Alzheimer's & Dementia: Translational Research & Clinical Interventions 3 (2017) 360-366

Alzheimer's & Dementia

Featured Article

Power analysis to detect treatment effects in longitudinal clinical trials for Alzheimer's disease

Zhiyue Huang[a,*], Graciela Muniz-Terrera[b], Brian D. M. Tom[a], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a]MRC Biostatistics Unit, University of Cambridge, UK
[b]Centre for Dementia Prevention, University of Edinburgh, UK

Abstract

Introduction: Assessing cognitive and functional changes at the early stage of Alzheimer's disease (AD) and detecting treatment effects in clinical trials for early AD are challenging.
Methods: Under the assumption that transformed versions of the Mini–Mental State Examination, the Clinical Dementia Rating Scale–Sum of Boxes, and the Alzheimer's Disease Assessment Scale–Cognitive Subscale tests'/components' scores are from a multivariate linear mixed-effects model, we calculated the sample sizes required to detect treatment effects on the annual rates of change in these three components in clinical trials for participants with mild cognitive impairment.
Results: Our results suggest that a large number of participants would be required to detect a clinically meaningful treatment effect in a population with preclinical or prodromal Alzheimer's disease. We found that the transformed Mini–Mental State Examination is more sensitive for detecting treatment effects in early AD than the transformed Clinical Dementia Rating Scale–Sum of Boxes and Alzheimer's Disease Assessment Scale–Cognitive Subscale. The use of optimal weights to construct powerful test statistics or sensitive composite scores/endpoints can reduce the required sample sizes needed for clinical trials.
Conclusion: Consideration of the multivariate/joint distribution of components' scores rather than the distribution of a single composite score when designing clinical trials can lead to an increase in power and reduced sample sizes for detecting treatment effects in clinical trials for early AD.

Keywords: Power analysis; Clinical trial; Sample size; Multivariate linear mixed-effects model; Composite score; Alzheimer's disease

1. Introduction

Much effort has been devoted to developing disease-modifying treatments that intervene in the pathobiologic processes involved in the early stage of Alzheimer's disease (AD). Any therapy that is effective at treating this early manifestation of the dementia process may provide an opportunity for managing the disease while patient function is relatively preserved [1]. Standard instruments used to quantify cognitive and functional decline in AD are relatively insensitive to the changes at early AD [2]. This raises challenges for assessing the early changes in cognition and function across the spectrum of AD [3] and makes detecting treatment effects in clinical trials for early AD even harder [2].

Power analysis is standard when designing clinical trials for detecting treatment effects. Ard *et al.* [4] provide a comprehensive review for clinical trials in AD. Misalignment of the power analysis can lead to possible errors in

[1]Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_ to_ apply/ADNI_ Acknowledgement_ List.pdf.

*Corresponding author. Tel.: +44 (0)1223 330300; Fax: +44 (0)1223 330365.

E-mail address: robin.huang@mrc-bsu.cam.ac.uk

http://dx.doi.org/10.1016/j.trci.2017.04.007

decisions regarding sample size. Too large samples may waste time, resources, and money and may unnecessarily expose some participants to inferior treatment if a treatment could have been shown to be more effective with fewer participants. Significant underestimation of the sample size may be a waste of time as it would unlikely lead to conclusive findings and therefore be unfair to all participants taking part in the trial. In this article, we are interested in the power/sample size to detect the treatment effects on the component scores in clinical trials for early AD.

In the literature of early AD, many researchers have used composite scores as single endpoints for performing power analysis [4]. A composite score is typically a linear combination of the scores of sensitive instruments. It provides a univariate summary of the component scores, avoids the multiple-hypothesis testing problem when each component score is considered separately, and reduces the impact of measurement error [5]. Furthermore, it may be more sensitive to the cognitive and functional decline than its separate components [6].

The construction of a composite score involves the selection and weighting of the component scores. Typically, the selection of the component scores may be based on a broad literature review regarding sensitivity to decline of candidate components [7], with equal weighting tending to be applied, possibly naively, to the chosen components. However, more statistically driven approaches can be used to derive the weights to construct more sensitive composite scores [2,6,8–12].

