Published in final edited form as:

J Alzheimers Dis. 2011 ; 26(Suppl 3): 369–377. doi:10.3233/JAD-2011-0062.

Power Calculations for Clinical Trials in Alzheimer's Disease

M. Colin Ard^a and Steven D. Edland^{a,b,*}

^aDepartment of Neuroscience, University of California, San Diego, La Jolla, CA, USA

^bDepartment of Family Preventive Medicine Division of Biostatistics, University of California, San Diego, La Jolla, CA, USA

Abstract

The Alzheimer research community is actively pursuing novel biomarker and other biologic measures to characterize disease progression or to use as outcome measures in clinical trials. One product of these efforts has been a large literature reporting power calculations and estimates of sample size for planning future clinical trials and cohort studies with longitudinal rate of change outcome measures. Sample size estimates reported in this literature vary greatly depending on a variety of factors, including the statistical methods and model assumptions used in their calculation. We review this literature and suggest standards for reporting power calculation results. Regardless of the statistical methods used, studies consistently find that volumetric neuroimaging measures of regions of interest, such as hippocampal volume, outperform global cognitive scales traditionally used in clinical treatment trials in terms of the number of subjects required to detect a fixed percentage slowing of the rate of change observed in demented and cognitively impaired populations. However, statistical methods, model assumptions, and parameter estimates used in power calculations are often not reported in sufficient detail to be of maximum utility. We review the factors that influence sample size estimates, and discuss outstanding issues relevant to planning longitudinal studies of Alzheimer's disease.

Keywords

Sample size; clinical trials; Alzheimer's disease; biostatistics

INTRODUCTION

There is increasing interest in the potential utility of biomarkers as outcomes in clinical trials. For example, the joint industry/NIH funded Alzheimer's Disease Neuroimaging Initiative (ADNI) was created expressly to investigate cerebrospinal fluid and volumetric neuroimaging measurements as diagnostic biomarkers of early Alzheimer's disease (AD) and as potential endpoints for monitoring clinical trial treatment effects [1]. ADNI has recruited and followed longitudinally approximately 200 AD cases, 400 mild cognitively impaired (MCI), and 200 age-matched cognitively normal controls [2]. Additional novel biomarker endpoints, including MR spectroscopy [3] and FDG-PET [4] have been proposed and are being actively pursued. The hope is that biomarkers will allow monitoring of treatment effects at earlier stages of disease, before traditional cognitive and functional endpoints are measurable. It is also becoming apparent that biomarker measurements, particularly volumetric neuroimaging measures, are substantially more precise than

traditional cognitive and functional measures, to the point that clinical trials using volumetric neuroimaging measures may be possible with a tenth or less of the sample size of current trials. Freely accessible ADNI data has provided a natural laboratory for exploring these issues. While this literature has consistently described the relative improvement in statistical power of imaging outcomes relative to cognitive outcomes, there is little consistency across reports in estimated sample size requirements for each particular outcome measure. To characterize these discrepancies, we have reviewed the ADNI power calculation publications with an eye to the influence of statistical methods on sample size projections.

LITERATURE REVIEW METHODS

Articles were identified based on a search of published papers listed on the ADNI website (adniinfo.org/Scientists/ADNIScientistsHome/ADNIPublications.aspx) as of February 4, 2011. All papers containing search terms “power” or “sample size” were reviewed for reported sample size calculations by one of the authors (MCA). Of 143 papers searched, 17 contained abstractable reports of previously unpublished analytic sample size calculations [5–21]. These papers report required sample size for a future clinical trial to observe a stated treatment effect, with the magnitude of the treatment effect described in terms of percentage slowing of disease progression relative to placebo. An additional six papers reported on sample sizes required for various analyses (e.g., detection of correlations, differences between dementia types, or the presence of atrophy) using a non-analytic sample-reduction method in which subjects were randomly discarded from the pilot data set until it was no longer possible to reject the hypotheses in question [22–27]. The remaining search hits were all due to papers reporting retrospective power calculations, papers that only reported relative gains in sample sizes, and papers that cited previously published estimates.

Among the 17 papers that presented prospective analytic calculations, most reported sample size required to detect a 25% reduction in the observed rate of change with 80% power. To facilitate comparisons across publications, sample size estimates based on a different percentage reduction were recalculated to a 25% reduction using the formula $n_{25} = (k/25)^2 n_k$, where k equals the percentage used in the original report. Sample size estimates for power other than 80% (typically 90%) were standardized to 80% using the formula $n_{0.8} = (z_{\alpha/2} + z_{0.2})^2 n_p / (z_{\alpha/2} + z_{1-p})^2$, where in this case the subscript p indicates the power of the trial expressed as a probability. The characterization of effect size as a fixed percent reduction in observed rate of change is useful for comparing power calculations across studies, but is not intended to serve as a model for research practice. In practice, effect sizes used in trial design should always be determined based on the plausibility and clinical significance of the hypothetical outcomes under consideration [28].

