Alzheimer's & Dementia: Translational Research & Clinical Interventions 5 (2019) 308-318

Featured Article

The relative efficiency of time-to-progression and continuous measures of cognition in presymptomatic Alzheimer's disease

Dan Li[a], Samuel Iddi[a,b], Paul S. Aisen[a], Wesley K. Thompson[c], Michael C. Donohue[a,*], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] *Alzheimer's Therapeutic Research Institute, Keck School of Medicine, University of Southern California, San Diego, CA, USA*
[b] *Department of Statistics, University of Ghana, Legon-Accra, Ghana*
[c] *Department of Psychiatry, University of California, San Diego, CA, USA*

Abstract **Introduction:** Clinical trials on preclinical Alzheimer's disease are challenging because of the slow rate of disease progression. We use a simulation study to demonstrate that models of repeated cognitive assessments detect treatment effects more efficiently than models of time to progression.

Methods: Multivariate continuous data are simulated from a Bayesian joint mixed-effects model fit to data from the Alzheimer's Disease Neuroimaging Initiative. Simulated progression events are algorithmically derived from the continuous assessments using a random forest model fit to the same data.

Results: We find that power is approximately doubled with models of repeated continuous outcomes compared with the time-to-progression analysis. The simulations also demonstrate that a plausible informative missing data pattern can induce a bias that inflates treatment effects, yet 5% type I error is maintained.

Discussion: Given the relative inefficiency of time to progression, it should be avoided as a primary analysis approach in clinical trials of preclinical Alzheimer's disease.

Keywords: Clinical trial simulations; Alzheimer's disease; Cox proportional hazards model; Longitudinal data; Mixed model of repeated measures (MMRM); Statistical power; Common close design; Bayesian joint mixed-effect model

1. Introduction

Presymptomatic (or preclinical) Alzheimer's disease (PAD) is defined by evidence of abnormal levels of fibrillar amyloid beta (Aβ) in brain as measured by positron emission tomography brain scan or cerebrospinal fluid (CSF) assay [1]. Clinical trials have been initiated in this early phase of disease with the hope that, as in other diseases, early interventions will be more successful in slowing progression [2–4].

In PAD, progression is typically measured by continuous assessments such as the Preclinical Alzheimer's Cognitive Composite (PACC), a cognitive performance assessment sensitive to amyloid-related decline [5]. An alternative measure of progression is transition from normal cognition to mild cognitive impairment (MCI). The diagnosis of MCI is not algorithmic. It is based on an expert clinician's subjective impression of clinical tests and interviews with participants or study partners. In contrast to cancer progression or death, the cognitive diagnosis (normal or MCI) can vary from one clinician to the next or from one study visit to the next. In a multicenter study, the diagnosis made by a clinician at a trial performance site may be confirmed by experts centrally based on review of assessments without the benefit of direct in-person assessment.

Conflicts of interest: The authors declare no potential conflicts of interest.

[1] 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

*Corresponding author. Tel.: +1 858 964 0790; Fax: +1 858 622 1904.
E-mail address: mdonohue@usc.edu

https://doi.org/10.1016/j.trci.2019.04.004
2352-8737/

Table 1
Descriptive statistics by baseline diagnosis, normal cognition (NC), and subjective memory concern (SMC) for the preclinical Alzheimer's disease population in the Alzheimer's Disease Neuroimaging Initiative

Variable	NC (N = 120)	SMC (N = 43)	Total (N = 163)
Age	75.21 (5.83)	72.77 (5.78)	74.57 (5.90)
APOE ε4 alleles			
0	52 (43)	23 (53)	75 (46)
≥1	111 (57)	140 (47)	88 (54)
ADAS delayed word recall	2.96 (1.79)	3.00 (2.08)	2.97 (1.86)
Logical memory—delayed recall	13.11 (3.15)	12.63 (3.19)	12.98 (3.16)
Trails B	93.40 (48.90)	89.10 (32.00)	92.30 (45.00)
MMSE	29.11 (1.13)	29.09 (0.89)	29.10 (1.07)
Category fluency (animals)	20.72 (5.32)	19.72 (5.60)	20.45 (5.40)
CDRSB			
0	111 (92)	36 (84)	147 (90)
0.5	8 (7)	7 (16)	15 (9)
1	1 (1)	0 (0)	1 (1)
FAQ			
0	108 (90)	32 (74)	140 (86)
1	7 (6)	8 (19)	15 (9)
2	2 (2)	0 (0)	2 (1)
3	2 (2)	3 (7)	5 (3)
5	1 (1)	0 (0)	1 (1)

NOTE. Values are given as count (%) or mean (SD).

Abbreviations: ADAS, Alzheimer's Disease Assessment Scale; APOE, apolipoprotein E; MMSE, Mini-Mental State Examination; CDRSB, Clinical Dementia Rating—Sum of Boxes; FAQ, functional assessment questionnaire; SD, standard deviation.

Some researchers prefer the inherent clinical meaningfulness of time-to-MCI analysis. Undoubtedly, for a given subject, a transition from normal cognition to MCI is more clinically meaningful than a point change in a continuous cognitive performance measure. However, in a clinical trial, we are still left to determine how large a randomized group difference in the rate of, or delay in, a clinically meaningful event is itself clinically meaningful.

The typical Alzheimer's clinical trial assesses cognition at clinic visits conducted every three or six months. With a continuous outcome, the primary contrast is estimated at the last scheduled visit, at approximately 4.5 years. Proponents of time to progression argue that the endpoint allows for a *common close design*, similar to oncology studies, in which follow-up can continue until the last subject enrolled reaches the 4.5-year visit. The Cox Proportional Hazards model [6] admits data collected under such a design. Linear mixed-effects models can also admit data from a common close design, but assumptions about the mean trend (e.g., quadratic time trends) are necessary, similar to the proportional hazards assumption.

Some related work has demonstrated the advantages of analyzing continuous outcomes, when available, over time-to-event outcomes in other contexts. Donohue et al. [7] reviewed the literature and provided an analytic demonstration that, under general conditions, a mixed-effect model comparison of rate of change on a continuous outcome is effectively always more powerful than an analysis of time to threshold. The authors also conducted simulations based on Alzheimer's Disease Neuroimaging Initiative (ADNI) MCI subjects and demonstrated that the marginal linear model and linear mixed models are more robust and efficient than the Cox model of time from MCI to dementia.