We therefore classify the statistical strategies used for the construction of a composite score into two major classes. The first is focused principally on selecting the most informative composite components and using prespecified weights not derived from statistical considerations; for example, Raghavan *et al.* [8] identify the informative component instruments based on standardized mean of 2-year change from baseline for a mild cognitive impairment (MCI) cohort and summed them to create a new composite score. The other is focused on "optimizing" the weights assigned to component scores based on an appropriate optimality criterion and is therefore more data driven; for example, some previous proposals find composite weights, which are sensitive to the clinical decline, by fitting linear mixed-effect models (LMMs) to the longitudinal composite scores [2,6,9]. Xiong *et al.* [6] propose composite weights that maximize the probability of observing a decline in one participant over a unit interval of time. Their weights can be considered as a special case of the composite weights proposed by Ard *et al.*, who use the power to detect the time effect in a clinical trial as their criterion and obtain the component weights by maximizing this criterion [2]. Ard *et al.*'s approach is applied to construct a composite atrophy index [9]. Another approach within this class is to base the estimation of the composite weights on a criterion that looks at the mean to standard deviation ratio of change over time [10,11]. Wang *et al.* [12] propose another composite score construct by using a linear clinical decline equation to select and reweight the component scores simultaneously.

In general, using composite scores as single endpoints may lose information to detect the changes in components [3]; for example, a large change in one component can be masked by small changes on other component scores. Data-driven composite scores have been further criticized [7]. Firstly, they may lose clinical interpretation. It is possible that a clinically meaningful component score has small weights in a data-driven composite score [7]. In addition, they may not be consistent across different data sets. Donohue *et al.* [7] apply cross-validation to quantify the out-of-sample performance of optimal composite scores and conclude that the overall performance of the optimal composite scores is worse than those composite scores derived without optimization.

A limited amount of the literature in AD has considered power analysis with multiple endpoints, although multiple endpoints are commonplace in AD. Under the assumption that the component scores are jointly from a multivariate linear mixed-effects model (MLMM), we compare three approaches with regard to their power to detect the treatment effects on component scores. Two of them are with multiple endpoints, whereas the other is with a single-composite endpoint.

2. Methods

2.1. MLMM for component scores

Mixed-effect models are from a class of useful statistical models for analyzing longitudinal data [13]. They allow a subset of the regression parameters (random effects) to vary randomly between participants and thereby characterize the natural heterogeneity in the target population in these parameters. Fixed effects are used to refer to the regression parameters, which are fixed but unknown and need to be estimated.

Assuming that all possible covariates are balanced (as would be assumed in a clinical trial through randomization), we model the component scores using an MLMM with a random intercept, fixed time, and time by treatment interaction effects. (The addition of further covariates can be easily incorporated if deemed necessary.) Such a model is able to simultaneously characterize the correlations between the component scores at each time t and the correlations across time for each component score.

Let Y_{ntj} be the j-th component score of the n-th participant at visit time t, where $n = 1,\ldots,N$, $t = 1,\ldots,T_n$, and $j = 1,\ldots,J$. Here, the number of visits T_n is a positive integer depending on the n-th participant, and the number of component scores J is prespecified. We use a linear function to link the component scores with the mixed effects

$$Y_{ntj} = \beta_{j0} + \gamma_j \times (\text{Treatment} \times \text{Time}) + \beta_{j2} \times \text{Time} + b_{nj} + \varepsilon_{ntj},$$

where γ_j is the j-th component treatment effect, b_{nj} is the random intercept that is unique to the j-th component score

of the n-th participant, and ε_{ntj} is the random error of the n-th participant on the j-th component score at time t. For each n, let $b_n = (b_{n1},\ldots,b_{nJ})^T$ independently follow a multivariate normal distribution with a mean vector 0 and a covariance matrix $\sum_b$. Here, for any matrix or vector A, the matrix A^{T} is the transpose of A. For each n and t, further let $\varepsilon_{nt}=(\varepsilon_{nt1},\ldots,\varepsilon_{ntJ})^T$ independently follow a multivariate normal distribution with the mean vector 0 and the covariance matrix $\sum_\varepsilon$. For each n and t, the error ε_{nt} and the random effects b_n are independent.

For each participant n and time t, the covariance matrix $\sum_\varepsilon$ characterizes the correlation structure between the component scores $Y_{nt1},\ldots,Y_{ntJ}$. For each participant n, the component scores $Y_{nt} = (Y_{nt1},\ldots,Y_{ntJ})^T$, $t = 1,\ldots,T_n$, are independent of each other through time conditional on the random effect b_n, but would be correlated marginally.

