Received: 30 April 2021 | Revised: 16 September 2021 | Accepted: 13 October 2021

DOI: 10.1002/sim.9244

RESEARCH ARTICLE

Statistics in Medicine WILEY

Limitations of clinical trial sample size estimate by subtraction of two measurements

Kewei Chen[1,2,3,4] | Xiaojuan Guo[4,5] | Rong Pan[2] | Chengjie Xiong[6,7] | Danielle J. Harvey[8] | Yinghua Chen[1,9] | Li Yao[4] | Yi Su[1,4] | Eric M. Reiman[1,9,10,11] | for the Alzheimer's Disease Neuroimaging Initiative

[1]Banner Alzheimer's Institute, Phoenix, Arizona, USA

[2]Department of Mathematics and Statistics, Arizona State University, Tempe, Arizona, USA

[3]Department of Neurology, University of Arizona, Phoenix, Arizona, USA

[4]Arizona Alzheimer's Consortium, Phoenix, Arizona, USA

[5]School of Artificial Intelligence, Beijing Normal University, Beijing, China

[6]Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, Beijing, China

[7]Knight Alzheimer's Disease Research Center, St. Louis, Missouri, USA

[8]Division of Biostatistics, Washington University School of Medicine in St. Louis, St. Louis, Missouri, USA

[9]Department of Public Health and Sciences, University of California, Davis, California, USA

[10]Division of Neurogenomics, Translational Genomics Research Institute, Phoenix, Arizona, USA

[11]Department of Psychiatry, University of Arizona, Tucson, Arizona, USA

Correspondence
Kewei Chen, Banner Alzheimer's Institute, 901 East Willetta Street, Phoenix, AZ 85006, USA.
Email: Kewei.chen@bannerhealth.com

In planning randomized clinical trials (RCTs) for diseases such as Alzheimer's disease (AD), researchers frequently rely on the use of existing data obtained from only two time points to estimate sample size via the subtraction of baseline from follow-up measurements in each subject. However, the inadequacy of this method has not been reported. The aim of this study is to discuss the limitation of sample size estimation based on the subtraction of available data from only two time points for RCTs. Mathematical equations are derived to demonstrate the condition under which the obtained data pairs with variable time intervals could be used to adequately estimate sample size. The MRI-based hippocampal volume measurements from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and Monte Carlo simulations (MCS) were used to illustrate the existing bias and variability of estimates. MCS results support the theoretically derived condition under which the subtraction approach may work. MCS also show the systematically under- or over-estimated sample sizes by up to 32.27% bias. Not used properly, such subtraction approach outputs the same sample size regardless of trial durations partly due to the way measurement errors are handled. Estimating sample size by subtracting two measurements should be treated with caution. Such estimates can be biased, the magnitude of which depends on the planned RCT duration. To estimate sample sizes, we recommend using more than two measurements and more comprehensive approaches such as linear mixed effect models.

KEYWORDS

linear mixed effects model, randomized clinical trial, sample size estimation, subtraction, two time point measurement

Kewei Chen and Xiaojuan Guo contributed equally to this work.
Data used in 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.

Funding information
National Institute of Mental Health, Grant/Award Number: R01MH057899; National Institute on Aging, Grant/Award Numbers: P30AG019610, R01AG031581, RF1AG041705, UF1AG046150; National Natural Science Foundation of China, Grant/Award Numbers: 61671066, 61731003, 62071051; National Science Foundation, Grant/Award Number: CMMI-1505260

1 INTRODUCTION

It is a common practice when planning randomized clinical trials (RCTs) to use an existing dataset for the estimation of sample size needed to detect the treatment efficacy. In RCTs for diseases such as Alzheimer's disease (AD), the longitudinal changes of an outcome variable are often obtained by subtracting measures from two time points: the baseline and the follow-up. Power analysis is always an important part of designing placebo controlled RCTs. Among a number of possible causes for the negative outcomes of multiple recent RCTs in patients with AD or mild cognitive impairment (MCI), the insufficient number of enrolled participants cannot be ruled out. This report examines the limitation of the power analysis that is based on the subtraction of available data from only two time points.

Statistical power analysis balances the likelihood of confirming a hypothesis (eg, with an 80% chance) and controlling the likelihood of a false positive (eg, with a 5% type-I error). A typical AD RCT tests the hypothesis that the efficacy outcome of an intervention will be better (eg, 25% more efficacious) on the treatment arm compared to the placebo arm. This co-called effect size could arise from interventions for disease modification, symptomatic treatment, or disease prevention, therefore it accounts for the drug's pharmacodynamic and/or pharmacokinetic profiles. For AD RCTs, the hypothesis is that the intervention introduced to the treatment arm slows or reverses the baseline-to-follow-up declines. Because estimated sample sizes for RCTs represent the minimum number of subjects needed to detect such effect size, under-estimates in sample size could adversely affect trial outcomes.