RESULTS

Tables 1 and 2 summarize reported sample size estimates for trials in AD and MCI. Table 1 summarizes estimates of sample size required to detect a 25% reduction in mean rate of decline on the standard cognitive outcome for AD treatment trials, the AD Assessment Scale - Cognitive Subscale (ADAS-Cog) [29]; Table 2 summarizes estimates of sample size required to detect a 25% reduction in atrophy rate for a likely MRI outcome, hippocampal volume. Reported sample size estimates for each measure are widely divergent. The differences in estimates may be explained by a number of factors, which we review in sequence below.

Trial design

Trial design, i.e. the length of the trial and the frequency of assessment, has a direct influence on statistical power. All other things being equal, longer trials, and to a lesser extent trials with finer assessment intervals, result in more precise estimates of rate of decline per arm and require fewer subjects to detect treatment effects. For example, Hua et al. [13] report a 6-fold decrease in required sample size for a 2-year compared to a 6-month long AD treatment trial using change on ADAS-Cog as the outcome variable (Table 1). Relatively noisy outcome measures, such as global cognitive scales represented here by the ADAS-Cog, experience more gain in precision and power by increased trial length or assessment frequency compared to relatively less noisy outcome measures such as volumetric imaging. For example, using change in hippocampal volume as the outcome measure, Hua et al. report only a 36% increase in sample size required for a 6-month compared to 2-year trial (Table 2). The influence of trial design on statistical power varies with different analysis plans. Within limits, longer trials and increased sampling frequency are associated with improved power for trials designed to detect changes in trajectory of disease under treatment, although there are diminishing returns with longer trials as dropout rates increase and linearity assumptions implicit in most statistical analysis plans become less tenable. Trials designed to detect acute, symptomatic treatment effects are unlikely to benefit from longer observation or increased frequency of sampling.

Magnitude of effect size

Effect size, the minimum treatment effect a trial is powered to detect, directly influences sample size requirements. For the power calculations reviewed here, the effect size is calculated as a percentage of the assumed mean rate of decline under the placebo condition. The various ADNI power calculation papers used different estimates of the placebo rate of decline in their calculations, and this explains to some degree the differences in required sample size reported. For example, for MCI treatment trials using the ADAS-Cog as the endpoint (Table 1), effects sizes used for power calculations range from 25% of 2.1 points per year [10] to 25% of 1.0 points per year [8, 14]. The sample size estimate in the former (1183 for a 1-year trial) is substantively smaller than the sample size estimates in the latter (4000+ and 2175 for a 1-year trial). Several papers did not report the effect size powered for [9, 13, 19], and we can only speculate on the extent to which differences in sample size projections reported in these papers are attributable to differences in assumed effect size. However, in general, when defining effect size as a percentage reduction in mean rate of decline, a smaller assumed mean rate of decline under the null hypothesis translates to smaller effects sizes powered for and larger required sample size.

Target population of planned clinical trial

A critical factor when setting the effect size for power calculations is the issue of defining the target population of the planned future clinical trial. For the most part the power calculations reviewed here used estimates of mean rate of decline within the ADNI cohorts as the assumed trajectory of disease under the placebo condition. The implicit assumption is that subjects recruited in future trials will look much like the subjects recruited into the ADNI cohort study, a reasonable assumption given that the ADNI recruitment network and methods parallel those used by many multicenter trials [5]. The differences in effect size (Tables 1 and 2) used by the various studies may follow in part from random variability in data obtained when the ADNI data were accessioned. Differences in effect size may also follow from differences in statistical methods used to calculate mean rate of decline, or differences in inclusion/exclusion criteria applied to the ADNI sample prior to estimation.

Regarding the effect of varying inclusion criteria, McEvoy et al. [17] describe the effect of inclusion criteria intended to enrich the study population for subjects more likely to have the

underlying neurodegenerative process that is the target of most planned therapies. For example, restricting recruitment to MCI subjects with baseline MRI atrophy patterns consistent with AD resulted in a cohort with mean trajectory of decline on the ADAS-Cog of 2.3 points per year compared to a mean decline of 1.5 points per year in the unrestricted cohort; sample size requirements correspondingly dropped by over one-half using this inclusion criterion, from 978 per arm to 458 per arm [17]. Similarly, restricting recruitment to subjects with the APOE $\epsilon 4$ risk allele increased the mean rate of ADAS-Cog decline to 1.7 points per year and reduced the required sample size to 774 persons per arm [17]. A limitation of trials with restrictive inclusion criteria is that findings only generalize to the subpopulation examined.