We can link the LMM for the composite scores to the MLMM for the components by letting $C_{nt}=\sum_{j=1}^J w_j Y_{ntj}$, $\alpha_0=\sum_{j=1}^J w_j\beta_{j0}$, $\gamma_w=\sum_{j=1}^J w_j\gamma_j$, $\alpha_2=\sum_{j=1}^J w_j\beta_{j2}$, $a_n=\sum_{j=1}^J w_j b_{nj}$, and $\delta_{nt}=\sum_{j=1}^J w_j\varepsilon_{ntj}$, where $w = (w_1,\ldots,w_J)^T$ is the vector of weights for the composite score [2]. The LMM for the composite score of the n-th participant at time t is therefore

$$C_{nt}=\alpha_0+\gamma_w\times(\text{Treatment}\times\text{Time})+\alpha_2\times\text{Time}+a_n+\delta_{nt},$$

where γ_w is the treatment effect on composite scores, and for each n, the random intercept, a_n, follows a normal distribution with mean 0 and variance $\sigma_a^2=w^T\sum_b w$, and for each n and t, the random error, δ_{nt}, follows a normal distribution with mean 0 and variance $\sigma_\delta^2=w^T\sum_\varepsilon w$.

2.2. Power analysis–hypothesis testing formulations

To detect the treatment effects on component scores, we consider three-hypothesis testing problems and their associated test statistics. Rejecting any of the null hypotheses suggests statistically significant component treatment effects.

The first hypothesis testing problem is to test the null hypothesis of no treatment effect in any of the components against the alternative that there is at least one non-zero treatment effect:

$$H_0:\gamma=0 \quad \text{vs} \quad H_A:\gamma\neq 0,$$

where $\gamma=(\gamma_1,\ldots,\gamma_J)^{\mathrm{T}}$ is the J-dimensional vector of treatment effects. The Wald statistic $\Xi_J=\widehat{\gamma}^{\mathrm{T}}\sum_\gamma^{-1}\widehat{\gamma}$ can be used, where $\widehat{\gamma}$ is the maximum likelihood estimator (MLE) of γ under the assumption of known covariance matrices for b_n and ε_{nt}, and $\sum_\gamma$ is the covariance matrix of $\widehat{\gamma}$. It follows that under the null hypothesis of no treatment effect for any of the components that the Wald test statistic will be distributed as a χ^2 distribution with J degrees of freedom, χ_J^2.

The second hypothesis testing problem considered is for the composite treatment effect, defined as a linear combination of the component treatment effects induced by the weights $w = (w_1,\ldots,w_J)^T$. Here, we test the null hypothesis of no composite treatment effect versus the alternative of a composite treatment effect. That is,

$$H'_0:\sum_{j=1}^J w_j\gamma_j=0 \quad \text{vs} \quad H'_A:\sum_{j=1}^J w_j\gamma_j\neq 0.$$

The Wald statistic, here, is $\Xi_{JC}(w)=(w^{\mathrm{T}}\sum_\gamma w)^{-1}(w^{\mathrm{T}}\widehat{\gamma})^2$, which is distributed as χ_1^2 under the null, H'_0.

The last hypothesis testing problem considers the case in which composite scores are used as single endpoints. It aims to test a single treatment effect on the composite scores

$$H''_0:\gamma_w=0 \quad \text{vs} \quad H''_A:\gamma_w\neq 0.$$

Given the variances σ_a^2 and σ_δ^2, let $\widehat{\gamma}_w$ be the MLE of γ_w and σ_γ^2 be its variance. We can use the Wald statistic $\Xi_C(w)=\sigma_\gamma^{-2}\widehat{\gamma}_w^2$, which follows the χ_1^2 distribution under H''_0, to test for this type of treatment effect.

The vector of weights w has different meanings under the last two hypotheses testing situations. The weights w are on the component treatment effects in the second, whereas the weights w reweight the component scores in the third. These testing approaches are equivalent only in the very special case of a linear link function, as is assumed in our setting.

