USC University of Southern California

Alzheimer's Therapeutic Research Institute

The relative efficiency of time-to-progression & continuous measures of cognition in Preclinical Alzheimer's

Dan Li[1], Samuel Iddi[1], Paul S. Aisen[1], Wesley K. Thompson[2], and Michael C. Donohue[1]

[1] Alzheimer's Therapeutic Research Institute, University of Southern California; [2]University of California, San Diego

ACTC ALZHEIMER'S CLINICAL TRIALS CONSORTIUM

BACKGROUND

Clinical trials in Preclinical Alzheimer's Disease (PAD) are challenging due to the slow rate of disease progression. We use a simulation study to demonstrate that models of repeated, continuous cognitive assessments detect treatment effects more efficiently compared to models of time-to-progression to Mild Cognitive Impairment (MCI) or dementia. We also explore the bias induced by hypothetical non-ignorable missingness. This is an extension of our previous work demonstrating analysis of continuous assessment scores is more powerful than time-to-progression from MCI to dementia in clinical trials in populations with MCI.[1]

METHODS

Simulation model for placebo group*.** Multivariate continuous data are simulated from a Bayesian ***joint mixed effects model (JMM) trained on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The ADNI PAD population is defined by a diagnosis of Cognitively Normal (CN) or Significant Memory Concern (SMC) at baseline, and florbetapir positron emission tomography (PET) standardized uptake value ratio (SUVR) above 1.11 or cerebrospinal fluid CSF amyloid beta (Aβ) below 950.6 pg/ml. The CSF threshold of 950.6 pg/ml was selected to yield the same proportion of PAD as the 1.11 SUVR threshold. We fit the model to the following 7 outcomes:

1. ADAS Delayed Word Recall (ADAS-DWR)
2. Logical Memory Paragraph Recall (LogMem)
3. Trail Making Test Part B (Trails B)
4. Mini-Mental State Examination (MMSE)

} 4 components of the modified PACC[2]

5. Category Fluency - Animals
6. Clinical Dementia Rating - Sum of Boxes (CDRSB)
7. Functional Assessment Questionnaire (FAQ)

} 3 additional assessments for diagnosis MCI or dementia

The multivariate mixed-effects model is specified as

$$y_{ijk} = \mathrm{x}'_{ijk}\beta_k + b_{0ik} + b_{1ik}t_{ijk} + \varepsilon_{ijk}$$

for subject i, time j, and outcome k; where β_k are fixed-effect regression coefficients, 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 0 and variance-covariance matrix Σ with dimension $2p$. The model with multivariate random effects has the advantage of reflecting the dependency within subjects and among outcomes. The $\varepsilon_{ijk} \sim N(0, \sigma_k^2)$ are residual errors.

Prior to modelling, raw scores are transformed to quantiles, then to approximate z-scores using inverse Gaussian distribution. Z-scores simulated from the fitted model are then back-transformed to the raw scale. Baseline covariates included age and carriage of an apolipoportein E4 (*APOE*ε4) allele. Subject- and outcome-specific random effects are assumed to follow a multivariate Gaussian distribution. Progressors to MCI or dementia were modelled separately from non-progressors. Bayesian estimation is performed via Markov Chain Monte Carlo (MCMC) sampling using the `mvmer` function in `R` package `Rstanarm`.[3]

***Random forest algorithm for MCI*.** In order 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 is an ensemble learning method for classification and regression.[4] 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 fitted using the R package `randomForest`. The fitted model is then applied to simulated continuous outcomes to predict a clinician's diagnosis.

***Simulated clinical trials*.** We simulated total sample sizes of either N=1000 or N=1500 participants randomized to either treatment or placebo in a 1:1 ratio. Visits were simulated every 6 months from 0 to 8 years. We simulated 1000 trials for each scenario. For the placebo group, no changes will be made to the JMM fit to ADNI. We simulate 24% of subjects from the model fit to progressors (to match the progression rate observed in 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 null (same as the control) treatment effects on all 7 continuous assessments.

To simulate non-ignorable 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 6 and 12 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 1.

In order to assess bias due to missing data, we simulate complete data for every subject. The complete data is 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.