Statistical analysis plan and assumptions

Power calculations are specific to the analysis plan of the planned trial. Sample size formulas for two-group comparisons under normality assumptions are of the form:

$$n/arm = 2(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma_d^2 / \Delta^2 \quad (1)$$

where Δ is the treatment effect size under the alternative, σ_d^2 is the within group variance of the outcome measure being compared across treatments, and $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the usual quantiles of the standard normal distribution, with α equal to the type I error rate of a two-sided test, typically set to 0.05, and $(1 - \beta)$ equal to the power of the trial, typically set to 0.8 or 0.9. Treatment effect Δ and variance σ_d^2 are defined in terms of the outcome measure to be used in the planned trial. For example, for a trial with two observations per subject and outcome measure of change from baseline to followup, Δ is the change in the treatment group minus the change in placebo, and σ_d^2 is the variance of change scores (e.g., Meinart [30], equation 9.14). In this example σ_d^2 can be estimated as the variance of change from baseline to follow-up observed in two-wave pilot data of comparable duration to the planned trial. For a trial with multiple observations per subject and outcome measure of least squares slope of longitudinal trajectories, Δ is the difference in expected slopes in treatment versus placebo and σ_d^2 is the within arm variance of least squares slopes [31]. In this example σ_d^2 can be estimated from the variance of least squares slopes observed in pilot data of comparable design to the planned trial. These are examples of two-stage “summary measures” analyses which require only the assumption that summary measures (i.e., change scores or least squares slopes) are independent, identically distributed asymptotically normal random variables. Several of the ADNI power calculation papers (Tables 1 and 2) use summary measures power formulas, although the exact statistical analysis and model assumptions used were not always stated in complete detail.

Several of the power calculation papers used formally parameterized longitudinal models and analysis plans as the basis of their power calculations. For example, McEvoy et al. [17] based power calculations on a linear mixed effects model analysis assuming longitudinal trajectories of decline are linear within subject and that the distribution of slopes and intercepts describing these trajectories is bivariate normal. Sample size requirements given this assumed model have been derived [32]. For a balanced design (with all subjects observed at the same time points), the required sample size per arm is:

$$n/arm = 2(z_{1-\alpha/2} + z_{1-\beta})^2 \times (\sigma_b^2 + \sigma_e^2 / \sum (t_i - \bar{t})^2) / \Delta^2 \quad (2)$$

where σ_b^2 and σ_e^2 are parameters of the linear mixed effects model, and $\sum (t_i - \bar{t})^2$ is the “design term”, where t_i indexes the times at which measures are made and $\bar{t}$ is the mean of

the times. For example, for a 12 months trial with observations at baseline, month 6 and month 12, $\Sigma(t_i - \bar{t})^2$ in units of years equals $(0-0.5)^2 + (0.5-0.5)^2 + (1-0.5)^2 = 0.5$. Here Δ is the difference in mean rate of decline in treatment versus control, and the parameters σ_b^2 and σ_e^2 from the mixed effects model are the person to person variability in random slopes and the residual error variance of model [32]. σ_b^2 and σ_e^2 can be estimated by fitting a linear mixed effects model to pilot data representative of the trial's target population. For balanced design pilot data, estimates by formula (2) are algebraically identical to estimates by the power formula for a summary measures analysis comparing the mean of least squares slopes of treatment to the mean of least squares slopes of controls (e.g. [33]).

An alternative mixed effects model power formula is:

$$n/arm = 2(z_{1-\alpha/2} + z_{1-\beta})^2 (\sigma_e^2 / \sum (t_i - \bar{t})^2) / \Delta^2 \quad (3)$$

This formula is appropriate assuming a mixed effects model in which the subjects have random intercepts but identical rates of decline within arm (or equivalently, a marginal model with compound symmetric covariance structure [34]). Formula (3) results in smaller sample size projections, but can be anti-conservative when the common within arm rate of decline assumption does not hold. Formulas (1), (2), and (3) assume equal sample size per arm. Some trials use unequal allocation ratios to increase the likelihood of assignment to the active treatment arm and make the trial more attractive to study participants (e.g., [35]). Unequal allocation trials are slightly less efficient and require a modest adjustment in total sample size [36].

Several of the papers reviewed here reported sample sizes using formula (3), either in lieu of [19], or in addition to [11, 17], formula (2). Sample size estimates derived using the random-intercepts model and formula (3) were generally smaller than estimates using the mixed effects model with random intercepts and random slopes and formula (2). Taken together these observations underscore the importance of model selection when powering trials.

Differences in the image processing methods

For volumetric imaging outcomes in particular, sample size estimates can vary depending on the method of image analysis used. Image processing can be based on manual tracings [33], semi-automated methods and fully automated methods (e.g., [17]). Even though each of these methods is measuring the same structure, they may have different signal-to-noise properties depending on the relative precision of the methods. For example, Leung et al. [16], calculated hippocampal volume change by two different image processing methods and calculated samples size requirements for each outcome measure. While both methods led to sample size estimates that were considerably smaller than estimates typical of global cognitive measures like the ADAS-Cog, they found that required sample size for the more efficient image processing method was between 32–54% smaller than the less efficient method (see Table 2). Characterizing the relative performance of various imaging technologies and processing techniques [10, 12, 13, 15, 16, 21, 23, 24], will be an important outcome of the ADNI exercise.

DISCUSSION

The results above suggest that the wide divergence of sample size estimates calculated from ADNI data can be explained by multiple factors beyond differences in trial design and target population, including differences in power calculation algorithms used, and, for neuroimaging outcomes, differences in the signal-to-noise profile of the different image

processing algorithms. Additional factors relevant to power calculations for AD trials, and general recommendations for improved reporting of power calculations, are discussed below.