Table 1 summarizes these three-hypothesis testing problem formulations. Under an alternative model, all the test statistics follow a noncentral χ^2 distribution and thereby determine the power to reject the associated null hypothesis. However, using less powerful test statistics will lead to larger sample sizes, which may be judged unethical. In the Supplementary document, we prove that for any given weights w, the test statistic $\Xi_{JC}(w)$ is no worse with regards to power than $\Xi_C(w)$. The test statistic Ξ_J does not uniformly outperform either $\Xi_{JC}(w)$ or $\Xi_C(w)$ over the range of w.

2.3. Power analysis–deriving the parameters required from analysis of MCI participants in Alzheimer's Disease Neuroimaging Initiative

For illustration, we conduct a power analysis for a two-arm randomized AD clinical trial with equal allocation probabilities. The component scores consist of the Mini–Mental

Table 1
Summary of the three hypothesis testing formulations to detect treatment effects

	Endpoints		
	Multivariate	Multivariate	Single composite
Statistical model	MLMM	MLMM	LMM
Null hypothesis	$\gamma=0$	$\sum_{j=1}^J w_j\gamma_j=0$	$\gamma_w=0$
Clinical interpretation	Component treatment effects	Composite treatment effect	Treatment effect on composite scores
Test statistic	Ξ_J	$\Xi_{JC}(w)$	$\Xi_C(w)$
Null distribution	χ_J^2	χ_1^2	χ_1^2

State Examination (MMSE), the Clinical Dementia Rating Scale-Sum of Boxes (CDR-SB), and the Alzheimer's Disease Assessment Scale–Cognition Subscale (ADAS-11) scores. We use data extracted from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.ucla.ca) to inform the specification of the various parameters required to perform the power analysis. This data set comprises 927 participants who are at MCI at baseline. The MMSE, the CDR-SB, and the ADAS-11 are recorded biannually for each participant over a total follow-up period of 10 years. To more closely satisfy the normality assumptions for the components in light of potential ceiling effects, we apply the Box-Cox transformation to the data and then rescaled them by their baseline standard deviation; see Supplementary Materials for details. The transformations applied are such that higher values of the transformed components indicate worse cognitive functioning.

We fit the MLMM to the three component scores; see the Supplementary Materials for details on how estimates of the rate of change parameters and the appropriate covariance structures necessary for us to perform the power analysis were obtained. The R function *mlmmm.em()* from the *mlmmm* package [14] was used to compute these estimates. The estimated annual rates of change on the transformed MMSE, the transformed CDR-SB, and the transformed ADAS-11 are 0.079 (95% confidence interval [CI]: 0.064, 0.095), 0.061 (95% CI: 0.045, 0.077), and 0.055 (95% CI: 0.040, 0.069), respectively. These annual rates of change correspond to small rates of change on the original untransformed scale and suggest that there is limited cognitive decline in those with MCI over the follow-up period. The estimated covariance matrices are

$$\widehat{\Sigma}_\varepsilon = \begin{bmatrix} 0.56 & 0.07 & 0.09 \\ 0.07 & 0.57 & 0.06 \\ 0.09 & 0.06 & 0.44 \end{bmatrix} \text{ and } \widehat{\Sigma}_b = \begin{bmatrix} 0.58 & 0.30 & 0.48 \\ 0.30 & 0.71 & 0.37 \\ 0.48 & 0.37 & 0.77 \end{bmatrix}.$$

We consider various designs for our clinical trial based on choosing different follow-up periods (i.e., 2, 3, 4, 5, and 6 years) and assuming that it is of interest to detect minimally clinically meaningful treatment effects corresponding to 25% reductions in the annual rates of change in the MMSE, CDR-SB, and ADAS-11 (transformed). These 25% reductions here also correspond approximately to 25% improvements in the treated versus control arms, if the components were considered on their original scales of measurement.

2.4. Power analysis–specifying the weights

We compare various weights for $\Xi_{JC}(w)$ and $\Xi_C(w)$ (optimal or otherwise) that can be used when performing a power analysis for the clinical trial designs mentioned in the early subsection. All the considered weight vectors are normalized by $\sum_{j=1}^{J} w_j^2 = 1$. The following weighting strategies are considered:

