HHS Public Access

Author manuscript

Biostat Epidemiol. Author manuscript; available in PMC 2018 March 13.

Published in final edited form as:

Biostat Epidemiol. 2017 ; 1(1): 36–58. doi:10.1080/24709360.2017.1331821.

Linear Combinations of Multiple Outcome Measures to Improve the Power of Efficacy Analysis ---Application to Clinical Trials on Early Stage Alzheimer Disease

Chengjie Xiong[1,6], **Jingqin Luo**[2,3], **John C Morris**[4,5,6], and **Randall Bateman**[4,6]

[1]Division of Biostatistics, Washington University, St. Louis, MO

[2]Division of Public Health, Department of Surgery, Washington University, St. Louis, MO

[3]Biostatistics Core, Siteman Cancer Center, Washington University, St. Louis, MO

[4]Department of Neurology, Washington University, St. Louis, MO

[5]Department of Pathology and Immunology, Washington University, St. Louis, MO

[6]Knight Alzheimer's Disease Research Center, Washington University, St. Louis, MO

Abstract

Modern clinical trials on Alzheimer disease (AD) focus on the early symptomatic stage or even the preclinical stage. Subtle disease progression at the early stages, however, poses a major challenge in designing such clinical trials. We propose a multivariate mixed model on repeated measures to model the disease progression over time on multiple efficacy outcomes, and derive the optimum weights to combine multiple outcome measures by minimizing the sample sizes to adequately power the clinical trials. A cross-validation simulation study is conducted to assess the accuracy for the estimated weights as well as the improvement in reducing the sample sizes for such trials. The proposed methodology is applied to the multiple cognitive tests from the ongoing observational study of the Dominantly Inherited Alzheimer Network (DIAN) to power future clinical trials in the DIAN with a cognitive endpoint. Our results show that the optimum weights to combine multiple outcome measures can be accurately estimated, and that compared to the individual outcomes, the combined efficacy outcome with these weights significantly reduces the sample size required to adequately power clinical trials. When applied to the clinical trial in the DIAN, the estimated linear combination of six cognitive tests can adequately power the clinical trial.

Keywords

Alzheimer disease; Cross-validation; Multivariate mixed model for repeated measures (MMMRM); Power; Sample size

Corresponding author: Chengjie Xiong, PhD, Division of Biostatistics, Washington University, St. Louis, MO 63110, (314) 362 3635 (Office), (314) 362 2693 FAX, chengjie@wustl.edu.

Introduction

Alzheimer disease (AD) is an age-related brain-damaging disorder that results in progressive cognitive and functional impairment and death. An estimated 5.4 million Americans are now living with AD, and that number will rise to as many as 16 million by 2050 [1]. Accumulating research suggests that neurodegenerative processes associated with AD may begin during middle age [2-4] and almost certainly many years prior to symptomatic onset [5-7]. Clinicopathologic studies also demonstrate that asymptomatic individuals can manifest the neuropathological changes of AD, notably senile plaques and neurofibrillary tangles [4, 8, 9]. These observations, coupled with the absence of treatments that alter the pathological processes of AD, have led to a major paradigm shift in the search for treatments of AD. The focus of modern AD clinical trials now is on individuals at the earliest clinical stages of the illnesses, variably labeled as Mild Cognitive Impairment (MCI, [10]), prodromal AD [11], and very mild dementia [12] that merits a Clinical Dementia Rating (CDR, [13]) of 0.5 (here together are termed early symptomatic AD [14]). Even the preclinical stage of AD [15], prior to the substantial development of symptoms, can be targeted as interventions in this stage may have the greatest chance of preserving brain function.

The major paradigm shift in randomized clinical trials (RCTs) on AD, however, poses a major challenge in designing such trials. Because symptomatic progression of the disease in these early stages typically is subtle, the rate of longitudinal change in the placebo arm is expected to be slow. Given that there have been no pharmaceutical treatments that can reverse the pathological processes of AD, a slow rate of change implies that there is not enough room for improvement even with an efficacious active treatment. As a result, the RCTs on early stage AD must be designed with very large sample sizes and very long follow-up to guarantee that meaningful statistical conclusions can be drawn. Large, long-duration RCTs are time-consuming and prohibitively costly, and present a major bottleneck to allow promising therapies to be fully tested on a timely manner [16]. Moreover, it remains an open question as to which cognitive test or tests provide the best power and should be used as the primary efficacy endpoint. Recent secondary preventive RCTs that enroll asymptomatic individuals with preclinical AD (and, in some, early symptomatic individuals) include the Anti-Amyloid Treatment in Asymptomatic Alzheimer's (A4) trial, the Dominantly Inherited Alzheimer Network (DIAN)-Trials Unit (DIAN-TU) trial, and the Alzheimer's Prevention Initiative (API) trial. They all propose to employ different cognitive endpoints for their efficacy analysis, typically in the form of some types of composite cognitive score with equal weights over several cognitive tests. For example, the A4 trial proposed a composite of four cognitive measures that are well established as showing sensitivity to decline in prodromal and mild dementia [17]: the total recall score from the Free and Cued Selective Reminding Test (FCSRT) (0-48 words, [18]), the delayed recall score on the Logical Memory IIa subtest from the Wechsler Memory Scale (0-25 story units, [19]), the Digit Symbol Substitution Test score from the Wechsler Adult Intelligence Scale–Revised (0-93 symbols, [20]), and the Mini-Mental State Examination (MMSE) total score (0-30 points, [21]). The composite score was determined from its components using an established normalization method [22]. Each of the 4 component change scores is divided by

the baseline sample standard deviation of that component, to form standardized z scores. These z scores are summed to form the composite. In the API trial, the proposed cognitive composite is also equally weighted among the following individual tests over multiple cognitive domains [23]: Consortium to Establish a Registry for Alzheimer's Disease (CERAD) Word List Recall, CERAD Boston Naming Test (high frequency items), MMSE Orientation to Time, CERAD Constructional Praxis, and Raven's Progressive Matrices (Set A) [24-27]. It is reported that the proposed composite is more sensitive than using the entire CERAD battery [23]. Whereas each of the proposed cognitive composite score weighs multiple tests equally, and has performed reasonably well, it remains unknown whether they provide the optimum power for designing RCTs on early stage AD. Given that recently revised Food and Drug Administration (FDA) guidelines for RCTs on early AD mandate that treatments be only approved if they demonstrate cognitive and functional benefits (FDA Guidance for Industry AD: Developing Drugs for the Treatment of Early Stage Disease [28]), it is crucial to search for the optimum cognitive endpoint that can best power these trials [16].

In this article, we propose a novel statistical method to combine multiple outcome measures in AD trials so that the power for testing the efficacy hypothesis can be improved. We employ a mixed model for repeated measures (MMRM) for each outcome measure and link the models across multiple outcome measures through a set of random effects. We derive the linear combination of multiple outcome measures such that the ratio between the mean change from the baseline and the standard deviation (SD) is maximized. We then provide estimates to the weights of the optimum linear combination. Because the power analysis for a real world clinical trial with the combined outcome measure requires estimates first for the weights, and then for the corresponding mean and SD on the combined outcome, we assess the impact of estimating the latter (i.e., mean and SD on the combined outcome) using the same data set for estimating the weights or an independent data set (i.e., cross-validation data set) through an extensive simulation study. Finally, we demonstrate the proposed methodology by combining multiple cognitive outcome measures across multiple domains using longitudinal data from the DIAN study [29], and assess the improvement in reducing the sample size of the future DIAN-TU trial when the optimum combination of multiple cognitive outcomes is used as the primary efficacy endpoint.