Our goal is to extend our earlier work in the MCI population [7] to the earlier biomarker-defined PAD population. Specifically, we aim to compare the performance of models of repeated measures of the PACC versus time to progression when evaluating treatment effects in randomized trials and to assess bias due to informative missingness. We also compare the common close design and the fixed follow-up design. We apply the mixed models of repeated measures (MMRMs) [8] for the analysis of change in the PACC score. Constrained longitudinal data analysis (cLDA) models [9] are also used to model the PACC scores, treating time as a continuous variable. Cox proportional hazards model is applied to the time-to-event endpoint.

2. Data

ADNI is a prospective observational cohort study, led by principal investigator Michael W. Weiner, MD, which is tracking cognitive, imaging, and biofluid markers of Alzheimer's in volunteers diagnosed as cognitively normal (CN), with subjective memory concern, MCI, and mild-to-moderate dementia. To simulate both longitudinal continuous markers and time to MCI for a PAD clinical trial, we first model the disease markers and clinical diagnosis using data from PAD ADNI participants. The PAD population is defined by a diagnosis of CN or subjective memory concern at baseline and florbetapir positron emission tomography standardized uptake value ratio above 1.11 [10] or CSF Aβ below 950.6 pg/ml. The CSF threshold of 950.6 pg/ml

Table 2
Missing data patterns assumed in simulations

Scenario		Missing data rate		
Group	Treatment	Perceived inefficacy	Intolerability	Completely at random
Active	Ineffective	15%	10%	5% per year
Active	Effective	8%	10%	5% per year
Placebo	Not applicable	15%	0%	5% per year

Participants having intolerability are simulated to drop out at month six, and those perceiving inefficacy drop out at twelve months.

was selected because it yields the same proportion of PAD as the 1.11 standardized uptake value ratio threshold. Follow-up observation reports, including a site clinician's diagnosis of CN, MCI, or dementia, are collected every three, six, or 12 months. For more information on the study design of ADNI, including protocols, see adni.loni.usc.edu.

Sensitive tests of cognition may show changes in PAD many years before the onset of functional decline [5,11]. In this work, we focus on the following seven cognitive outcomes in the PAD population:

1. Alzheimer's Disease Assessment Scale delayed word recall (ADASDWR) [12],
2. Logical memory paragraph recall (LogMem) [13],
3. Trail making test part B (Trails B) [14],
4. Mini-Mental State Examination (MMSE) [15],
5. Category fluency—animals,
6. Clinical Dementia Rating—Sum of Boxes (CDRSB) [16], and
7. Functional assessment questionnaire (FAQ).

Baseline covariates considered include age and carriage of an apolipoprotein E4 (APOE $\varepsilon4$) allele. The PAD population includes a total of $N = 163$ individuals, in which $N = 39$ (23.9%) were observed to progress to MCI over a median follow-up time of 4.0 years (interquartile range: 2.1 to 5.6 years; maximum: 11.5 years). Baseline characteristics of the modeled PAD cohort are presented in Table 1.

3. Methods

3.1. Joint mixed-effects model for longitudinal data

To derive a model to simulate plausible data, we first fit a model to observed ADNI data. We apply a joint (or multivariate) mixed-effects model (JMM) to simultaneously model continuous longitudinal data for disease markers in the PAD population. The model respects the within-subject correlation over time and among the battery outcomes.

Linear mixed-effects models are commonly used to model continuous longitudinal data. The multivariate mixed-effects model is specified as $y_{ijk} = \mathbf{x}'_{ijk}\boldsymbol{\beta}_k + b_{0ik} + b_{1ik}\, t_{ijk} + \varepsilon_{ijk}$ for subject i, time j, and outcome k, where $\boldsymbol{\beta}_k$ are fixed-effect regression coefficients, and b_{0ik} and b_{1ik} are the subject- and outcome-specific random intercept and slope. The random effects are assumed to follow a multivariate Gaussian distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, with dimension $2p$, that is, $(b_{0i1},\cdots,b_{0ip},b_{1i1},\cdots,b_{1ip})' \sim N(0,\boldsymbol{\Sigma})$. The model with multivariate random effects has the advantage of reflecting the dependency within subjects and among outcomes. The $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma_k^2)$ is the residual error.

Because the outcomes are on different scales, we transform the raw outcome measures into a quantile scale ranging from 0 to 1 (least impaired to most severe dementia). Quantiles are calculated using the empirical cumulative distribution function using weights that are inversely proportional to the number of observations from each diagnostic category for each outcome. The quantiles were then transformed by the inverse Gaussian quantile function, resulting in an approximate Z-score before submitting to the model. When simulating data from these models, the simulated Z-scores can then be transformed back to the original scale, which can be integer valued.

Bayesian estimation is performed via Markov Chain Monte Carlo (MCMC) sampling using the stan_mvmer function in R package Rstanarm [17].

3.2. Random forest algorithm for diagnosis of MCI

To simulate a clinician's diagnosis of MCI or dementia, we first use ADNI data to learn an algorithm to approximate this decision. The random forest algorithm [18] is an ensemble learning method for classification and regression. In our application, clinician diagnosis of normal cognition versus MCI or dementia is the binary outcome variable, and the seven continuous markers, age, and education are the predictors. The model is fit using the R package randomForest [19]. The fitted model is then applied to simulated continuous outcomes to predict a clinician's diagnosis.

3.3. Competing clinical trial models for continuous and time-to-event outcomes in simulation study

The simulated treatment effect on time to progression is modeled by the Cox proportional hazards model. For the PACC, we consider MMRM and the cLDA proposed by Liang and Zeger [9]. Similar to most likelihood-based approaches for longitudinal data, all three models assume any missing data are missing at random (MAR).

The version of the PACC used in the study is a composite of four assessments: ADASDWR, LogMem, log transformation of Trails B, and MMSE. Each of the four component scores is first centered by subtracting the baseline sample mean and then divided by the baseline sample standard deviation of that component, to form standardized Z scores. These Z scores are averaged to form the composite.

The MMRM models treat change from baseline in the PACC score as the outcome and baseline PACC as a predictor. It treats time as a categorical variable, which allows general mean trends in each group. MMRM has been extensively used for testing treatment effects at specific time points in clinical trials because participants are often evaluated at a fixed and relatively small number of time points [20]. In our simulation study, the within-subject dependence is modeled by a first-order autoregressive covariance structure.