1 The equal weights vector $w_Z = (3^{-1/2}, 3^{-1/2}, 3^{-1/2})^{\mathrm{T}}$ assumes that the component treatment effects are equally important or that the treatment effect on the average of the component scores is of interest. Typically this strategy may be adopted in practice and therefore provides a benchmark to compare the other weighting strategies.
2 The unit vectors $w_{(1)} = (1,0,0)^{\mathrm{T}}$, $w_{(2)} = (0,1,0)^{\mathrm{T}}$, and $w_{(3)} = (0,0,1)^{\mathrm{T}}$ consider the situations in which either only one of the component treatment effects or the treatment effect on a single component is of interest.
3 The optimal weights vector for $\Xi_{JC}(w)$, denoted by w^*_{JC}, is optimal in the sense that $\Xi_{JC}(w^*_{JC})$ has the greatest power to reject H'_0 under a given alternative. In the Supplementary Materials, it is proven that $\Xi_{JC}(w^*_{JC})$ is always more powerful than Ξ_J in rejecting the associated null hypothesis given same conditions. The optimal weights w^*_{JC} are the eigenvector associated with the largest eigenvalue of $\sum_\gamma^{-1}\gamma^*\gamma^{*\mathrm{T}}$, which is proportional to $\sum_\gamma^{-1}\gamma^*$, where γ^* is the treatment effect vector under the alternative. In Table 2, we list the optimal weights for $\Xi_{JC}(w)$ for the different trial duration scenarios.
4 The optimal weights vector for $\Xi_C(w)$, denoted by w^*_C, maximizes the power of $\Xi_C(w)$ to detect the treatment effects under a given alternative over all possible normalized w; see the Supplementary document for the algorithm to calculate w^*_C. The composite score induced by w^*_C is the most sensitive for detecting a treatment effect on the composite score. The optimal weights w^*_C for different trial scenarios are listed in Table 2.

3. Results

Table 3 presents the sample sizes required for each of the aforementioned weighting specifications and under the different trial duration scenarios when the statistical power is specified at 80% and the significance level is set at 5%. Also reported are the calculated sample sizes when each component is considered separately for powering the trial, and a Bonferroni correction is applied. Here, the maximum of the three calculated sample sizes based on the three components is chosen as the sample size to be specified for the trial.

Table 2
The optimal weights for $\Xi_{JC}(w)$ and $\Xi_C(w)$ in each trial duration

		Trial duration				
Weights	Component	2 years	3 years	4 years	5 years	6 years
w^*_{JC}	MMSE	0.7670	0.7641	0.7576	0.7511	0.7451
	CDR-SB	0.4961	0.4958	0.4964	0.4971	0.4978
	ADAS-11	0.4069	0.4128	0.4238	0.4344	0.4438
w^*_C	MMSE	0.7151	0.7104	0.7061	0.7026	0.6999
	CDR-SB	0.5052	0.5050	0.5048	0.5046	0.5044
	ADAS-11	0.4832	0.4902	0.4966	0.5017	0.5057

From the table, we observe that the test statistic $\Xi_{JC}(w^*_{JC})$ gives the smallest sample sizes (numbers highlighted in bold) for each of the clinical trial design scenarios considered. Moreover, we make the following points after examining Table 3.

A substantial number of participants may be required when a trial for early AD only lasts for 2 years, under our assumptions. We estimate that at least 17,000 participants would need to be recruited in a 2-year AD trial in an MCI population to have sufficient power (i.e., 80%) to detect a 25% reduction in the annual rate of change on each of the transformed component scores. Recruitment of such numbers may be infeasible for a 2-year duration clinical trial in early AD with four biannual follow-up visits and even if feasible failure rates could potentially be high for early AD populations. Note that the required sample sizes will decrease with increasing trial duration, assuming biannual visits.

The required sample sizes to detect the treatment effect on the transformed MMSE are much smaller than the ones to detect the treatment effect on the transformed CDR-SB or ADAS-11 (comparing $w_{(1)}$ rows to $w_{(2)}$ and $w_{(3)}$ rows in Table 3). Let us consider a clinical trial of 3 years duration as an example. The required sample sizes obtained by $\Xi_{JC}(w_{(1)})$ is 55.0% of the ones obtained by $\Xi_{JC}(w_{(2)})$ and 54.6% of the ones obtained by $\Xi_{JC}(w_{(3)})$. This implies that the transformed MMSE is the more sensitive measure for detecting a treatment effect for early AD than transformed CDR-SB and the ADAS-11 measures [15-17].