Sensitivity analysis

The validity of a power calculation is dependent in large part upon the accuracy of the (assumed known) parameter values used in its calculation. In practice these values are almost always calculated from pilot data, as is the case in the ADNI papers reviewed here, and hence contain some degree of random variability. The practical consequence of this randomness is potentially significant, especially when the pilot study used for parameter estimation is small. Several of the reviewed papers reported statistical tests or confidence intervals to characterize the variability inherent in sample size estimates ([10, 11, 13, 15–17, 21, 26, 27], see also [6]). For example, McEvoy et al. [17] used a bootstrap procedure to calculate 95% confidence intervals around sample size estimates based on ADNI data. They found that even with the relatively large ADNI pilot data set these confidence intervals can be large [17], demonstrating the importance of confidence interval calculation as a sensitivity analysis when powering trials.

Treatment target (disease-specific versus non-disease specific)

Some age-related cognitive decline and brain atrophy is experienced even within cognitively normal elderly. This is potentially relevant to the design of trials, as treatments that target the Alzheimer neurodegenerative process specifically may have no effect on non-Alzheimer related decline, and non-Alzheimer related decline may comprise a substantial fraction of the total decline that the sample size calculations described in Tables 1 and 2 are powered to detect. ADNI includes an age-matched, cognitively normal healthy control cohort from which the potential influence of non-treatment responsive age-associated decline can be estimated. We illustrate this with sample power calculations for hypothetical trials of MCI subjects powered to detect a 50% slowing of disease progression as measured by various neuroimaging measures (Tables 3 and 4, adapted from [32]).

Table 3 summarizes estimated sample size requirements assuming a treatment that is effective at slowing both disease-specific atrophy and non-disease-specific age-associated atrophy. Table 3 is analogous to estimates summarized in Table 2 except that effect size is set to 50% slowing of progression. Power calculations are by formula (2) with parameter estimation using longitudinal ADNI data [32]. Ventricular volume was the most efficient outcome measure under this scenario, requiring an estimated 83 subjects per arm to detect a difference in rate of atrophy equal to one half the rate of atrophy observed in the ADNI pilot data. Mid-temporal cortical thinning, whole brain atrophy, and right hippocampal atrophy were slightly less efficient as potential endpoints (Table 3).

Table 4 summarizes samples size requirements to detect a 50% slowing of disease-specific atrophy, where disease-specific atrophy is defined as the atrophy experienced by MCI subjects that is above and beyond the atrophy experienced by age-matched cognitively normal ADNI subjects, and the effect size Δ is calculated as 50% of the difference between the normal and MCI rate of atrophy. For trials powered to detect a slowing of disease-specific atrophy (Table 4), middle temporal cortical thinning was the most statistically efficient outcome measure, requiring 252 (left mid-temporal cortex) to 319 (right mid-temporal cortex) subjects per arm to detect a difference in rate of atrophy equal to one half the rate attributable specifically to the Alzheimer degenerative process. Ventricular volume, the most efficient outcome for detecting non-disease-specific atrophy (Table 3), was the least efficient volumetric outcome for detecting Alzheimer's disease-specific atrophy (Table 4).

Which sample size algorithm is most appropriate for a given trial? Table 3 is appropriate for treatments presumed to target both non-specific age-associated atrophy and Alzheimer's disease-associated atrophy. Table 4 is appropriate for treatments presumed to target only Alzheimer's disease-associated atrophy. Table 4 is conservative if you presume that some age-associated atrophy observed in cognitively intact elderly is due to a preclinical Alzheimer's disease neurodegenerative process, in which case sample sizes intermediate between Tables 3 and 4 would be sufficient. Further examples and discussion of this issue can be found in references [8, 11, 15–17, 21, 32].

Minimum reporting standards for power calculations

As noted above, a number of the ADNI publications did not report the magnitude of treatment effect being powered or did not explicitly state the statistical analysis plan upon which power calculations were based. We suggest that, as a minimum standard for reporting power calculation findings, these two items be reported. Estimates of minimum sample size requirements are of little utility to readers if the algorithm used for power calculations and the methods for calculating parameter estimates used in those calculations are not reported. Furthermore, if the power calculation formula and parameter estimates are published (e.g., McEvoy et al. [17]), then outside investigators can use this information to inform sample size calculations for alternative designs (e.g., longer trials or trials with greater sampling frequency) or alternative treatment effect sizes.

Additional considerations

Consideration of several additional issues can greatly improve the value of power calculation reports. Power calculation estimates are valid only if implicit model assumptions are true. Pilot data (e.g., ADNI data) can be used to test these implicit assumptions, and describing diagnostics to justify the proposed analysis plan and power calculation algorithm would greatly improve power calculation reports. As discussed above, parameters used in power calculations are estimated with some uncertainty, and a sensitivity analysis (reporting confidence intervals around sample size estimates) is also an important qualification of power calculation findings. Detailed descriptions of the cognitive and demographic characteristics of the pilot data increases the utility of power calculation reports as well. Covariate adjustment was not discussed in this review, but may be a means of improving the efficiency of clinical trials and deserves further consideration [9]. Finally, we have also not addressed pragmatic issues such as adjusting sample size calculations to accommodate study subject dropout or loss to follow-up [13], which may vary as a function of the research protocol requirements of the various measurement methods.

