory reliability of test results
2. Diagnostic Accuracy	2. Clinical validity	Sensitivity/specificity vs. gold standard test or vs. some other standard
3. Diagnostic Impression		Change in presumptive diagnosis following introduction of new test results
4. Diagnostic Action		Initiation or cessation of treatment; impact on use of additional diagnostic studies
5. Patient Outcomes	3. Clinical utility	Cognitive/functional decline, time to institutionalization, side effects of treatment driven by test results, mortality
6. Societal Outcomes		Cost-effectiveness of testing

At the first level of the hierarchy are studies of technical efficacy, which evaluate whether the tests in question are able to measure accurately and reliably what they purport to measure. Examples of this kind of study include a test of whether a PET scan suggesting a certain level of amyloid in the brain is found to be accurate upon autopsy, or whether an MRI scan of the hippocampus can be repeated on different machines and interpreted by two different readers with the same result. This first level in the evidence hierarchy corresponds exactly with the "analytic validity" category of evidence used in the laboratory testing community and popularized by the Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Workgroup, a multidisciplinary academic group convened by the Centers for Disease Control in 2004 to establish a systematic, evidence-based process for assessing genetic tests and other applications of genomic technology.

Level 2 tests evaluate test accuracy when compared to a "gold standard" approach in a given population. This kind of study would include those that evaluate the correspondence between "positive" imaging or CSF studies and findings of dementia that meet AD core clinical criteria; it could also include studies of emerging diagnostic techniques that use imaging or CSF test results as the reference standard. This level also corresponds exactly with an EGAPP category: clinical validity.

Level 3 studies in the evidence hierarchy, however, have no specific analogue in the EGAPP model. Level 3 studies cross the step identified in the analytic framework from issues of diagnostic performance to the question of the impact of test results on the diagnostic impression of clinicians. Studies at this level can evaluate the impact of test results on the "confidence" of clinicians in their diagnosis, or, more usefully, try to measure the impact on actual presumptive diagnoses, tracking how often test results change clinicians' diagnosis and/or plans for treatment. This is a logical precursor to level 4 studies, which not only evaluate the reported plans for treatment but also capture the effect of the diagnostic test on the actual actions taken in treatment or management. This level in the evidence hierarchy also does not correspond to a specific category in the EGAPP model. Studies at this level often focus on the treat/no-treat choice, but they can also evaluate selection of treatment when there are multiple options or they can evaluate any change in further diagnostic workup following the receipt of test results.

Level 5 studies are those that are able to measure and compare actual patient outcomes experienced by patients who receive a diagnostic test and patients who do not. The outcomes of studies at this level would include potential risks of the test itself (e.g. headache after lumbar puncture) and also capture downstream effects of the treatments that are guided by test results. This level corresponds to the final EGAPP category, clinical utility, which they define as follows: "evidence of improved measurable clinical outcomes, and the test's usefulness and added value to patient management decision-making compared with current management without testing" (Teutsch, 2009). Level 6 studies in the evidence hierarchy introduce a "societal" perspective by examining not only clinical outcomes but also cost and cost-effectiveness considerations. There is no corresponding category in the EGAPP model.

It is important to note that, unlike the evidence hierarchies for studies of therapeutic interventions, neither the Fryback and Thornbury evidence hierarchy nor the EGAPP genetic test evidence categories specify the study design of the studies in each level/category. Thus, for example, studies that would provide level 5/clinical utility evidence could be RCTs in which patients are randomized to get the test or not, but studies in this level could also be prospective or retrospective cohort designs, or even well-conducted case series with historical controls. Judgments about whether

there is "adequate" evidence to demonstrate that diagnostic tests lead to improved patient outcomes thus require a two-part process: determining whether the level/category of evidence is adequate, and evaluating whether the quality of the evidence in each level is adequate to be persuasive.

When robust evidence at level 5/clinical utility is available, evaluation for insurance coverage is usually relatively straightforward. Similarly, level 1/analytic validity evidence provides too little assurance of evidence linking test results to patient outcomes to warrant serious consideration for coverage. It is therefore worth exploring further here some considerations regarding evidence levels 2, 3, and 4: evidence that spans between tests of diagnostic accuracy to those evaluating the impact of testing on treatment selection. Level 2 evidence, in particular, is the most common type of study performed and the interpretation of the results for insurance coverage often proves controversial. In most cases a new diagnostic test is being compared to a gold standard test, such as when non-invasive colorectal cancer screening tests are compared to optical colonoscopy as the standard to determine their sensitivity and specificity. But the development of biomarker tests for diagnosing AD and predicting outcomes will most likely present a completely different scenario and raise different key questions. The more recent treatment trials for AD therapeutics have been designed to include potential imaging and CSF biomarker tests for all patients as part of baseline testing, and in some cases as part of the eligibility criteria. The outcomes of all patients will be followed with the putative biomarker results as potential predictors of treatment response. Any positive impact of treatment will be identified in relationship to these biomarker results, and therefore there will always be direct evidence of the outcomes of treatment among these cases. In the scenario in which positive treatment outcomes are achieved, it seems most likely that treatment success will be identified only within a subset of patients with particular biomarker characteristics at baseline. In a sense, future patients with these biomarker results will have a new disease: "treatable MCI" or "treatable AD dementia," and the set of biomarker results associated with positive treatment response might then serve as a new gold standard against which future tests would be tested.

If and when this new gold standard exists for the diagnosis of patients who, as a class, will have positive treatment outcomes, then AD diagnostics will enter a second phase in which alternatives to the gold standard diagnostic test(s) will be evaluated. Researchers may develop new tests with potential advantages including greater sensitivity and specificity, or tests that may have equivalent diagnostic accuracy but other advantages, including less invasiveness or lower cost. In this future phase technology assessment groups and insurers will continue to focus most scrutiny on tests whose primary advantage is a purported greater sensitivity for "disease," since these tests will identify more patients who will be considered for treatment but whose response to treatment will not be known from the original trials. Imagine, for example, that the original trial showing treatment benefit found this benefit only among patients with a certain CSF tau protein level. If a new blood test is shown to identify all these same patients but, in addition, suggests additional patients with "positive" results indicating the same underlying pathological process, it will be difficult to know whether treating these additional patients would offer them the same balance of risks and benefits documented for patients identified only with the gold standard test. Level 2 evidence is unlikely to be sufficient in this case, and even level 3 and level 4 evidence, showing that test results change diagnostic impressions and treatment choices for patients receiving the new test, will prove unpersuasive to technology assessment groups and insurers who will require evidence on the treatment outcomes for the new patients identified by the new test.

For the immediate future, however, the primary issue for the field of AD diagnostics research revolves around the validation of initial biomarkers as surrogate outcomes and as diagnostic tests. We will return to the issue of study design to enhance the evidence on biomarker clinical utility in a later section of this white paper. The following section provides a review of the current literature on AD biomarker testing, including an assessment of where the current evidence falls on the Fryback and Thornbury evidence hierarchy, and a summary of those studies that have attempted to move beyond diagnostic accuracy to assess clinician- and patient-related effects of testing.

Review of Current Evidence

Diagnostic Outcomes

In order to characterize the state of current literature on diagnostic testing for AD, a literature search was conducted for English-language studies published from January 2000 to March 2012 using the MEDLINE and PsycInfo databases. A total of 755 articles were initially identified; after elimination of clinical reviews, editorials, etc., a total of 621 articles were evaluated. As shown in Table 2 below, nearly all of these studies focused on technical efficacy and diagnostic accuracy alone, and 5 studies examined the impact of testing on diagnostic impression. There were no studies identified that evaluated the effects of testing on subsequent clinical decision-making. In addition, while many studies have documented the ability of biomarkers to predict key patient outcomes such as cognitive decline, quality of life, requirements for assistance in activities of daily living, or institutionalization, we identified no studies that documented improvement in these outcomes in patients receiving such testing as compared to those tested using clinical criteria alone.

Table 2. Studies of diagnostic testing for AD, by Fryback and Thornbury hierarchy.

Study Level	Number of studies
Technical Efficacy	17
Diagnostic Accuracy	553
Diagnostic Impression	5
Diagnostic Action	None
Patient Outcomes	None
Societal Outcomes	None

Diagnostic studies varied substantially in design, populations, protocols for testing and interpretation, and the gold standard used for diagnosis. Evaluations of inter-lab reliability were rare for CSF tests in our literature review, but there were several examining test-retest reliability of imaging modalities, with most results indicating reliability in the 90%-96% range. Most of the studies identified, however, involved 1-2 clinical centers only, limiting their ability to examine the impact of variability in test vendor and interpretation software, test reader skill, and testing method or protocol on diagnostic accuracy. Not surprisingly, clinical diagnosis was the reference standard in most studies, as definitive autopsy findings are often impractical to use in many situations. However, the framework for differential diagnosis (and the resulting population spectrum evaluated) was highly variable, and included comparisons of AD vs. no dementia, vs. other forms of dementia, vs. various levels of MCI by severity, and so on.

Other information was found to be lacking in many studies. For example, few studies documented the frequency with which a failed test or a test with equivocal or indeterminate results occurred; those that did include this information tended to exclude such patients from analysis, thereby overestimating accuracy. In actual clinical practice, these results would most often trigger retesting or further diagnostic testing of some kind. In addition, data on the threshold for positivity were often limited or absent, particularly in imaging studies. While this information was more frequently employed in analyses of CSF biomarkers, thresholds were highly variable, as has been documented in other systematic reviews and meta-analyses (Bloudek, 2011; van Harten, 2011).

With all these limitations in mind, it is not surprising that studies of diagnostic accuracy demonstrated a wide range of findings for sensitivity and specificity. As has been described in recent systematic reviews (Bloudek, 2011), our literature review also found that the range of sensitivity of CSF amyloid and tau tests has ranged from approximately 60-100%, with some of the variability arising from different thresholds for what defines a positive test. The sensitivity of various imaging modalities has been found to range from approximately 40-100%, with analogous issues of variable thresholds for what constitutes a positive result. Specificity for CSF tests range from 60-100%, and from 20-100% for imaging modalities.

A systematic review was recently undertaken to support consideration by the European Medicines Agency (EMA) of validation of CSF and PET-amyloid biomarkers for enrichment of clinical trial populations (EMA, 2012). In their review of CSF data they focused on 11 recent studies reporting sensitivity and specificity results on CSF tests using autopsy data as the gold standard. Based on their survey of this literature, EMA found that CSF amyloid-β_{42} alone did not always differentiate AD from other non-AD dementias. Use of either CSF amyloid-β_{42} or tau alone provided a modest improvement in likelihood ratios, but the combined use of CSF amyloid-β_{42} and total-tau improved both specificity and positive likelihood ratios substantially. The EMA review of the data on PET imaging using radiotracers binding to brain amyloid also concluded that elevated amyloid burden, as determined by these tests, increases the probability that patients classified as AD on clinical criteria do indeed have existing amyloid pathology. Based on this review the EMA has formally "qualified" CSF amyloid-β_{42} and PET-amyloid imaging as biomarkers for enrichment of clinical trials. The EMA did not, however, qualify these tests as surrogate outcome measures or diagnostic tools, a distinction we will return to in a later section of this paper.

The 5 studies that we found in our literature review that went beyond diagnostic accuracy to measure impact on clinical impressions or clinical actions are described in detail in Appendix B. Two of these studies assessed the impact of sequential use of imaging following clinical diagnosis, but measured only the change in clinicians' "confidence" in their diagnoses after imaging, not any actual changes to presumptive diagnosis (Heckemann, 2008; Raji, 2010). Three other studies did evaluate the impact of biomarker testing on diagnostic impression. In a prospective cohort, Kester and colleagues (2010) evaluated patients with AD, MCI, other dementias (including frontotemporal dementia or FTD) and no dementia, and evaluated the impact of additional CSF biomarker data following clinical diagnosis. In 11 of 109 cases (10%), the clinicians changed their initial diagnoses. The majority of these changes were from MCI to AD (n=4) or from AD to other dementias (although the paper does not specify) (n=4). Importantly, however, these were changes in working diagnoses only, as disease was not pathologically confirmed and therefore it is impossible to know whether the changes in diagnosis were correct or not.

Another study evaluating the impact of biomarker results on clinical impression gave clinicians FDG-PET test results after the clinicians had used clinical exam and symptom checklist results to provide a presumptive diagnosis for patients in a retrospective cohort who had autopsy-confirmed AD or FTD (Foster, 2007). Of 270 diagnoses, 42 (16%) were changed after receipt of FDG-PET data. Of the 42 changed diagnoses, 34 (81%) were changes from an incorrect to a correct diagnosis, whereas the remaining 8 diagnostic changes were from a correct to an incorrect diagnosis.