The approaches that use the optimal weights could require at least 60% fewer participants than the ones using $w_{(2)}$ or $w_{(3)}$. In our analysis, the performances of $\Xi_{JC}(w)$ and $\Xi_C(w)$ with w_Z are comparable to the ones using the optimal weights. This is a consequence of the estimated parameters obtained from the analysis of the ADNI data giving rise to optimal weights that are close to w_Z (Table 2). Comparable performances across these three statistics will not in general be expected when using other component outcomes.

The sample sizes calculated under $\Xi_{JC}(w)$ are always smaller than the ones calculated under $\Xi_C(w)$ for fixed weights, although the reduction may not be significant; for example, there is a 3% reduction in sample sizes when $\Xi_{JC}(w)$ is used with $w=w^*_{JC}$. Such gain in efficiency is obtained by specifying the correlation structure among the component scores in the MLMM.

Table 3
The sample sizes calculated by each approach with 80% statistical power and 5% significance level by trial duration

Test statistic	Weights	Trial duration				
		2 years	3 years	4 years	5 years	6 years
Ξ_J	-	23,714	7041	2983	1550	908
$\Xi_{JC}(w)$	$w_{(1)}$	24,934	7447	3192	1678	994
	$w_{(2)}$	45,259	13,548	5789	3030	1786
	$w_{(3)}$	45,844	13,635	5789	3014	1769
	w_Z	17,672	5242	2216	1149	672
	w^*_{JC}	**17,072**	**5069**	**2148**	**1116**	**654**
	w^*_C	17,139	5090	2156	1120	656
$\Xi_C(w)$	$w_{(1)}$	26,851	8059	3451	1809	1067
	$w_{(2)}$	46,524	13,929	5943	3105	1827
	$w_{(3)}$	47,654	14,189	6017	3126	1831
	w_Z	17,881	5306	2242	1162	679
	w^*_{JC}	17,625	5236	2214	1147	671
	w^*_C	17,549	5212	2205	1143	669
		2 years	3 years	4 years	5 years	6 years
Bonferroni correction		63,563	18,926	8025	4170	2443

NOTE. Numbers given in bold indicates the test statistic $\Xi_{JC}(w^*_{JC})$ that gives the smallest sample sizes for each of the considered clinical trial design scenarios.

4. Discussion

We have described three approaches for performing power analysis to detect treatment effects in clinical trials for early AD. From our investigations, we found that jointly modeling the component scores and then constructing sensitive test statistics or composite scores based on optimal weights will improve the efficiency of clinical trials. Under our model assumptions, testing based on the optimal composite treatment effect will lead to the smallest required sample sizes and therefore should be recommended when powering clinical trials in AD if treatment effects on multiple components are of interest.

We end the article with the following discussion points.

4.1. Model assumptions

We assume that the component scores are jointly from an MLMM. This may be too strong an assumption for analyzing some cognitive and function scores in AD, because the component scores usually are discrete with strong ceiling or floor effects. Consider the CDR-SB as an example. The CDR-SB is the sum of six component scores, including the Memory Score, the Orientation Score, the Judgement and Problem Solving Score, the Community Affairs Score, the Home and Hobbies Score, and the Personal Care Score. The component scores except the Personal Care Score have the discrete range 0, 0.5, 1, 2, and 3, whereas the Personal Care Score has the range 0, 1, 2, and 3. From the ADNI data, over 30% of individuals have 0 in each component score of the CDR-SB, which would indicate strong floor effects (zero-heavy data). Therefore, it may not be appropriate to use an MLMM with CDR-SB on its original scale or even after transformation as done in this article. The use of other models, which take account of zero-heavy data may be appropriate; see Farewell et al. [18] for a comprehensive review.

In our power analysis results, we took the covariance matrices of ε_{nt} and b_n to be known when fitting the MLMM. This allowed us to obtain explicit formulas for the MLEs and their covariance, which enabled us to compare the powers of the test statistics and calculate the optimal composite scores. In practice, these covariance matrices would need to be estimated. They may be obtained from

previous investigations or through a pilot study. However, note that without considering the variability in the estimated covariance matrices, there would be a tendency to underestimate the required sample sizes. Monte Carlo studies can be applied to obtain more accurate sample sizes [19]. However, these would require intensive computational work to compute the optimal weights.