We also explore models that treat time as a continuous variable. In cLDA, the baseline outcome is treated as a response variable rather than a covariate, and the two randomized groups are constrained to have the same mean at baseline [21,22]. We explore models with linear or quadratic time trends for each group.

3.4. Simulation setup

We conducted a simulation study to evaluate the performance of the competing models described in Section 3.3. In each of 1000 simulated clinical trials with visits every 6 months from 0 to 8 years, a total of 1000 and 1500 patients are, respectively, randomized to either treatment or placebo in 1:1 ratio. We also assume the proportion of MCI progressors to be 24% (based on ADNI data, as noted previously).

For the placebo group, no changes will be made to the JMM fit to ADNI. For the treatment group, we will impose large (40% improvement on rate of change over the control), moderate (30% improvement), small (20% improvement), and no (same as the control) treatment effects on all outcomes.

To simulate nonignorable missing data, three dropout categories are considered: intolerability, inefficacy, and missing completed at random (MCAR). Participants having intolerability or inefficacy drop out from the study immediately after six and twelve months, respectively. For MCAR, we assume linear attrition rate of 5% per year for both the treatment and placebo groups. The simulated dropout rates are described in Table 2.

To assess bias due to missing data, we simulate complete data for every subject. The complete data are appropriately censored for the analysis of "observed" data and left uncensored for analysis of the "complete" data. Completers and MCAR dropouts are assumed to have the same longitudinal mean profile within each treatment arm. Dropouts due to intolerability are simulated to have the expected benefit, on average, until dropout, followed by an "unobserved" benefit that is diminished by a factor of 15%. Dropouts due to inefficacy are simulated to have no benefit.

The four competing clinical trial models are MMRM, cLDA^1 (linear) and cLDA^2 (quadratic) for continuous PACC scores, and Cox for time to progression, with two baseline covariates namely age at baseline and carriage of the APOE ε4 allele. The Cox model will use all data observed up to 8 years until the last subject reaches the final scheduled visit under the common close design. We assume a linear enrollment rate such that enrollment is completed in 4 years and about half the subjects contribute "extra" common close follow-up in the 4.5- to 8-year range to the Cox model. The MMRM, cLDA^1, and cLDA^2 will only use data up to last scheduled visit, that is, from 0 to 4.5 years.

We focus on "treatment policy" estimands of interest. The estimand will be the difference between randomized groups in the intention-to-treat population in terms of (1) rate (hazard ratio) of progression to MCI/dementia (Cox); (2) group difference in PACC at the final study time point (MMRM and cLDA^1); or (3) area between mean PACC curves (cLDA^2). We show how to carry out the hypothesis test of case (III) in the Supplementary Material. Let Y_{ijk} denote the simulated PACC scores for subject i randomized to group j at time point k, where $i = 1,\cdots,n_j$, $j = D,P$, and $k = 1,\cdots,T$. And $k = 0$ represents the baseline time point, D is the treatment group, and P is the placebo group. Under MMRM and cLDA^1, for example, the objective is to estimate the between-treatment difference $\delta = \mu_P - \mu_D$, where $\mu_j = E(Y_{ijT} - Y_{ij0})$. A two-tailed test $H_0{:}\delta = 0$ versus $H_1{:}\delta \neq 0$ is carried out to evaluate whether treatment is different from placebo.

For each simulated data set, we apply all four competing models to calculate point estimates of δ using the observed data (i.e., δ_{obs}) and the complete data (i.e., δ_{comp}). For each model, "bias" is calculated as the median of the 1000 point estimates of δ_{obs} minus δ_{comp}; "bias in percent" is computed as the median of the 1000 point estimates of δ_{obs} minus δ_{comp} and then divided by δ_{comp}. The interquartiles Q_1 and Q_3 are also summarized.

In a real clinical trial, the endpoint is measured for completers but is missing for those who either drop out from the study either because of inefficacy or intolerability or those who remain in the study after initiating rescue medication. Mehrotra et al. [23] discussed that the commonly used MMRM with the embedded MAR assumption can deliver an exaggerated estimate of the aforementioned estimand of interest, in favor of the drug. This happens, in part, due to implicit imputation of an overly optimistic mean for dropouts in the treatment group. To remedy this, they proposed a formula-based two-step approach by treating the true endpoint distribution for treatment group as a mixture of distributions (one each for the completers and dropouts) rather than a single distribution. Their approach reduces the bias associated with the traditional MMRM while maintaining power. To increase the precision in estimating δ, we apply their method to MMRM, cLDA^1, and cLDA^2 models in the simulation study.

4. Results

4.1. JMM and random forest fit to ADNI data

We fit a JMM for PAD participants who were observed to progress to MCI and a separate JMM for those who did not progress. Seven outcome measures described in Section 2 are included in the model. Fixed-effect covariates for each outcome include age at baseline and carriage of the APOE ε4 allele. Three parallel Markov chains are run for 4000

Table 3
Posterior estimates (means and 95% CIs) of the fixed-effect covariates for the joint mixed-effect model fit to seven outcomes for stable and MCI progressor subpopulations