The final study in this category examined the potential impact of amyloid imaging on both diagnosis and intended management in patients with progressive cognitive decline and suspicion of AD

(Grundman, 2012, in press). A total of 229 patients recruited at 19 sites were imaged with florbetapir PET scanning. The site physician provided a provisional diagnosis and intended management plan both before and after the scan. The scan interpretations (113 positive and 116 negative) were associated with a change in reported diagnosis in 54.6% of cases and a high rate of change in intended management (> 85%) for at least one aspect of the management plan. For example, intended use of cholinesterase inhibitor or memantine treatment increased by 17.7% in cases having positive scans and decreased by 23.3% in cases with negative scans. Among subjects who were enrolled and had not yet completed a diagnostic workup (n=119), obtaining the results of the florbetapir scan was associated with absolute reductions of 24.4% and 32.8% in planned anatomic imaging and neuropsychological testing respectively. Key limitations of the study, as noted by the authors, included measurement of *intended* rather than *actual* changes in management as well as recruitment of physicians who were experts in memory disorders.

While all of these studies extend beyond evaluation of diagnostic accuracy, they provide only limited evidence on the impact of biomarker testing on intended clinician action, and do not include any evidence of the impact of testing on actual changes in further diagnostic testing or treatment. Nor, obviously, do they provide data on whether the additional information gleaned from biomarker testing leads to improved patient outcomes. Although the current clinical benefits of treatment for AD are quite limited, there are potential effects of testing on patient and family psychological outcomes and the ability to plan for the future. We provide a summary of existing data on these outcomes in the section below.

Psychological/Social Outcomes

It has been suggested that early testing for AD may positively or negatively impact psychological well-being, changes in health behaviors, future planning, and medical resource utilization for patients and their families (Iliffe, 2009; Illes, 2007). Such issues are neither novel nor unique to AD, as they have been evaluated in multiple therapeutic areas (Heshka, 2008) including genetic screening for hereditary breast and ovarian cancer (Claes, 2005; Hamilton, 2009) as well as testing for conditions with no known effective treatment such as Huntington's disease (Decruyenaere, 2003; Meiser, 2000; Richards(a), 2004; Richards(b), 2004; Timman, 2004), and amyotrophic lateral sclerosis (ALS) (Fanos, 2011).

The most widely published study of the outcomes associated with disclosure of AD risk was carried out in a population of adult children of parents who were diagnosed with late-onset AD. The Risk Evaluation and Education for Alzheimer's Disease (REVEAL) study was a randomized control trial funded by the National Institute on Aging and the Ethical, Legal, and Social Implications Research Program (NIA/ELSI) in which participants were randomized to a group that had the results of genetic testing for apolipoprotein E (APOE) disclosed to them and a group that did not. Primary outcomes of interest were measures of psychological impact as described above, as well as modification to health behaviors and changes to insurance plans.

At multiple time points over a one-year period, there was no evidence of statistically-significant differences between the disclosure and non-disclosure groups in measures of anxiety, depression or test-related distress (Green, 2009). The subgroup of participants in the disclosure group who learned they were APOE-positive had a period of heightened distress relative to those with negative test results. After 6 months, however, differences began to diminish and were nonsignificant by

one year, indicating the potential for individuals to assimilate the information and accommodate. Those who were APOE-negative demonstrated significantly less distress at 6 months than those in either the APOE-positive or non-disclosure groups; again, however, differences between APOE-positive and APOE-negative patients were nonsignificant after this timepoint.

Similar effects have been noted in studies of genetic testing for Huntington's disease and hereditary breast and ovarian cancers (e.g., BRCA 1/2) (Timman, 2004; Decruyenaere, 2003; Hamilton, 2009); those identified as carriers tended to have heightened general anxiety and stress over the short term, but levels either reverted to baseline or did not differ from noncarriers by 6-12 months of follow-up. Noncarriers generally reported short-term improvement in general psychological status or a status that was unaltered from baseline. However, findings from a meta-analysis of studies of emotional distress according to BRCA 1/2 status suggested that, for measures of *cancer-specific* distress, noncarriers reported significant reductions in distress up to one year following test results, while changes among carriers were nonsignificant.

In REVEAL, more participants with known APOE-positive status reported changes in a composite measure of AD "prevention activities" (i.e., medication/vitamin use, diet, and exercise) as compared to their APOE-negative and non-disclosure counterparts after one year (Chao, 2008). APOE-positive status was also significantly correlated with participants "thinking about" or making actual changes to long-term care insurance and with "thinking about" changes to life insurance (Zick, 2005). A recent survey indicated that almost two-thirds of individuals are willing to pay out of pocket to have a genetic test for AD (Kopits, 2011). Even in the absence of an effective treatment option, up to 75% of individuals in one sample speculated that they would be willing to pay for a definitive diagnostic test (Neumann, 2012), suggesting some perceived benefit "of knowing."

Guidelines and Insurer Coverage Policies for Alzheimer's Disease Diagnosis

Clinical Guidelines

Guidelines from multiple organizations in the U.S. and Europe universally support the initial diagnosis of AD with a comprehensive clinical and cognitive assessment including neuropsychological testing. Most of these organizations also support the use of structural imaging with either MRI or CT scans in conjunction with clinical examination to provide information on changes in brain anatomy as well as to differentiate AD from other neurodegenerative diseases and rule out conditions such as cerebral hemorrhage. Advanced imaging techniques such as FDG-PET and SPECT are considered potentially appropriate in cases where the clinical diagnosis of AD is questionable. In contrast, CSF-based biomarker data is considered investigational and not recommended for routine use by three of the four organizations listed below; only the European Federation of Neurological Sciences (EFNS) supports their use in cases of atypical AD presentation. Note: the American Academy of Neurology (AAN) is currently updating its 2001 guidelines; the update is not yet published.

American Psychiatric Association, 2007
http://psychiatryonline.org/content.aspx?bookid=28&sectionid=1679489

European Federation of Neurological Societies, 2010
http://onlinelibrary.wiley.com/doi/10.1111/j.1468-1331.2010.03040.x/abstract

National Institute for Health and Clinical Excellence with the Social Care Institute for Excellence, 2006 & 2011
http://guidance.nice.org.uk/CG42/Guidance/1-7/pdf/English

American College of Radiology, Appropriateness Criteria®, 2010
http://www.acr.org/Quality-Safety/Appropriateness-Criteria/Diagnostic/Neurologic-Imaging

Insurer Coverage Policies

As with clinical guidelines, the Centers for Medicare and Medicaid Services (CMS) as well as major private insurers in the U.S. tend not to restrict coverage for clinical examination and neuropsychological testing in the diagnosis of AD. One notable exception is restriction of coverage of neuropsychological testing to AD diagnosis only, and exclusion of coverage for such testing in the diagnosis of MCI, citing a lack of data on the impact of such testing on clinical decision-making (Cigna). In contrast, laboratory biomarkers (including CSF and urinary evaluation) are universally considered to be investigational and not covered, with most payers citing a lack of evidence on test reliability and appropriate interpretation, no evidence of improved accuracy over use of clinical criteria alone, and a paucity of data on changes in clinical management or patient outcomes that would result from the use of such biomarkers.

Similarly, advanced imaging techniques, including FDG-PET, SPECT and fMRI, are considered to be investigational and unproven modalities for the diagnosis of AD, for the reasons listed above. CMS

has made a National Coverage Determination (NCD) for FDG-PET, however, that provides limited coverage for its use in (1) the differential diagnosis of FTD vs. AD, or (2) a CMS-approved clinical trial focused on the utility of FDG-PET in the diagnosis or treatment of dementing neurodegenerative diseases. In this case, it was felt that the evidence was sufficient on the use of FDG-PET to differentially diagnose FTD vs. AD, provided that cognitive decline of at least six months' duration had been documented and other causes of dementia had been ruled out.

Centers for Medicare and Medicaid Services (CMS)
http://www.cms.gov/medicare-coverage-database/details/ncd-details.aspx?NCDId=288&ncdver=3&NCAId=64&NcaName=Positron+Emission+Tomography+%28FDG%29+for+Alzheimer%2527s+Disease%2fDementia&CoverageSelection=National&KeyWord=alzheimer&KeyWordLookUp=Title&KeyWordSearchType=And&bc=gAAAABAAIAAA&

Selected National and Regional Private Payers

Aetna
http://www.aetna.com/cpb/medical/data/300_399/0349.html

CIGNA
http://www.cigna.com/healthcareprofessionals/resources-for-health-care-professionals/clinical-payment-and-reimbursement-policies/medical-categories.html#cat6

Humana
http://apps.humana.com/tad/tad_new/home.aspx?type=provider

Regence
http://blue.regence.com/trgmedpol/contents/a.html

United
https://www.unitedhealthcareonline.com/b2c/CmaAction.do?channelId=ca174ccb4726b010VgnVCM100000c520720a____

Wellpoint/Anthem/UniCare
http://www.anthem.com/cptsearch_shared.html

Key Ongoing and Planned Studies of Alzheimer's Disease Diagnostics

Details on major ongoing studies focused specifically on diagnostic tests for AD are provided in Appendix C. A total of six cohort studies were identified through clinicaltrials.gov, three of which are evaluations of the recently approved radiotracer florbetapir (Amyvid). One of these is assessing acute safety issues associated with administration. Two additional studies are academic-based evaluations of florbetapir related to safety and expansion of an imaging database. In addition to the Amyvid studies, two industry-funded studies are examining the sensitivity and specificity of other amyloid-β-targeting compounds used with PET imaging (florbetaben and flutemetamol). In the latter of these studies, patients with a clinical diagnosis of amnestic MCI are being followed for up to two years to measure time to conversion from MCI to clinically probable AD in patients with normal vs. abnormal patterns of flutemetamol uptake. Finally, an investigator-initiated study is focused on the potential correlation of the results of oculomotor testing with other biomarkers (e.g., hippocampal volume or central atrophy).

While these studies are attempting to address some evidence gaps, such as standardization of biomarkers, inclusion of a broad spectrum of disease, and explicit comparison to standard diagnosis, several gaps remain. Study follow-up is limited, and with one exception (rate of cognitive decline), study designs are not focused on the impact of test results on clinical decision-making or patient outcomes. An RCT has recently been published on clinicaltrials.gov that is attempting to address these concerns in part. In this study, sponsored by Avid Radiopharmaceuticals, a total of 600 patients at 56 sites worldwide with MCI or dementia symptoms will receive PET-amyloid imaging. Physicians treating patients in the intervention arm will have immediate access to scan results, while those managing control patients will be blinded to scan results for 12 months. The primary outcomes of interest include changes in patient management at months 3 and 12 of follow-up and patient prognosis (i.e., association between scan status and rate of cognitive decline). Secondary outcomes include changes in clinical diagnosis, physician confidence in diagnosis, patient and caregiver advice and counseling, and caregiver self-efficacy for managing dementia. The study is expected to be completed by December 2014.

In addition, on the horizon are three important clinical trials in various stages of development and planning, each of which is intended to provide information on the progression of disease as well as the impact of testing and treatment in asymptomatic individuals. These are the Anti-Amyloid Treatment of Asymptomatic Alzheimer's Disease (A4) trial; the Dominantly Inherited Alzheimer Network (DIAN) – Therapeutic Trials Unit (TTU) trial; and the Alzheimer's Prevention Initiative (API). Each of these studies is enrolling participants who are clinically normal but with specific risk factors for development of AD.

The A4 study will be a randomized, double-blind placebo-controlled study of cognitively normal patients who demonstrate a positive amyloid-β burden through PET imaging (Sperling, 2012; Alzheimer's Association, 2012; Ryan, 2012). Potential subjects will be at least 70 years of age with CDR = 0 (no impairment), and MMSE 27-30 (no cognitive impairment). Projected enrollment is 1000 patients, with 500 randomized each to treatment and placebo. An additional 500 patients who test negative on PET-amyloid imaging will be followed as a natural history study. The

treatment to be evaluated has yet to be determined, and will be finalized later this year. The primary outcome is the rate of change on a cognitive composite measure, based on memory and executive function, over a 3-year timeframe. Secondary outcomes will include various biomarkers including CSF and MRI measures. An important substudy will examine the impact of disclosure of PET amyloid imaging results to patients and caregivers. The A4 is awaiting final funding approval from NIH and will potentially begin enrollment as early as 2013.

DIAN-TTU is designed to evaluate 3 different drugs for potential prevention of AD (Bateman(a), 2012; Alzheimer's Association, 2012). Patients enrolled in this study will have a known pathogenic mutation for development of AD, and will be allocated to one of three treatment arms or to placebo (n=160, with 40 patients/arm). Eighty additional patients without the genetic mutation will receive placebo. The primary outcome of the first phase of the trial, lasting two years, is determination of the interaction between drug and its potential targets. Continuation into the second phase of the study will depend on whether an interaction between treatment and testing is found, and will evaluate cognitive outcomes over three years.

Coordinated through the Banner Alzheimer's Institute, the API is funded through grants from NIA, Genentech, and private donations (Reiman, 2012; Alzheimer's Association, 2012). The first clinical prevention trial to be conducted through API will enroll predominantly Colombian patients who are cognitively healthy but possess a known pathogenic mutation for development of AD. Patients will be randomized to treatment and placebo arms (n=216, with 108 patients/arm). An additional 108 patients without the mutation will also receive placebo. The trial is a double-blind study, lasting up to 60 months. The treatment arm will evaluate an anti-amyloid monoclonal antibody, crenezumab (Genentech), which binds to various forms of amyloid-β42. The primary outcome is the change in the API composite cognitive score; however, an interim analysis will be conducted at 24 months on clinical biomarkers, such as FDG-PET, CSF assays and cognitive tests, and their correlation with projected treatment effects.