CONCLUSIONS

We emphasize that we have focused exclusively on statistical issues in comparing published ADNI power calculation papers. A number of issues beyond statistical considerations are critical to planning clinical trials. Not the least of these is establishing the relative feasibility and practical significance of a given percentage slowing of progression on cognitive versus proposed volumetric imaging measures. The current Food and Drug Administration standard for approving Alzheimer treatments is demonstrated effectiveness in slowing of cognitive and functional decline. The utility of neuroimaging outcomes, e.g., to demonstrate biological effect in phase 2 trials, or, ultimately, the acceptance of these biomarker measures as outcome measures for phase 3 trials, is yet to be established [37]. Nonetheless, the papers reviewed here consistently demonstrate the potential utility of these outcomes from the statistical efficiency perspective. The increased statistical efficiency translates to shorter trials with substantially smaller sample sizes, meaning more drugs could be effectively tested for the same cost in terms of dollars and human subject burden. Shorter, smaller trials

may also be amenable to adaptive trial designs, which would open new avenues for potential gain in trial efficiency.

Acknowledgments

Supported by NIH/NIA AG010483 (SDE), AG005131 (SDE, MCA), and AG034439 (SDE).

References

1. Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, Trojanowski JQ, Toga AW, Beckett L. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement*. 2005; 1:55–66. [PubMed: 17476317]
2. Weiner MW, Aisen PS, Jack CR Jr, Jagust WJ, Trojanowski JQ, Shaw L, Saykin AJ, Morris JC, Cairns N, Beckett LA, Toga A, Green R, Walter S, Soares H, Snyder P, Siemers E, Potter W, Cole PE, Schmidt M. The Alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimers Dement*. 2010; 6:202–211. [PubMed: 20451868]
3. Ashford, JW.; Adamson, M.; Beale, T.; La, D.; Hernandez, B.; Noda, A.; Rosen, A.; O'Hara, R.; Fairchild, JK.; Spielman, D.; Yesavage, JA. MR spectroscopy for assessment of memantine treatment in mild to moderate Alzheimer dementia. In: Ashford, JW., et al., editors. *Handbook of Imaging the Alzheimer Brain*. IOS Press; Amsterdam: 2011. p. 599-604.
4. Förster, S.; Buschert, VC.; Buchholz, HG.; Teipel, SJ.; Frieze, U.; Zach, Ca; Fougere, C.; Rominger, A.; Drzezga, A.; Hampel, H.; Bartenstein, P.; Buerger, K. Effects of a 6-month cognitive intervention program on brain metabolism in amnesic MCI and mild Alzheimer's disease. In: Ashford, JW., et al., editors. *Handbook of Imaging the Alzheimer Brain*. IOS Press; Amsterdam: 2011. p. 605-616.
5. Aisen PS, Petersen RC, Donohue MC, Gamst A, Raman R, Thomas RG, Walter S, Trojanowski JQ, Shaw LM, Beckett LA, Jack CR Jr, Jagust W, Toga AW, Saykin AJ, Morris JC, Green RC, Weiner MW. Clinical Core of the Alzheimer's disease neuroimaging initiative: progress and plans. *Alzheimers Dement*. 2010; 6:239–246. [PubMed: 20451872]
6. Beckett LA, Harvey DJ, Gamst A, Donohue M, Kornak J, Zhang H, Kuo JH. The Alzheimer's disease neuroimaging initiative: Annual change in biomarkers and clinical outcomes. *Alzheimers Dement*. 2010; 6:257–264. [PubMed: 20451874]
7. Carrillo MC, Sanders CA, Katz RG. Maximizing the Alzheimer's disease neuroimaging initiative II. *Alzheimers Dement*. 2009; 5:271–275. [PubMed: 19362888]
8. Chen K, Langbaum JB, Fleisher AS, Ayutyanont N, Reschke C, Lee W, Liu X, Bandy D, Alexander GE, Thompson PM, Foster NL, Harvey DJ, de Leon MJ, Koeppe RA, Jagust WJ, Weiner MW, Reiman EM. Twelve-month metabolic declines in probable Alzheimer's disease and amnesic mild cognitive impairment assessed using an empirically pre-defined statistical region-of-interest: findings from the Alzheimer's Disease Neuroimaging Initiative. *Neuroimage*. 2010; 51:654–664. [PubMed: 20202480]
9. Fleisher AS, Donohue M, Chen K, Brewer JB, Aisen PS. Applications of neuroimaging to disease-modification trials in Alzheimer's disease. *Behav Neurol*. 2009; 21:129–136. [PubMed: 19847051]
10. Ho AJ, Hua X, Lee S, Leow AD, Yanovsky I, Gutman B, Dinov ID, Lepore N, Stein JL, Toga AW, Jack CR Jr, Bernstein MA, Reiman EM, Harvey DJ, Kornak J, Schuff N, Alexander GE, Weiner MW, Thompson PM. Comparing 3 T and 1.5 T MRI for tracking Alzheimer's disease progression with tensor-based morphometry. *Hum Brain Mapp*. 2010; 31:499–514. [PubMed: 19780044]
11. Holland D, Brewer JB, Hagler DJ, Fennema-Notestine C, Dale AM. Subregional neuroanatomical change as a biomarker for Alzheimer's disease. *Proc Natl Acad Sci U S A*. 2009; 106:20954–20959. [PubMed: 19996185]
12. Hua X, Lee S, Yanovsky I, Leow AD, Chou YY, Ho AJ, Gutman B, Toga AW, Jack CR Jr, Bernstein MA, Reiman EM, Harvey DJ, Kornak J, Schuff N, Alexander GE, Weiner MW, Thompson PM. Optimizing power to track brain degeneration in Alzheimer's disease and mild cognitive impairment with tensor-based morphometry: an ADNI study of 515 subjects. *Neuroimage*. 2009; 48:668–681. [PubMed: 19615450]