In the MLMM for component scores, it is assumed that, for each n, the errors ε_{nt}, $t = 1,\ldots,T_n$, are independent across time. This implies that the time correlation of Y_{nt}, $t = 1,\ldots,T_n$, is induced only through the random intercepts b_n. This can be generalized so as to introduce the auto correlations between ε_{nt}, $t = 1,\ldots,T_n$. Such generalization would raise computational challenges, and a bespoke program would be needed. (We were unable to find a statistical software package that would allow us to fit this more generalized model).

4.2. Wald statistics

The considered Wald statistics are used to detect the component treatment effect, but they do not make distinction between beneficial effects and deleterious effects. However, because currently in early AD, there may be an expectation that any treatment brought forward for confirmatory testing in a phase III trial has undergone rigorous assessment at phase II to ensure that it does not confer harm, it may be of interest to investigate rejecting H_0 under the alternative that all the component treatment effects γ are non-negative. In this situation, the Wald statistic Ξ_J follows a mixture of χ^2_p distribution, $P = 0,\ldots,J$, where χ^2_0 distribution is the distribution with mass 1 at point 0. In general, it is challenging to calculate the weights that combine the χ^2_p distribution, $P = 0,\ldots,J$, [20].

When the weights w in $\Xi_{JC}(w)$ and $\Xi_C(w)$ are non-negative elementwise, we may modify the alternatives against H'_0 and H''_0 to

$$H'_A : \sum_{j=1}^{J} w_j\gamma_j > 0$$

and

$$H''_A : \gamma_w > 0,$$

respectively. We can use the Z-statistics, $\Xi^{1/2}_{JC}(w)$ and $\Xi^{1/2}_{C}(w)$, for the one-sided tests. They follow the standard normal distribution under their associated null hypothesis. However, the elements of the optimal weights w^*_{JC} and w^*_C may not always be non-negative.

4.3. Parameters necessary for powering clinical trials

It is crucial to obtain plausible values of the parameters needed for the power analysis, including the annual change rates, the covariance matrix of random effects, and the covariance matrix of errors. These parameter values can be informed from a pilot study or existing studies [21]. However, there always exists the concern whether the specified alternative truly represents the clinical trial target population effect of interest and how the variability of the alternatives will affect the calculated sample sizes, sensitivity analysis is recommended [4]. McEvoy et al. [22] compute 95% CIs on the sample sizes through bootstrapping. We also present the 95% bootstrap CIs for the calculated sample sizes in our Supplementary document.

The effect sizes must be determined based on rationale and justification from theory and clinical experiences [4]. When the effect sizes are set to be the percentages of the annual rate of change, they are approximately invariant to the transformation on the component scores if the term $\gamma_j \times (\text{Treatment} \times \text{Time}) + \beta_{j2} \times \text{Time}$ in the MLMM is around zero.

The derivation and use of optimal weights w^*_{JC} and w^*_C here were for the clinical purpose of powering a trial. We did not propose a new composite score to be used as an endpoint but constructed the most powerful test statistics with the optimal weights w^*_{JC} and the most sensitive composite score with the weights w^*_C to detect treatment effects. We further argued that no extra information or no further model assumption than what is typically needed is required to calculate them given the alternatives. Therefore, it is helpful to compute and use the optimal weights in power analysis. For other clinical purposes, the optimal weights w as defined and clinically meaningful weights may conflict. In such situations, we suggest modifying the criterion for determining the optimal weights to take account clinical meaningfulness.

Acknowledgments

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking EPAD grant agreement no. 115736 and MRC programme grant (MC_UP_1302/3).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda

Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.trci.2017.04.007.

RESEARCH IN CONTEXT

1. Systematic review: The authors reviewed the literature on constructing composite scores sensitive to the early changes in cognition and function and for detecting treatment effects in clinical trials for early AD. Under the assumption that the component scores are jointly from an MLMM, three approaches are compared with regard to their power to detect treatment effects. The authors calculate sample sizes based on these three approaches.
2. Interpretation: Jointly modeling the component scores and using data-driven optimal weights will improve the efficiency of clinical trials for early AD. Power analysis based on using the optimal composite treatment effect requires the smallest sample sizes.
3. Future directions: It is required to study more flexible statistical models and develop associated software to power a study for early AD.