While the availability of data on the natural history of cognitive decline in asymptomatic patients with risk factors for AD will be useful, the real potential power of these planned trials is to connect measures of pathologic and genetic abnormality to an appropriate target population for treatment. While this will not answer all questions about the appropriate populations for drug treatment, the potential is there for these trials to produce "level 5" evidence (i.e., impact on patient outcomes), thereby validating the use of both the test and the treatment in one study.

In addition to the studies described above, other studies have recently received funding for examination of diagnostic approaches to AD. These include validation of a blood-based screening tool utilizing data from the Texas Alzheimer's Research & Care Consortium and the Mayo Clinic in Jacksonville, FL; and development of a neuroimaging paradigm, combining memory tasks with hippocampal function, to distinguish between normal aging and that seen in preclinical AD (NIH RePORTER, 2012).

Biomarker Validation Research Designs

When considering the current state of evidence on AD diagnostic tests from the perspective of an analytic framework or from the number of studies at higher levels of established evidence hierarchies, it is easy to understand why technology assessment groups and insurers have judged that current evidence is not adequate to demonstrate that AD diagnostic tests improve patient outcomes. It is obviously critical to this judgment that the current treatments for AD are considered to provide minimal, short-term clinical benefits, and that there have not been any studies to date evaluating the impact that positive or negative test results have on outcomes such as subsequent test ordering, selection of treatment, and patient quality of life. Without an effective treatment for positive test results or adequate evidence of net health benefits provided by "negative" test results, it is impossible to validate the clinical utility for any diagnostic test. It is important to note the limitations in the current evidence that exist at nearly every step in the analytic framework for AD diagnostics, corresponding to each level of the Fryback and Thornbury evidence hierarchy. Even at the level of technical efficacy, there remain significant questions about the standardization of testing modalities across different machines, readers, laboratories, and institutions. Meanwhile, diagnostic accuracy data are plagued by the difficulty in establishing a practical gold standard other than autopsy results. And data are notably sparse or absent on the impact of test results on changes in diagnosis, treatment or management selection, and patient outcomes.

Given that there are no significant "clinical" benefits of current treatment options, it has been proposed by some that assessment of the utility of AD diagnostic tests should include consideration of psychological and family outcomes, such as the potential benefits of reassurance provided by a negative test, or of the ability to make financial plans in light of a diagnosis of early AD dementia. As previously noted, one study's findings suggests that a majority of Americans would be willing to pay for an AD diagnostic test even without an effective treatment (Neumann, 2012), but the data supporting actual psychological or planning benefits are sparse and inconsistent. In addition, the standard policy of insurers in the U.S., including Medicare, has been to provide insurance coverage only for diagnostic tests that are believed to produce clinical benefits. "Reassurance" or information for information's sake has never been enough for coverage.

All this will change if and when a treatment is found to slow the progression of AD or to treat its symptoms more effectively. Within the pivotal trial or trials that demonstrate clinical effectiveness there will be criteria used to define eligible patients, criteria that will be based on neuropsychological test results, biomarker test results, or some combination thereof. Successful treatment outcomes will make these eligibility criteria a de facto diagnostic test to identify "treatment-responsive" MCI or AD dementia. The level of evidence in the hierarchy will have jumped overnight from diagnostic accuracy studies to those in which patient outcomes can be directly linked to treatment of patients meeting a particular diagnostic approach.

But it is in the design of these new trials seeking to test AD therapeutics in minimally symptomatic and pre-symptomatic patients that key elements should be considered in order to "build in" the means to generate evidence that will be most persuasive to technology assessment groups and insurers in the future. One critical issue that may arise is the question of how to interpret biomarker results that are found – retrospectively – to be associated with positive treatment outcomes. Will new prospective trials using these same biomarker results as eligibility criteria be

required in order to validate them as useful diagnostic criteria? In addition, it must be anticipated that the moment one diagnostic approach has been found to be able to identify patients with a positive treatment outcome, there will be immediate interest in evidence on alternative, possibly more practical or less expensive diagnostic approaches. How to validate additional biomarkers embedded in clinical therapeutic trials for use as diagnostic tools will be the key question.

Biomarkers for AD have received intense interest in recent years, but not due primarily to their potential to serve as diagnostic tests. The greater interest has arisen because the AD research community recognizes that without validated biomarkers it will be impossible to study potential treatments for MCI and early AD, where the new paradigm of AD suggests treatments should be targeted in order to interrupt the downstream effects that, so far, have proven irreversible. As pointed out by Holtzman, "biomarkers will need to be used to select people who are clinically normal but at high risk for near-term cognitive decline. Without such an approach, trial size will be enormous and cost prohibitive, and individuals may be subjected to treatments that have the potential for toxicity with no clear benefit" (Holtzman, 2012).

But, as mentioned earlier in this white paper, there are other important potential uses of biomarkers in AD drug development research and clinical care. Biomarkers that facilitate the identification of the biochemical effects of a drug in short-term pilot studies may identify those drug candidates derived from animal models that affect the disease process in patients. Biomarkers could aid in patient population selection and assessment of near-term drug effects on disease progression, as mentioned by Holtzman. In industry-led drug discovery and development, biomarkers could also facilitate selection of drug candidates, verify the mechanism of action, define dose effects, and enable clinical trials to be shortened and run with reduced sample size. In pivotal trials biomarkers could serve as true surrogate endpoints for clinical outcomes to guide regulatory decision-making. And, lastly, biomarkers could serve as prognostic and diagnostic tools in clinical practice.

In the field of clinical oncology, the challenge of validating biomarkers that can identify patients likely to benefit from targeted therapies has produced a rich literature on study design considerations, many of which are relevant for studies of biomarkers for AD (Mandrekar, 2009; Simon, 2009). Prospectively designed clinical trials are the gold standard approach to validating biomarkers, and most prospective RCT designs for this purpose fall into one of four categories: targeted, or enrichment designs; sequential testing strategy designs; and marker-based designs, which include marker-based strategy designs and marker-by-treatment-interaction designs (Mandrekar, 2009).

A full discussion of all of these design options, and of hybrids between them, is beyond the scope of this paper but can be found in several academic sources (Mandrekar, 2009; Scheibler, 2012). The two designs most relevant for AD biomarker validation are enrichment designs and marker-by-treatment-interaction designs. Enrichment designs are particularly important for AD research because some manner of enhancing the recruitment of patients who are at higher risk of rapid progression of symptoms is necessary to make treatment trials of early AD feasible within a 3-4 year timeline. In an enrichment design the marker is used to enrich the sample *before* randomization. For example, the A4 study described earlier will use PET-amyloid imaging to enrich the sample of patients with patients whose positive test results are presumed to mean that they have a higher likelihood of progressing relatively rapidly from normal cognition to MCI or early AD dementia. If

this study finds a difference in patient-relevant outcomes, a PET-amyloid guided therapy plan will have proven to be beneficial, and PET-amyloid will become the gold standard diagnostic test for "treatment-responsive" AD.

The other very relevant biomarker validation study design type for AD is the marker-by-treatment-interaction design. This design is similar to an RCT of any new drug or other intervention. Patients are randomized to receive either the new drug or the conventional treatment (or a placebo). The potential biomarker tests are performed before the randomization but are not used as eligibility criteria to enrich the patient sample. Instead, in this design the biomarker test results should ideally be kept masked from investigators as well as patients. Analyses of subsequent patient outcomes can then determine whether treatment is particularly effective, or not effective, for given (ideally pre-specified) subgroups. This study design is especially important for the development of information about a wide range of potential biomarkers because it can be nested within an enrichment design study. For example, a study using current CSF or PET-amyloid imaging biomarkers as criteria for enrichment of the patient sample could also measure many other putative biomarkers at baseline and evaluate later whether these other biomarkers can predict response to treatment among patients with positive CSF and/or PET-amyloid results.

Prospectively designed clinical trials are the gold standard approach to validating biomarkers, but special consideration should be given to the issue of whether biomarkers for AD could be validated retrospectively from data gathered in an RCT or large cohort study. AD takes such a long time to progress that prospective validation studies may prove impractical. Therefore it is generally viewed as reasonable to test the predictive ability of a marker using data from previous high-quality RCTs evaluating therapies for which a marker is proposed to be predictive (Khlief, 2010; Simon, 2009). Some commentators have noted that the distinction between prospective and retrospective studies is more a matter of semantics, and that the real issue in determining whether data are useful for validating biomarkers is whether they come from experimental or observational studies (Simon, 2009).

Enrichment designs and marker-by-treatment-interaction designs seem almost certain to become the standard for trials of interventions targeting MCI, mild AD dementia, or even completely asymptomatic individuals. Adaptive marker-by-treatment-interaction designs can also be developed which follow a Bayesian approach in which positive treatment results are sought sequentially at several pre-determined biomarker result thresholds and, if found, determine subsequent patient accrual and randomization plans. The European drug regulator, the EMA, has already moved to formally qualify both PET-amyloid imaging and CSF amyloid-β_{42} as biomarkers for the use of enrichment of patient study populations (EMA, 2012). Enrichment designs have the potential to create a very clear, persuasive body of evidence linking specific biomarker tests to patient outcomes from new treatments. If they can be combined with marker-by-treatment-interaction approaches to evaluate multiple other potential biomarkers simultaneously, it may be possible to create an efficient platform for validating biomarkers not only for research design purposes but as diagnostic tools for identifying patients who are most likely to have a positive response to treatment.

Evidence and Coverage Decision-Making

What Insurers Will Be Looking For: Evidence on Clinical Effectiveness

It cannot be repeated too often that insurers, and the technology assessment groups that provide evidence evaluations for them, will be looking for persuasive evidence that diagnostic tests for AD improve patient outcomes. All the other potential roles that biomarker tests can play in drug development and research design are viewed as important by payers, but coverage determinations will be determined by insurers' assessment of whether the evidence is adequate to demonstrate that all the step-by-step links in the analytic framework are fulfilled, and that the use of a diagnostic test will improve patient outcomes.

How will insurers judge whether the body of evidence on a particular diagnostic test is "adequate"? As discussed earlier, this standard will be more difficult to meet in the current era of treatments with limited effectiveness. But even in an alternative future scenario in which treatments are developed with substantial positive benefits, insurers will require that the clinical utility of diagnostic tests be adequately demonstrated. First, insurers will consider whether the test comes under the jurisdiction of the FDA and, if so, they will be extremely unlikely to provide coverage unless the FDA has approved the test. Even if the FDA has approved the test, however, this will provide no guarantee of coverage by insurers. FDA approval of diagnostic tests, particularly if the underlying technology is based on imaging techniques that have been approved in the past, may not require clinical trials in order to receive approval through the 510k pathway (FDA, 2004). Even if the FDA has required one or more clinical trials, the trials may be trials of diagnostic performance only, with no requirement to evaluate outcomes related to diagnostic impression, subsequent diagnostic or therapeutic actions, or patient outcomes. For example, the FDA recently approved florbetapir, the new radiopharmaceutical agent used in conjunction with positron-emission tomographic (PET) imaging of the brain in cognitively impaired adults undergoing evaluation for AD and other causes of cognitive decline (Yang, 2012). As FDA staff wrote in an article describing the rationale for their regulatory approval of florbetapir, "the FDA did not require clinical data assessing the effect of florbetapir imaging on clinical management or patients' health. The FDA code of regulations (in 21 CFR 315.5[a]) mandates that the effectiveness of a diagnostic radiopharmaceutical agent should be determined by an evaluation of the ability of the agent to provide useful clinical information related to the proposed indications for use. FDA guidance further recognizes that imaging information may in some instances 'speak for itself' with respect to clinical value and that diagnostic approval may therefore not require assessment of the effects on clinical management or health outcomes" (Yang, 2012).

Insurers will not apply this same evidentiary threshold when making coverage determinations. Evidence linking diagnostic performance to the impact on patient outcomes will be necessary. In the absence of more effective treatments for AD, it will be difficult to demonstrate that existing or new diagnostic tests improve patient outcomes. Studies could be conducted to evaluate whether positive test results provide some advantages in patient/family planning, but this type of benefit has not been routinely judged a "health" benefit by insurers in the past. But what about the possible clinical benefits of a negative test result? In order to judge the clinical benefits of negative test results insurers would expect that real-world studies measure the impact of negative test

results on subsequent diagnostic testing and treatment decisions. The possible benefits of a negative test result include:

- Reduction in the number of patients begun on AD treatments, potentially reducing harms and costs from unnecessary medication use.
- Ability to identify patients currently on AD treatments who can be taken off treatment, thereby potentially reducing treatment-related harms and costs.
- More targeted use of additional diagnostic evaluations that may identify other causes of dementia more rapidly, potentially improving outcomes, reducing risks, and/or reducing costs. Clinical scenarios that may suggest this benefit include but are not limited to:
    1. Patients presenting with a small cerebral hemorrhage that may be due to amyloid angiopathy. A negative AD diagnostic test would suggest the need for an angiogram in order to identify a treatable surgical aneurysm.
    2. Patients meeting criteria for MCI/early AD but who have an unusual presentation. A negative AD test would quickly trigger alternative diagnostic evaluations.
    3. Patients presenting with dementia and signs somewhat suggestive of normal pressure hydrocephalus (NPH). A negative AD diagnostic test would lead to further evaluation for NPH, a treatable condition.
- Patients presenting with complaints from self or family of possible early MCI. A negative test result would lead to reassurance and possible improved quality of life.