Method

We consider the case when M disease outcome measures (or markers) are to be combined to achieve the smallest possible sample sizes required to adequately power a clinical trial. We assume a clinical trial on AD **with** baseline (t=0) **and** t=1, 2, …, T, scheduled post-baseline assessments. In analyzing the clinical trial data on AD, it has been a routine practice to first compute the change from baseline on efficacy outcomes for each individual and then compare the mean change across individuals between the active treatment arm and the placebo arm. For marker m, m=1, 2, …, M, we use y_{ijt}^{m} (in the following, a super index m does not mean a power function with the exception of variance which is always squared in notation) to denote the change from baseline at post-baseline time t=1, 2, …, T, for a

randomly selected individual *j* from *i*-th treatment arm (*i*=0=placebo arm, *i*=1=active therapy). We further assume a multivariate MMRM [**MMMRM**, 30-32] as follows:

$$y_{ijt}^{m}(t)=\mu_{it}^{m}+p_{ij}^{m}+e_{ijt}^{m},$$

where μ_{it}^{m} is the mean of *m*-th marker for *i*-th treatment at time *t*, p_{it}^{m} is the subject-level random effects on the *m*-th marker, and *e*'s are within-subject errors across different time points. Here, for each individual marker *m*=1, 2, …, *M*, we follow the standard assumptions that subject-level random effects *p*'s are independent of within-subject errors *e*'s, and that $e_{ij}^{m}=\left(e_{ij1}^{m},e_{ij2}^{m},\ldots,e_{ijT}^{m}\right)^{t}$ (here super index *t* means matrix transpose) follows a multivariate normal distribution with mean 0 and covariance matrix $\sum_{e}^{m}$ (structured or unstructured) whose entries are σ_{mtt}^{2} (i.e., the variances) on the diagonal for *t*=1, 2, …, *T*, and $\sigma_{mt_1t_2}$ off the diagonal when $1 \le t_1 \ne t_2 \le T$. To introduce the correlation between different markers from the same subjects, we further assume that the vector of random effects, $p_{ij}=\left(\begin{array}{cccc} p_{ij}^{1} & p_{ij}^{2} & \cdots & p_{ij}^{M} \end{array}\right)^{t}$, follows another multivariate normal distribution with mean 0 and a *M* by *M* covariance matrix

$$\sum_{p}=\left(\begin{array}{cccc} \sigma_{p1}^{2} & \sigma_{p12} & \cdots & \sigma_{p1M} \\ & \sigma_{p2}^{2} & \cdots & \sigma_{p2M} \\ & & \cdots & \cdots \\ & & & \sigma_{pM}^{2} \end{array}\right).$$

We also assume that, conditioning on p_{ij}, e_{ij}^{m}'s are independent of each other for *m*=1, 2, …, *M*. The ultimate efficacy goal in a clinical trial of AD is to estimate and compare the change from the baseline at the final time point, *T*. The statistical power to compare the drug and the placebo depends not only on how efficacious the drug is, but also independently on how sensitive the outcome measures are. In fact, if the effect size of the active treatment is expressed as a percentage of improvement on the mean change from the baseline, in comparison to the placebo arm, the statistical power is a function of the ratio between the mean (i.e., the magnitude) and the standard deviation on the change from the baseline for the placebo arm. The power analysis can be based on a single outcome measure. However, the fact that essentially all individual outcome measures show only subtle changes during early stages of AD in the placebo arm, coupled with a relatively high heterogeneity in early disease progression (i.e., large variance on the change from baseline), implies that a power analysis using a single outcome measure will likely result in a formidable sample size for these trials. Therefore, it is reasonable to hypothesize that use of multiple outcome measures may help to reduce the sample sizes required to adequately power RCTs on early stage AD by maximizing the ratio between the mean (i.e., the magnitude) and the standard deviation on the change from the baseline for the placebo arm. Here we provide the linear combination of multiple outcome measures so that the sample size to detect a fixed effect size of the active drug (in the percentage of improvement on mean change from the baseline as

compared to the placebo arm) can be minimized with a reasonable statistical power (say, 80%). For the placebo arm, let $Y_{0jt}=\begin{pmatrix} y_{0jt}^1 & y_{0jt}^2 & \cdots & y_{0jt}^M \end{pmatrix}^t$ be the vector (on change from baseline) of all M outcome measures for a random subject j at time t. Let $\mu_{0T}=\begin{pmatrix} \mu_{0T}^1 & \mu_{0T}^2 & \cdots & \mu_{0T}^M \end{pmatrix}^t$ be the mean (change from baseline) vector at last time point T where the efficacy comparison will be conducted. Let Ω be the covariance matrix of $Y_{0jT}=\begin{pmatrix} y_{0jT}^1 & y_{0jT}^2 & \cdots & y_{0jT}^M \end{pmatrix}^t$. The entries in the diagonal of the matrix are variance $\sigma^2_{y^m_{0jT}}=(\sigma^2_{pm}+\sigma^2_{mTT})$. The entries off the diagonal of the matrix are the covariance $\mathrm{Cov}(y^{m_1}_{0jT}, y^{m_2}_{0jT})=\sigma_{pm_1m_2}$ for $1 \le m_1 \ne m_2 \le M$.

Let $W=(w_1, w_2, \ldots, w_M)^t$ be a weight vector to be assigned to the outcome measures $m=1, 2, \ldots, M$. Let $U=W^tY_{0jT}=\sum_{m=1}^M w_m y^m_{0jT}$, be the linear combination of these outcome measures at last time T. The mean of U is $\mu^U_{0T}=\sum_{m=1}^M w_m\mu^m_{0T}$, and the variance is $\sigma^2_U=W^t\Omega W$. One optimum choice of the weight vector is to make the mean at last time T (on change from baseline) of the combined marker U (i.e., $\mu^U_{0T}=W^t\mu_{0T}$) as large as possible, whereas at the same time, the variance of U (i.e., $\sigma^2_U=W^t\Omega W$) as small as possible. Because the power function for comparing the means between the active treatment arm and the placebo arm at time T ultimately depends on the ratio of the mean and the standard deviation (SD) of U, maximizing the ratio (**in absolute value**) between the mean and standard deviation over all the possible choices of weight vectors will lead to improvement in designing clinical trials, i.e., the most reduced sample sizes. Let

$$R=\frac{W^t\mu_{0T}}{\sqrt{W^t\Omega W}}.$$

We assume here that the individual outcome measures are all oriented so that $\mu_{0T}<0$. We further assume that the weight vector W is chosen so that $W^t\mu_{0T}<0$. Maximizing R (**in absolute value**) is equivalent to maximizing

$$R^2=\frac{W^t\mu_{0T}\mu^t_{0T}W}{W^t\Omega W}.$$

The maximum is achieved when W is an eigenvector W_0 corresponding to the largest eigenvalue of $\Omega^{-1}\mu_{0T}\mu^t_{0T}$ [33]. The maximizing value of R^2 is the largest eigenvalue λ_0 of $\Omega^{-1}\mu_{0T}\mu^t_{0T}$. Because $\Omega^{-1}\mu_{0T}\mu^t_{0T}$ is a matrix of rank 1, the largest eigenvalue of the matrix is its trace, i.e., $\lambda_0=\mu^t_{0T}\Omega^{-1}\mu_{0T}$. Further, it is easy to verify that $W_0=-\Omega^{-1}\mu_{0T}$ is an eigenvector corresponding to the largest eigenvalue (unique up to a constant). Therefore, the