To estimate the number of subjects needed for future RCTs, researchers typically use existing longitudinal datasets.[1-3] The existing data are often times observational only, and the observed changes from the progression of the underlying disease are conceptually equivalent to what would be expected from the placebo arm in the planned RCT. It is not uncommon that the existing longitudinal data have only measurements from two time points—the baseline and one follow-up, and the sample size estimation is based on their difference.[4-7] Other methods exist, with data available from additional data points beyond only two time points, such as linear mixed effects (LME) models[8,9] and mixed models for repeat measures (MMRM).[10,11] This study discusses the limitation of power analyses related to the situation where data from only two time points are available.

The subtraction approach suffers from a major weakness that is counter intuitive: the estimated number of subjects is the same regardless of the duration of the proposed RCT (unless with additional conditions, see Discussion). This report examines this weakness and derives the conditions under which longitudinal data from only two time points could be used to adequately estimate sample size. Finally, we used the LME model and Monte Carlo simulations (MCS) to analyze longitudinal hippocampal volumetric data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to investigate the bias and variability of sample size estimates.

2 MATERIALS AND METHODS

2.1 Theoretical considerations

We assume the use of the LME model to derive the condition under which the simple subtraction approach can be adequate for the two-time-point longitudinal data analysis from multiple subjects. The LME model jointly explains the population-level fixed effects and the subject-specific random effects to evaluate the relationship between average responses and time measurements based on a linear regression paradigm.[8] Such model generally has three assumptions: (1) the change of response over time is linear, (2) the observation at each time point is contaminated by an additive Gaussian error (residual term), which is independently and identically distributed (i.i.d) for all visits (this can be loosened, but is outside the scope of this report), and (3) the random effect terms are independent of the residual term.

We estimate the sample size per arm for a balanced RCT that would last for T_p years by first assuming no subject dropout (T_p is the number of years for a planned trial, and the subscript p stands for planned). For an existing dataset that was collected independent of and before the proposed trial, we assume there are N subjects ($N>1$) and subject j ($j=1, 2, \ldots, N$) has the baseline observation x_j^b and a follow-up observation x_j^f with the observation time length T_j, where T_j is the time duration (in years) between the first and second visits for subject j. We recognize the baseline-to-follow-up time variability among subjects and that the average baseline-to-follow-up time duration does not necessarily equal to the planned RCT duration.

2.2 | Simple subtraction

A straightforward subtraction procedure to estimate sample size with two-time-point measurements is as follows (Figure 1). First, the per-year change for each subject is found; that is, for subject j it is $\frac{x_j^b - x_j^f}{T_j} = \frac{\Delta_j}{T_j}$. Under the assumption of linear change, the estimated T_p-year change for subject j becomes $T_p \times \frac{x_j^b - x_j^f}{T_j}$. The mean change over T_p-years over all N subjects is $T_p \times \Delta$, and the standard deviation (*std*) is $T_p \times std_\Delta$. Here, Δ is the one-year mean change over N subjects,$\Delta = \frac{1}{N}\sum_j \frac{\Delta_j}{T_j}$, and std_Δ is the within arm standard deviation, $std_\Delta = \sqrt{\frac{1}{N-1}\sum_j \left(\frac{\Delta_j}{T_j} - \Delta\right)^2}$. For planning clinical trials, we assume Δ is the change in the placebo arm without any intervention. Note that the durations (the time interval lengths) between the two consecutive time points for individual subjects, the annualized average over subjects (Δ above) and the effect size are for the existing dataset, therefore are as given. The inadequacy of using such preacquired data (with the given durations) to estimate the sample size for a planned clinical trial with its own duration is the focus of this study. See more on this important issue in the Discussion section.

For sample size estimation, we assume the statistical power to be 80% and the two-tailed type-I error to be 0.05. Also, we assume the treatment effect to be 25%.[3,5,12-15]In other words, the change in the treatment arm of the planned clinical trial is 0.75×Δ due to the treatment. As a side note and discussed below, our conclusions are independent of these parameters. We also assume that the treatment has no effects on the variability or both trial arms have the same standard deviation equivalently. With these parameters and assumptions, the sample size estimation is straightforward, having a closed mathematical expression. The estimated sample size N_p depends only on the ratio of the standard deviation over the mean difference between the two arms$\left(\left(T_p \times std_\Delta\right) / \left(T_p \times 0.25 \times \Delta\right) = std_\Delta / (0.25 \times \Delta)\right)$ based on the Gaussian distribution assumption: $N_p = 2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \times (std_\Delta / (0.25\Delta))^2$, where the appearance of 2 is related to the fact that the std of the mean difference is equal to the std of the placebo arm (or equivalently the treatment arm) multiplied by $\sqrt{2}$, $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the corresponding z-score at $1-\alpha/2$ and $1-\beta$ separately, only determined by the given statistical power $(1-\beta)$ and type-I error (α).[1,16] Note, subscript p in N_p is again for planned (trial) and N_p is the estimated sample size per arm for a balance design. Thus, given these statistical power, the type-I error and the treatment effect, the sample size is only dependent on the ratio std_Δ/Δ, independent of the trial duration, which is counter intuitive. Also, this provides us the rational for our discussion below to focus on the ratio std_Δ/Δ only.

