Supplementary Material

Influence of Subject-Specific Effects in Longitudinal Modelling of Cognitive Decline in Alzheimer's Disease

Supplementary Methods

Development of synthetic datasets

To complement the model evaluation for the meta-database and test generalizability of the influence of subject-specific effects, validation datasets were generated using simulation of 500 separate synthetic cohorts each with 400 participants in 6-month increments out to 60 months of evaluation. Cohorts first sampled the population-level covariates used by the CPath parameterization before generating panels of simulated ADAS-Cog subscale scores expected from subjects with equivalent demographic characteristics. Simulated population-level covariates were baseline age, sex, *APOE4* allele counts, and baseline MMSE and were generated from the meta-database to create cohorts similar in disposition to a representative population expected for studies in cognitive decline. Ages were randomly sampled from the observed meta-database cohort with additional demographics synthetically generated using classification and regression trees (CART) to create similar marginal combinations of covariates with assistance from the `synthpop` package in R [1]. Final evaluation timepoints for each synthetic subject were randomly permuted to simulate a 15% dropout rate followed by a row-wise deletion of 15% of all remaining timepoints to simulate reasonably anticipated missingness in a real-world study. To create the longitudinally correlated ADAS-Cog panel data, Gower's distance was first calculated among the actual subjects in the meta-database according to the population-level covariates described above. This distance was used to cluster the meta-database subjects using weighted median spheroid distance to create 20 distinct similarity clusters. Simulated participants were assigned to the nearest cluster according to their generated demographics and randomly linked to the ADAS-Cog measures of an actual meta-database subject within the same similarity cluster. A mixed-effects beta regression model for each cluster was created using these linked ADAS-Cog measures with cubic polynomial time as fixed effects with random intercepts and slopes using unstructured covariance. Each synthetic subject then had new ADAS-Cog measures generated according to their corresponding cluster-specific model with fixed and random effects randomly generated from the model covariance matrices using multivariate normal sampling. To accommodate the extended 60-month timeframe and generalization to other datasets, the covariance matrices were relaxed to allow for more varied ADAS-Cog scores at later timepoints. This process generated unique panels of ADAS-Cog scores for each simulated participant while retaining serial correlation and within-subject covariance structure expected from real-world subjects with similar population-level demographics and characteristics.

Model designs

First described in 2012, the CPath model for AD was developed from a variety of literature reported values and cohort studies to describe progression of the ADAS-Cog in both natural history and randomized clinical trial setting and create a framework to generate representative simulation cohorts [2]. Additionally, subject-specific effects can be randomly sampled using model covariance matrices for both intercept and slope. Model parameters were developed using both summary-level and patient-level data using a Bayesian implementation to adjust meta-data from the literature with individual-level effects. Further details about the CPath model can be found in Rogers et. al. [2] as well as an implementation in R using the `adsim` package [3], including coefficient values for population-level covariates effects along with covariance measures used to generate subject-specific effects.

In addition to the pre-specified parameterizations of the CPath model, novel beta regression mixed-effects models were developed *de novo* directly from the datasets. The same set population-level demographics used by the CPath model were selected but coefficient and covariance values were generated dynamically from each dataset to provide a comparison point to the pre-defined parameterizations from the meta-study. Use of ad hoc models also created model covariances which could be used for subject-specific effect imputation and fitted values for intercepts and slopes for use in observation forecasting of ADAS-Cog scores for modeled individuals.

The other *de novo* model design used the supervised machine learning method of mixed-effects random forests (MERF) [4]. Random forest models are ensemble methods which improve upon standard decision tree designs by allowing for “feature bagging” to randomly select a subset of model features and generate a forest of partial feature set trees. Tree outputs are averaged across the forests to improve overall predictive accuracy. MERF models extend the random forest by including mixed-effects models in terminal nodes to accommodate the serial correlation inherent in repeated measures data by generating subject-specific effects and updating the population-level effects in the random forests stochastically. This study used a modification of the design presented by Capitaine et al. in the `longituRF` package in R [4] to either use known subject-specific effects for intercepts and slopes for observation forecasting of ADAS-Cog scores, to impute subject-specific effects by sampling from the covariance matrices of the correlation models developed in the terminal nodes in a fashion similar to the CPath generation of subject-specific effects, or suppress random effects altogether so prediction relied solely on the fixed effects of the random forests.

REFERENCES

[1] Nowok B, Raab GM, Dibben C (2016) synthpop: Bespoke creation of synthetic data in R. *J Stat Software* **74**, 1-26.

[2] Rogers JA, Polhamus D, Gillespie WR, Ito K, Romero K, Qiu R, Stephenson D, Gastonguay MR, Corrigan B (2012) Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: a beta regression meta-analysis. *J Pharmacokinet Pharmacodyn* **39**, 479-498.

[3] Polhamus D (2013) adsim: Simulate Alzheimer's disease clinical trials. R package version 3.0.

[4] Capitaine L, Genuer R, Thiebaut R (2021) Random forests for high-dimensional longitudinal data. *Stat Methods Med Res* **30**, 166-184.