	Progressor (N = 39)	Stable (N = 124)
Parameter	Mean (95% CI	Mean (95% CI
ADAS delayed word recall		
Intercept	−8.244 (−15.39, −1.451)	−4.913 (−7.755, −2.003)
Year	0.330 (0.189, 0.464)	0.064 (0.021, 0.108)
Age	0.110 (0.021, 0.201)	0.062 (0.023, 0.100)
APOE ε4	0.572 (−0.319, 1.437)	0.218 (−0.247, 0.670)
Logical memory paragraph recall		
Intercept	−6.897 (−15.425, 0.905)	−1.840 (−4.983, 1.350)
Year	0.261 (0.136, 0.395)	0.033 (−0.084, 0.016)
Age	0.096 (−0.005, 0.206)	0.020 (−0.023, 0.062)
APOE ε4	0.039 (−0.959, 1.099)	0.465 (−0.044, 0.985)
Trails B		
Intercept	−9.458 (−14.898, −3.918)	−6.364 (−9.020, −3.792)
Year	0.353 (0.252, 0.445)	0.022 (−0.028, 0.073)
Age	0.124 (0.051, 0.193)	0.084 (0.050, 0.119)
APOE ε4	0.141 (−0.540, 0.858)	0.622 (0.187, 1.087)
MMSE		
Intercept	0.852 (−191.780, 185.973)	−1.385 (−75.020, 72.568)
Year	0.009 (−3.918, 4.011)	0.022 (−2.590, 2.698)
Age	0.007 (−2.432, 2.436)	0.020 (−0.903, 0.944)
APOE ε4	0.040 (−1.116, 11.346)	0.115 (−5.683, 5.900)
Category fluency—animals		
Intercept	1.430 (−127.590, 130.195)	0.942 (−96.958, 98.426)
Year	0.047 (−2.910, 2.786)	0.025 (−3.399, 3.798)
Age	−0.009 (−1.658, 1.606)	−0.011 (−1.224, 1.211)
APOE ε4	0.036 (−8.234, 8.775)	−0.118 (−7.911, 7.920)
CDRSB		
Intercept	−6.537 (−364.967, 344.177)	1.094 (−82.421, 76.732)
Year	0.082 (−7.263, 6.390)	0.006 (−2.853, 2.947)
Age	0.081 (−4.230, 4.517)	−0.011 (−1.006, 1.027)
APOE ε4	−0.224 (−20.697, 19.566)	0.117 (−5.925, 6.358)
FAQ		
Intercept	3.458 (−380.068, 367.151)	0.261 (−32.960, 32.991)
Year	0.023 (−7.838, 7.140)	0.0007 (−1.1420, 1.1710)
Age	−0.002 (−4.487, 4.718)	−0.003 (−0.410, 0.449)
APOE ε4	0.343 (−22.127, 22.506)	0.014 (−2.667, 2.525)

Abbreviations: ADAS, Alzheimer's Disease Assessment Scale; APOE, apolipoprotein E; MMSE, Mini-Mental State Examination; CDRSB, Clinical Dementia Rating—Sum of Boxes; FAQ, functional assessment questionnaire; SD, standard deviation; CI, credible interval; MCI, mild cognitive impairment.

iterations, and the first 2000 warm-up iterations are discarded. Every fourth value of the remaining part of each chain is stored to reduce correlation, yielding a total of 1500 samples for posterior analysis. Table 3 shows the posterior means and 95% credible intervals of the covariate-effect parameters. Fig. 1 shows the subject-level observations and predictions according to time in years of the seven markers for all individuals, in which the blue and red curves are estimated using the locally estimated scatter plot smoother. The bottom panel shows that the predictions provide reasonable trends of the observations. The posterior estimates from JMM will be later used as the true parameter values to simulate the panel of continuous markers.

For the random forest, 500 trees are fitted, and the number of variables selected at each split is 3. The node impurity of each tree is measured by the Gini index. The results show that CDRSB, LogMem, and FAQ are three most important outcomes for determining the diagnosis of MCI. The model has a 6.19% out-of-bag error rate and 93.81% out-of-bag accuracy rate. Using the fitted random forest, the simulated cognitive status can be obtained from the simulated continuous markers. Fig. 2 shows the Kaplan-Meier estimated progression rate of the ADNI-PAD population (black solid line) along with the progression rate from one large simulated placebo group (red dots). The simulated progression yields closer concordance with the Kaplan-Meier estimates at the earlier stage. Although we observe discrepancies between the two lines in the middle and the right tail, the red line still lies within the 95% confidence intervals. Both the subject-level trajectories and the progression rate illustrate that the simulated data plausibly mimic the observed data.

4.2. Simulation results

Fig. 3 shows the results of one simulated clinical trial with a 20% treatment effect and sample size n = 1000. The figure