2.3 | When simple subtraction is adequate for sample size estimation

As we noticed just above, the sample size estimation for the planned clinical trial is solely based on the ratio std_Δ/Δ of the existing dataset (viewed as the placebo arm in the trial). Using an LME model, the baseline and follow-up measures for subject j in the existing dataset are:

$$
\begin{aligned}
x_j^b &= b + kt_j^b + b_j + k_j t_j^b + \varepsilon_j^b, \\
x_j^f &= b + kt_j^f + b_j + k_j t_j^f + \varepsilon_j^f, \qquad (1)
\end{aligned}
$$

where b and k are the fixed intercept and slope for all subjects, respectively. The index j corresponds to subject j, and b_j and k_j are the random intercept and slope for subject j, respectively. The baseline and follow-up times are t_j^b and t_j^f, respectively, and $T_j = t_j^b - t_j^f$. Also, the term ε_j^b or ε_j^f is the measurement noise, referred to as the residual error for baseline time and followup time, respectively. As mentioned earlier, the residual error is i.i.d, with zero expectation and standard deviation σ_e. Thus, the measurement change in subject j is:

$$\Delta_j = x_j^b - x_j^f = T_j(k + k_j) + \varepsilon_j^b - \varepsilon_j^f. \tag{2}$$

Note that the residual error difference $\varepsilon_j^b - \varepsilon_j^f$ is a random variable, cannot be canceled out by the subtraction (though the difference has expectation 0). We denote the standard deviation of residual error difference as σ_d. Under the i.i.d. assumption, $\sigma_d = \sqrt{2}\sigma_e$, where σ_e is the standard deviation of the error term at a given visit. Note that under no circumstances we can assume zero σ_d.

The sample size for an RCT, which is planned to last for T_p years, can be estimated by using baseline and follow-up measurements for either an ideal or real-world scenario. Both scenarios utilize Equation (2).

Ideal scenario: The data pairs of baseline and follow-up measurements used for the sample size estimation of an RCT of T_p-year duration are acquired exactly T_p years apart for all subjects; that is, $T_j = T_p$ for all j. In this case, for the j^{th} subject, we can directly use baseline and follow-up data to calculate the variance of change over trial period such as.

$$var\left(T_p\frac{\Delta_j}{T_j}\right) = var\left(T_p\frac{\Delta_j}{T_p}\right) = var(\Delta_j) = T_p^2\sigma_\beta^2 + \sigma_d^2, \tag{3}$$

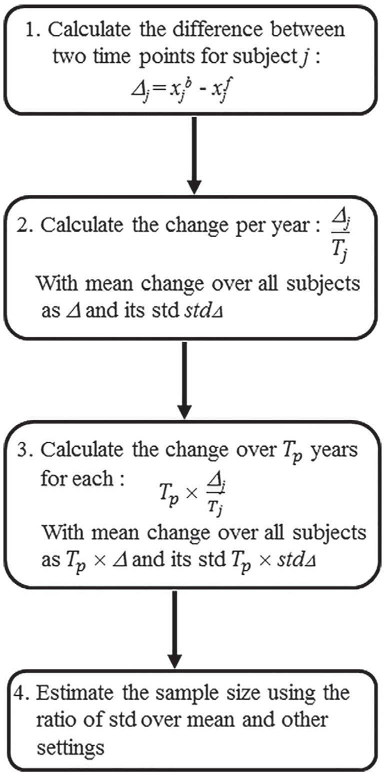
1. Calculate the difference between two time points for subject j :
$\Delta_j = x_j^b - x_j^f$
2. Calculate the change per year : $\frac{\Delta_j}{T_j}$
With mean change over all subjects as Δ and its std $std\Delta$
3. Calculate the change over T_p years for each : $T_p \times \frac{\Delta_j}{T_j}$
With mean change over all subjects as $T_p \times \Delta$ and its std $T_p \times std\Delta$
4. Estimate the sample size using the ratio of std over mean and other settings

FIGURE 1 The flowchart of estimating sample size with the two-time point subtraction procedure.

where σ_β^2 is the variance of random slope k_j over subjects. The variance of the sum of $T_p \frac{\Delta_j}{T_j}$ over N subjects for total change during T_p years is:

$$var\left(\sum_j T_p \frac{\Delta_j}{T_j}\right) = var\left(\sum_j T_p \frac{\Delta_j}{T_p}\right) = var\left(\sum_j \Delta_j\right) = NT_p^2\sigma_\beta^2 + N\sigma_d^2. \qquad (3')$$