mean (change from baseline at last time T) for the optimally combined marker $U_0=W_0^tY$ is given by

$$\mu_0=-\mu_{0T}^t\Omega^{-1}\mu_{0T}.$$

The variance of the optimally combined marker $U_0=W_0^tY$ at the final time point T is

$$\sigma_0^2=\mu_{0T}^t\Omega^{-1}\mu_{0T}=-\mu_0.$$

Notice that our proposed analytic approach with MMMRM is very different from some of the traditional clinical trial analyses on later stage AD [34-36] in which a cross sectional analysis was conducted on the cognitive change from baseline, typically on the last-observation-carried-forward (LOCF) endpoint. We choose MMMRM because almost all reported clinical trials on AD are longitudinal (designed with cognitive measurements per subject at multiple time point including intermediate time points in addition to baseline and the end of trial), and almost all have considerable missing data and dropouts. Although the main efficacy test in clinical trials on AD is on change from two time points (baseline to the end of the trial), and the approach of LOCF has been used in the literature to impute the missing data and hence reduce the efficacy test into a simple cross sectional analysis, it has been well documented in the clinical trial literature [37] that the cross sectional analysis on the LOCF endpoint provides biased estimates to the true effect size. Therefore, appropriate longitudinal models such as MMRM that can adequately handle missing data and utilize the entire longitudinal data (including those at intermediate time point between the baseline and the end of the trial) must be used to provide valid efficacy test. In addition, the entire mean vector μ_{0T}, as well as the covariance matrix Ω in the MMMRM, are both needed in the derived optimal weights, $W_0 = -\Omega^{-1}\mu_{0T}$. The maximum likelihood estimation (MLE) of these parameters, especially those in Ω, must rely on the MMMRM under the standard assumptions in the model. Further, because the variance-covariance parameters (part of matrix Ω) in the subject-level random effects, $p_{ij}=\begin{pmatrix} p_{ij}^1 & p_{ij}^2 & \cdots & p_{ij}^M \end{pmatrix}^t$, are shared by all the time points (including the intermediate time points) within the same subjects, the MMMRM provides a much better estimation to all these important parameters because the model allows the use of all longitudinal data in the estimation, including those from subjects whose cognitive data are available only on some, but not all, of the cognitive tests, as well as from those who drop out in the middle of the trial. As a matter of fact, this optimal approach of using all available data to estimate the parameters in Ω is exactly what we will implement in the real world application of our proposed methodology (see below: **Designing Clinical Trials on Autosomal Dominant Alzheimer Disease**).

Although we have focused on a methodology that compared treatment arms at the last time point, similar analytic approaches can be used to power secondary efficacy comparisons such as those at any post baseline time point, which is especially appealing, given that the longitudinal cognitive decline in the progression of AD is **likely** not a linear pattern, and

different drugs may have different efficacy for specific time windows of the disease progression.

It is important to point out that different treatments to be tested could have very different efficacy profiles across multiple cognitive outcomes, depending on their specific mechanisms of action. Because nobody knows the individual efficacy specific to each cognitive outcome for a novel treatment to be tested, it is essentially impossible to obtain the entire efficacy profiles across multiple cognitive outcomes at the design stage of a clinical trial during which a well defined primary cognitive efficacy endpoint has to be proposed and used to power the trial. Hence, our proposed methodology does not intend to identify the optimum combination of multiple cognitive outcomes for a specific efficacy profile from a specific treatment. Instead, our proposed methodology offers a general purpose approach that has the potential to provide a single composite score of multiple outcomes to improve the design of any clinical trial on early stage AD, which is consistent with the fact that essentially all published clinical trials on later stage AD thus far have all been based on the same single cognitive endpoint [34-36], i.e., the Alzheimer's Disease Assessment Scale for Cognition (ADAS-Cog, [38]). Just like the way ADAS-cog was developed, our proposed optimum combination of cognitive outcomes can be readily estimated from observational studies on subjects without receiving any active treatments. These estimates can be obtained from large observational studies such as the National Alzheimer's Coordinating Center's (NACC) Uniform Data Set (UDS, [39]) and the Alzheimer's Disease Neuroimaging Initiative (ADNI, [40]).

Simulation Results

In designing a real world clinical trial by using the proposed methodology, the optimum weights have to be first estimated using existing or pilot data that are collected in a sample comparable to the targeted population for the therapeutic intervention. These weights can then be used to form the combined outcome measure and then to power the clinical trials. It is important to assess how well these weights can be estimated, assuming different sizes of the sample for the pilot data. Notice that the power analysis also requires the estimates to the mean changes from the baseline on the combined outcome as well as the associated SD, which can be based on the same data or an independent sample (**i.e., the cross-validation sample**). We posit that, if these estimates are based on the same samples that are used to derive the optimum weights for combining the multiple outcome measures, some bias could be introduced in the power analysis. We therefore design a simulation study to examine the validity of our proposed methodology as well as the amount of bias if both weights and parameters (i.e., mean and SD) for the subsequent power analyses are estimated from the same samples. We consider the case when 3 outcome measures are to be combined across a set of chosen parameters in the trivariate mixed effects model representing a wide range of effect sizes across individual outcome measures. Specifically, for m=1, 2, 3, and t=1, 2, 3 (i.e., 3 post-baseline follow-ups) and the model for the placebo, $y^m_{0jt}(t)=\mu^m_{0t}+p^m_{0j}+e^m_{0jt}$, we assume a non-linear pattern of the mean for individual outcome measures over time,

$$\mu_{0t}^{m} = -(t^2+m)/6.$$

We also assume that $e_{ij}^{m} = \left(e_{ij1}^{m}, e_{ij2}^{m}, e_{ij3}^{m} \right)^{t}$ follows a multivariate normal distribution with mean 0 and auto-regressive covariance matrix

$$\sum\nolimits_{e}^{m} = \begin{pmatrix} 0.4 & \rho\sqrt{0.24} & \rho^2\sqrt{0.32} \\ & 0.6 & \rho\sqrt{0.48} \\ & & 0.8 \end{pmatrix}$$

Notice here we allow an increasing variance as a function of time, which is very common when cognitive changes from baseline are used as the primary efficacy endpoint in clinical trials on AD. We further assume that $p_{ij} = \left(p_{0j}^{1}, p_{0j}^{2}, p_{0j}^{3} \right)^{t}$ follows a trivariate normal distribution with mean 0 and covariance matrix

$$\sum\nolimits_{p} = 2 \begin{pmatrix} 1 & r & r \\ & 1 & r \\ & & 1 \end{pmatrix}.$$

We simulated 100 pairs of data sets from the model. Each pair contains two independently simulated data sets: one is used to estimate the optimum weights for combining three markers (called the training data set), and the other is used to estimate the mean changes (from baseline) as well as the related variance on the combined marker for the power analysis (called the validation data set). Table 1 reports the true weights, the mean of 100 sets of estimated weights, the bias and mean square error (MSE) on the estimated weights (averaged over 100 estimates for each weight and then over 3 weights), and the true and estimated ratio between the mean and SD at t=3 (the endpoint of the trial where efficacy comparison will be done) for the true and estimated combination of 3 outcome measures for ρ=0.2, r=0.2. Table 1A to Table 1H in the Appendix report the similar results for other options of ρ=0.2, 0.5, 0.8, and r=0.2, 0.5, 0.8. Table 2 presents the results for the subsequent power analyses for a clinical trial with two arms and a 1:1 sample size ratio, i.e. the average total sample sizes (over 100 power analyses from 100 simulated pairs of data sets) required for 80% power to detect specified effect sizes (assuming $\rho = r = 0.5$). 40 subjects were assumed for the training data set (to estimate the optimum weights) as well as for the validation data set (to estimate the mean and SD on the combined outcome) in Table 2. Table 2A to Table 2E in Appendix report the similar power analysis for other choices of sample sizes in the training/validation data sets (n=the shared sample size in both training and validation data sets=60, 80, 100, 120, and 140).