We consider two outcome measure: (1) the 4-item modified Preclinical Alzheimer's Cognitive Composite (PACC) or (2) time-to-MIC or dementia as determined by random forest applied to all 7 simulated assessments. The four competing clinical trial models are (1) Mixed Model of Repeated Measures (MMRM), (2) constrained Longitudinal Data Analysis-linear (cLDA-linear) and (3) cLDA-quadratic for continuous PACC scores; and (4) Cox for time-to-progression, with two baseline covariates: age at baseline and carriage of the APOEε4 allele. The Cox model will use all data observed out 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-linear and cLDA-quadratic only use data up to last scheduled visit, i.e., from 0 to 4.5 years.

The efficacy estimates will be the difference between randomized groups in the intention-to-treat population in terms of either: (1) Group difference in PACC at final study time point (MMRM and cLDA-linear); (2) Area between mean PACC curves (cLDA-quadratic); or (3) Rate (hazard ratio) of progression to MCI/Dementia (Cox).

METHODS (continued)

***Missing Data Bias Adjustment*.** Mehrotra, et al.[5] proposed a method for a bias adjustment to account for data missing not at random. They proposed a formula-based two-step approach which assumes the endpoint distribution for the treatment group to be a mixture of distributions (one each for the completers and dropouts). Separate models are fit to (1) all placebo, (2) all active, (3) active completers, and (4) active dropout groups. The model fit to all active, under the assumption of data Missing at Random (MAR), is adjusted using estimates from the models fit to subgroups. We apply their method to MMRM and cLDA models in the simulation study.

Scenario		Missing data rate		
Group	Efficacy	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

Table 1. Missing data patterns assumed in simulations. Participants experiencing intolerability are simulated to drop out at month six, and those perceiving inefficacy drop out at twelve months.

RESULTS

Figure 1 shows the subject-level JMM fits over time for the seven modeled assessments. The random forest found CDRSB, LogMem and FAQ to be three most important assessments for determining the diagnosis of MCI. The model had a 6.19% out-of-bag error rate and 93.81% out-of-bag accuracy rate. Figure 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 mimics the observed data.

Figure 3 shows the results of one simulated clinical trial with a 20% treatment effect and sample size N=1000. The figure illustrates the group trends obtained by fitting the four different models. Simulated power and Type I error are summarized in Figure 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 expect 5% error rate. The linear cLDA model consistently exhibits the greatest power of the four models, followed by quadratic cLDA, MMRM, and Cox model. For example, with a trial of size N=1000 subjects of a drug with a 30% treatment effect, the simulated power is 96% for cLDA-linear, 86% for cLDA-quadratic, 79% for MMRM, and 33% for Cox model. In comparing analysis of complete versus observed data, it seems the missing data does not increase Type I error, but it does inflate power. This suggests the bias is only an issue with an effective drug, in which case the effectiveness might appear inflated.

Table 2 further summarizes the bias induced by the missing data pattern. The table summarize the bias as a percent of effect seen in complete data. 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 Mehrotra, et al. successfully shrinks the magnitude of bias, e.g. 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.