In considering the potential benefits of negative test results, insurers will want studies that can measure not only benefits but potential harms as well. The possible harms of a negative test result include:

- The possibility that the negative test is a false negative result. A false negative result may lead to false reassurance and a failure to seek other diagnostic testing for treatable causes.
- Alternatively, false negative results could lead to aggressive additional diagnostic testing with attendant unnecessary harms and costs.
- Negative test results may incur costs and some risks while not changing treatment decisions in an era of treatments with limited effectiveness.

Even when the clinical value of a diagnostic test is assumed to lie primarily in the ability of a negative test result to exclude AD as a possible diagnosis, evaluation of the net health benefit of the test must take into consideration the potential benefits and harms of positive tests, including false positive tests, as well. In the current era of minimally-effective treatments, the burden will be much higher to show robust evidence that diagnostic test results convey overall net health benefits. Thus insurers will want to see randomized controlled trials in which a diagnostic test is used as it would be in clinical practice among representative patient populations and clinicians. All patients, including those with indeterminate findings, should be followed long enough to measure and compare the impact of the diagnostic test on subsequent diagnostic and therapeutic actions. Studies long enough to evaluate the ultimate impact on long-term patient outcomes are not feasible, but robust studies on downstream diagnostic and therapeutic actions will be necessary to help insurers judge whether the body of evidence is "adequate" to demonstrate a positive impact on patient outcomes.

Looking to the future, the evidence on the impact on patient outcomes of diagnostic tests for AD, particularly biomarkers, will emerge from the incorporation of biomarkers in pivotal trials of new therapeutic agents. Although enrichment and marker-by-treatment-interaction designs will embed

biomarker test results into the design of these trials, there are still important study characteristics that will be looked at carefully by insurers should the studies demonstrate "positive" treatment results. First, how representative are the patients? Patients recruited into these trials should be as representative as possible of the broader patient population who will be considered for treatment, and therefore diversity in age, sex, ethnic background, and clinical comorbidities will be important in order to support claims for the generalizability of study results.

Second, biomarker tests that are used, particularly those used as eligibility criteria in enrichment designs, should be tests that already have robust evidence of their standardization and reliability in the clinical community. If this is not the case, then it will be very difficult to have confidence that any evidence of positive treatment benefit among marker-positive patients will occur when the tests are used in real-world clinical practice.

Third, it will be important for all clinical studies involving biomarkers to include the outcomes of patients with borderline, indeterminate, or conflicting test results. Data on these patients are needed in order to gain a full picture of the potential impact of introducing the test into wider use in the clinical community.

Finally, researchers and manufacturers should always remember that the underlying question about the evidence on diagnostic tests for AD will remain whether *meaningful* patient outcomes are shown to improve with treatment linked to test results. Outcomes must be patient-relevant, clinically significant, and of meaningful duration. These outcomes include cognitive function impacting abilities of patients to perform daily activities; institutionalization; quality of life; and mortality. Technology assessment groups and insurers will be aware that some of these outcomes will not be relevant or feasible in studies of drugs to delay or modify the course of AD in pre-symptomatic patients. Biomarkers may therefore need to be used as surrogate outcome measures, but in order to be recognized as valid surrogate outcomes there must be evidence of a strong, independent, and consistent association between the treatment-induced change in the biomarker and the meaningful clinical outcome measure (Baker, 2003; Fleming, 1996). None of the existing AD biomarkers have yet been shown to meet these criteria, so the work of gaining evidence linking biomarker evolution with clinical outcomes must continue, both within enrichment and marker-by-treatment-interaction designs and through large, population-based natural history studies.

Although the first studies of successful therapeutic agents that will automatically validate some biomarker profile as diagnostic of "treatment-responsive" AD will be randomized controlled trials, subsequent studies of additional biomarkers or other diagnostic approaches may take the form of retrospective analyses as discussed earlier in the section on biomarker validation study designs. In addition, some assessments of alternative diagnostic approaches to "gold standard" biomarker approaches may be done through combinations of rigorous assessments of diagnostic performance and modeling studies that evaluate the potential longer-term impact on treatment decisions and patient outcomes. In all these cases technology assessment groups and insurers will look for studies that are designed to avoid the possibility of a statistical hunting expedition for positive findings. Evaluators will cast a wary eye on studies that ignore outcomes linked to incorrect and indeterminate test results, extrapolate optimal test performance from results obtained among select clinicians and patient groups, or assume that test information is automatically applied appropriately in all situations. Ultimately, insurers will judge as "adequate" only a cumulative body

of evidence that provides firm evidence of the overall beneficial impact of diagnostic testing on significant patient outcomes.

What Insurers Will Be Looking For: Evidence on Value

The most important consideration for insurers in the United States will be determinations of the clinical utility of diagnostic tests for AD. However, even though insurance coverage determinations in this country are not driven directly by considerations of economic impact, any time there are significant costs involved in the adoption of a new test or treatment, evidence on the "value" of a new intervention is weighed in the overall judgment of how medical policy should be framed to maximize potential benefits of the intervention while minimizing both patient risks and overall costs. Since diagnostic testing for AD may involve expensive tests such as imaging and/or radionuclide tests, and since future therapies for AD may themselves be quite expensive, certainly on the cumulative, population-based level, evidence on value should be included as a goal of the research agenda for AD diagnostics.

There is wide variation in the approach to evaluating "value" by insurers, both within the US and internationally. For some international insurers and health systems the usual metric with which to judge value is the incremental cost of a new service divided by the net health gain it provides. The most common form of this type of metric is the incremental cost per quality-adjusted life year (QALY). But the cost per QALY is rarely used in the US in either medical service or pharmaceutical coverage decisions. And for diagnostic tests, estimating a cost per QALY usually requires hypotheses of long-term outcomes and their impact on length and quality of life that can be extremely uncertain given the short-term data usually available on the impact of diagnostic tests in clinical practice.

Thus in the US insurers will be more likely to favor evidence on value that captures several short-term aspects. First, insurers will use incremental cost information linked not to long-term quality-adjusted life years but to more tangible short-term results. For example, insurers are likely to welcome evidence comparing the "cost per additional true positive result" for a test with a superior specificity than another test, or the "cost per avoided false negative test" for a test with superior sensitivity.

Second, although nearly all insurers will be interested in the price of the test itself and in comparisons of incremental costs for short-term diagnostic performance outcomes, they will also want to see evidence on the comparative costs of the entire "diagnostic pathway" introduced by a new diagnostic test. Evidence of the entire diagnostic pathway would include the costs for all clinician visits, test, and treatments up to the point when a diagnosis is made. In order to be useful in judgments of comparative value, the costs of the new diagnostic pathway would need to be compared directly to the costs of the diagnostic pathway in "usual care." Real-world costs of this type will not come from pivotal clinical trials of new therapeutic agents but the data from these trials could be used to create simulation models that can provide estimates of the comparative overall costs.

In the current era in which existing therapies for AD are very inexpensive, insurers will be interested in the costs attributed directly to any new diagnostic tests. But if new, more expensive treatments are developed, the costs of the diagnostic pathway should be extended to include the downstream costs of the initial treatments that are triggered by the testing. Insurers will be looking for evidence

that includes the experience and cost associated with false positive, false negative, and indeterminate results as well.

Through assessment of the comparative value of different diagnostic pathways insurers will be able to determine whether an expensive test creates cost offsets by limiting the use of other diagnostic testing and/or by narrowing the number of patients viewed as appropriate for treatment with expensive therapies. Today a diagnostic test that costs $2,000 would be viewed as low value given the limited impact on treatment costs and patient outcomes. But if a new, more effective therapy emerges at a cost of $20,000 per year, it is quite possible that insurers will judge a diagnostic test that costs $2,000 as high value if it more accurately targets a smaller set of patients who are most likely to benefit from treatment.

The evidence on comparative value is likely to require a combination of data from short-term clinical trials and estimates based on simulation models. Given the increased uncertainty introduced by the limitations of these techniques, insurers will look for studies and analyses performed by independent investigators that have a high degree of transparency, including the ability to replace Medicare or other generic cost data with insurer-specific costs in the simulation models. Insurers will also hope to find models of cost impact that evaluate the differential impact of using a new diagnostic test in populations with varying AD prevalence so that they can fully understand the potential clinical and economic impact of the test when used in populations with different underlying rates of "true" disease.

As stated earlier, considerations of economic impact and comparative value do not play a role in the evaluation of evidence on clinical effectiveness, nor do they drive coverage determinations. But it is essential that insurers understand the potential clinical and economic trade-offs involved in using a diagnostic test in different patient populations, for looking at both sets of information may guide their thinking in determining appropriateness criteria or other medical policies intended to maximize the appropriate, cost-effective use of a new medical technology.

What Insurers Will Be Looking For: Contextual Considerations

Even after considering the evidence on clinical effectiveness and comparative value, insurers must also integrate other potential "contextual" considerations into the decision-making leading to a coverage determination. Contextual considerations can include a wide range of other issues, including the following:
- Guidelines and other opinions of professional clinical societies
- Precedents set by prior coverage determinations for similar technologies/conditions
- Coverage determinations of other insurers
- Relative ability to "manage" the introduction of the new diagnostic test in order to maximize its appropriate use through identification of the service in claims data and application of medical management policies
- Patient advocacy
- Legislative coverage mandates
- Relative degree of financial risk shared with globally budgeted entities such as Accountable Care Organizations
- Legal and ethical considerations

Any number of these issues could be viewed as salient in a coverage determination for a new diagnostic test for AD. Insurer rationales for coverage determinations should be transparent and explicit when any of these considerations affects the final decision.

Recommendations for Future Research and Research Design

Based on this perspective on evidence held by technology assessment groups and insurers, and the earlier overview presented in this white paper on the state of evidence regarding AD diagnostic tests, the current paradigm for the pathophysiology of AD, and the study designs that can be considered in the validation of biomarkers, we now present a set of targeted recommendations for the development of further research that will help generate the type of evidence that can meet all relevant evidentiary standards. Included here are recommendations intended to frame all types of research needed to establish on more solid ground our understanding of the relationship between diagnostic tests and the course of AD, as well as recommendations focused on clinical trial design of studies designed to measure the effectiveness of test-and-treat strategies for AD.

Broad Research Agenda Recommendations

1. ***In the current era of AD treatments of limited effectiveness, randomized controlled trials should be performed to evaluate diagnostic tests with potential overall net health benefits.*** Given the limited clinical benefit of current treatments, any claims that diagnostic tests provide overall net health benefits will need to be supported by results from rigorous randomized controlled trials of the use of a new diagnostic approach in clinical practice. Insurers will want to see randomized controlled trials in which a diagnostic test is used as it would be in clinical practice among representative patient populations and clinicians. All patients, including those with indeterminate findings, should be followed long enough to measure and compare the impact of the diagnostic test on subsequent diagnostic and therapeutic actions. Tests that do not impact these outcomes and only provide some form of "prognostic" information are unlikely to be viewed as contributing substantially to overall net health benefit. Studies long enough to evaluate the ultimate impact on long-term patient outcomes are not feasible, but robust studies on downstream diagnostic and therapeutic actions will be necessary to help insurers judge whether the body of evidence is "adequate" to demonstrate a positive impact on patient outcomes.

2. ***Develop a framework for assessing the social and economic impact of diagnosis of pre-clinical AD.*** As described previously, there are limited data suggesting that the effects of early diagnosis in AD and similar conditions on anxiety/stress, financial planning, relationships, etc. are relatively short-lived. However, there is also recognition that the science on the "psychic value" of testing is relatively early in its evolution. Research is needed to develop better instruments for capturing psychological and family-related outcomes related to receiving test results. Studies on therapeutic agents should seek to evaluate these outcomes when subjects' biomarker results will not be blinded.

3. ***Looking to a future when there are more effective treatments for AD, continue to conduct biomarker studies in selected populations as well as in large population-based cohorts to evaluate the natural history of AD as well as the prognostic value of multiple combinations of neuropsychological testing and biomarkers.*** While prospective cohort studies (e.g., Alzheimer's Disease Neuroimaging Initiative) that have recruited convenience samples of patients are ongoing to evaluate the performance of multiple biomarkers, there remains a need to conduct large, more broadly representative cohort studies to evaluate the correlation of various neuropsychological tests and biomarkers (including imaging) with the development of cognitive symptoms and progression of cognitive decline in AD. It may well be that, as with coronary artery disease, there are multiple risk factors for cognitive decline, and longitudinal cohort studies are the best way to gain deeper understanding into how best to identify patients at greatest risk and those most likely to benefit from active intervention. Advances in the prognostic value of neuropsychological testing have been observed; these tests should therefore be included in all studies and not minimized in favor of more invasive and expensive biomarker tests. Populations considered should represent the spectrum of individuals at risk for AD, including individuals with no cognitive symptoms, those with MCI, and individuals at varying stages of dementia. Special consideration should be made to oversample populations typically underrepresented in these studies, such as those of certain racial or ethnic backgrounds as well as persons with lower socioeconomic status and educational attainment.