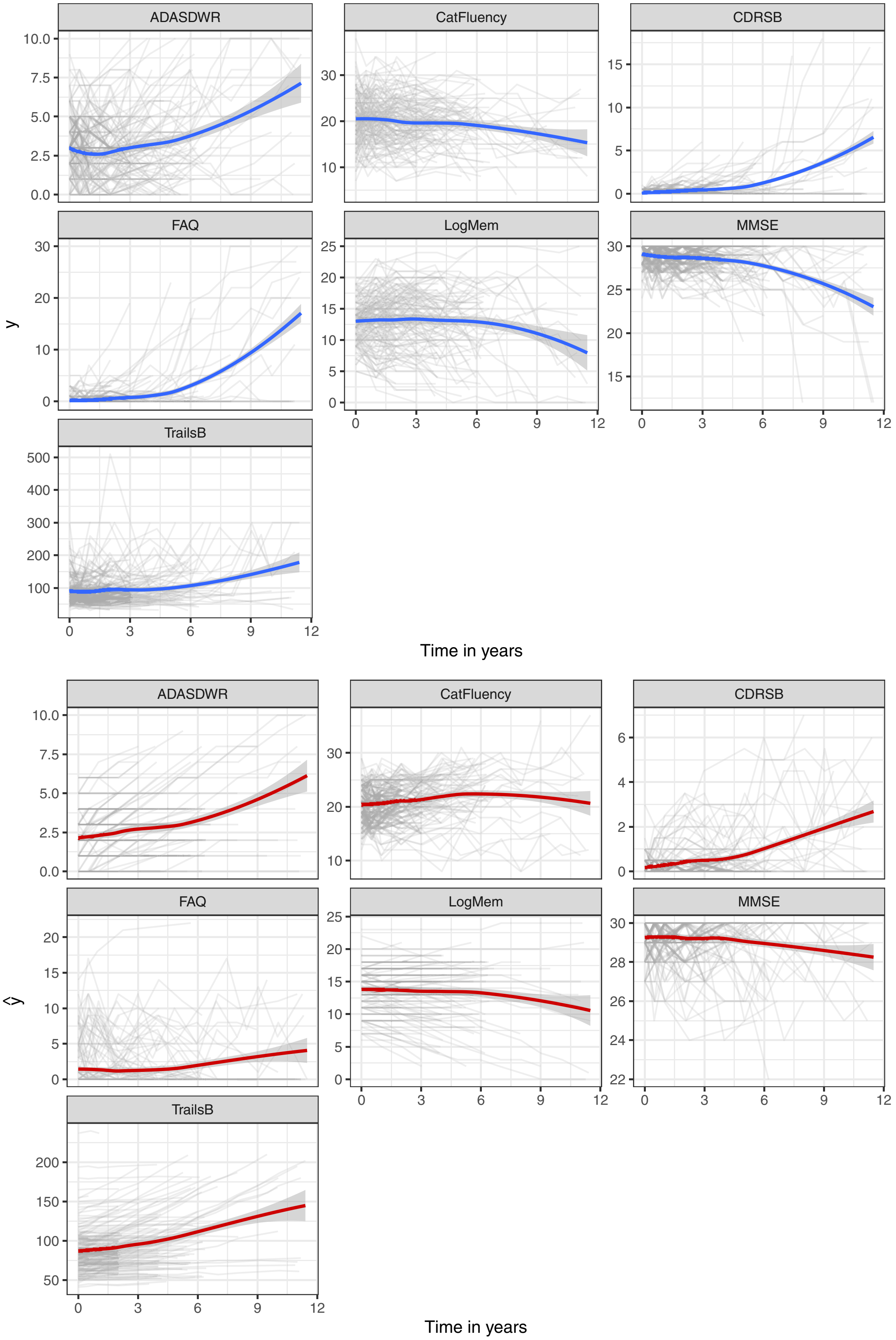

Fig. 1. Observed (upper panel) and predicted (lower panel) longitudinal profiles of the seven markers for all individuals. Bold lines are locally estimated scatter plot smoother. Abbreviations: ADASDWR, Alzheimer's Disease Assessment Scale delayed word recall; MMSE, Mini-Mental State Examination; CDRSB, Clinical Dementia Rating—Sum of Boxes; FAQ, functional assessment questionnaire.

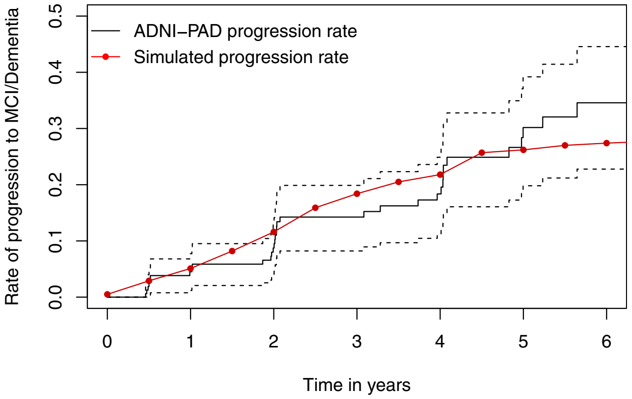

Fig. 2. Kaplan-Meier estimated rate of progression to MCI or dementia. Abbreviations: MCI, mild cognitive impairment; ADNI-PAD, Alzheimer's Disease Neuroimaging Initiative—presymptomatic (or preclinical) Alzheimer's disease.

illustrates the group trends obtained by fitting the four different models.

Simulated power and type I error are summarized in Table 4. Under the null hypothesis (no treatment effect), the MMRM exhibits smaller-than-expected type I error (about 2%), whereas the other models are closer to the expected 5% error rate. The Cox model consistently exhibits the weakest power of the four models. MMRM has the next best performance, followed by the quadratic (cLDA^2) and linear (cLDA^1) models. For example, with a trial of sample size $N = 1000$ subjects of drug with a 30% treatment effect, the simulated power is 33% for Cox, 79% for MMRM, 86% for cLDA^2, and 96% for cLDA^1. In comparing analysis of complete versus observed data, it seems the missing data do not increase type I error, but they do inflate power. This suggests the bias is only an issue with an effective drug, in which case the effectiveness might appear inflated. Fig. 4 shows the powers in all scenarios.

Tables 5 and 6 further examine the bias induced by the missing data pattern. The tables summarize the median and interquartile ranges (Q_1, Q_3) of the bias on the PACC scale (Table 5) and as a percent of effect seen in complete data [6]. The Cox model seems to have smaller bias with 20% treatment effect, but as the treatment grows, the bias is comparable for all models. The method proposed by