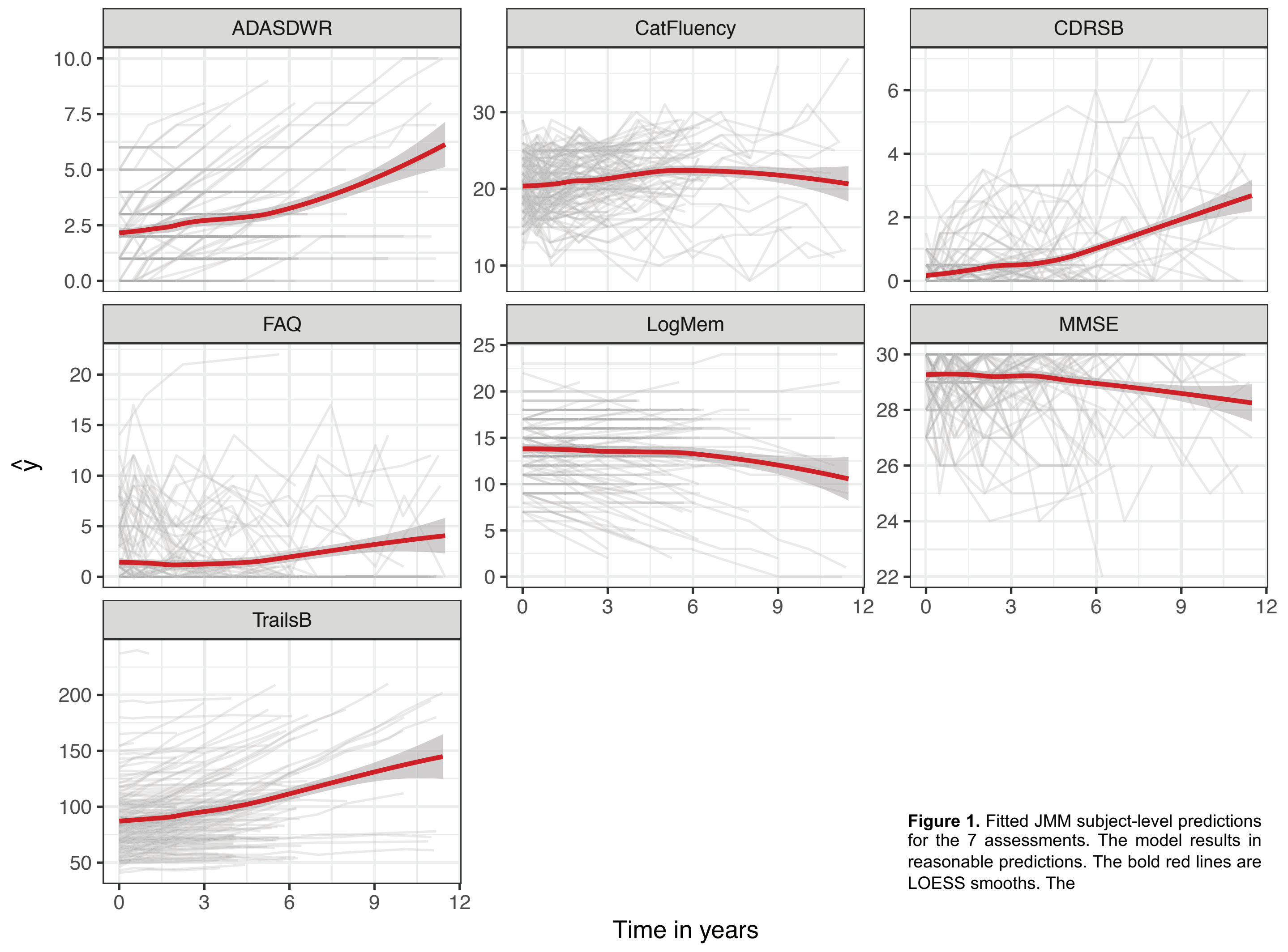

Figure 1. Fitted JMM subject-level predictions for the 7 assessments. The model results in reasonable predictions. The bold red lines are LOESS smooths. The

RESULTS (continued)