4. ***Develop consensus standards for biomarker test deployment and interpretation.*** The performance of any biomarker test or combination will be severely hampered by a lack of standards for how the test should be performed as well as for evaluation and interpretation of test results. For example, in a recent meta-analysis, 89% of the heterogeneity across studies of CSF testing of amyloid-β and tau was found to be due to different cutoffs for positivity across studies (Bloudek, 2011). For imaging tests, standards could be developed to address concerns such as test duration, protocols for reducing motion artifacts and limiting radiation exposure, minimum image resolution, etc. Many of these standards will require participation by industry to agree on procedures for image-building in scanner software and other related concerns; the Quantitative Imaging Biomarkers Alliance (QIBA) of the Radiological Society of North America can play a key role in gaining consensus across stakeholders. For laboratory tests, standards could be developed around specimen collection and storage, laboratory processing, threshold for positivity, etc., with the recognition that values may change as more is discovered about the correlation between test results and disease progression. To this point, a new study, funded through the NIH and with manufacturer support, will focus on harmonization of image reconstruction among different PET/CT vendors in an effort to decrease variability in quantitative imaging results (Paul Kinahan, PhD, personal communication, October 25, 2012). While the study will initially focus on imaging to target cancer therapeutics, the methodology is expected to have applicability to multiple disciplines.

5. ***As certain biomarkers gain validation for use as predictive of progression of disease, it will be important to study their predictive accuracy across the full spectrum of AD.*** If the first AD biomarkers gain validation based on their predictive accuracy in MCI or early AD, it cannot be assumed that they serve equally well as predictors of progression or of treatment response in later stages of the illness. Biomarkers will therefore need to be tested among patients across the spectrum of AD.

6. ***In studies that have used positive biomarker tests as inclusion criteria (enrichment design studies), include in baseline tests other potential biomarkers that can also be evaluated (nested marker-by-treatment-interaction studies). Ideally, always include additional test options that would be simpler, more accessible, and less expensive than the "gold standard" set of biomarkers used to qualify for inclusion.*** While CSF- and imaging-based biomarkers show promise in AD diagnosis, there are challenges to their use in typical clinical practice. Access to PET scanners may be limited, for example, as they are not uniformly distributed across the country. In addition, collection of CSF is a` procedure that is not without potential harm to the patient. Development of inexpensive early-detection tests highly accessible through primary care providers would be of interest to patients, clinicians, and insurers. Evidence on such potential "first-line" diagnostic alternatives should be sought within pivotal trials assessing "gold standard" biomarkers. The most highly prized diagnostic performance characteristic of a first-line diagnostic would be reliably high negative predictive value to be able to exclude illness. High specificity would obviously be preferred as well, but in a population-based diagnostic strategy it would not be unreasonable to accept lower specificity and to send only those patients with positive findings on first-line testing for further tests that have higher specificity.

7. ***Given that diagnostic testing for AD may involve expensive tests such as imaging, radionuclide tests, and CSF biomarkers, and that future therapies for AD may themselves be quite expensive, certainly on the cumulative, population-based level, evidence on comparative value should be included as a goal of the research agenda for AD diagnostics.*** Evidence on comparative value is likely to require a combination of data from short-term clinical trials and estimations based on simulation models. Given the increased uncertainty in the results introduced by the limitations of these techniques, insurers will look for studies and analyses performed by independent investigators that have a high degree of transparency, including the ability to replace Medicare or other generic cost data with insurer-specific costs in the simulation models. Insurers will also hope to find models of cost impact that evaluate the differential impact of using a new diagnostic test in populations with varying AD prevalence so that they can fully understand the potential clinical and economic impact of the test when used in populations with different underlying rates of "true" disease.

8. ***Given that many important clinical and economic outcomes occur years after diagnostic testing, a broad research agenda will benefit from the use of simulation modeling (decision analysis).*** Modeling can be very useful in combining high quality data gathered in observational studies with data from RCTs to determine where major sources of uncertainty exist and the magnitude of their relative contribution to key patient outcomes.

Trial Design

1. ***Design clinical trial protocols to enhance the generalizability of results to typical clinical practice.*** Even in a well-conducted enrichment design study the direct linkage of biomarker-based efficacy data from the trial to effectiveness in clinical practice will <u>not</u> be clear if a) alternative diagnostic criteria to identify the population for testing are used in clinical practice; and/or b) additional patients are considered eligible for treatment based on more generous thresholds for a "positive test result" (e.g., lower levels of amyloid-β on PET imaging than used in the trial). While the preference would be to use study entry and treatment criteria that could be replicated at all levels of clinical practice, it might also be acceptable to conduct sensitivity analyses on diagnostic accuracy in a broader population to evaluate the potential impact on patient outcomes.

2. ***Use a common set of consensus-based diagnostic test measurement thresholds and patient outcome measures.*** In order to evaluate biomarker and neuropsychological test results as potential diagnostic tests it will be helpful for the research community to arrive at common definitions of "positive" and "negative" test thresholds. Similarly, future trials, particularly large studies of asymptomatic, at-risk individuals, should use a common core set of consensus-driven measurement instruments for cognitive and functional outcomes. By using a common core set of outcome measures, applied at the same time points, the results of multiple studies will be able to be synthesized and evaluated more effectively. Outcome measures for each trial will need to be tailored to the expected mechanism of action of the active agent, and therefore outcome measures will not be entirely uniform, but a common core set will facilitate comparison of data across studies necessary for comparative effectiveness reviews of alternative diagnostic approaches. Standards should be set for the use of the test, interpretation of results, instruments to be employed for measuring cognitive and other outcomes, the frequency of measurement, the duration of follow-up, and other such concerns, in order to combine and compare data from multiple studies. There are no absolute criteria by which to select a single "best" outcome measure for various cognitive and other outcomes; indeed, determining neuropsychological cultural equivalence across different populations will be challenging. This process will therefore require broadly-constituted consensus methods, ideally convened and managed by an independent and trustworthy entity such as the National Institutes of Aging.

3. ***Complement unusual enriched populations in early studies with studies that enroll representative patient populations in order to enhance the generalizability of results to real-world clinical practice.*** Studies have been initiated in populations with rare genetic mutations (e.g., DIAN, API) that put them at much higher risk of AD and which may not reflect the same underlying pathophysiological process as spontaneous AD in the general population. Direct extrapolation of the risks and benefits of testing and treatment from unusually enriched population studies to patients in broader populations is not recommended. Instead, positive results from enriched population studies should be used as a springboard toward attempts to replicate study results in more generalizeable study populations.

4. ***Broaden the potential treatment population in trials.*** Manufacturers may be tempted to use highly selective entry criteria to identify a narrow "pure" cohort of patients most likely to have AD. However, if the treatment is shown to be effective, there will be pressure from clinicians, patients, and other stakeholders to use the treatment in patients whose MCI or dementia is less well characterized. It will therefore be helpful to attempt to address this in trial design by including patients with "mixed" forms of dementia. Subgroup analyses could be pre-specified to assess whether the treatment appears to work in multiple populations. This would be preferable to having stakeholders guess at efficacy beyond the pure cohort if the trial population was more limited.

5. ***For effective therapeutic agents developed through enrichment designs, consider further analyses to evaluate whether the original enrichment criteria were so narrow that less stringent enrichment criteria might identify many other patients who would benefit from treatment.*** Because enrichment designs will be predominantly used in secondary prevention trials, when a successful therapeutic agent is found the diagnostic quandary will be whether to generalize the diagnostic criteria to include other patients who may not meet fully the original enrichment standard. For example, early phase proof-of-principle studies may benefit from enrolling only the most highly selected AD study samples in which both CSF and PET-amyloid biomarker features are present. If positive treatment results were found for patients with this profile it would still provide no information on the response to treatment of patients who had one or the other but not both of these biomarkers. Thus strict enrichment designs maximize the specificity of the enrichment/diagnostic criteria for AD but impose a potentially substantial loss of sensitivity that would need to be re-addressed in later stages of clinical studies. Understanding how other, broader sets of diagnostic criteria "fit" with enrichment criteria will be very important, and where feasible, a follow-on marker-based strategy design may be appropriate to compare the enrichment criteria to another (more broad) diagnostic strategy, or to conduct subsequent enrichment studies with expansive entry criteria.

When follow-on prospective studies are felt to be impractical due to their duration it may be feasible to provide a supplementary "linked evidence" approach so that indirect comparisons can be made of the diagnostic accuracy of biomarker vs. standard or other diagnostic strategies, although resulting patient outcomes from these alternatives will only be estimates. For example, an indirect approach might involve linking evidence from studies of test accuracy vs. the current standard with data suggesting that the test result changes treatment practice. Note that a fully-linked evidence approach will *only* be meaningful when it is clinically feasible to link the multiple sets of data—in other words, the evidence for the proposed test and the evidence for the proposed treatment have been generated in similar patient populations. If the test identifies patients earlier or with a different spectrum of disease than the patients in whom the drug has been tested, then it is not clinically sensible to link this evidence. In such circumstances direct evidence is needed.

6. ***Retrospective assessment of a prognostic biomarker can only be done using data from well-conducted randomized controlled trials and with prospectively stated hypotheses, analysis techniques, and patient populations, with a pre-defined and standardized assay and scoring system for "positive" results. In other words: data mining should not be done to search retrospectively for combinations of clinical characteristics and biomarker results that are correlated with positive treatment outcomes.*** Prospective clinical trials are the "gold standard" approach to validate a predictive biomarker, and both enrichment designs and marker-by-treatment-interaction designs can provide high quality evidence. Even within prospective marker-by-treatment-interaction designs, however, it may be tempting to use statistical methods to search for combinations of biomarkers that are correlated with positive treatment outcomes. When old samples from previous RCTs exist, it may also seem prudent to test new hypothetical biomarkers by returning to these samples and retrospectively determining whether they were associated with positive treatment outcomes. Any such attempt at retrospective biomarker validation should only use data from well-conducted RCTs as opposed to a cohort or single-arm study, as it assures that the patients who were treated with the drug for whom the marker is purported to be predictive are comparable to those who were not (Mandrekar, 2009). Retrospective analyses of potential biomarker combinations can always be done to generate hypotheses, but validation requires either new prospective studies (preferable), or at a minimum, the use of data from RCTs in which there were prospectively stated hypotheses, analysis techniques, target patient populations, and pre-defined assays and scoring systems for thresholds that determine "positive" biomarker test results. If validation analyses meet these criteria and, importantly, if results can be replicated across multiple studies, populations, and databases, then technology assessment groups and insurers are likely to have confidence in the retrospective validation of a biomarker as a diagnostic and/or prognostic tool.

7. ***Consider clinically-equivalent but lower-cost diagnostic strategies in translating trial results to clinical practice.*** As mentioned previously, the biomarkers currently under investigation may pose challenges in broader populations because of limited access, cost, invasiveness, or other concerns. Wherever possible, more accessible and lower-cost alternatives to the "gold standard" approaches used in clinical trials should be included in pivotal trials and in subsequent studies. However, if the treatment examined poses a risk of significant toxicity, any alternative diagnostic strategy that involves a specificity tradeoff (i.e., higher rates of false positives) will need to be tested prospectively before such a strategy can be put into practice. For example, several biomarkers are currently being examined in both plasma and CSF, but current evidence is equivocal regarding the links between them. Further research is needed to establish the links between changes in plasma-based and CSF measures to validate use of the former in a diagnostic strategy.

References

1. Aetna. Clinical Policy Bulletin: Alzheimer's disease: diagnosis. http://www.aetna.com/cpb/medical/data/300_399/0349.html. Accessed July, 2012.

2. Agüero-Torres H, Fratiglioni L, Guo Z, Viitanen M, von Strauss E, Winblad B. Dementia is the major cause of functional dependence in the elderly: 3-year follow-up data from a population-based study. *Am J Public Health*. 1998;88(10):1452-1456.

3. Albert MS, DeKosky ST, Dickson D, et al. 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. *Alzheimers Dement*. 2011;7(3):270-279.

4. Alzheimer's Association. AAIC press release, July 17, 2012. http://www.alz.org/aaic/releases/tues_1030amct_ivig_trial.asp. Accessed July, 2012.

5. Alzheimer's Disease International. World Alzheimer Report 2010: the global economic impact of dementia. http://www.alz.co.uk/research/files/WorldAlzheimerReport2010.pdf. Accessed July, 2012.

6. American College of Radiology (ACR). ACR Appropriateness Criteria®: Dementia and movement disorders. http://www.acr.org/~/media/ACR/Documents/AppCriteria/Diagnostic/DementiaAndMovementDisorders.pdf. Accessed July, 2012.