13. Hua X, Lee S, Hibar DP, Yanovsky I, Leow AD, Toga AW, Jack CR Jr, Bernstein MA, Reiman EM, Harvey DJ, Kornak J, Schuff N, Alexander GE, Weiner MW, Thompson PM. Mapping Alzheimer's disease progression in 1309 MRI scans: power estimates for different inter-scan intervals. *Neuroimage*. 2010; 51:63–75. [PubMed: 20139010]
14. Landau SM, Harvey D, Madison CM, Koeppe RA, Reiman EM, Foster NL, Weiner MW, Jagust WJ. Associations between cognitive, functional, and FDG-PET measures of decline in AD and MCI. *Neurobiol Aging*. 2009;10.1016/j.neurobiolaging.2009.07.002
15. Leung KK, Clarkson MJ, Bartlett JW, Clegg S, Jack CR Jr, Weiner MW, Fox NC, Ourselin S. Robust atrophy rate measurement in Alzheimer's disease using multi-site serial MRI: tissue-specific intensity normalization and parameter selection. *Neuroimage*. 2010; 50:516–523. [PubMed: 20034579]
16. Leung KK, Barnes J, Ridgway GR, Bartlett JW, Clarkson MJ, Macdonald K, Schuff N, Fox NC, Ourselin S. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. *Neuroimage*. 2010; 51:1345–1359. [PubMed: 20230901]
17. McEvoy LK, Edland SD, Holland D, Hagler DJ Jr, Roddey JC, Fennema-Notestine C, Salmon DP, Koyama AK, Aisen PS, Brewer JB, Dale AM. Neuroimaging enrichment strategy for secondary prevention trials in Alzheimer disease. *Alzheimer Dis Assoc Disord*. 2010; 24:269–277. [PubMed: 20683184]
18. Nestor SM, Rupsingh R, Borrie M, Smith M, Accomazzi V, Wells JL, Fogarty J, Bartha R. Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. *Brain*. 2008; 131:2443–2454. [PubMed: 18669512]
19. Schuff N, Woerner N, Boreta L, Kornfield T, Shaw LM, Trojanowski JQ, Thompson PM, Jack CR Jr, Weiner MW. MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain*. 2009; 132:1067–1077. [PubMed: 19251758]
20. Wolz R, Heckemann RA, Aljabar P, Hajnal JV, Hammers A, Lotjonen J, Rueckert D. Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. *Neuroimage*. 2010; 52:109–118. [PubMed: 20382238]
21. Yushkevich PA, Avants BB, Das SR, Pluta J, Altinay M, Craige C. Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: an illustration in ADNI 3 T MRI data. *Neuroimage*. 2010; 50:434–445. [PubMed: 20005963]
22. Chou YY, Lepore N, Avedissian C, Madsen SK, Parikshak N, Hua X, Shaw LM, Trojanowski JQ, Weiner MW, Toga AW, Thompson PM. Mapping correlations between ventricular expansion and CSF amyloid and tau biomarkers in 240 subjects with Alzheimer's disease, mild cognitive impairment and elderly controls. *Neuroimage*. 2009; 46:394–410. [PubMed: 19236926]
23. Ho AJ, Stein JL, Hua X, Lee S, Hibar DP, Leow AD, Dinov ID, Toga AW, Saykin AJ, Shen L, Foroud T, Pankratz N, Huentelman MJ, Craig DW, Gerber JD, Allen AN, Corneveaux JJ, Stephan DA, DeCarli CS, DeChairo BM, Potkin SG, Jack CR Jr, Weiner MW, Raji CA, Lopez OL, Becker JT, Carmichael OT, Thompson PM. A commonly carried allele of the obesity-related FTO gene is associated with reduced brain volume in the healthy elderly. *Proc Natl Acad Sci U S A*. 2010; 107:8404–8409. [PubMed: 20404173]
24. Hua X, Leow AD, Lee S, Klunder AD, Toga AW, Lepore N, Chou YY, Brun C, Chiang MC, Barysheva M, Jack CR Jr, Bernstein MA, Britson PJ, Ward CP, Whitwell JL, Borowski B, Fleisher AS, Fox NC, Boyes RG, Barnes J, Harvey D, Kornak J, Schuff N, Boreta L, Alexander GE, Weiner MW, Thompson PM. Alzheimer's Disease Neuroimaging I. 3D characterization of brain atrophy in Alzheimer's disease and mild cognitive impairment using tensor-based morphometry. *Neuroimage*. 2008; 41:19–34. [PubMed: 18378167]
25. Morra JH, Tu Z, Apostolova LG, Green AE, Avedissian C, Madsen SK, Parikshak N, Hua X, Toga AW, Jack CR Jr, Schuff N, Weiner MW, Thompson PM. Automated 3D mapping of hippocampal atrophy and its clinical correlates in 400 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. *Hum Brain Mapp*. 2009; 30:2766–2788. [PubMed: 19172649]
26. Stein JL, Hua X, Lee S, Ho AJ, Leow AD, Toga AW, Saykin AJ, Shen L, Foroud T, Pankratz N, Huentelman MJ, Craig DW, Gerber JD, Allen AN, Corneveaux JJ, DeChairo BM, Potkin SG,