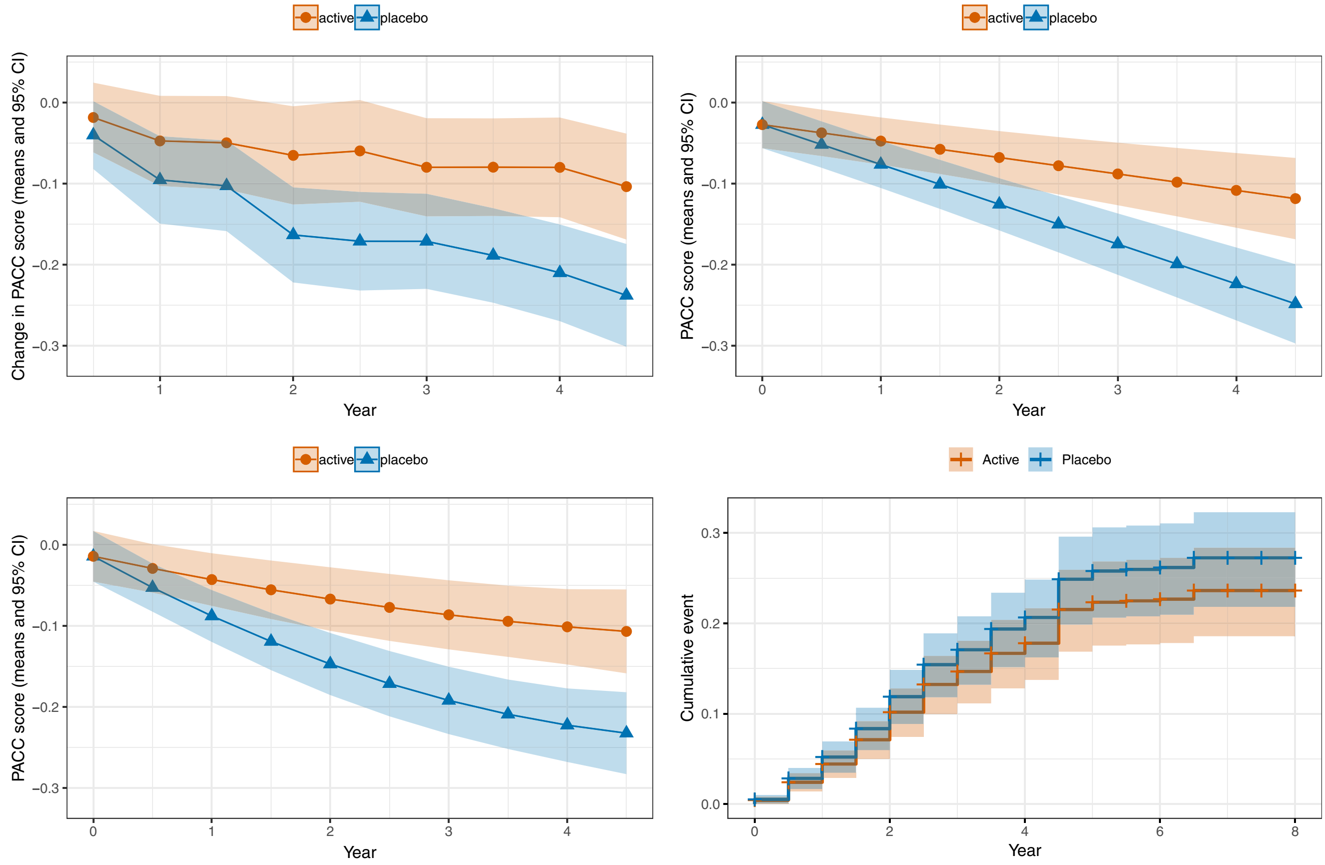

Fig. 3. Results of one simulated clinical trial with 20% treatment effect from (A) analysis of change from baseline using a categorical time MMRM of the PACC; (B) a cLDA model of PACC with linear time trends; (C) a cLDA model of PACC with quadratic time effects; and (D) Kaplan-Meier curves comparing the time-to-progression to mild cognitive impairment or dementia for the two groups. Abbreviations: MMRM, mixed models of repeated measures; PACC, Preclinical Alzheimer's Cognitive Composite; cLDA, constrained longitudinal data analysis.

Table 4
Power and type I error from 1000 simulated clinical trials

Sample size	Treatment	Observed data MMRM	Observed data cLDA[1]	Observed data cLDA[2]	Observed data Cox PH	Completed data MMRM	Completed data cLDA[1]	Completed data cLDA[2]	Completed data Cox PH
1000	0%	0.021	0.051	0.053	0.040	0.027	0.049	0.057	0.046
	20%	0.404	0.702	0.502	0.188	0.298	0.564	0.402	0.159
	30%	0.794	0.957	0.856	0.322	0.666	0.897	0.751	0.274
	40%	0.970	0.999	0.981	0.496	0.907	0.990	0.947	0.425
1500	0%	0.024	0.042	0.054	0.058	0.014	0.048	0.051	0.055
	20%	0.560	0.843	0.660	0.261	0.454	0.722	0.550	0.232
	30%	0.927	0.996	0.954	0.452	0.847	0.973	0.907	0.392
	40%	1.000	1.000	1.000	0.653	0.994	1.000	0.996	0.573

NOTE. The rows with 0% treatment effect simulate the type I error, which we expect to be near 5%.
Abbreviations: MMRM, mixed models of repeated measures; cLDA, constrained longitudinal data analysis; PH, proportional hazards.

Mehrotra et al. [23] successfully shrinks the magnitude of bias, for example, from 27% in favor of treatment to −4.4% in favor of placebo for MMRM with 20% treatment effect. The method appears to overcorrect the bias in favor of placebo in these simulations.

5. Discussion

We use Bayesian JMM fit using ADNI data to simulate correlated longitudinal data that might plausibly arise in a PAD clinical trial. We used a random forest algorithm, also fit using ADNI, to algorithmically diagnose MCI in the simulated data so that we could compare models of the PACC to the Cox model of time to progression. The models of PACC consistently provide at least twice the power of the Cox model even when the Cox model has the benefit of considerably more follow-up visits under a common close design. Given this inefficiency, the time-to-progression analysis should be avoided in PAD.

Some might still argue that the clinical meaningfulness of the time to progression is worth the cost of a larger, longer trial. However, given that the random forest provided a purely algorithmic diagnosis with 93.81% out-of-bag accuracy, it is suggested that there is minimal additional value in the diagnosis. And again, while the progression outcome

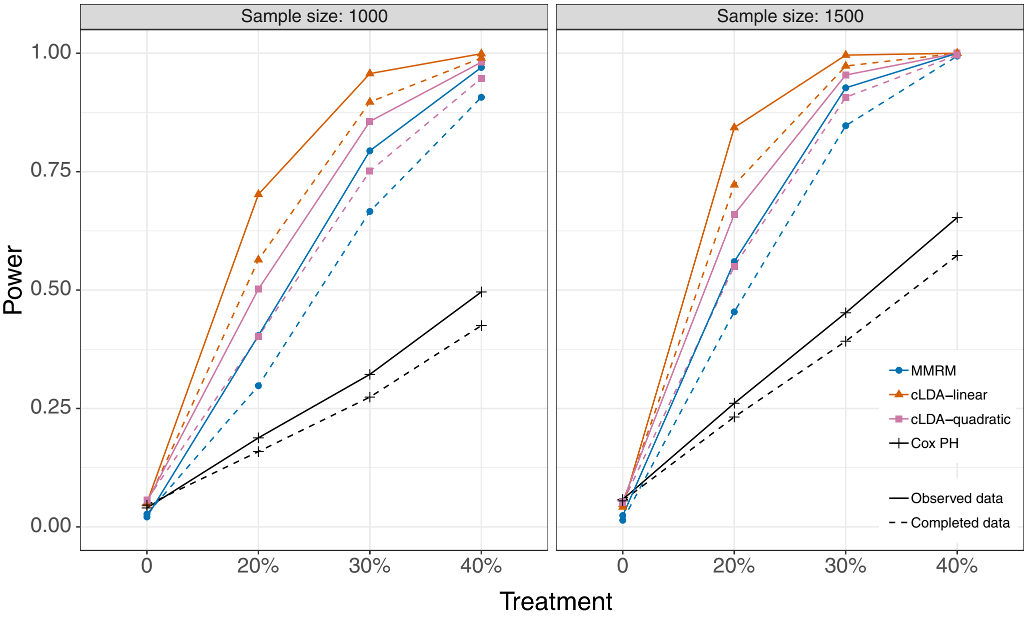

Fig. 4. Statistical power for the MMRM, cLDA, and Cox proportional hazards model for treatment effects 0% (type I error), 20%, 30%, and 40% for sample sizes of n = 1000 (left panel) and n = 1500 (right panel). Solid lines indicate power estimates for data observed after simulated nonignorable missingness, and dashed lines indicate power that would be achieved with complete data (including observations that would be unobserved in reality). The observed data show greater power with fewer observations because the nonignorable missingness induces a bias in favor of the treatment. Abbreviations: MMRM, mixed models of repeated measures; cLDA, constrained longitudinal data analysis; PH, proportional hazards.