Real-world scenario: T_j varies among subjects according to Equation (2). We first convert a change to a change per-year, and then we write Equation (4) for the T_p-year change:

$$T_p \frac{\Delta_j}{T_j} = T_p(k + k_j) + T_p \frac{\varepsilon_j^b - \varepsilon_j^f}{T_j}. \qquad (4)$$

Under the assumption of independent subject specific random effect and residual error, for the j^{th} subject, we have

$$var\left(T_p \frac{\Delta_j}{T_j}\right) = T_p^2\sigma_\beta^2 + \frac{T_p^2\sigma_d^2}{T_j^2}. \qquad (5)$$

And for N subjects:

$$var\left(\sum_j T_p \frac{\Delta_j}{T_j}\right) = NT_p^2\sigma_\beta^2 + T_p^2\sigma_d^2 \sum_j \frac{1}{T_j^2}. \qquad (5')$$

Because both the ideal and real-world scenarios have the same goal of estimating the sample size for an RCT with a duration of T_p years, we can identify under what condition the estimated sample sizes are the same for the two scenarios. We only need to consider the variance, or the standard deviation to be used in the sample size estimation, because the expected mean change over T_p years is simply $T_p \times k$, owing to the zero expectation of the k_j term in both scenarios. By equating Equation (3') with Equation (5'), Equation (6) is obtained which shows the condition under which the two scenarios are equivalent, that is, they provide equivalent sample size estimation:

$$\sum_j \frac{1}{T_j^2} = \frac{N}{T_{p0}^2}. \qquad (6)$$

In this expression, we use T_{p0} specially to indicate it is the trial duration that can make use of the existing data to provide adequate sample size. It is interesting to note that T_{p0} is not the simple arithmetic mean over T_j. Instead, the reciprocal of the squared T_{p0} equals the mean of the reciprocals of squared individual time lengths T_j as Equation (6) can be rewritten as: $\frac{1}{N}\sum_j \frac{1}{T_j^2} = \frac{1}{T_{p0}^2}$.

2.4 | Data

While there are a number of different well-established AD biomarkers,[7,17-21] we opt to use volumetric hippocampus measurements from structural magnetic resonance imaging (MRI) data.[22-25] Any existing dataset with variable follow-up time points could be used to illustrate the limitations of the simple subtraction procedure. We use the MRI data from the ADNI (adni.loni.usc.edu) to illustrate a general point, relevant to sample size estimates for all clinical trials, not only AD. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), and clinical and neuropsychological assessments can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org. The ADNI study was approved by the institutional review boards of the participating institutions. Informed written consent was obtained from all participants.

The participants include 182 persons with clinical diagnosis of MCI (male/female: 108/74; age range: 57.2 to 90.7 years; age mean: 72.9± 7.0 years; mini-mental state examination (MMSE) range: 22 to 30; mean MMSE: 28.1±1.7). Each person has at least two longitudinal structural MRI observations over the time interval from 1.50 to 2.37 years. Data from all visits are used to estimate the parameters in the LME model. For simple subtraction, we use the baseline data and the data from one follow-up visit that occurred on average approximately 2 years after the baseline (mean: 1.995 ± 0.119 years).

The structural MRI data are acquired at multiple sites. (http://adni.loni.usc.edu/methods/documents/mri-protocols/) and are gradwarped, intensity corrected and scaled for gradient drift. They are preprocessed with FreeSurfer 5.3 with images intensity normalization and skull stripping, and cortical and subcortical regions are labeled. The automated segmentations have also been manually inspected and corrected as needed (surfer.nmr.mgh.harvard.edu/fswiki/).[26,27] We obtain the relative hippocampal volume by dividing hippocampal volume by intracranial cavity volume (ICV) for each subject. The mean relative hippocampal volume of all subjects is 0.00467 ± 0.00082 at baseline and 0.00451 ± 0.00096 at the 1.995-year follow-up. The average change of relative hippocampal volume over this period is 0.00017 ± 0.00036.

2.5 | Monte Carlo simulations based on LME model

The MCS has been conducted based on the LME modeling of the ADNI data. We simulate the relative hippocampus volume at different relative ages. The relative age is calculated as the difference between the ages at the time of scanning and the time of conversion to AD for subjects who progressed to AD at the end of the follow-up period (MCI-c). For subjects who maintained a diagnosis of MCI (MCI-nc), their relative ages are calculated as the difference between the scanning time and the average conversion time for MCI-c subjects.

For the 182 subjects, we fit data from 438 visits to the LME model, which includes more time points than the selected two-time points. Here, we estimate the following model parameters—fixed parameters, random effect covariance matrix and the random error term variance matrix, and we treat these estimated parameter values as true model values in the subsequent MCS. Then we compare the sample size output according to Equation (3) to the sample size generated from the simple subtraction procedure.