- Weiner MW, Thompson P. Voxelwise genome-wide association study (vGWAS). *Neuroimage*. 2010; 53:1160–1174. [PubMed: 20171287]
27. Stein JL, Hua X, Morra JH, Lee S, Hibar DP, Ho AJ, Leow AD, Toga AW, Sul JH, Kang HM, Eskin E, Saykin AJ, Shen L, Foroud T, Pankratz N, Huentelman MJ, Craig DW, Gerber JD, Allen AN, Corneveaux JJ, Stephan DA, Webster J, DeChairo BM, Potkin SG, Jack CR Jr, Weiner MW, Thompson PM. Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease. *Neuroimage*. 2010; 51:542–554. [PubMed: 20197096]
 28. Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch Gen Psychiatry*. 2006; 63:484–489. [PubMed: 16651505]
 29. Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. *Am J Psychiatry*. 1984; 141:1356–1364. [PubMed: 6496779]
 30. Meinert, CL. *Clinical Trials Design, Conduct and Analysis*. Oxford University Press, Inc; New York: 1986. p. 84
 31. Schlesselman JJ. Planning a longitudinal study: II. *Frequency of measurement and study duration*. *J Chron Dis*. 1973; 26:561–570. [PubMed: 4759581]
 32. Edland, SD. Which MRI measure is best for Alzheimer's disease prevention trials: Statistical considerations of power and sample size. 2009 Joint Stat Meeting Proceedings; 2009. p. 4996-4999.
 33. Jack CR Jr, Shiung MM, Gunter JL, O'Brien PC, Weigand SD, Knopman DS, Boeve BF, Ivnik RJ, Smith GE, Cha RH, Tangalos EG, Petersen RC. Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. *Neurology*. 2004; 62:591–600. [PubMed: 14981176]
 34. Diggle, P.; Heagerty, P.; Liang, K-Y.; Zeger, S. *Analysis of Longitudinal Data*. 2. Oxford University Press; Oxford: 2002.
 35. Aisen PS, Schneider LS, Sano M, Diaz-Arrastia R, van Dyck CH, Weiner MF, Bottiglieri T, Jin S, Stokes KT, Thomas RG, Thal LJ. High-dose B vitamin supplementation and cognitive decline in Alzheimer disease: a randomized controlled trial. *JAMA*. 2008; 300:1774–1783. [PubMed: 18854539]
 36. Vozdolska R, Sano M, Aisen P, Edland SD. The net effect of alternative allocation ratios on recruitment time and trial cost. *Clinical Trials*. 2009; 6:126–132. [PubMed: 19342464]
 37. Aisen PS, Andrieu S, Sampaio C, Carrillo M, Khachaturian ZS, et al. Report of the task force on designing clinical trials in early (predementia) AD. *Neurol*. 2011; 76:280–286.

Table 1

Sample size required to detect a 25% reduction in annual rate of change for ADAS-Cog scores in AD and MCI (80% power and two-sided $\alpha = 0.05$)

Paper	Trial design Yrs (#Obs)	AD		MCI	
		ROC	n/arm	ROC	n/arm
Aisen et al. (2010) ^a	1(n.s.)	4.3	407	1.1	4099
Beckett et al. (2010) ^a	2(5)	4.37	–	1.05	375
Chen et al. (2010) ^a	1(2)	3.8	[353, 505] ^b	1.0	[4026, 4219] ^b
Fleisher et al. (2009) ^a	1(n.s.)	n.s.	[474, 612] ^c	–	–
	2(n.s.) ^d	–	–	n.s.	854
Ho et al. (2010) ^a	1(2)	3.29	583	2.43	1183
Holland et al. (2009) ^e	1(3)	4.08	624	1.19	4167
	2(5)	4.84	324	1.44	1232
Hua et al. (2010) ^a	0.5(2)	n.s.	1371	n.s.	16645
	1(2)	n.s.	483	n.s.	8212
	1.5(2)	–	–	n.s.	1381
	2(2)	n.s.	215	n.s.	1013
Landau et al. (2009) ^f	1(3)	3.8	312	1.0	2175
McEvoy et al. (2010) ^e	2(5)	–	–	1.47	978
Nestor et al. (2008) ^{a,d}	0.5(2)	4.8	769	1.2	9500+
Schuff et al. (2009) ^{d,g}	0.5(2)	n.s.	557	n.s.	3484
	1(2)	n.s.	609	n.s.	6985
	1(3)	n.s.	426	n.s.	6241

Yrs (#Obs) = length of trial in years (number of evenly spaced observations); ROC = annual rate-of-change; n.s. = not specified; – = not applicable (scenario not investigated in the manuscript).