7. American Psychiatric Association (APA). APA Practice Guidelines: Treatment of patients with Alzheimer's disease, and other dementias, second edition. http://psychiatryonline.org/content.aspx?bookid=28&sectionid=1679489. Accessed July, 2012.

8. Amyvid™ [package insert]. Indianapolis, IN: Eli Lilly and Company; 2012.

9. Anthem Blue Cross and Blue Shield. Medical Policies and Clinical UM Guidelines. http://www.anthem.com/cptsearch_shared.html. Accessed July, 2012.

10. Ashida S, Koehly LM, Roberts JS, Chen CA, Hiraki S, Green RC. The role of disease perceptions and results sharing in psychological adaptation after genetic susceptibility testing: the REVEAL study. *Eur J Hum Genet*. 2010;18(12):1296-1301.

11. Bäckman L, Jones S, Berger AK, Laukka EJ, Small BJ. Cognitive impairment in preclinical Alzheimer's disease: a meta-analysis. *Neuropsychology*. 2005;19(4):520-531.

12. Baker SG, Kramer BS. A perfect correlate does not a surrogate make. *BMC Med Res Methodol*. 2003;3:16.

13. Bateman R (a). Alzheimer's Association International Conference, 2012: Dominantly Inherited Alzheimer's Network (DIAN) trials. http://www.youtube.com/watch?v=U32pFgYVoUs. Accessed July, 2012.

14. Bateman RJ (b), Xiong C, Benzinger TLS, et al. Clinical and biomarker changes in dominantly inherited Alzheimer's disease [published online ahead of print July 11, 2012]. *N Engl J Med*. doi: 10.1056/NEJMoa1202753.

15. Blennow K. Biomarkers in Alzheimer's disease drug development. *Nat Med*. 2010;16(11):1218-1222.

16. Blennow K. CSF biomarkers for mild cognitive impairment. *J Intern Med*. 2004;256(3):224-234.

17. Bloudek LM, Spackman DE, Blankenburg M, Sullivan SD. Review and meta-analysis of biomarkers and diagnostic imaging in Alzheimer's disease. *J Alzheimers Dis*. 2011;26(4):627-645.

18. Brandt J, Aretouli E, Neijstrom E, et al. Selectivity of executive function deficits in mild cognitive impairment. *Neuropsychology*. 2009;23(5):607-618.

19. Brookmeyer R, Evans DA, Hebert L, et al. National estimates of the prevalence of Alzheimer's disease in the United States. *Alzheimers Dement*. 2011;7(1):61-73.

20. Cedazo-Minguez A, Winblad B. Biomarkers for Alzheimer's disease and other forms of dementia: clinical needs, limitations and future aspects. *Exp Gerontol*. 2010;45(1):5-14.

21. Chang Y-L, Bondi MW, McEvoy LK, et al. for the Alzheimer's Disease Neuroimaging Initiative. Global clinical dementia rating of 0.5 in MCI masks variability related to level of function. *Neurology*. 2011;76(7):652-659.

22. Chao S, Roberts JS, Marteau TM, Silliman R, Cupples LA, Green RC. Health behavior changes after genetic risk assessment for Alzheimer disease: the REVEAL study. *Alzheimer Dis Assoc Disord.* 2008;22(1):94-97.

23. Cigna. Clinical Payment and Reimbursement Policies: Medical categories. http://www.cigna.com/healthcareprofessionals/resources-for-health-care-professionals/clinical-payment-and-reimbursement-policies/medical-categories.html#cat6. Accessed July, 2012.

24. Claes E, Evers-Kiebooms G, Denayer L, et al. Predictive genetic testing for hereditary breast and ovarian cancer: psychological distress and illness representations 1 year following disclosure. *J Genet Couns.* 2005;14(5):349-363.

25. Clark LR, Schiehser DM, Weissberger GH, Salmon DP, Delis DC, Bondi MW. Specific measures of executive function predict cognitive decline in older adults. *J Int Neuropsychol Soc*. 2012;18(1):118-127.

26. Colburn WA. Optimizing the use of biomarkers, surrogate endpoints, and clinical endpoints for more efficient drug development. *J Clin Pharmacol*. 2000;40(12 Pt 2):1419-1427.

27. Cummings JL. The Neuropsychiatric Inventory: assessing psychopathology in dementia patients. *Neurology*.1997;48(suppl 6):S10-S16.

28. Davis DG, Schmitt FA, Wekstein DR, Markesbery WR. Alzheimer neuropathologic alterations in aged cognitively normal subjects. *J Neuropathol Exp Neurol*. 1999;58(4):376-388.

29. Decruyenaere M, Evers-Kiebooms G, Cloostermans T, et al. Psychological distress in the 5-year period after predictive testing for Huntington's disease. *Eur J Hum Genet*. 2003;11(1):30-38.

30. Devanand DP, Liu X, Tabert MH, et al. Combining early markers strongly predicts conversion from mild cognitive impairment to Alzheimer's disease. *Biol Psychiatry*. 2008;64(10):871-879.

31. European Medicines Agency. Qualification opinion of Alzheimer's disease novel methodologies/biomarkers for the use of CSF AB 1-42 and t-tau signature and/or PET-amyloid imaging (positive/negative) as a biomarkers for enrichment, for use in regulatory clinical trials – in mild and moderate of Alzheimer's. http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2011/12/WC500118365.pdf. Accessed August, 2012.

32. Fanos JH, Gronka S, Wuu J, Stanislaw C, Andersen PM, Benatar M. Impact of presymptomatic genetic testing for familial amyotrophic lateral sclerosis. *Genet Med*. 2011;13(4):342-348.

33. Feldman HH, Jacova C, Robillard A, et al. Diagnosis and treatment of dementia: 2. Diagnosis. *CMAJ*. 2008;178(7):825-836.

34. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med*. 1996;125(7):605-613.

35. Foster NL, Heidebrink JL, Clark CM, et al. FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease. *Brain*. 2007;130(Pt 10):2616-2635.

36. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making*. 1991;11(2):88-94.

37. Gomar JJ, Bobes-Bascaran MT, Conejero-Goldberg C, Davies P, Goldberg TE for the Alzheimer's Disease Neuroimaging Initiative. *Arch Gen Psychiatry*. 2011;68(9):961-969.

38. Gorelick PB, Scuteri A, Black SE, et al. Vascular contributions to cognitive impairment and dementia: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2011;42(9):2672-2713.

39. Green RC, Roberts JS, Cupples LA, et al. Disclosure of *APOE* genotype for risk of Alzheimer's disease. *N Engl J Med*. 2009;361(3):245-254.

40. Grundman M, Pontecorvo MJ, Salloway SP, et al. for the AV45-A17 Study Group. Potential impact of amyloid imaging on the diagnosis and intended management in patients with progressive cognitive decline. *Alzheimer Dis Assoc Disord*. In press.

41. Haense C, Herholz K, Jagust WJ, Heiss WD. Performance of FDG PET for detection of Alzheimer's disease in two independent multicentre samples (NEST-DD and ADNI). *Dement Geriatr Cogn Disord*. 2009;28(3):259-266.

42. Hamilton JG, Lobel M, Moyer A. Emotional distress following genetic testing for hereditary breast and ovarian cancer: a meta-analytic review. *Health Psychol*. 2009;28(4):510-518.

43. Hampel H, Frank R, Broich K, et al. Biomarkers for Alzheimer's disease: academic, industry and regulatory perspectives. *Nat Rev Drug Discov*. 2010;9(7):560-574.

44. Heckemann RA, Hammers A, Rueckert D, Aviv RI, Harvey CJ, Hajnal JV. Automatic volumetry on MR brain images can support diagnostic decision making. *BMC Med Imaging*. 2008;8:9.

45. Heister D, Brewer JB, Magda S, Blennow K, McEvoy LK for the Alzheimer's Disease Neuroimaging Initiative. *Neurology*. 2011;77(17):1619-1628.

46. Helzner EP, Scarmeas N, Cosentino S, Tang MX, Schupf N, Stern Y. Survival in Alzheimer disease: a multiethnic, population-based study of incident cases. *Neurology*. 2008;71(19):1489-1495.

47. Heshka JT, Palleschi C, Howley H, Wilson B, Wells PS. A systematic review of perceived risks, psychological and behavioral impacts of genetic testing. *Genet Med*. 2008;10(1):19-32.

48. Holtzman DM. CSF biomarkers for secondary prevention trials: why markers of amyloid deposition and neurodegeneration are both important. *Arch Neurol*. 2012;69(6):691-692.

49. Hort J, O'Brien JT, Gainotti G, et al. EFNS guidelines for the diagnosis and management of Alzheimer's disease. *Eur J Neurol*. 2010;17(10):1236-1248.

50. Humana. Medical Coverage Policies. http://apps.humana.com/tad/tad_new/home.aspx?type=provider. Accessed July, 2012.

51. Humpel C. Identifying and validating biomarkers for Alzheimer's disease. *Trends Biotechnol*. 2011;29(1):26-32.

52. Iliffe S, Robinson L, Brayne C, et al. Primary care and dementia: 1. Diagnosis, screening and disclosure. *Int J Geriatr Psychiatry*. 2009;24(9):895-901.

53. Illes J, Rosen A, Greicius M, Racine E. Prospects for prediction: ethics analysis of neuroimaging in Alzheimer's disease. *Ann N Y Acad Sci*. 2007;1097:278-295.

54. Jack CR Jr (a), Albert MS, Knopman DS, et al. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*. 2011;7(3):257-262.

55. Jack CR Jr (b), Barkhof F, Bernstein MA, et al. Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. *Alzheimers Dement*. 2011;7(4):474-485.

56. Jack CR Jr, Knopman DS, Jagust WJ, et al. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol*. 2010;9(1):119-128.

57. Jack CR Jr, Lowe VJ, Weigand SD, et al. Serial PIB and MRI in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease. *Brain*. 2009;132(Pt 5):1355-1365.

58. Jagust W, Reed B, Mungas D, Ellis W, Decarli C. What does fluorodeoxyglucose PET imaging add to a clinical diagnosis of dementia? *Neurology*. 2007;69(9):871-877.

59. Jedynak BM, Lang A, Liu B, et al. for the Alzheimer's Disease Neuroimaging Initiative. A computational neurodegenerative disease progression score: method and results with the Alzheimer's Disease Neuroimaging Initiative cohort. *Neuroimage*. 2012;63(3):1478-1486.

60. Jobst KA, Barnetson LP, Shepstone BJ. Accurate prediction of histologically confirmed Alzheimer's disease and the differential diagnosis of dementia: the use of NINCDS-ADRDA and DSM-III-R criteria, SPECT, x-ray CT, and Apo E4 in medial temporal lobe dementias. Oxford Project to Investigate Memory and Aging. *Int Psychogeriatr*. 1998;10(3):271-302.

61. Kantarci K. 1H magnetic resonance spectroscopy in dementia. *Br J Radiol*. 2007; 80 Spec No 2:S146-S152.

62. Katz R. Biomarkers and surrogate markers: an FDA perspective. *NeuroRx*. 2004;1(2):189-195.

63. Kester MI, Boelaarts L, Bouwman FH, et al. Diagnostic impact of CSF biomarkers in a local hospital memory clinic. *Dement Geriatr Cogn Disord*. 2010;29(6):491-497.

64. Khleif SN, Doroshow JH, Hait WN for the AACR-FDA-NCI Cancer Biomarkers Collaborative. AACR-FDA-NCI Cancer Biomarkers Collaborative consensus report: advancing the use of biomarkers in cancer drug development. *Clin Cancer Res*. 2010;16(13):3299-3318.

65. Klatka LA, Schiffer RB, Powers JM, Kazee AM. Incorrect diagnosis of Alzheimer's Disease: a clinicopathologic study. *Arch Neurol*. 1996;53(1):35-42.

66. Knopman DS, DeKosky ST, Cummings JL, et al. Practice parameter: diagnosis of dementia (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*. 2001;56(9):1143-1153.

67. Knopman DS, Parisi JE, Salviati A, et al. Neuropathology of cognitively normal elderly. *J Neuropathol Exp Neurol*. 2003;62(11):1087-1095.

68. Kopits IM, Chen C, Roberts JS, Uhlmann W, Green RC. Willingness to pay for genetic testing for Alzheimer's disease: a measure of personal utility. *Genet Test Mol Biomarkers*. 2011;15(2):871-875.

69. Landau SM, Harvey D, Madison CM, et al. for the Alzheimer's Disease Neuroimaging Initiative. *Neurology*. 2010;75(3):230-238.

70. Machulda MM, Ward HA, Borowski B, et al. Comparison of memory fMRI among normal, MCI and Alzheimer's patients. *Neurology*. 2003;61(4):500-506.

71. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *J Clin Oncol*. 2009;27(24):4027-4034.

72. Mathers C, Leonard M. Global burden of dementia in the year 2000: summary of methods and data sources (2003). http://www.who.int/entity/healthinfo/statistics/bod_dementia.pdf. Accessed July, 2012.