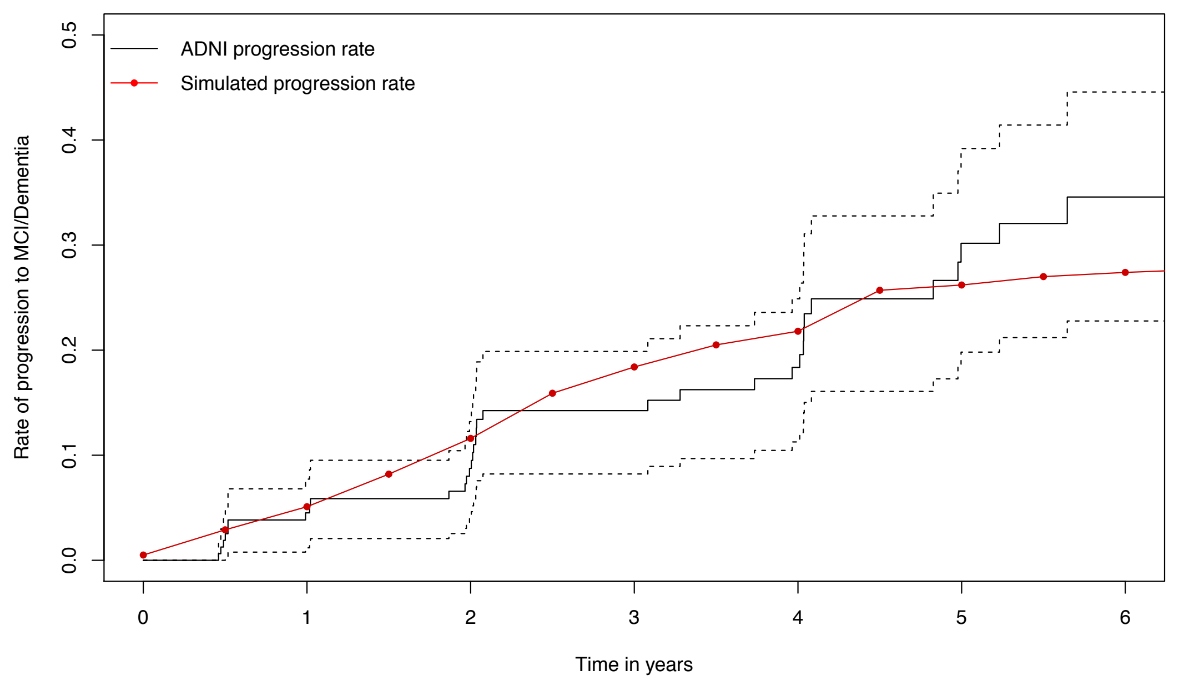

Figure 2. The progression rate in simulated data (red) was similar to that estimated in ADNI by Kaplan-Meier.

Sample size	Analysis Method	20% Effect		30% Effect		40% Effect	
		Median	(Q1, Q3)	Median	(Q1, Q3)	Median	(Q1, Q3)
1000	MMRM	27.1	(7.0, 52.3)	29.9	(16.3, 46.8)	29.6	(19.4, 42.3)
	cLDA1	29.6	(12.4, 51.9)	29.8	(18.9, 43.7)	29.7	(21.4, 39.7)
	cLDA2	24.5	(5.5, 50.2)	26.5	(13.7, 42.6)	26.2	(16.5, 37.9)
	CoxPH	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-L-Mehrotra	-1.7	(-16.2, 15.4)	-1.7	(-11.3, 9.1)	-2.0	(-9.2, 5.7)
	cLDA-Q-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)
	cLDA1	29.1	(15.7, 48.4)	29.2	(19.9, 40.9)	29.3	(22.2, 37.8)
	cLDA2	24.8	(8.8, 45.6)	25.4	(15.2, 38.2)	25.5	(17.8, 34.6)
	CoxPH	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-L-Mehrotra	-2.1	(-13.5, 11.4)	-2.3	(-9.8, 5.7)	-2.4	(-7.9, 3.5)
	cLDA-Q-Mehrotra	-6.1	(-18.8, 12.7)	-5.5	(-13.9, 5.7)	-5.5	(-11.5, 2.8)

Table 2. Bias induced by missing data as a percentage of the effect estimated from complete data. Median and interquartile range are based on 1000 simulated trials for the given sample size, treatment effect, and analysis method.