Superscripts:

^a analysis plan or longitudinal statistical model not fully specified;

^b ranges reflect different values for nested subsets of the data;

^c range reflects values from two cited reports;

^d N/arm in table adjusted from values reported in manuscript to reflect 25% effect size and 80% power, see Methods for details;

e linear mixed effects model with random slopes and intercepts assumed;
 f model assumptions alternatively cited as linear mixed effects with random-intercepts only and linear mixed effects with random slopes and intercepts;
 g random-intercepts only model assumed.

Table 2

Sample size required to detect a 25% reduction in annual rate of hippocampal atrophy in AD and MCI (80% power and two-sided $\alpha = 0.05$)

Paper	Trial design Yrs (#Obs)	AD		MCI	
		Atrophy rate	n/arm	Atrophy rate	n/arm
Aisen et al. (2010) ^a	1(n.s.)	n.s.	99	n.s.	208
Beckett et al. (2010) ^a	1(n.s.)	(116 mm ³)	–	(80 mm ³)	202
Holland et al. (2009) ^b	1(3)	3.42%	111	2.10%	235
Hua et al. (2010) ^a	2(5)	3.28%	67	1.96%	179
	0.5(2)	n.s.	114	n.s.	143
Leung et al. (2010) ^a	1(2)	n.s.	68	n.s.	125
	1.5(2)	n.s.	–	n.s.	117
	2(2)	n.s.	84	n.s.	103
	1(2) ^{c1}	4.57%	78	2.86%	196
McEvoy et al. (2010) ^b	1(2) ^{c2}	4.58%	170	3.68%	285
	2(5)	–	–	1.93%	186
Schuff et al. (2009) ^{d,e}	0.5(2)	3.3% (53.5 mm ³)	346	2.0% (37.7 mm ³)	709
	1(2)	4.4% (72.0 mm ³)	189	2.6% (47.5 mm ³)	522
Wolz et al. (2009) ^a	1(3)	n.s.	191	n.s.	503
	1(2)	3.85%	67	2.34%	206
	2(3) ^f	3.37%	46	2.25%	121
Yushekevich et al. (2010) ^a	1(2)	–	–	2.04%	220 ^g

Yrs (#Obs) = length of trial in years (number of evenly spaced observations); Atrophy Rate reported as % change/year, or when indicated in parenthesis, reported in units of (mm³/year); n.s. = not specified; – = not applicable (scenario not reported in the manuscript).

Superscripts:

^a analysis plan or longitudinal statistical model not fully specified;

^b linear mixed effects model with random slopes and intercepts assumed;

^{c1} measurement by multiple-atlas propagation and segmentation-hippocampal boundary shift integral (MAPS-HBSI);

^{c2} measurement by Surgical Navigation Technologies;

- ^d random-intercepts only model assumed;
- ^e N/arm in table adjusted from values reported in manuscript to reflect 25% effect size and 80% power, see Methods for details;
- ^f Atrophy rates adjusted to % change/year, rates in original manuscript given as % change/2 years;
- ^g N/arm in table calculated directly from MCI atrophy rate. N/arm presented in original manuscript was for disease-specific treatment effect adjusted for normal age-related atrophy.

Table 3

Sample size required to detect a 50% slowing of overall atrophy (trial of 12 months, with observation at 0, 6 and 12 months, equal allocation to arms, and 90% power)

Outcome variable	MCI mean slope	σ_b^2	σ_e^2	n/arm
Whole brain ^a	−3345	1613	2168	90
Ventricles ^a	1975	1033	1183	83
Left hippocampus ^a	−34.118	27.974	15.943	115
Right hippocampus ^a	−34.188	28.952	19.432	93
Left mid-temporal cortex ^b	−0.022	0.006	0.013	84
Right mid-temporal cortex ^b	−0.020	0.012	0.013	94

^a volume in mm³;

^b cortical thickness in mm.

Sample size required to detect a 50% slowing of disease-specific atrophy, defined as atrophy above and beyond that experienced by age-matched non-impaired elderly (trial of 12 months, with observation at 0, 6 and 12 months, equal allocation to arms, and 90% power)

Table 4

Outcome variable	Non-impaired annual decline	Disease specific annual decline	σ_b^2	σ_3^2	n/arm
Whole brain ^a	-1724	-1625	1613	2168	384
Ventricles ^a	1201	772	1033	1183	544
Left hippocampus ^a	-17.141	-16.977	27.974	15.943	376
Right hippocampus ^a	-16.427	-17.761	28.952	19.432	425
Left mid-temporal cortex ^b	-0.009	-0.013	0.006	0.013	252
Right mid-temporal cortex ^b	-0.009	-0.011	0.012	0.013	319

^a volume in mm³;
^b cortical thickness in mm.