Designing Clinical Trials on Autosomal Dominant Alzheimer Disease

Autosomal dominant Alzheimer's disease (ADAD) has informed the field of AD research about the molecular and biochemical mechanisms that are believed to underlie the pathological basis of AD. The DIAN study since 2008 has established an international, multicenter registry of individuals at risk or with a known causative mutation of AD in the amyloid precursor protein (APP), presenilin 1 (PSEN1), or presenilin 2 (PSEN2) genes [29]. Interim cross-sectional analyses indicate a cascade of AD biomarker changes that begin at least 20 years before the symptomatic onset of AD [6]. The DIAN study evaluates participants at entry and longitudinally thereafter with clinical and cognitive batteries, structural, functional, metabolic, and amyloid imaging protocols, and biological fluid (blood and cerebrospinal fluid) collection with the goal of determining the sequence of changes in asymptomatic gene carriers who are destined to develop AD. Early analysis of DIAN longitudinal study has provided a suggestive algorithm of biomarker change in the ADAD population over time before the estimated age of onset of AD [41]. This reinforces the rationale that biomarker changes are ongoing in asymptomatic ADAD participants and allows the identification of drugs that demonstrate biomarker efficacy before they are tested in a large scale critical trial for cognitive efficacy. Building on this information, the DIAN-TU **was** formally launched in December, 2012, to conduct clinical trial in the ADAD population by evaluating two anti-amyloid monoclonal antibodies, in comparison with placebo, for AD biomarker target engagement [29]. This trial measures the effects of the drugs on a comprehensive set of AD biomarkers (e.g., amyloid deposition [42], cerebrospinal fluid (CSF) Aβ and tau [41], magnetic resonance imaging (MRI) brain atrophy [43], and positron emission tomography (PET) imaging with 2- [18F] fluoro-2-deoxy-D-glucose (FDG PET, [44]) that would be used to determine if a drug is likely to have a cognitive benefit in a subsequent cognitive endpoint trial.

An adaptive design has been implemented to seamlessly transition the biomarker phase of DIAN-TU trial to the Phase III trial in which the primary goal is to establish the efficacy of selected active drugs in slowing the rate of cognitive decline. A large cognitive battery has been administered in DIAN participants [45], covering a wide range of domains. Based on recent reports on cognitive domains that may exhibit early progression from the ongoing A4 trial [17] and the API trial [23] as well as a preliminary analysis on the DIAN database, to demonstrate the application of our proposed methodology, we have chosen the following six candidates of the cognitive tests that can be considered for powering the upcoming Phase III cognitive endpoint trial:

1. the total score of the MMSE [21], a widely used screening test that evaluates overall cognitive function by examining orientation to place, orientation to time, attention, concentration, language, ability to recall previously given words and visual-spatial skills. The Score of MMSE ranges from 0 to 30 [21];
2. the total number of 16 words that are recalled (WORDDEL) by participants after a delay of approximately 20-30 minutes from the time they were read out loud to the participants. The WORDDEL score ranges from 0 to 16 [45];

3. the total score (PAPER) from a multiple choice test of spatial problem solving ability, requiring participants to mentally manipulate pieces of paper. The PAPER score ranges from 0 to 12 [46];

4. the score on the Digit Symbol (DIGSYM) subscale of the Wechsler Adult Intelligence Scale (WAIS) which, by evaluating the amount of time it takes participants' to recode pairs of digit and symbol items, examines several cognitive abilities including attention, psychomotor speed, complex scanning, visual tracking as well as working memory. The WAIS DIGSYM score ranges from 0 to 93 [20];

5. the total score on a subscale on the Pair Binding task (PAIRBINDING), which assesses associative recall accuracy. The score of PAIRBINDING ranges from 0 to 12 [47];

6. the total score on the 30-items Boston Naming test (BOSTON), examining word retrieval performance by requiring participants to name objects presented in drawings. The total score of BOSTON ranges from 0 to 30 [48].

The cognitive data used for the power analysis are from the sixth data freeze of the DIAN database, which had a data cutoff date of June 30, 2013. A total of 53 individuals met the inclusion and exclusion criteria for the ongoing DIAN TU trial on biomarkers. All had a pathogenic mutation for AD, and baseline cognitive assessment and at least one (and up to 3) post-baseline follow-up measurements where each may occur either within 1, 2 or 3 years after the baseline. Only participants with a CDR of 0, 0.5, and 1 whose expected years to onset (computed as the participant's age minus the reported age of symptom onset from the affected parent) is between -15 to 10 years at baseline [49] are included in the analysis. Table 3 presents the baseline demographic, cognitive, and functional information of subjects in the data set.

For all six cognitive tests discussed above, higher scores indicate better cognitive performance. For each outcome measure and each individual, z-scores are derived by subtracting the mean of the entire sample at the baseline from the individual's raw scores (both at baseline and at follow-ups) and then dividing the difference by the standard deviation (SD) computed at baseline. Subsequently, the z-scores are used to compute change scores (i.e. change from baseline to each of the three annual follow-ups) which are then utilized as the outcome measures for power analyses.

Because of the relatively small sample size from the pilot data set, a stepwise process is implemented to combine the cognitive outcomes. First, a univariate mixed model for repeated measures [30] is used to model the change of z-scores over the longitudinal courses for each cognitive outcome measures. An unstructured covariance matrix for the within-subject errors is assumed. The mean change from baseline for **each** cognitive outcome, as well as the variance/covariance parameters are estimated from the model. These estimates are used to conduct the power analysis for a future clinical trial using each cognitive outcome as the primary efficacy endpoint. Table 4 presents the sample sizes needed, for each of the six cognitive measures, for detecting various effect sizes (expressed as a percentage improvement relative to the estimated year 3 change from baseline) with 80% power.

Second, the cognitive outcome measure that results in the smallest sample size from the univariate power analysis is chosen to combine with each of the remaining cognitive outcome measures. A sequence of bivariate mixed models for repeated measures [BMMRM, 31-32] is used to model the z-scores over the longitudinal courses between 2 outcome measures. An unstructured covariance matrix for the within-subject errors is assumed for each outcome measure, and another unstructured covariance matrix for the random effects across all outcome measures is also assumed. The optimum weights are first estimated from each BMMRM using the estimated mean change from baseline for both cognitive outcomes, as well as the variance/covariance parameters. The mean change from baseline for the combined cognitive outcome and the SD are then estimated from another univariate MMRM on the combined outcome, which are then used to conduct the power analysis. Because of the small sample size for the pilot data, no cross-validation procedure is implemented in these power analyses. Table 5 presents the sample sizes needed, for each combined outcome of two cognitive measures, for detecting various effect sizes (expressed as a percentage improvement relative to the estimated year 3 change from baseline) with 80% power.