73. Matsuda H, Mizumura S, Nagao T, et al. An easy Z-score imaging system for discrimination between very early Alzheimer's disease and controls using brain perfusion SPECT in a multicentre study. *Nucl Med Comm*. 2007;28(3):199-205.

74. Mayeux R, Saunders AM, Shea S, et al. Utility of the apolipoprotein E genotype in the diagnosis of Alzheimer's disease. Alzheimer's Disease Centers Consortium on Apolipoprotein E and Alzheimer's Disease. *N Engl J Med*. 1998;338(8):506-511.

75. McKhann GM, Knopman DS, Chertow H, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*. 2011;7(3):263-269.

76. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology. 1984;34(7):939-944.

77. Mehta KM, Yaffe K, Pérez-Stable EJ, et al. Race/ethnic differences in AD survival in US Alzheimer's Disease Centers. *Neurology*. 2008;70(14):1163-1170.

78. Meiser B, Dunn S. Psychological effect of genetic testing for Huntington's disease: an update of the literature. *J Neurol Neurosurg Psychiatry*. 2000;69(5):574-578.

79. Mickes L, Wixted JT, Fennema-Notestine C, et al. Progressive impairment on neuropsychological tasks in a longitudinal study of preclinical Alzheimer's disease. *Neuropsychology*. 2007;21(6):696-705.

80. Mitsumoto T, Ohya N, Ichimiya A, et al. Diagnostic performance of Tc-99m HMPAO SPECT for early and late onset Alzheimer's disease: a clinical evaluation of linearization correction. *Ann Nucl Med*. 2009;23(5):487-495.

81. Montine TJ, Phelps CH, Beach TG, et al. National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. *Acta Neuropathol*. 2012;123(1):1-11.

82. Morris JC. Clinical Dementia Rating: on-line training system. http://www.alz.washington.edu/NONMEMBER/cdr2.html. Accessed July, 2012.

83. Nasreddine ZS, Phillips NA, Bedirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc. 2005;53(4):695-699.

84. National Institute for Health and Clinical Excellence (NICE). National clinical practice guideline: Dementia: a NICE-SCIE guideline on supporting people with dementia and their carers in health and social care. http://www.nice.org.uk/nicemedia/live/10998/30320/30320.pdf. Accessed July, 2012.

85. National Institutes of Health. Research Portfolio Ongoing Reporting Tools (RePORT). http://projectreporter.nih.gov/reporter.cfm. Accessed October, 2012.

86. Neumann PJ, Cohen JT, Hammitt JK, et al. Willingness-to-pay for predictive tests with no immediate treatment implications: a survey of US residents. *Health Econ*. 2012;21(3):238-251.

87. O'Bryant SE, Xiao G, Barber R, et al. for the Texas Alzheimer's Research & Care Consortium and the Alzheimer's Disease Neuroimaging Initiative. A blood-based screening tool for Alzheimer's disease that spans serum and plasma: findings from TARC and ADNI. *PLoS One*. 2011;6(12):e28092.

88. Pearl GS. Diagnosis of Alzheimer's Disease in a community hospital-based brain bank program. South Med J. 1997;90(7):720-722.

89. Peña-Casanova J. Alzheimer's Disease Assessment Scale – cognitive in clinical practice. *Int Psychogeriatr*. 1997;9 suppl 1:105-114.

90. Petersen RC, Doody R, Kurz A, et al. Current concepts in mild cognitive impairment. *Arch Neurol*. 2001;58(12):1985-1992.

91. Price JL, Morris JC. Tangles and plaques in nondemented aging and "preclinical" Alzheimer's disease. *Ann Neurol*. 1999;45(3):358-368.

92. Raji CA, Lee C, Lopez OL, et al. Initial experience in using continuous arterial spin-labeled MR imaging for early detection of Alzheimer disease. *AJNR Am J Neuroradiol*. 2010;31(5):847-855.

93. Rasmusson DX, Brandt J, Steele C, Hedreen JC, Troncoso JC, Folstein MF. Accuracy of clinical diagnosis of Alzheimer Disease and clinical features of patients with non-Alzheimer neuropathology. *Alzheimer Dis Assoc Disord*. 1996;10(4):180-188.

94. Reiman E. Alzheimer's Association International Conference, 2012: The Alzheimer's Prevention Initiative. http://www.youtube.com/watch?v=6qJKLOj_vt4&feature=relmfu. Accessed July, 2012.

95. Richards F (a). Couples' experiences of predictive testing and living with the risk or reality of Huntington disease: a qualitative study. *Am J Med Genet A*. 2004;126A(2):170-182.

96. Richards F (b), Williams K. Impact on couple relationships of predictive testing for Huntington disease: a longitudinal study. *Am J Med Genet A*. 2004;126A(2):161-169.

97. Rinne JO, Någren K. Positron emission tomography in at risk patients and in the progression of mild cognitive impairment to Alzheimer's Disease. *J Alzheimers Dis*. 2010;19(1):291-300.

98. Roberts JS, Cupples LA, Relkin NR, Whitehouse PJ, Green RC and for the REVEAL (Risk Evaluation and Education for Alzheimer's Disease) Study Group. Genetic risk assessment for adult children of people with Alzheimer's disease: the Risk Evaluation and Education for Alzheimer's Disease (REVEAL) study. *J Geriatr Psychiatry Neurol*. 2005;18(4):250-255.

99. Ryan L. Update on Alzheimer's disease clinical trials. https://www.alz.washington.edu/NONMEMBER/SPR12/Ryan.pdf. Accessed July, 2012.

100. Salmon DP, Thomas RG, Pay MM, et al. Alzheimer's disease can be accurately diagnosed in very mildly impaired individuals. *Neurology*. 2002;59(7):1022-1028.

101. Scheibler F, Zumbé P, Janssen I, et al. Randomized controlled trials on PET: a systematic review of topics, design, and quality. *J Nucl Med*. 2012;53(7):1016-1025.

102. Schneider LS, Kennedy RE, Cutter GR, Alzheimer's Disease Neuroimaging Initiative. Requiring an amyloid-β1-42 biomarker for prodromal Alzheimer's disease or mild cognitive impairment does not lead to more efficient clinical trial. *Alzheimers Dement*. 2010;6(5):367-377.

103. Shaw LM, Vanderstichele H, Knapik-Czajka M, et al. Cerebrospinal fluid biomarker signature in Alzheimer's Disease Neuroimaging Initiative subjects. *Ann Neurol*. 2009;65(4):403-413.

104. Shulman KI. Clock-drawing: is it the ideal cognitive screening test? *Int J Geriatr Psychiatry*. 2000;15(6):548-561.

105. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J Natl Cancer Inst*. 2009;101(21):1446-1452.

106. Small GW, Bookheimer SY, Thompson PM, et al. Current and future uses of neuroimaging for cognitively impaired patients. *Lancet Neurol*. 2008;7(2):161-172.

107. Sperling RA. Alzheimer's Association International Conference, 2012: Anti-Amyloid Treatment of Asymptomatic Alzheimer's Disease (A4) Trial. http://www.youtube.com/watch?v=f3Q_yUAGdYQ. Accessed July, 2012.

108. Sperling RA (a), Aisen PS, Beckett LA, et al. 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. *Alzheimers Dement*. 2011;7(3):280-292.

109. Sperling RA (b), Jack CR Jr, Aisen PS. Testing the right target and right drug at the right stage. *Sci Transl Med*. 2011;3(111):111cm33.

110. Temple R. Are surrogate markers adequate to assess cardiovascular disease drugs? JAMA. 1999;282(8):790-795.

111. Teutsch SM, Bradley LA, Palomaki GE, et al. on behalf of the EGAPP Working Group. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genet Med*. 2009;11(1):3-14.

112. The Regence Group. Medical Policy Manual. http://blue.regence.com/trgmedpol/contents/a.html. Accessed July, 2012.

113. Thepharmaletter. Roche and Lilly drugs chosen for major Washington University Alzheimer's prevention trial. http://www.thepharmaletter.com/file/117140/roche-and-lilly-drugs-chosen-for-major-washington-university-alzheimers-prevention-trial.html. Accessed October, 2012.

114. Timman R, Roos R, Maat-Kievit A, Tibben A. Adverse effects of predictive testing for Huntington disease underestimated: long-term effects 7-10 years after the test. *Health Psychol.* 2004;23(2):189-197.

115. Tombaugh TN. Trail Making Test A and B: normative data stratified by age and education. *Arch Clin Neuropsychol*. 2004;19(2):203-214.

116. Twamley EW, Ropacki SA, Bondi MW. Neuropsychological and neuroimaging changes in preclinical Alzheimer's disease. *J Int Neuropsychol Soc*. 2006;12(5):707-735.

117. UnitedHealthcare. Policies, Protocols and Guides. https://www.unitedhealthcareonline.com/b2c/CmaAction.do?channelId=ca174ccb4726b010VgnVCM100000c520720a____. Accessed July, 2012.

118. U.S. Department of Health and Human Services. National plan to address Alzheimer's disease. http://aspe.hhs.gov/daltcp/napa/NatlPlan.pdf. Accessed July, 2012.

119. U.S. Department of Health & Human Services, Centers for Medicare & Medicaid Services. National Coverage Determination (NCD) for FDG PET for dementia and neurodegenerative diseases (220.6.13). http://www.cms.gov/medicare-coverage-database/details/ncd-details.aspx?NCDId=288&ncdver=3&NCAId=64&NcaName=Positron+Emission+Tomography+%28FDG%29+for+Alzheimer%2527s+Disease%2fDementia&CoverageSelection=National&KeyWord=alzheimer&KeyWordLookUp=Title&KeyWordSearchType=And&bc=gAAAABAAIAAA&. Accessed July, 2012.

120. U.S. Department of Health & Human Services, Food and Drug Administration. Guidance for industry; Developing medical imaging drug and biological products; Part 2: Clinical indications (2004). http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM071603.pdf. Accessed October, 2012.

121. van Harten AC, Kester MI, Visser P-J, et al. Tau and p-tau as CSF biomarkers in dementia: a meta-analysis. *Clin Chem Lab Med*. 2011;49(3):353-366.

122. Wattjes MP, Henneman WJP, van der Flier WM, et al. Diagnostic imaging of patients in a memory clinic: comparison of MR imaging and 64-detector row CT. *Radiology*. 2009;253(1):174-183.

123. Yang L, Rieves D, Ganley C. Brain amyloid imaging – FDA approval of florbetapir F18 injection. *N Engl J Med*. 2012;367(10):885-887.

124. Zick CD, Mathews CJ, Roberts JS, Cook-Deegan R, Pokorski RJ, Green RC. Genetic testing for Alzheimer's disease and its impact on insurance purchasing behavior. *Health Aff.* 2005;24(2):483-490.

Appendices

APPENDIX A: Potential Conflicts of Interest and Other Influences on Judgment

APPENDIX B: Studies of Diagnostic Impression

APPENDIX C: Ongoing Clinical Studies

APPENDIX A: Potential Conflicts of Interest and Other Influences on Judgment

PDG Participant Name	Potential COI/Other Influences
Mark Bondi, PhD, ABPP/CN	No conflicts
Robin Cisneros	No conflicts
Joshua T. Cohen, PhD	Consultant to Pfizer, Inc.; involved in project work funded by Janssen Pharmaceuticals, Inc.
Michele DiPalo	Director of Health Services Evaluation for Blue Cross/Blue Shield of Massachusetts
Howard Feldman, MD, FRCP	Consultant to Pfizer, Inc.; Novartis AG; and GlaxoSmithKline PLC. Former employee of Bristol-Myers Squibb Co.
G. Scott Gazelle, MD, MPH, PhD	Consultant to GE Healthcare, Inc.
David Holtzman, MD	Co-founded C2N Diagnostics and is on the Scientific Advisory Board of C2N Diagnostics. Consultant to Astra Zeneca, Bristol-Myers Squibb, and Genentech
Bradley T. Hyman, MD, PhD	No direct conflicts. Spouse is employed by Novartis AG
Kejal Kantarci, MD, MSc	Consultant to Takeda Pharmaceutical Ltd.
Jason Karlawish, MD	No conflicts
Robert McDonough, MD	Senior Medical Director, Clinical Research and Policy Development, Aetna, Inc.
Mark Mintun, MD	Chief Medical Officer, Avid Radiopharmaceuticals, Inc. (owned by Eli Lilly)
Jeffrey R. Petrella, MD	Consultant to Vivo, Inc. (software company/developer) and Janssen Pharmaceuticals, Inc.
Creighton H. Phelps, PhD	No conflicts
Murray Raskind, MD	Member, Scientific Advisory Board, Janssen Alzheimer Immunotherapy Research & Development, LLC
James Rollins, MD	Employee of Centers for Medicare & Medicaid Services, sponsor of 2013 MedCAC meeting on use of beta amyloid PET in dementia and neurodegenerative disease
Alan Rosenberg, MD	Vice President, WellPoint, Inc.