To examine how the size of the existing dataset affected the estimated sample size, we run MCS on datasets consisting of N=182, N = 182 × 2 and N = 182 × 3 subjects, which correspond to the original, doubled and tripled sample sizes of the existing dataset, respectively. We compute the mean bias and associated std for a given planned trial duration over 500 repetitions.

Each simulation run *i, (i=1, 2, ..., 500)* has the following steps:

1. For each subject *j, (j=1, 2, ...;, N)*:
   a. Select the two-time points that are T_j years apart.
   b. Generate the subject's random intercept and slope following the joint Gaussian distribution with the given parameters from the fitted model.
   c. Generate the Gaussian measurement error under the assumption it is independent from the random effect terms.
   d. Compute the hippocampal volumes x_j^b and x_j^f at the two time points.
   e. Repeat steps 1a–d for all N subjects.
2. Apply the simple subtraction procedure to the simulated data from the N subjects.
3. Estimate the sample size, N_{pi}, for simulation i with 80% power, a two-tailed type-I error of 0.05 and a 25% treatment effect in the treatment arm relative to the placebo arm. The different N_{pi} estimates are computed separately for the planned trial durations of T_p=1, 1.98 and 5 years, with the original, doubled, and tripled existing dataset sizes, respectively. Note that 1.98 is the T_p value that satisfies Equation (6).
4. Compute the percentage bias of the relative true sample size based on the LME model and the simple subtraction approach.

3 | RESULTS

3.1 | Sample size estimation with simple subtraction

For a given treatment effect, the simple subtraction method estimates the same number of subjects required regardless of trial duration. The simple subtraction procedure fails because the standard deviation of measured changes is not linear with respect to time. More specifically, the noise term in Equation (2), the residual error of the data, is unrelated to the trial duration, T_j. Though the random intercepts canceled out between two visits, two residual error terms do not cancel out because they are assumed to be i.i.d.

3.2 | Condition under which simple subtraction is adequate

Sample size estimation using simple subtraction is adequate when the subject-dependent time intervals between baseline and follow-up satisfy Equation (6), or equivalently,

$$T_{p0} = \sqrt{\frac{N}{\sum_j \frac{1}{T_j^2}}}. \tag{6'}$$

TABLE 1 The relative sample size bias (percentage) and the variations (std) from simple subtraction for different planned RCT durations

	$T_p = 1$		$T_p = T_{p_0} = 1.98$		$T_p = 5$	
Data size	**Mean**	**std**	**Mean**	**std**	**Mean**	**std**
$N = 182$	−30.85	24.06	4.02	36.20	21.74	42.36
$N = 182 \times 2$	−30.76	16.66	4.15	25.07	21.89	29.34
$N = 182 \times 3$	−32.27	12.68	1.89	19.08	19.24	22.33

Note: T_p is the number of years for a planned trial; T_{p0} is the value of T_p satisfying Equation (6); std represents the standard deviation.

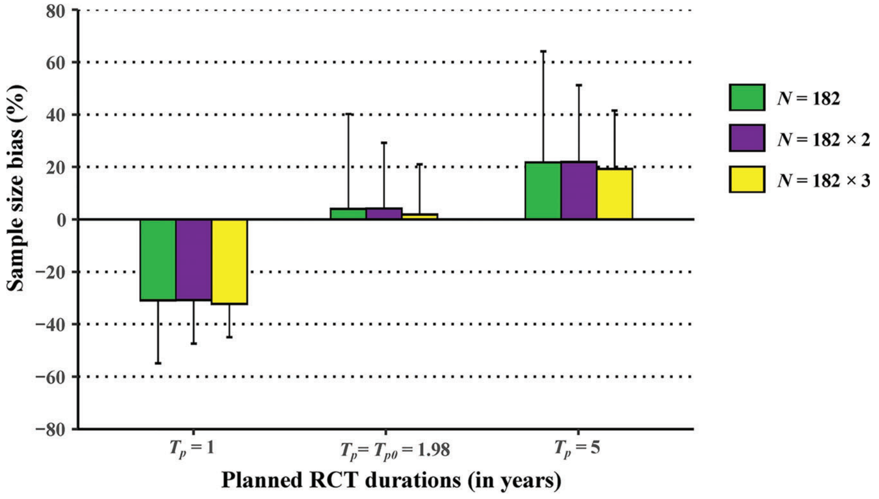

FIGURE 2 The relative sample size bias (percentage) based on subtraction of the existing dataset from two-time points that are average approximately two years apart. Error bars are standard deviations, and T_p is the trial duration. When $T_p = T_{p0}$ =1.98 years, the sample size can be estimated using simple subtraction. When $T_p = 1$ years (<1.98), the sample size obtained with the simple subtraction is under-estimated. When $T_p = 5$ years (>1.98), the sample size is over-estimated. As shown by the error bars, the variability of the sample size bias would decrease when N increases [Colour figure can be viewed at wileyonlinelibrary.com]

We note that a common T_p for all subjects is a special case that satisfies Equations (6) or (6'). Also, when the standard deviation of residual error is small relative to the standard deviation of random slope, these residual error terms can be ignored or treated as close to zero, then the residual errors would cancel out with subtraction.