The above process is repeated to form a composite of three cognitive outcome measures using the two cognitive measures whose combination yields the smallest sample sizes in Table 5. In order to derive the optimum weights, the combination of two cognitive measures, computed at the previous step, is treated as a single outcome variable which will be further combined with each of the remaining cognitive outcome measures through another sequence of bivariate mixed models for repeated measures (BMMRM). This “forward selection” algorithm for combining multiple outcome measures, bearing some resemblance to forward model selection in standard regression analyses, is one approach to deal with a situation when a researcher faces a potentially large pool of candidate outcome measures with a relatively small sample size. In which case, the simultaneous estimation of weights for many outcome measures, in a single high dimensional multivariate mixed model for repeated measures, may not be computationally feasible and in our experience, create frequent problems when the MMMRM do not converge likely because of a limited sample size. Table 6 presents the sample sizes needed, for each combined outcome of three cognitive measures, for detecting various effect sizes (expressed as a percentage improvement relative to the estimated year 3 change from baseline) with 80% power.

In our simulation studies as well as our application to the DIAN-TU trial, all the estimated weights were positive. However, it is mathematically possible that some of the estimated weights could be negative. This is because that, even if the individual outcome measures are all oriented so that $\mu_{0T} < 0$ component wise and the weight vector W is chosen so that $W^t\mu_{0T} < 0$, the optimal weight vector $W_0 = -\Omega^{-1}\mu_{0T}$ are not all guaranteed to be positive component wise. If some weights are indeed estimated as negative in our application to DIAN-TU trial, they would make the interpretation of the combined cognitive outcome less straightforward. For example, the estimated negative weights may point to a much deeper insight of the multi-domain cognitive progression of early stage AD, namely, the fastest rate of cognitive progression at early stage AD is not from a simple (positively) weighted average of multiple cognitive domains, but some type of contrast between domains that are positively weighted and those negatively weighted. Whereas the possible negative weights

and their potential implications to our understanding of the early cognitive progression on AD is not the focus of the current manuscript, future research is needed in this direction.

Discussion

There is currently a major paradigm shift in the search for treatments of AD, that is, the focus of modern AD clinical trials now is on individuals at the earliest clinical stages, prior to the substantial development of clinical symptoms. The subtle disease progression at the early stages, however, poses a major challenge in designing such RCTs as a huge sample size is required to adequately power such trials. The lack of detection of progression by individual cognitive outcomes makes the sample size for RCTs on early stage AD a formidable task to achieve.

We tackled this challenge by combining multiple cognitive outcomes across several domains to optimize the rate of progression (divided by the SD) in individuals at an early stage or preclinical stage of AD. Whereas ongoing clinical trials on early stage AD such as the A4 trial and the API trial have proposed differing cognitive composites using the same weights across multiple cognitive domains [17, 23], we sought to improve these by mathematically optimizing the weights so that the sample sizes for adequately powering such trials could be minimized. We proposed a multivariate mixed model for repeated measures to jointly model the longitudinal progression of multiple outcomes. The model amounts to individual MMRM for each cognitive outcome, but allows the correlation of multiple outcome measures from the same individuals through a set of random effects. We mathematically derived the optimum weights and the mean change from baseline on the combined outcome. We further provided estimates to the optimum weights as well as the estimates to the mean change from baseline and SD on the combined outcome using the maximum likelihood estimates from the standard statistical software such as PROC MIXED/SAS [50].

We conducted a simulation study to examine how accurate the optimum weights can be estimated as a function of sample sizes and under various parameter settings. Because the power analysis on efficacy outcome ultimately depends on the ratio of the mean change (from baseline) and the corresponding SD, we also assessed how the maximized ratio can be accurately estimated. Furthermore, we examined how the proposed methodology can be best implemented in designing real world clinical trials. Because our approach requires two sequential steps: first to estimate the weights to combine multiple outcomes, and then to form the combination and estimate the mean and SD on the combined outcome for the power analysis, a naive approach is to implement both steps in the same pilot data set for the power analysis. We proposed to estimate the weights to combine multiple outcomes and the mean and SD on the combined outcome for the power analysis on two independent samples (i.e., a training sample and a cross-validation sample), and assessed the difference between the two approaches on the results of subsequent power analyses.

Results in Table 1 indicate that the optimum weights can be accurately estimated, even when the sample sizes are as small as 40. The estimated bias is close to 0 for a wide range of sample sizes starting from 40 and other covariance parameters. The mean square error (MSE) is also small, and as expected, decreases as the sample size increases. Most

importantly, the estimated ratio of the mean and SD on the combined outcome is also close to the true ratio, and progressively closer as the sample size increases.

Results in Table 2 indicate a significant reduction of the sample size in powering a 1:1 two-arm clinical trial by using the combined outcome measures, as compared to each individual outcome measure. Furthermore, the naive approach of estimating both weights to combine multiple outcomes and the resulting mean and SD on the combined outcome **from** the same pilot data set produces an under-estimate to the sample sizes to adequately power clinical trials, the under-estimation is more severe when the effect sizes are relatively small.

Finally, we applied the proposed method to the DIAN-TU trial on early stage AD by combining six cognitive tests from the DIAN cognitive battery. We found that the sample sizes to adequately power the DIAN-TU Phase III trial on cognitive endpoint are significantly reduced when the combined outcomes are used, in comparison to individual ones. We also found that the sample size improvement is more profound with the combination of three cognitive tests than two. Because the set of all linear combinations of more cognitive outcomes are a bigger set than all linear combinations from a subset of the outcomes (with zero weights for those outcomes not in the subset), adding more tests will always do better (i.e., smaller sample sizes) in theory. An important question in practical applications is how many cognitive tests are needed in the combinations until no appreciable improvement can be achieved. Further research is needed in this area. We also point out that there seem to be some conflicting results in Table 5 and Table 6 (i.e., the estimated sample size using MMSE and PAPER and BOSTON is actually larger than that using only MMSE and PAPER). This is due to the fact that all the optimum weights reported in Table 5 and 6 to combine multiple cognitive tests have to be estimated from the existing database from DIAN, each time with different data sets due to missing data in the added marker as well as dropouts. Different data sets in general result in different parameter estimates, which then implies different weights for the combined cognitive composite, leading to potentially conflicting sample sizes in power analyses.

Additionally, FDA and other regulatory authorities have traditionally required a novel treatment on AD only be approved with evidence of a beneficial impact on both cognitive outcome and functional outcome. Whereas FDA's new draft guidelines on RCTs on early stage AD seem more flexible in the functional impact due to the fact that subjects with early stage AD (preclinical stage or prodromal stage) have essentially no functional impairment [28], our proposed methodology can also be applied to any set of functional outcomes, (i.e., Activity of Daily Living (ADL), CDR sum-of-box) to derive the optimal combination of multiple functional outcomes for optimizing the efficacy comparison on the functional outcome in RCTs on early stage AD.