References

[1] Morris JC. Mild cognitive impairment and preclinical Alzheimer's disease. Geriatrics 2005;Suppl:9–14.
[2] Ard MC, Raghavan N, Edland SD. Optimal composite scores for longitudinal clinical trials under the linear mixed effects model. Pharm Stat 2015;14:418–26.
[3] Snyder PJ, Kahle-Wrobleski K, Brannan S, Miller DS, Schindler RJ, DeSanti S, et al. Assessing cognition and function in Alzheimer's disease clinical trials: do we have the right tools? Alzheimers Dement 2014;10:853–60.
[4] Ard MC, Edland SD. Power calculations for clinical trials in Alzheimer's disease. J Alzheimers Dis 2011;26:369–77.
[5] Crane PK, Carle A, Gibbons LE, Insel P, Mackin RS, Gross A, et al. Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). Brain Imaging Behav 2012;6:502–16.
[6] Xiong C, Van Belle G, Chen K, Tian L, Luo J, Gao F, et al. Combining multiple markers to improve the longitudinal rate of progression: application to clinical trials on the early stage of Alzheimer's disease. Stat Biopharm Res 2013;5:54–66.
[7] Donohue MC, Sun CK, Raman R, Insel PS, Aisen PS. Cross-validation of optimized composites for preclinical Alzheimer. Alzheimers Dement (N Y) 2017;3:123–9.
[8] Raghavan N, Samtani MN, Farnum M, Yang E, Novak G, Grundman M, et al. Alzheimer's Disease Neuroimaging Initiative. the ADAS-Cog revisited: novel composite scales based on ADAS-Cog to improve efficiency in MCI and early AD trials. Alzheimers Dement 2013;9:S21–31.
[9] Edland SD, Ard MC, Sridhar J, Cobia D, Martersteck A, Mesulam MM, et al. Proof of concept demonstration of optimal composite MRI endpoints for clinical trials. Alzheimers Dement (N Y) 2016;2:177–81.
[10] Langbaum JB, Hendrix SB, Ayutyanont N, Chen K, Fleisher AS, Shah RC, et al. An empirically derived composite cognitive test score with improved power to track and evaluate treatments for preclinical Alzheimer's disease. Alzheimers Dement 2014;10:666–74.
[11] Ayutyanont N, Langbaum JB, Hendrix SB, Chen K, Fleisher AS, Friesenhahn M, et al. 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;75:652–60.
[12] Wang J, Logovinsky V, Hendrix SB, Stanworth SH, Perdomo C, Xu L, et al. ADCOMS: a composite clinical outcome for prodromal Alzheimer's disease trials. J Neurol Neurosurg Psychiatry 2016;87:993–9.
[13] Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. Hoboken, NJ: John Wiley & Sons; 2012.
[14] Yucel RM. R mlmmm Package: fitting Multivariate Linear Mixed Effects Models with Missing Values. Turkiye Klinikleri J Biostat 2015;7:11–24.
[15] Amieva H, Goff ML, Millet X, Orgogozo JM, Peres K, Barberger-Gateau P, et al. Prodromal Alzheimer's disease: successive emergence of the clinical symptoms. Ann Neurol 2008;64:492–8.
[16] Bateman RJ, Xiong C, Benzinger TL, Fagan AM, Goate A, Fox NC, et al. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. N Engl J Med 2012;367:795–804.
[17] Fleisher AS, Chen K, Quiroz YT, Jakimovich LJ, Gomez MG, Langois CM, et al. Associations between biomarkers and age in the presenilin 1 E280A autosomal dominant Alzheimer disease kindred: a cross-sectional study. JAMA Neurol 2015;72:316–24.
[18] Farewell VT, Long DL, Tom BD, Yiu S, Su L. Two-part and related regression models for longitudinal data. Anal Ref Stcd Its Apon 2016;72:316–24.
[19] Muthén LK, Muthén BO. How to use a Monte Carlo study to decide on sample size and determine power. Struct Equ Modeling 2002;9:599–620.
[20] Silvapulle MJ, Sen PK. Constrained Statistical Inference: Order, Inequality, and Shape Constraints. Hoboken, NJ: John Wiley & Sons; 2011.
[21] Edland SD, Ard MC, Li W, Jiang L. Design of pilot studies to inform the construction of composite outcome measures. Alzheimers Dement (N Y) 2017;3:213–8.
[22] McEvoy LK, Edland SD, Holland D, Hagler DJ Jr, Roddey JC, Fennema-Notestine C, et al. Neuroimaging enrichment strategy for secondary prevention trials in Alzheimer's disease. Alzheimer Dis Assoc Disord 2010;24:269–77.