3.3 | Sample size bias and variability from simple subtraction

The theoretical sample sizes are calculated based on the LME model. We attain the sample sizes of 1581, 1051, and 898 for the planned RCT with durations T_p of 1, 1.98, and 5 years, respectively. To calculate the relative sample size bias, we generate the ratios of the difference between the estimated sample size from simple subtraction and the theoretical sample size over the theoretical value for different planned RCT durations. The biases and variations have been all assessed using the 500 times MCS for different trial durations. Table 1 and Figure 2 show the relative sample size biases (percentage) and the variations from simple subtraction for the three different planned RCT durations.

As shown in Table 1 and Figure 2, the sample size bias is almost zero when the trial duration T_p equals to T_{p0} ($T_{p0} = 1.98$), consistent with our theoretical conclusion. These biases are 4.02%, 4.15% and 1.89% for $N = 182$, N = 182 × 2 and N = 182 × 3, respectively. When T_p is less than T_{p0}, (e.g., $T_p = 1$) the sample size obtained with the simple subtraction procedure is under-estimated by as much as 32.27%. When T_p is larger than T_{p0} (eg, when $T_p = 5$), the sample size is over-estimated by 21.89%. As shown by the error bars in Figure 2, the variability of the sample size bias decreases when N increases (eg, when $T_p = T_{p0} = 1.98$, std = 36.20% , 25.07% , 19.08% , for $N = 182$, N = 182 × 2, and N = 182 × 3, respectively).

4 | DISCUSSION

We have examined the sample size estimation for a planned RCT via subtraction of two-time-point measurements, discussed why the estimated sample size should in general depend on trial duration, and identified the special cases when the sample size estimation from simple subtraction might be correct. We note that the use of two-time-point measures, which is common for expensive studies like those that use neuroimaging-based biomarkers, differs from studies with more frequent measures, such as safety or clinical outcomes. Our findings do not apply to studies with more frequent measures or to studies that are analyzed with statistical procedures like LME models[8,9] or MMRM.[10,11] Estimated sample sizes generated from more than two observations have been reported extensively in the literature.[9,14,15,28-30] We note that it is important to account for measurement variability appropriately across time for power analysis.

We would like to characterize this report as with the nature of reporting a problem rather than resolving a problem. Nevertheless, mixed in the following discussions and in *italic fonts*, we offered some words of advice on how to deal with the limitations when data from only two time points are available. These pieces of advice are intended for some "cheap" solutions for the problem we reported in this investigation.

To account for both between-subject random effects and within-subject residual errors and to understand the issues related to simple subtraction approach, we used the LME model.[8] It is noticed that the simple subtraction has to be properly scaled in order to account for the inter-subject two-visit time length variation (see Equation (2)). If not linear, dividing T_j may not get you a term that is independent of T_j. In the nonlinear case, the simple subtraction is practically not feasible unless we assume that the magnitude of changes does not depend on the time length over which the changes occur.

When T_p was equal to T_{p0}, simple subtraction is adequate (Table 1 and Figure 2). *In this case, we advise it is safe to use the simple subtraction method.* For longer RCT durations, the required sample size is generally smaller. Simple subtraction underestimates the sample size when $T_p < T_{p0}$ and overestimates the sample size when $T_p > T_{p0}$ (Figure 2).*When the overestimation is affordable, then the simple subtraction approach can be used.* Such trends are observed based on the average estimated sample size bias from MCS. The absolute difference between the theoretical sample size and the estimated sample size from simple subtraction is larger when $T_p < T_{p0}$ (nearly 30%) than when $T_p > T_{p0}$ (nearly 20%). For each RCT duration, the standard deviation of the relative sample size bias shows a decrease trend as the number of subjects in the existing dataset increase (Table 1 and Figure 2).

The simple subtraction approach worked if one of the following conditions is met: (1) the measurement error is ignorable or linearly related to trial duration; (2) the proposed RCT duration T_p and the individual T_j satisfy Equations (6); and (3) if between-subject variability of the annualized change is low, the std_Δ/Δ ratio is roughly proportional to $1/T_p$ and the corresponding sample size decreases for longer T_p duration. Of course, our findings can be generalized to other outcome measures in AD or trials for any other diseases/conditions.

Because it uses data from just two time points, the simple subtraction method lacks the information on how the change occurs between the two measurements. If the change happens to be linear, it is possible for simple subtraction to accurately estimate sample sizes when the follow-up time is equal to the planned trial duration. If the change is nonlinear, more studies are needed for the use of the simple subtraction method. It is possible that the simple subtraction method could be amended to account for the variance in the data, but such a modification is outside the scope of this paper. Our math derivations can potentially be used in future studies to generate an accurate subtraction method based on two time points and LME models.