Finally, the MMMRM assumes that all random effects as well as the random errors are all normally distributed. It is well known that many of the cognitive tests used in early stage AD studies are subject to ceiling and floor effects and may not follow normal distributions. The validity of our proposed optimum weights for combining multiple cognitive outcomes against the distributional assumption need to be evaluated before the proposed methodology can be fully utilized in real world clinical trials on early stage AD. As a matter of fact, under

the multivariate MMRM, $y_{ijt}^{m}(t)=\mu_{it}^{m}+p_{ij}^{m}+e_{ijt}^{m}$, our derivation of the optimum weights over the linear combinations is never based on the assumption of normal distributions for *p*'s and *e*'s. The only assumptions used are the existence of the covariance matrix, the conditional independence of e_{ij}^{m}'s for *m*=1, 2, …, *M* (given p_{ij}), and the independence between *p*'s and *e*'s. Hence, our proposed optimum weights are robust against departure from the normal distributions. Further, even without the assumption of normal distributions, these weights still maximize the absolute value of $R=W^{t}\mu_{0T}/\sqrt{W^{t}\Omega W}$ over all possible weights. Notice that *R* is related directly to the asymptotic Wald test statistic on the combined cognitive outcome. Hence, assuming that the distributions of *p*'s and *e*'s meet standard regularity conditions on the smoothness of the density functions in the model but are not normal, the asymptotic property of the maximum likelihood estimators still implies that our estimated optimum weights minimize the same size required to power the efficacy comparison that is based on the asymptotic Wald's test.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the referees and the editor for their constructive comments. This study was supported by National Institute on Aging (NIA) grant R01 AG034119 and R01 AG053550 (Dr. Xiong). Additional support was provided by NIA grant UF1 AG03243807 (Dr. Bateman) and P50 AG005681 (Dr. Morris). The authors thank the Genetics Core (Alison Goate, DPhil, Core Leader) of the DIAN for the genetic data, and the Clinical Core (Dr. Morris and Dr. Jason Hassenstab) of the DIAN for the clinical and cognitive data. The corresponding author (Dr. Xiong) had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. Alzheimer's Association. http://www.alz.org/downloads/Facts_Figures_2014.pdf
2. Braak H, Braak E. Frequency of stages of Alzheimer-related lesions in different age categories. Neurobiol Aging. 1997 Jul-Aug;18(4):351–7. [PubMed: 9330961]
3. Price JL, Morris JC. Tangles and plaques in nondemented aging and “preclinical” Alzheimer's disease. Ann Neurol. 1999; 45:358–368. [PubMed: 10072051]
4. Kok E, Haikonen S, Luoto T, Huhtala H, Goebeler S, Haapasalo H, Karhunen PJ. Apolipoprotein E-dependent accumulation of Alzheimer disease-related lesions begins in middle age. Ann Neurol. 2009 Jun; 65(6):650–7. [PubMed: 19557866]
5. Morris JC, Price JL. Pathologic correlates of nondemented aging, mild cognitive impairment, and early stage Alzheimer's disease. J Mol Neurosci. 2001; 17:101–118. [PubMed: 11816784]
6. Bateman RJ, Xiong C, Benzinger TL, Fagan AM, Goate A, Fox NC, Marcus DS, Cairns NJ, Xie X, Blazey TM, Holtzman DM, Santacruz A, Buckles V, Oliver A, Moulder K, Aisen PS, Ghetti B, Klunk WE, McDade E, Martins RN, Masters CL, Mayeux R, Ringman JM, Rossor MN, Schofield PR, Sperling RA, Salloway S, Morris JC. Dominantly Inherited Alzheimer Network. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. N Engl J Med. 2012 Aug 30; 367(9):795–804. [PubMed: 22784036]
7. Katzman R. Editorial: The prevalence and malignancy of Alzheimer disease. A major killer. Arch Neurol. 1976; 33:217–218. [PubMed: 1259639]
8. Bennett DA, Schneider JA, Arvanitakis Z, Kelly JF, Aggarwal NT, Shah RC, Wilson RS. Neuropathology of older persons without cognitive impairment from two community-based studies. Neurology. 2006 Jun 27; 66(12):1837–44. [PubMed: 16801647]

9. Price JL, McKeel DW Jr, Buckles VD, Roe CM, Xiong C, Grundman M, Hansen LA, Petersen RC, Parisi JE, Dickson DW, Smith CD, Davis DG, Schmitt FA, Markesbery WR, Kaye J, Kurlan R, Hulette C, Kurland BF, Higdon R, Kukull W, Morris JC. Neuropathology of nondemented aging: presumptive evidence for preclinical Alzheimer disease. Neurobiol Aging. 2009 Jul 30.(7):1026–36. [PubMed: 19376612]
10. Petersen RC, Doody R, Kurz A, Mohs RC, Morris JC, Rabins PV, Ritchie K, Rossor M, Thal L, Winblad B. Current concepts in mild cognitive impairment. Arch Neurol. 2001 Dec; 58(12):1985–92. [PubMed: 11735772]
11. Dubois B, Feldman HH, Jacova C, et al. Revising the definition of Alzheimer's disease: a new lexicon. Lancet Neurol. 2010; 9:1118–27. [PubMed: 20934914]
12. Morris JC. Revised Criteria for Mild Cognitive Impairment May Compromise the Diagnosis of Alzheimer Disease Dementia. Arch Neurol. 2012; 69(6):700–708. [PubMed: 22312163]
13. Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology. 1993 Nov; 43(11):2412–4.
14. Morris JC, Blennow K, Froelich L, Nordberg A, Soininen H, Waldemar G, Wahlund LO, Dubois B. Harmonized diagnostic criteria for Alzheimer's disease: recommendations. J Intern Med. 2014; 275:204–213. [PubMed: 24605805]
15. Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, Iwatsubo T, Jack CR Jr, Kaye J, Montine TJ, Park DC, Reiman EM, Rowe CC, Siemers E, Stern Y, Yaffe K, Carrillo MC, Thies B, Morrison-Bogorad M, Wagster MV, Phelps CH. 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 May; 7(3):280–92. [PubMed: 21514248]
16. Ringman JM, Grill J, Rodriguez-Agudelo Y, Chavez M, Xiong C. Commentary on “a roadmap for the prevention of dementia II: Leon Thal Symposium 2008.” Prevention trials in persons at risk for dominantly inherited Alzheimer's disease: opportunities and challenges. Alzheimers Dement. 2009 Mar; 5(2):166–71. [PubMed: 19328453]
17. Donohue MC, Sperling RA, Salmon DP, Rentz DM, Raman R, Thomas RG, Weiner M, Aisen PS. Australian Imaging Biomarkers, and Lifestyle Flagship Study of Ageing; Alzheimer's Disease Neuroimaging Initiative; Alzheimer's Disease Cooperative Study. The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. JAMA Neurol. 2014 Aug; 71(8):961–70. [PubMed: 24886908]
18. Grober E, Hall CB, Lipton RB, Zonderman AB, Resnick SM, Kawas C. Memory impairment, executive dysfunction and intellectual decline in preclinical Alzheimer's disease. J Int Neuropsychol Soc. 2008; 14(2):266–278. [PubMed: 18282324]
19. Wechsler, D. WMS-R: Wechsler Memory Scale–Revised: Manual. San Antonio, TX: Psychological Corporation; 1987.
20. Wechsler, D. Wechsler Adult Intelligence Scale–Revised. San Antonio, TX: Psychological Corporation; 1981.
21. Folstein MF, Folstein SE, McHugh PR. “Mini-mental state”: a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975; 12(3):189–198. [PubMed: 1202204]
22. Cutter GR, Baier ML, Rudick RA, et al. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain. 1999; 122(pt 5):871–882. [PubMed: 10355672]
23. Ayutyanont N, Langbaum JB, Hendrix SB, Chen K, Fleisher AS, Friesenhahn M, Ward M, Aguirre C, Acosta-Baena N, Madrigal L, Muñoz C, Tirado V, Moreno S, Tariot PN, Lopera F, Reiman EM. The Alzheimer's prevention initiative composite cognitive test score: sample size estimates for the evaluation of preclinical Alzheimer's disease treatments in presenilin 1 E280A mutation carriers. J Clin Psychiatry. 2014; 75(6):652–60. [PubMed: 24816373]
24. Acosta-Baena N, Sepulveda-Falla D, Lopera-Gómez CM, et al. Pre-dementia clinical stages in presenilin 1 E280A familial early-onset Alzheimer's disease: a retrospective cohort study. Lancet Neurol. 2011; 10(3):213–220. [PubMed: 21296022]