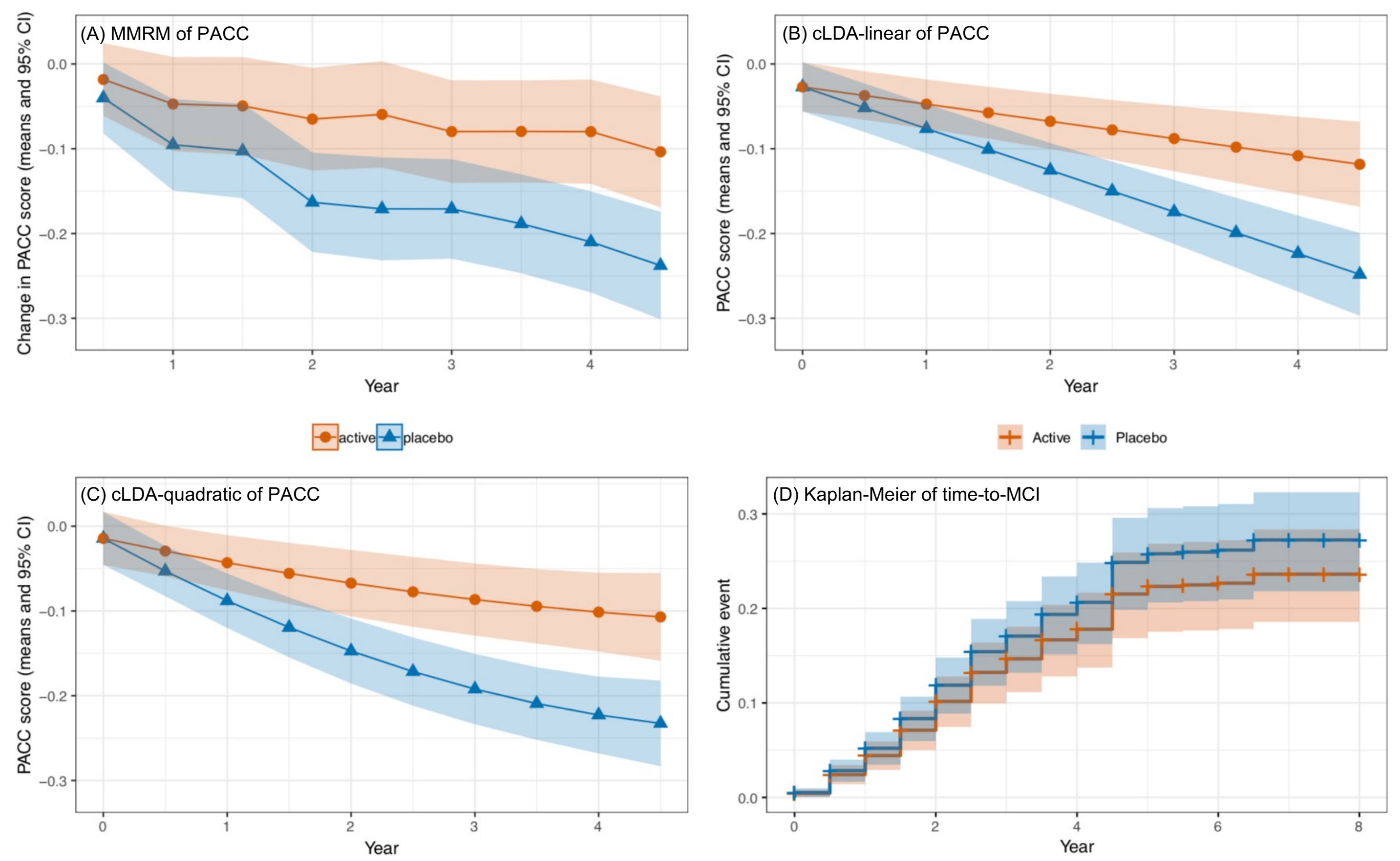

Figure 3. Results of one simulated clinical trial with 20% treatment effect and N=1000 from (a) analysis of change from baseline using a categorical time MMRM of the PACC; (b) a cLDA model of PACC with linear time trends; (c) a cLDA model of PACC with quadratic time effects; and (d) Kaplan-Meier curves comparing the time-to-progression to MCI or dementia.

RESULTS (continued)