Though the sample size estimation equation does not include trial duration explicitly, it is implicitly included via the effect size, in the sense that one would expect the effect size corresponding to the treatment effect to be smaller for short duration than for long ones. To take the exact advantage of bigger effect size in the existing dataset, researchers should select only samples whose baseline-followup time differences satisfy Equation (6) with the corresponding larger T_p, or samples whose baseline-followup time differences are close to T_p (longer duration means bigger effect size). These suggestions given here reflected the limitation of the subtraction of two consecutive time points in that the effect size and the duration are preset due to the fact that dataset used were already in existence. The only flexibility left is then the selection of subcohort from this existing dataset to match the planned trial duration, recognizing the fact that the quantitative and yet implicit relationship between effect size and the trial duration was not considered appropriately by this simple subtraction approach. With more longitudinal time points available (>2), this effect size/duration relationship can be quantified, and the sample size more adequately estimated for varying duration of a planned trial.

The fact that the simple subtraction approach caused the estimated sample sizes to be the same regardless of trial duration in general (but see possible exceptions for condition 3) needs better understanding. Note that the absolute change, which corresponds to the final treatment effect, could be different depending on the trial duration. If the relative treatment effect was 25% for a one-year clinical trial and the change in the placebo arm was 1.0, then the difference between the placebo and treatment arms would be 0.75. If the relative treatment effect was still 25% but the trial duration was two years, then the two-year absolute change without treatment would be 2.0, which does rely on the assumption of linearity with time. The difference between the placebo and treatment arms with the 25% treatment effect will be 1.5. Thus, even though the one-year and two-year clinical trials would have had the same common relative treatment effect of 25%, the absolute differences between the two arms would not be the same depending on the length of time.

There are limitations of this study. The first limitation of this study is primarily related to the assumptions we included in the LME model. We assumed the noise in the data was Gaussian and that the measured changes were linear with time. Secondly, we only derived the conditions under which simple subtraction might work and identified the issues with this method. Procedures to address the inadequacy of the simple subtraction method when data are only available from two-time points should be topics for future research. Briefly, such procedures may 1) require prior knowledge of the LME model or 2) consider the possibility of combining bias information to correct sample size estimates after more careful examinations of the effects of intersubject variability in the time to follow-up measurements on the bias and variability of sample size estimation.

In conclusion, the use of simple subtraction on two time points for the estimation of RCT sample sizes should be used with caution because this method can be biased when the trial duration is longer or shorter than the observed measurement intervals and there are substantial individual variations in measurement interval.

ACKNOWLEDGEMENTS

This work was supported by the National Institute on Aging, USA (P30AG019610; R01AG031581; UF1AG046150; RF1AG041705); the National Institute of Mental Health (R01MH057899); Alzheimer's Disease Core Center and the state of Arizona; the National Science Foundation, USA (CMMI-1505260); the National Natural Science Foundation (NNSF), China (61671066, 62071051); Key Program of NNSF, China (61731003) and the China Scholarship Council, China.

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.; Cogstate; 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 Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Kewei Chen and **Eric M. Reiman**: Designed the study. **Kewei Chen** and **Yinghua Chen**: Performed the data analysis. **Xiaojuan Guo**, **Kewei Chen**, **Chengjie Xiong**, **Danielle J. Harvey**, **Li Yao**, and **Eric M. Reiman**: Interpreted and discussed the results. **Rong Pan**, **Yi Su**, **Xiaojuan Guo**, and **Kewei Chen**: Contributed to refining the ideas. **Xiaojuan Guo** and **Kewei Chen**: Drafted and revised the manuscript. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

ORCID

Kewei Chen https://orcid.org/0000-0001-8497-3069
Rong Pan https://orcid.org/0000-0001-5171-8248