25. Lopera F, Ardilla A, Martínez A, et al. Clinical features of early-onset Alzheimer disease in a large kindred with an E280A presenilin-1 mutation. JAMA. 1997; 277(10):793–799. [PubMed: 9052708]
26. Arango Lasprilla JC, Iglesias J, Lopera F. Neuropsychological study of familial Alzheimer's disease caused by mutation E280A in the presenilin 1 gene. Am J Alzheimers Dis Other Demen. 2003; 18(3):137–146. [PubMed: 12811988]
27. Rosselli MC, Ardila AC, Moreno SC, et al. Cognitive decline in patients with familial Alzheimer's disease associated with E280a presenilin-1 mutation: a longitudinal study. J Clin Exp Neuropsychol. 2000; 22(4):483–495. [PubMed: 10923058]
28. Food and Drug Administration. Guidance for industry Alzheimer's disease: Developing drugs for the treatment of early stage disease Draft Guidance. US Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER); 2013. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM338287.pdf Updated January 16, 2013
29. Mills SL, Mallmann JA, Santacruz AM, Fuqua A, Carril M, Aisen PS, Althage MC, Belyew S, Benzinger TL, Brooks WS, Buckles VD, Cairns NJ, Clifford D, Danek A, Fagan AM, Farlow MR, Fox N, Ghetti B, Goate A, Heinrichs D, Hornbeck RD, Jack C, Jucker M, Klunk WE, Marcus DS, Oliver AM, Ringman JM, Rossor MN, Salloway S, Schofield PR, Snider BJ, Snyder PJ, Sperling R, Stewart CR, Thomas RG, Xiong C, Bateman R. Preclinical Trials in Autosomal Dominant Alzheimer's Disease: Implementation of the DIAN-TU. Revue Neurologique. 2013; 169(10):737–743. [PubMed: 24016464]
30. Diggle, PJ., Heagerty, P., Liang, KY., Zeger, SL. Analysis of Longitudinal Data (2nd ed). New York: Oxford University Press; 2002.
31. Shah A, Laird NM, Schoenfeld DA. A random-effects model for multiple characteristics with possibly missing data. J Amer Statist Assn. 1997; 92:775–9.
32. Fieuws S, Verbeke G. Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach. Stat Med. 2004 Oct 30; 23(20):3093–104. [PubMed: 15449333]
33. Noble, B., Daniel, JW. Applied Linear Algebra. Englewood Cliffs, NJ: Prentice-Hall Inc; 1977.
34. Rafii MS, Walsh S, Little JT, Behan K, Reynolds B, Ward C, Jin S, Thomas R, Aisen PS. A phase II trial of huperzine A in mild to moderate Alzheimer disease. Neurology. 2011; 76:1389–1394. [PubMed: 21502597]
35. Rogers SL, Farlow MR, Doody RS, Mohs R, Friedhoff LT. Donepezil Study Group. A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease. Neurology. 1998; 50:136–145. [PubMed: 9443470]
36. Sano M, Bell KL, Galasko D, Galvin JE, Thomas RG, van Dyck CH, Aisen PS. A randomized, double-blind, placebo-controlled trial of simvastatin to treat Alzheimer disease. Neurology. 2011; 77:556–563. [PubMed: 21795660]
37. Mallinckrodt CH, Sanger TM, Dubé S, DeBrota DJ, Molenberghs G, Carroll RJ, Potter WZ, Tollefson GD. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. Biol Psychiatry. 2003; 15:754–60.
38. Mohs RC, Knopman D, Petersen RC, Ferris SH, Ernesto C, Grundman M, Sano M, Bieliauskas L, Geldmacher D, Clark C, Thal LJ. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. The Alzheimer's Disease Cooperative Study. Alzheimer Dis Assoc Disord. 1997; 11(2):S13–21. [PubMed: 9236948]
39. Morris JC, Weintraub S, Chui HC, Cummings J, Decarli C, Ferris S, Foster NL, Galasko D, Graff-Radford N, Peskind ER, Beekly D, Ramos EM, Kukull WA. The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. Alzheimer Dis Assoc Disord. 2006 Oct-Dec;20(4):210–6. [PubMed: 17132964]
40. Jack CR Jr, Wiste HJ, Vemuri P, Weigand SD, Senjem ML, Zeng G, Bernstein MA, Gunter JL, Pankratz VS, Aisen PS, Weiner MW, Petersen RC, Shaw LM, Trojanowski JQ, Knopman DS. Alzheimer's Disease Neuroimaging Initiative. Brain beta-amyloid measures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to Alzheimer's disease. Brain. 2010 Nov; 133(11):3336–48. [PubMed: 20935035]

41. Fagan AM, Xiong C, Jasielec MS, Bateman RJ, Goate AM, Benzinger TL, Ghetti B, Martins RN, Masters CL, Mayeux R, Ringman JM, Rossor MN, Salloway S, Schofield PR, Sperling RA, Marcus D, Cairns NJ, Buckles VD, Ladenson JH, Morris JC, Holtzman DM. Dominantly Inherited Alzheimer Network. Longitudinal change in CSF biomarkers in autosomal-dominant Alzheimer's disease. Sci Transl Med. 2014 Mar 5.6(226):226ra30.

42. Benzinger TL, Blazey T, Jack CR Jr, Koeppe RA, Su Y, Xiong C, Raichle ME, Snyder AZ, Ances BM, Bateman RJ, Cairns NJ, Fagan AM, Goate A, Marcus DS, Aisen PS, Christensen JJ, Ercole L, Hornbeck RC, Farrar AM, Aldea P, Jasielec MS, Owen CJ, Xie X, Mayeux R, Brickman A, McDade E, Klunk W, Mathis CA, Ringman J, Thompson PM, Ghetti B, Saykin AJ, Sperling RA, Johnson KA, Salloway S, Correia S, Schofield PR, Masters CL, Rowe C, Villemagne VL, Martins R, Ourselin S, Rossor MN, Fox NC, Cash DM, Weiner MW, Holtzman DM, Buckles VD, Moulder K, Morris JC. Regional variability of imaging biomarkers in autosomal dominant Alzheimer's disease. Proc Natl Acad Sci U S A. 2013 Nov 19; 110(47):E4502–9. [PubMed: 24194552]

43. Mintun MA, LaRossa GN, Sheline YI, Dence CS, Lee SY, Mach RH, Klunk WE, Mathis CA, DeKosky ST, Morris JC. [^{11}C] PIB in a nondemented population: Potential antecedent marker of Alzheimer disease. Neurology. 2006; 67:446–452. [PubMed: 16894106]

44. Jagust WJ, Reed B, Mungas D, Ellis W, Decarli C. What does fluorodeoxyglucose PET imaging add to a clinical diagnosis of dementia? Neurology. 2007; 69:871–877. [PubMed: 17724289]

45. Storandt M, Balota DA, Aschenbrenner AJ, Morris JC. Clinical and psychological characteristics of the initial cohort of the Dominantly Inherited Alzheimer Network (DIAN). Neuropsychology. 2014 Jan; 28(1):19–29. [PubMed: 24219606]

46. Salthouse TA, Mitchell DR, Skovronek E, Babcock RL. Effects of adult age and working memory on reasoning and spatial abilities. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1989; 15:507–516.