Reisa A. Sperling, MD, MMSc	Consultant to Avid Radiopharmaceuticals, Inc. Planning a prevention trial that is industry sponsored and will partner with commercial interests. Consultant to many other companies, including Bayer AG
William H. Thies, PhD	No conflicts
Kathleen Welsh-Bohmer, PhD, ABCN	Consultant to Takeda and Zinfandel Pharmaceuticals. Planning a prevention trial that is industry sponsored and will partner with commercial interests
Kristine Yaffe, MD	Consultant to Novartis AG; serves on a Data and Safety Monitoring Board for Takeda Pharmaceuticals
Eric Yuen, MD	Vice President, Head of Clinical Development, Johnson & Johnson, Inc. – Janssen Alzheimer Immunotherapy, LLC

APPENDIX B: Studies of Diagnostic Impression

Author	Year	Study Design	Treatment Arms	Diagnostic Criteria (Gold Standard)	Disease Severity (at baseline)	Interventions/ Assessments	Procedure	Threshold for positivity	Findings
Foster NL	2007	Retrospective cohort	AD, n=31 FTLD, n=14 Controls, n=33	Neuropathological criteria (NIA-Reagan)	AD MMSE: 14.0 ± 8.7 range, 0-27	Comparison of clinical and imaging diagnoses: (1) clinical scenario (2) symptom checklist (3) 1 + 2 (4) transaxial FDG-PET (5) SSP FDG-PET	Evaluation by 6 neurologists of each approach & of adding FDG-PET to clinical approaches	(1) Scan abnormality (overall) (2) Abnormal metabolism (3) Symmetry between hemispheres	SSP FDG-PET superior to clinical evaluation; adding FDG-PET to clinical diagnosis increased diagnostic accuracy and confidence, particularly in cases of clinical uncertainty (71 to 90%)
Heckemann RA	2008	Retrospective cohort	AD, n=7 Normal controls, n=7 Reference cohort (normal study subjects), n=22	Clinical criteria (NINCDS-ADRDA)	NR	Evaluation of MRI data using grey-scale images initially, followed by use of a size rank color overlay	Review by neuroradiologist and general radiologist diagnostic confidence recorded after grey-scale, then color overlay	NR	Color overlay was useful in 18 of 28 cases; significant impact of color overlay on diagnostic confidence(p<0.02)

Author	Year	Study Design	Treatment Arms	Diagnostic Criteria (Gold Standard)	Disease Severity (at baseline)	Interventions/ Assessments	Procedure	Threshold for positivity	Findings
Kester MI	2010	Prospective cohort	AD, n=47 MCI, n=18 Other dementia, n=26 No dementia, n=18	Clinical criteria (NINCDS-ADRDA, Petersen)	AD MMSE: 22 ± 5 MCI MMSE: 25 ± 3	Use of CSF biomarkers when combined w/clinical diagnosis	Clinical diagnosis first made by consensus of multidisciplinary team; separate clinician interpretation of CSF results (no knowledge of clinical diagnosis)	Abnormal cutoff levels: CSF-tau >375 pg/ml CSF-Aβ42 <550 pg/ml CSF-ptau-181 >52 pg/ml	After receiving CSF profiles, 10% of diagnoses changed and clinicians became 32% more confident in AD diagnosis (from 51% to 83%)
Raji CA	2010	Retrospective analysis of prospective cohort	AD, n=13 Normal cognition, n=19	Clinical criteria (NINCDS-ADRDA)	Probable AD Modified MMSE: 88.9 ± 6.31 range, 76-97	Evaluation of CASL and SGPR MRI data (perfusion and volume changes)	4 neuroradiologists evaluated scans as "normal" or "abnormal" based on separate sets of criteria for perfusion and volume	Not explicitly stated	Inter-rater reliability was superior w/CASL; CASL MR data had higher mean sensitivity (85%) and accuracy (70%); Significant association between confidence and correct classification w/CASL in 3/4 readers

Abbreviations: AD: Alzheimer's disease; CASL: continuous arterial spin labeling; CSF: cerebrospinal fluid; FDG: fluorodeoxyglucose; FTLD: frontotemporal lobe dementia; MCI: mild cognitive impairment; MMSE: mini-mental state examination; MRI: magnetic resonance imaging; NIA: National Institute on Aging; NINCDS-ADRDA: National Institute of Neurological and Communicative Disorders and Stroke - Alzheimer's Disease and Related Disorders Association; PET: positron emission tomography; SPGR: spoiled gradient recalled; SSP: stereotactic surface projection

NOTE: Grundman, *in press* will be incorporated into this table when available online

APPENDIX C: Ongoing Clinical Studies (Phase II, III and IV)

ClinicalTrials Identifier	Study Sponsor	Study Dates	Clinical Phase	Interventions	Estimated Patient Enrollment	Reference Standard	Inclusion/ Exclusion Criteria	Outcomes
NCT01325259	University Hospital, Tours	Apr. 2009 – Dec. 2012	2	Clinical evaluation of [18F]AV-45 PET with FDG-PET, MRI and neuropsycho-logical assessment	N=65 No cognitive impairment, n=15 MCI, n=20 AD, n=30	Clinical diagnosis	Inclusion 1) Age ≥ 60 years 2) AD diagnosis (NINCDS-ADRDA) or amnesic MCI diagnosis 3) MMSE >18 and ≤ 28 for MCI/AD patients 4) Controls, MMSE ≥ 28 5) Study period > 7 years 6) French as native language Exclusion 1) History of alcoholism 2) Diabetes 3) Hypertension (≥ 180/100) 4) Chronic pulmonary disease 5) Cranial trauma with LOC > 15 min. 6) Severe depression or anxiety 7) History of psychiatric disease	Primary Standard uptake value ratios in specific regions of interest

ClinicalTrials Identifier	Study Sponsor	Study Dates	Clinical Phase	Interventions	Estimated Patient Enrollment	Reference Standard	Inclusion/ Exclusion Criteria	Outcomes
NCT01238458	Chang Gung Memorial Hospital	Nov. 2009 – Jun. 2012	2	Clinical evaluation of Florbetapir F 18 (18F-AV-45)	N=150 Cognitively normal, n=30 MCI, n=60 AD, n=60	Clinical diagnosis	Inclusion 1) Age ≥ 50 years 2) Cognitively normal patients, MMSE > 24 3) AD diagnosis by NINCDS-ADRDA criteria 4) MCI diagnosis, undefined Exclusion 1) Current or planned pregnancy 2) Modified Hachinski score > 4 or evidence of vascular dementia (NINDS-AIREN criteria) 3) Clinically significant abnormal lab values and/or medical or psychiatric illness 4) History of drug/alcohol abuse 5) Evidence of neurodegenerative disease other than AD, cognitive impairment from trauma, brain damage, brain infarction, epilepsy, clinically significant psychiatric disease	Primary Expansion of database of [18F]AV-45 PET imaging in AD and MCI for definition of a positive scan, over 1 year Secondary Expansion of safety database of [18F]AV-45 PET imaging with respect to adverse event count, lab parameters, vital signs and ECG Prevalence of Aβ positivity in AD and MCI patients

ClinicalTrials Identifier	Study Sponsor	Study Dates	Clinical Phase	Interventions	Estimated Patient Enrollment	Reference Standard	Inclusion/ Exclusion Criteria	Outcomes
NCT01020838	Bayer	Nov. 2009 – Aug. 2014	3	Clinical evaluation of Florbetaben (BAY 94-9172) – single injection	N=216	Postmortem diagnosis	Inclusion 1) Age ≥ 21 years 2) Willingness to donate brain for examination in case of death Exclusion 1) Current or planned pregnancy 2) Severe cerebral macrovascular disease or brain tumor 3) Severe cardiovascular instability requiring ICU care and/or therapeutic intervention	Primary Sensitivity and specificity of visual assessment of tracer uptake, on Day 1 – one scanning period after injection Secondary Sensitivity and specificity of the composite "whole brain" regional visual assessment in detecting/excluding cerebral β-amyloid plaques Sensitivity and specificity of quantitative assessment of regional tracer uptake in BAY94-9172 PET images Sensitivity and specificity of visual assessment of BAY94-9172 images on the subject level in detecting/excluding cerebral β-amyloid Safety and tolerability of a single dose of BAY94-9172, over 8 days

ClinicalTrials Identifier	Study Sponsor	Study Dates	Clinical Phase	Interventions	Estimated Patient Enrollment	Reference Standard	Inclusion/ Exclusion Criteria	Outcomes
NCT01028053	GE Healthcare	Dec. 2009 – Jan. 2013	3	Clinical evaluation of Flutemetamol (18F) injection	N=225 Patients with aMCI	Clinical diagnosis	<u>Inclusion</u> 1) Age ≥ 60 years 2) Amnestic MCI diagnosis by Petersen criteria 3) Modified Hachinski Ischemic Scale score ≤ 4 4) MMSE score 24-30 5) aMCI not due to structural causes <u>Exclusion</u> 1) Significant neurologic disease other than aMCI 2) Major depression, bipolar disease within 1 year 3) History of schizophrenia 4) Psychotic features, agitation or behavioral problems leading to potential compliance issues within 3 months	<u>Primary</u> Time to conversion to clinically probable AD in aMCI patients with abnormal and normal patterns of flutemetamol uptake based on visual assessment of PET scan, up to 2 years after initial injection <u>Secondary</u> Time to conversion to clinically probable AD in aMCI patients below and above a threshold of brain levels of flutemetamol based on semi-quantitative assessment of PET scan, up to 2 years after initial injection

ClinicalTrials Identifier	Study Sponsor	Study Dates	Clinical Phase	Interventions	Estimated Patient Enrollment	Reference Standard	Inclusion/ Exclusion Criteria	Outcomes
NCT01518374	Avid Radiopharmaceuticals	Dec. 2009 – Dec. 2015	2	Clinical evaluation of Florbetapir F 18 (18F-AV-45)	N=600	N/A	Inclusion 1) Age ≥ 18 years 2) Able to tolerate the PET scan procedure Exclusion 1) Clinically significant hepatic, renal, pulmonary, metabolic or endocrine disturbances 2) Clinically significant CVD 3) No current or planned pregnancy 4) History of drug/alcohol abuse	Primary Safety assessment: adverse event frequency related to florbetapir administration, over 24 hours Secondary Florbetapir F 18 tracer uptake as determined by standard uptake value ratio, over 90 minutes

ClinicalTrials Identifier	Study Sponsor	Study Dates	Clinical Phase	Interventions	Estimated Patient Enrollment	Reference Standard	Inclusion/ Exclusion Criteria	Outcomes
NCT01577394	Assistance Publique – Hôpitaux de Paris	Jun. 2011 – Jun. 2014	3	Evaluation of oculomotor testing	N=100 Patients with AD or DLB	N/A	Inclusion 1) Age ≥ 65 years 2) DLB diagnosis by Consortium on DLB criteria (McKeith et al 2005) 3) AD diagnosis by DSM IV and NINCDS-ADRDA criteria 4) No major sensory deficits 5) MMSE > 20 Exclusion 1) Parkinson syndrome progressing over more than 1 year 2) Use of AChEI medication 3) Use of anti-Parkinson drugs 4) Neuroleptic drug use within 3 months 5) Geriatric Depression Scale score > 10 6) Medication that may interfere with measuring dopamine transporter 7) Survival of less than 1 year 8) Lack of French fluency	Primary Variability of reflexive saccades latencies, at 1 year Secondary Correlation between oculomotor records and neuropsychological examination Correlation between hippocampal volume and neuropsychological examination in DLB Central atrophy differences between DLB and AD Percentage of CSF alpha-synuclein to distinguish DLB and AD Variations at 1 year in oculomotor test scores and neuropsychological tests scores Various assessments of saccades

ClinicalTrials Identifier	Study Sponsor	Study Dates	Clinical Phase	Interventions	Estimated Patient Enrollment	Reference Standard	Inclusion/ Exclusion Criteria	Outcomes
NCT01703702	Avid Radiopharmaceuticals	Oct. 2012 – Dec. 2014	4	Evaluation of effectiveness of florbetapir (F18) in patient management; association of scan status and cognitive decline	N=600 Patients with MCI or AD	Clinical diagnosis	Inclusion 1) Age 50 – 90 years 2) For MCI arm: MMSE 24-30 3) For AD arm: diagnosis by clinical criteria; MMSE 16-24, lack of clinical indication for non-neurodegenerative cause for cognitive impairment 4) Current evaluation for cognitive decline, or within previous 18 months 5) Lack of high confidence in diagnosis of cognitive decline Exclusion 1) Current or planned pregnancy 2) Current serious or unstable illness 3) Known results of previous amyloid imaging scan 4) Known brain lesion or alternate pathology explaining clinical presentation 5) Previous investigational trial participation within 30 days 6) Previous receipt of investigational amyloid-targeting agent 7) Radiopharmaceutical imaging or treatment within 7 days of study scan	Primary Proportion of patients with a change in management from baseline to 3 months for patients receiving immediate scan results versus those receiving results in 12 months Association between scan status and cognitive decline between baseline and 12 months Secondary Change in patients' clinical diagnoses in whom scan results were not predicted by initial clinical diagnosis Change in clinician's confidence regarding clinical diagnosis, based on scan results Change in caregiver and patient advice and counseling Evaluation of caregiver self-efficacy for management of dementia based on scan information