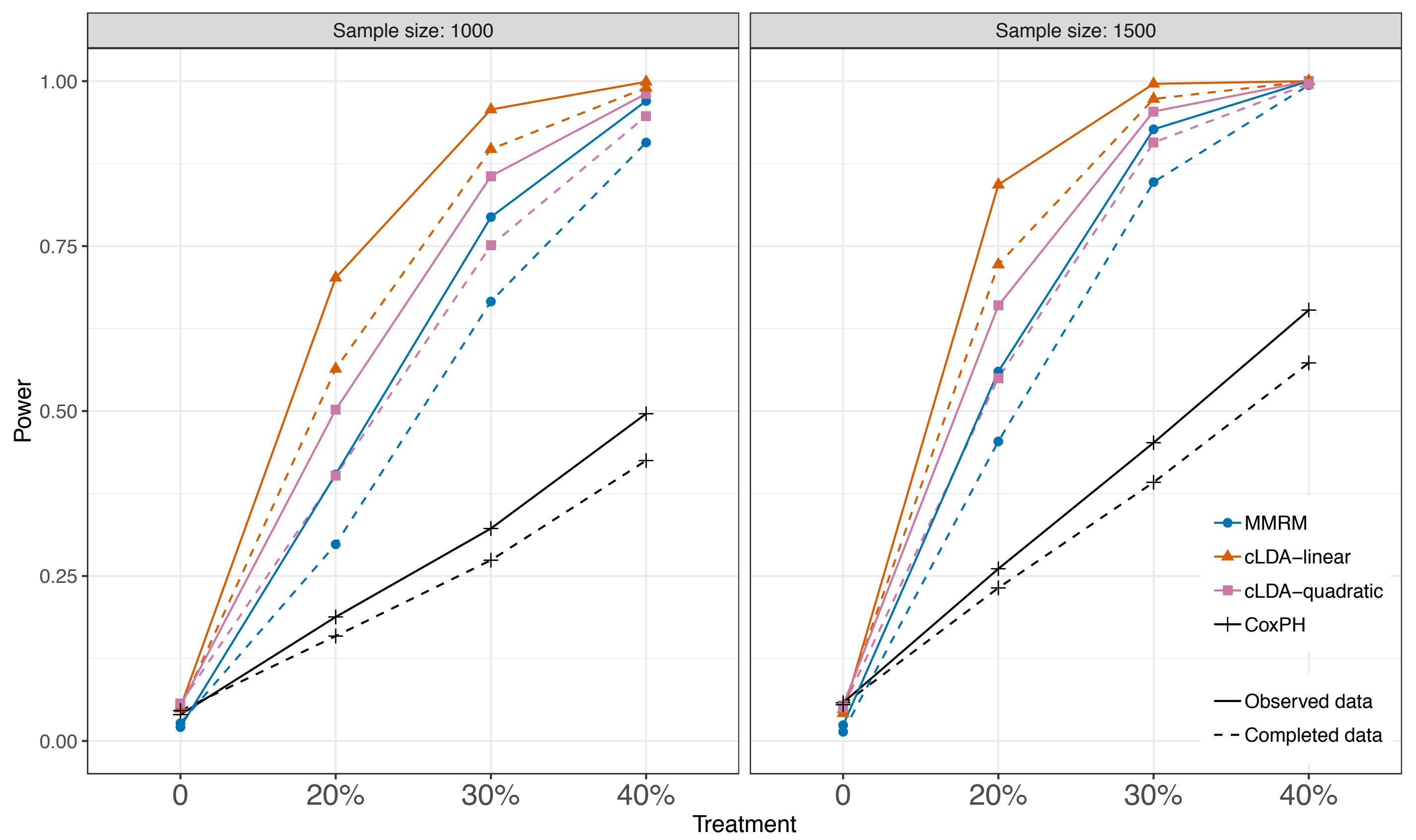

Figure 4. Statistical power for MMRM, cLDA and Cox PH with sample sizes of N=1000 (left) and N=1500 (right). Solid lines indicate power estimates for data observed after simulated non-ignorable missingness; and dashed lines indicate power that would be achieved with complete data (including observations that would be unobserved in reality). The observed data shows greater power with fewer observations because the non-ignorable missingness induces a bias in favor of the treatment.

Discussion

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 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 suggests that there is minimal additional value in the diagnosis. And again, while the progression outcome 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 argued that clinical diagnosis cannot be adequately modelled algorithmically using trial data. That is, clinical assessment and diagnosis by a trial site clinician may consider information not captured by trial measures. But 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 which ignore the correlations among important outcomes will likely not provide reliable comparisons.

All of 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 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 over-correction 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 tipping point analyses.

ACKNOWLEDGEMENTS

We are grateful to the ADNI study volunteers and their families. This work was supported by National Institute on Aging grant R01-AG049750.

REFERENCES

1. Donohue MC et al. The relative efficiency of time-to-threshold and rate of change in longitudinal data, *Contemporary clinical trials.* 32 (5) (2011) 685–693.
2. Donohue MC, et al. The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. *JAMA Neurology*. *71* (8) (2014) 961-970.
3. Goodrich B et al. `rstanarm`: Bayesian applied regression modeling via Stan. `R` package version 2.17.4 (2018). http://mc-stan.org.
4. Breiman L. Random forests. *Machine Learning.* 45 (1) (2001) 5–32.
5. Mehrotra DV, et al. Missing data in clinical trials: control-based mean imputation and sensitivity analysis. *Pharmaceutical Stat.* 16 (5) (2017) 378–392.

For preprint of working paper and pdf of poster: tinyurl.com/y57vm3cp