47. Naveh-Benjamin M. Adult age differences in memory performance: Tests of an associative deficit hypothesis. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2000; 26:1170–1187.

48. Mack WJ, Freed DM, Williams BW, Henderson VW. Boston Naming Test: Shortened versions for use in Alzheimer's disease. Journal of Gerontology: Psychological Sciences. 1992; 45:P154–P158.

49. Ryman DC, Acosta-Baena N, Aisen PS, et al. Symptom onset in autosomal dominant Alzheimer disease: A systematic review and meta-analysis. Neurology. 2014; 83:1–8.

50. Littell, R., Milliken, GA., Stroup, W., et al. SAS SYSTEM FOR MIXED MODELS. Cary NC: SAS Institute Inc; 1996.

Table 1

True and estimated weights (averaged over 100 simulation samples) to combine three markers for the maximum ratio of mean and SD (efficacy comparison assumed at the end of trial i.e., time=3, N=sample size for both training and validation data sets ρ=0.2, r=0.2)

N	True weights	Estimated weights	Bias (MSE) on weights	True Mean/SD	Estimated Mean/SD
40	0.4412 0.5079 0.5787	0.4502 0.5246 0.5739	0.0070 0.0136	-1.6804	-1.6476
60	0.4412 0.5079 0.5787	0.4756 0.5206 0.5960	0.0215 0.0106	-1.6804	-1.6973
80	0.4412 0.5079 0.5787	0.4644 0.5266 0.6079	0.0237 0.0084	-1.6804	-1.6514
100	0.4412 0.5079 0.5787	0.4499 0.5228 0.6023	0.0158 0.0057	-1.6804	-1.6964
120	0.4412 0.5079 0.5787	0.4464 0.5145 0.5800	0.0044 0.0045	-1.6804	-1.6802
140	0.4412 0.5079 0.5787	0.4422 0.5286 0.5965	0.0132 0.0036	-1.6804	-1.6711

Table 2

Total sample sizes (averaged over 100 simulated pairs of data sets) required for 80% power to detect specified effect sizes in a 1:1 two-arm clinical trial ($\rho = r = 0.5$). weights estimated from training data set with 40 subjects; mean and SD on the combined outcome estimated from either the same data set or another independent sample of the same size. Efficacy comparison assumed at time=3. N_i =total sample size using i-th efficacy outcome alone; N =total sample size using the true weight for the combined outcome; NP =total sample size using the combined outcome with weights estimated from training data sets and mean/SD estimated from validation data sets; NS = total sample size using the combined outcome with weights and SD estimated from the same data sets. ES=effect size as a percentage of improvement from the estimated mean for the placebo at t=3)

ES	N_1	N_2	N_3	N	NP	NS	$NP - NS$
20%	837	711	564	370	384	342	42
25%	536	456	362	237	246	220	27
30%	373	317	251	165	171	153	19
35%	274	233	185	121	126	112	14
40%	210	179	142	93	97	86	11
45%	166	141	112	74	77	68	9
50%	135	115	91	60	62	56	7

Table 3
Baseline and longitudinal characteristics of the DIAN sample (N=53)

Age, mean (SD)	42.2 (8.6)
Female, N (%)	31 (58.5)
Education, years, mean (SD)	14.2 (2.3)
Estimated years to symptom onset mean (SD)	-2.8 (5.3)
APOE genotype, N (%)	
23	3 (5.7)
24	2 (3.8)
33	31 (58.4)
34	15 (28.3)
44	2 (3.8)
MMSE: Baseline mean (SD),	28.0 (2.3)
Scale range	0 - 30
MMSE: 3-year change from baseline mean (SD)	-4.5 (4.9)
WORDDEL: Baseline mean (SD)	1.9 (2.0)
Scale range	0 - 16
WORDDEL: 3-year change from baseline mean (SD)	-0.8 (1.1)
PAPER: Baseline mean (SD)	5.9 (2.8)
Scale range	0 - 12
PAPER: 3-year change from baseline mean (SD)	-1.1 (1.5)
WAIS DIGSYM: Baseline mean (SD)	53.3 (15.2)
Scale range	0 - 93
WAIS DIGSYM: 3 year change from baseline mean (SD)	-4.0 (11.3)
PAIRBINDING: Baseline mean (SD)	9.8 (2.7)
Scale range	0 - 12
PAIRBINDING: 3 year change from baseline mean (SD)	-0.5 (3.3)
BOSTON: Baseline mean (SD)	26.4 (3.2)
Scale range	0 - 30
BOSTON: 3 year change from baseline mean (SD)	-1.5 (5.2)

Table 4

Total sample sizes required for 80% power to detect specified effect sizes (ES) at Year3 in each individual marker from the DIAN data.

ES	MMSE	WORDDEL	PAPER	DIGSYM	PAIRBINDING	BOSTON
20%	909	1329	1521	1630	20040	22858
25%	582	851	973	1044	12826	14630
30%	404	591	676	725	8907	10160
35%	297	434	497	533	6544	7464
40%	228	333	381	408	5010	5715
45%	180	263	301	322	3959	4516
50%	146	213	244	261	3207	3658

Table 5

Weights and total sample sizes required for 80% power to detect specified effect sizes (ES) at Year 3 in each combined marker composed of MMSE (individual marker associated with smallest sample sizes) and each remaining individual marker from the DIAN data.

	MMSE PAPER	**MMSE PAIRBINDING**	**MMSE WORDDEL**	**MMSE DIGSYM**	**MMSE BOSTON**
Estimated weights	0.7283 0.2692	0.5397 0.1218	0.9507 0.4532	0.7462 0.1803	0.4992 0.0763
ES					
20%	536	576	583	1133	1137
25%	343	369	373	725	728
30%	239	256	259	504	505
35%	175	188	191	370	371
40%	134	144	146	284	285
45%	106	114	116	224	225
50%	86	93	94	182	182

Table 6

Weights and total sample sizes required for 80% power to detect specified effect sizes (ES) at Year 3 in each combined marker composed of MMSE & PAPER (composite of two markers associated with smallest sample sizes) and each remaining individual marker from the DIAN data.

	MMSE PAPER WORDDEL	**MMSE PAPER PAIRBINDING**	**MMSE PAPER DIGSYM**	**MMSE PAPER BOSTON**
Estimated weights	2.6552 0.9815 2.5862	0.6648 0.2458 0.2295	1.0049 0.3714 0.6970	0.3315 0.1225 0.1739
ES				
20%	218	532	750	764
25%	140	340	480	489
30%	97	237	334	340
35%	72	174	245	250
40%	55	133	188	191
45%	43	105	149	151
50%	35	85	120	123