Table 5
Bias of the treatment effect due to missingness

		20%	30%	40%
Sample size	Analysis method	Median (Q_1, Q_3)	Median (Q_1, Q_3)	Median (Q_1, Q_3)
1000	MMRM	0.018 (0.006, 0.031)	0.028 (0.015, 0.040)	0.037 (0.024, 0.049)
	cLDA^1	0.019 (0.009, 0.029)	0.028 (0.018, 0.038)	0.038 (0.028, 0.048)
	cLDA^2	0.038 (0.011, 0.065)	0.058 (0.030, 0.084)	0.077 (0.050, 0.104)
	Cox PH	−0.033 (−0.074, 0.010)	−0.045 (−0.086, −0.001)	−0.059 (−0.102, −0.017)
	MMRM-Mehrotra	−0.001 (−0.015, 0.012)	−0.002 (−0.016, 0.011)	−0.003 (−0.017, 0.010)
	cLDA^1-Mehrotra	−0.001 (−0.011, 0.008)	−0.001 (−0.012, 0.007)	−0.003 (−0.012, 0.007)
	cLDA^2-Mehrotra	−0.006 (−0.034, 0.022)	−0.010 (−0.038, 0.018)	−0.014 (−0.042, 0.014)
1500	MMRM	0.018 (0.006, 0.028)	0.027 (0.016, 0.037)	0.036 (0.025, 0.047)
	cLDA^1	0.018 (0.010, 0.026)	0.027 (0.019, 0.036)	0.037 (0.028, 0.045)
	cLDA^2	0.037 (0.013, 0.061)	0.056 (0.032, 0.080)	0.075 (0.052, 0.099)
	Cox PH	−0.028 (−0.064, 0.005)	−0.042 (−0.076, −0.009)	−0.055 (−0.090, −0.021)
	MMRM-Mehrotra	−0.002 (−0.012, 0.009)	−0.003 (−0.013, 0.008)	−0.004 (−0.015, 0.007)
	cLDA^1-Mehrotra	−0.001 (−0.009, 0.006)	−0.002 (−0.010, 0.005)	−0.003 (−0.011, 0.004)
	cLDA^2-Mehrotra	−0.008 (−0.028, 0.015)	−0.012 (−0.032, 0.011)	−0.016 (−0.035, 0.007)

Abbreviations: MMRM, mixed models of repeated measures; cLDA, constrained longitudinal data analysis; PH, proportional hazards.

is more qualitative than the PACC on the subject level, the group-level result is still quantitative (e.g., a hazard ratio) and requires additional interpretation to assign clinical meaning.

One might also argue that clinical diagnosis cannot be adequately modeled algorithmically using trial data. That is, clinical assessment and diagnosis by a trial-site clinician may consider information not captured by trial measures. However, the cognitive, clinical, and functional assessments are designed to capture the relevant information, and clinicians generally rely on similar information obtained through less structured assessments. It seems questionable that a site clinician will gain much reliable information beyond the assessments; indeed, this is the justification for central expert panel adjudication of site diagnoses.

The Bayesian joint models are well suited to simulating plausible panels of correlated longitudinal data necessary to compare clinical trial designs. This approach could be useful in many other contexts where one is interested in a fair comparison of different outcome measures, different combinations of correlated outcomes, or different models of treatment effect. Simulations that ignore the correlations among important outcomes will likely not provide reliable comparisons.

All the models considered were susceptible to bias induced by a plausible missing data pattern. However, this bias seemed to only affect scenarios with an effective

Table 6
Bias in percent (%) of the treatment effect due to missingness based on 1000 simulated trials for the given sample size, treatment effect, and analysis method

		20%	30%	40%
Sample size	Analysis method	Median (Q_1, Q_3)	Median (Q_1, Q_3)	Median (Q_1, Q_3)
1000	MMRM	27.1 (7.0, 52.3)	29.9 (16.3, 46.8)	29.6 (19.4, 42.3)
	cLDA^1	29.6 (12.4, 51.9)	29.8 (18.9, 43.7)	29.7 (21.4, 39.7)
	cLDA^2	24.5 (5.5, 50.2)	26.5 (13.7, 42.6)	26.2 (16.5, 37.9)
	Cox PH	17.4 (−16.1, 55.0)	22.2 (−4.5, 52.7)	25.5 (5.2, 50.4)
	MMRM-Mehrotra	−4.4 (−23.2, 20.6)	−2.9 (−15.9, 13.3)	−2.8 (−12.7, 8.6)
	cLDA^1-Mehrotra	−1.7 (−16.2, 15.4)	−1.7 (−11.3, 9.1)	−2.0 (−9.2, 5.7)
	cLDA^2-Mehrotra	−6.0 (−21.2, 15.9)	−4.5 (−15.5, 9.4)	−4.7 (−13.0, 5.2)
1500	MMRM	27.5 (9.7, 52.8)	28.2 (16.6, 43.3)	28.3 (19.6, 39.3)
	cLDA^1	29.1 (15.7, 48.4)	29.2 (19.9, 40.9)	29.3 (22.2, 37.8)
	cLDA^2	24.8 (8.8, 45.6)	25.4 (15.2, 38.2)	25.5 (17.8, 34.6)
	Cox PH	18.0 (−8.2, 46.9)	22.7 (3, 46.3)	24.3 (8.6, 44.6)
	MMRM-Mehrotra	−3.0 (−19.4, 17.6)	−3.0 (−13.8, 9.7)	−3.1 (−11.2, 6.2)
	cLDA^1-Mehrotra	−2.1 (−13.5, 11.4)	−2.3 (−9.8, 5.7)	−2.4 (−7.9, 3.5)
	cLDA^2-Mehrotra	−6.1 (−18.8, 12.7)	−5.5 (−13.9, 5.7)	−5.5 (−11.5, 2.8)

Abbreviations: MMRM, mixed models of repeated measures; cLDA, constrained longitudinal data analysis; PH, proportional hazards.

treatment and did not inflate type I error under the null hypothesis. The Mehrotra method shows promise in correcting this bias, but it might overcorrect in favor of placebo, and it would be impossible to detect this overcorrection in practice. Given that type I error is not inflated, we are inclined to suggest no change to the status quo approach in which the primary analysis is based on likelihood-based methods which are robust to MAR and applying appropriate MNAR sensitivity analyses such as the delta method [24].