REFERENCES

1. Ard MC, Edland SD. Power calculations for clinical trials in Alzheimer's disease. *J Alzheimers Dis*. 2011;26:369-377.
2. Brookmeyer R, Abdalla N. Design and sample size considerations for Alzheimer's disease prevention trials using multistate models. *Clin Trials*. 2019;16:111-119.
3. Grill JD, Di L, Lu PH, et al. Estimating sample sizes for predementia Alzheimer's trials based on the Alzheimer's disease neuroimaging initiative. *Neurobiol Aging*. 2013;34:62-72.
4. Fox NC, Cousens S, Scahill R, Harvey RJ, Rossor MN. Using serial registered brain magnetic resonance imaging to measure disease progression in Alzheimer disease: power calculations and estimates of sample size to detect treatment effects. *Arch Neurol*. 2000;57:339-344.
5. Chen K, Langbaum JB, Fleisher AS, et al. Twelve-month metabolic declines in probable Alzheimer's disease and amnestic 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.
6. Tabrizi SJ, Scahill RI, Durr A, et al. Biological and clinical changes in premanifest and early stage Huntington's disease in the TRACK-HD study: the 12-month longitudinal analysis. *Lancet Neurol*. 2011;10:31-42.
7. Beckett LA, Harvey DJ, Gamst A, et al. The Alzheimer's disease neuroimaging initiative: annual change in biomarkers and clinical outcomes. *Alzheimers Dement*. 2010;6:257-264.
8. Bernal-Rusiel JL, Greve DN, Reuter M, Fischl B, Sabuncu MR. Statistical analysis of longitudinal neuroimage data with linear mixed effects model. *NeuroImage*. 2013;66:249-260.
9. Fujishima M, Kawaguchi A, Maikusa N, Kuwano R, Iwatsubo T, Matsuda H. Sample size estimation for Alzheimer's disease trials from Japanese ADNI serial magnetic resonance imaging. *J Alzheimers Dis*. 2017;56:75-88.
10. Donohue MC, Sperling RA, Salmon DP, et al. The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. *JAMA Neurol*. 2014;71:961-970.

11. Lu K, Luo X, Chen P. Sample size estimation for repeated measures analysis in randomized clinical trials with missing data. *Int J Biostat*. 2008;4:1-16.
12. Beckett LA. Community-based studies of Alzheimer's disease: statistical challenges in design and analysis. *Stat Med*. 2000;19:1469-1480.
13. Hua X, Lee S, Yanovsky I, et al. 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.
14. Hua X, Lee S, Hibar DP, et al. Mapping Alzheimer's disease progression in 1309 MRI scans: power estimates for different inter-scan intervals. *NeuroImage*. 2010;51:63-75.
15. Hua X, Gutman B, Boyle CP, et al. Accurate measurement of brain changes in longitudinal MRI scans using tensor-based morphometry. *NeuroImage*. 2011;57:5-14.
16. Fox NC, Ridgway GR, Schott JM. Algorithms, atrophy and Alzheimer's disease: cautionary tales for clinical trials. *NeuroImage*. 2011;57:15-18.
17. Langbaum JB, Fleisher AS, Chen K, et al. Ushering in the study and treatment of preclinical Alzheimer disease. *Nature Rev Neurol*. 2013;9:371-381.
18. Reiman EM, Jagust WJ. Brain imaging in the study of Alzheimer's disease. *NeuroImage*. 2012;61:505-516.
19. Jack CR, Holtzman DM. Biomarker modeling of Alzheimer's disease. *Neuron*. 2013;80:1347-1358.
20. Wang HF, Shen XN, Li JQ, et al. Clinical and biomarker trajectories in sporadic Alzheimer's disease: a longitudinal study. *Alzheimers Dement*. 2020;12:e12095.
21. Meyer P-F, Binette A, Gonneaud J, Breitner J, Villeneuve S. Characterization of Alzheimer disease biomarker discrepancies using cerebrospinal fluid phosphorylated Tau and AV1451 positron emission tomography supplemental content. *JAMA Neurol*. 2020;77:508-516.
22. Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol*. 2010;6:67-77.
23. Resnick SM, Scheltens P. MRI-based biomarkers of preclinical AD: an Alzheimer signature. *Neurology*. 2012;78:80-81.
24. Hampel H, Bürger K, Teipel SJ, Bokde AL, Zetterberg H, Blennow K. Core candidate neurochemical and imaging biomarkers of Alzheimer's disease. *Alzheimers Dement*. 2008;4:38-48.
25. Marizzoni M, Ferrari C, Jovicich J, et al. Predicting and tracking short term disease progression in amnestic mild cognitive impairment patients with prodromal Alzheimer's disease: structural brain biomarkers. *J Alzheimers Dis*. 2019;69:3-14.
26. Fischl B, Salat DH, Busa E, Albert M, Dale AM. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*. 2002;33:341-355.
27. Fischl B, Salat DH, van der Kouwe AJW, et al. Sequence-independent segmentation of magnetic resonance images. *NeuroImage*. 2004;23:S69-S84.
28. Hua X, Hibar DP, Ching CRK, et al. Unbiased tensor-based morphometry: improved robustness and sample size estimates for Alzheimer's disease clinical trials. *NeuroImage*. 2013;66:648-661.
29. Gutman BA, Hua X, Rajagopalan P, et al. Maximizing power to track Alzheimer's disease and MCI progression by LDA-based weighting of longitudinal ventricular surface features. *NeuroImage*. 2013;70:386-401.
30. Ithapu VK, Singh V, Okonkwo OC, Chappell RJ, Dowling NM, Johnson SC. Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment. *Alzheimers Dement*. 2015;11:1489-1499.

How to cite this article: Chen K, Guo X, Pan R, et al. Limitations of clinical trial sample size estimate by subtraction of two measurements. *Statistics in Medicine*. 2021;1-11. doi: 10.1002/sim.9244