Acknowledgments

The authors are grateful to the ADNI study volunteers and their families.

This work was supported by National Institute on Aging grant R01-AG049750. Data collection and sharing for this project was funded by the 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 organizations: 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 Inc. The Canadian Institutes of Health Research provided 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.

Supplementary Data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.trci.2019.04.004.

RESEARCH IN CONTEXT

1. Systematic review: Donohue et al. (2011) explored the relative efficiency of time-to-event versus continuous outcomes, reviewed the literature, derived an analytic calculation of the relative efficiency, and simulated trials in mild cognitive impairment (MCI) populations. The current work extends this earlier work to the preclinical Alzheimer's population. We reviewed trials in preclinical Alzheimer's disease on clinicaltrials.gov and found that most use a continuous primary outcome but at least one is using time to MCI.
2. Interpretation: The simulation study confirms that continuous outcomes provide about twice the statistical power to detect treatment effects compared with time to MCI. Plausible scenarios of attrition due to intolerability and perceived lack of efficacy inflate estimates of treatment benefit, although type I error is not inflated.
3. Future directions: The novel simulation methodology using hierarchical Bayesian mixed-effect models of multiple outcomes and random forests can be used to optimize preclinical Alzheimer's clinical trial efficiency and power and to minimize bias.

References

[1] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Demen 2011;7:280–92.

[2] Sperling RA, Rentz DM, Johnson KA, Karlawish J, Donohue M, Salmon DP, et al. The A4 study: Stopping AD before symptoms begin? Sci Translational Med 2014;6:228fs13.

[3] ClinicalTrials.gov. An efficacy and safety study of atabecestat in participants who are asymptomatic at risk for developing Alzheimer's dementia (EARLY), Tech. rep., National Library of Medicine (US), Bethesda, MD (2015 Oct 6 - 2019 Jan 21), https://clinicaltrials.gov/ct2/show/NCT02569398. Accessed January 2019.

[4] Caputo A, Racine A, Paule I, Martens EP, Tariot P, Langbaum JB, et al. Rationale for selection of primary endpoints in the Alzheimer Prevention Initiative Generation study in cognitively healthy APOE4 homozygotes, Alzheimer's & Dementia. J Alzheimer's Assoc 2017; 13:P1452.

[5] Donohue MC, Sperling RA, Salmon DP, Rentz DM, Raman R, Thomas RG, et al. The preclinical Alzheimer cognitive composite: Measuring amyloid-related decline. JAMA Neurol 2014;71:961–70.

[6] Cox DR. Regression models and life-tables. J R Statis-tical Soc Ser B (Methodological) 1972;34:187–202.

[7] Donohue M, Gamst A, Thomas R, Xu R, Beckett L, Petersen R, et al. The relative efficiency of time-to-threshold and rate of change in longitudinal data. Contemp Clin Trials 2011; 32:685–93.

[8] Mallinckrodt CH, Sanger TM, Dube S, DeBrota DJ, Molenberghs G, Carroll a J, et al. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. Biol Psychiatry 2003; 53:754–60.

[9] Liang K-Y, Zeger SL. Longitudinal data analysis of continuous and discrete responses for pre-post designs, Sankhya. Indian J Stat Ser B 2000;62:134–48.

[10] Landau SM, Mintun MA, Joshi AD, Koeppe RA, Petersen RC, Aisen PS, et al. Amyloid deposition, hypometabolism, and longitudinal cognitive decline. Ann Neurol 2012;72:578–86.

[11] Donohue MC, Sperling RA, Petersen R, Sun C-K, Weiner MW, Aisen PS. Association between elevated brain amyloid and subsequent cognitive decline among cognitively normal persons. JAMA 2017; 317:2305–16.

[12] Mohs RC, Cohen L. Alzheimer's Disease Assessment Scale (ADAS). Psychopharmacol Bull 1988;24:627.

[13] Wechsler D. WMS-R: Wechsler Memory Scale-Revised, The Psychological Corporation, San Antonio, TX; 1987.

[14] Tombaugh TN. Trail Making Test A and B: Normative data stratified by age and education. Arch Clin Neuropsychol 2004;19:203–14.

[15] Folstein MF, Folstein SE, McHugh PR. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189–98.

[16] Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology 1993;43:2412–4.

[17] Goodrich B, Gabry J, Ali I, Brilleman S. Rstanarm: Bayesian applied regression modeling via Stan, R package version 2.17.4, http://mc-stan.org/, 2018. Accessed January 2019.

[18] Breiman L. Random forests. Machine Learn 2001;45:5–32.

[19] Liaw A, Wiener M. Classification and regression by randomForest. R News:18–22, https://CRAN.R-project.org/doc/Rnews/.

[20] Siddiqui O, Hung HJ, O'Neill R. MMRM vs. LOCF: A comprehensive comparison based on simulation study and 25 nda datasets. J Biopharm Stat 2009;19:227–46.

[21] Liu GF, Lu K, Mogg R, Mallick M, Mehrotra DV. Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? Stat Med 2009;28:2509–30.

[22] Lu K. On efficiency of constrained longitudinal data analysis versus longitudinal analysis of covariance. Biometrics 2010;66:891–6.

[23] Mehrotra DV, Liu F, Permutt T. Missing data in clinical trials: Control-based mean imputation and sensitivity analysis. Pharm Stat 2017; 16:378–92.

[24] Rubin DB. Formalizing subjective notions about the effect of nonrespondents in sample surveys. J Am Stat Assoc 1977;72:538–43.