Generate Synthetic Data in R for a Hypothetical Alzheimer's Disease Trial

Ron Handels[1,2], Linus Jönsson[2], Lars Lau Raket[3]

[1] Alzheimer Centrum Limburg, School for Mental Health and Neuroscience, Maastricht University, Maastricht, Netherlands; [2] Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden; [3] Lund University, Department of Clinical Sciences, Lund University, Lund, Sweden

Introduction

Individual-level data of recent Alzheimer's Disease (AD) trials are difficult to obtain. Synthetic/simulated data could be used for preparatory, training or explorative research with low risk of privacy breach.

Aim: generate a synthetic version of an original real-world observational dataset, subsequently apply a plausible AD treatment effect, and make our method open-source available.

Method part 1 – synthetic data

Original data:

1. Obtain original real-world data from the ADNI study on demographic (age, sex, education), clinical (*cognition*: MMSE and ADAS; *function*: FAQ; *composite cognition/function*: CDR, ADCOMS) and biological (genetics: APOE4; cerebrospinal fluid: ABeta, Tau; imaging: PET-SUVR-centiloid) outcomes at baseline, 6, 12 and/or 18-month follow-up (35 variables), with missing data multiple-imputed to obtain 10 sets of 537 individuals.
2. Estimate (theoretical) minimum and maximum (all continuous variables) and proportions (all categorical variables).
3. Rescale to 0-1 range (continuous).
4. Estimate beta distribution shape parameters (method of moments; continuous).
5. Transform to cumulative density function (using shape parameters; continuous) and to cumulative probability (categorical).
6. Convert to a normal distribution.
7. Estimate variance-covariance matrix.

Synthetic data:

8. Generate random correlated normal data using Cholesky decomposition of variance-covariance.
9. Transform to cumulative density function.
10. Transform to inverse cumulative density function of beta distribution (using beta distribution shape parameters; continuous).
11. Rescale to original range (using minimum and maximum and proportions from step 2).

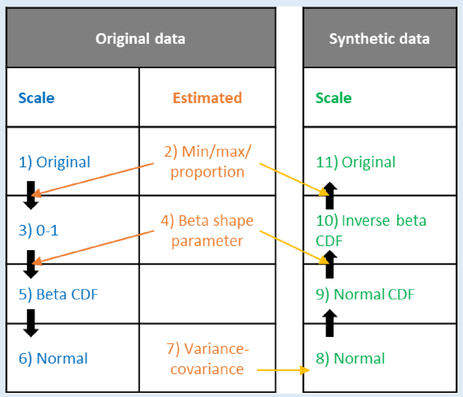
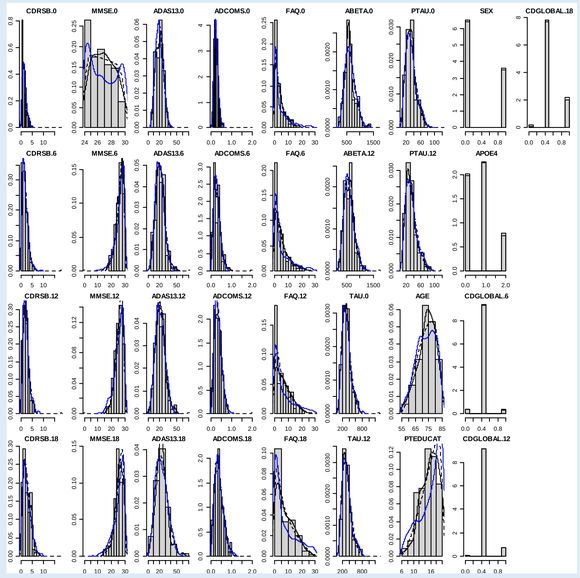


Figure: steps to generate synthetic data from original data

Code availability

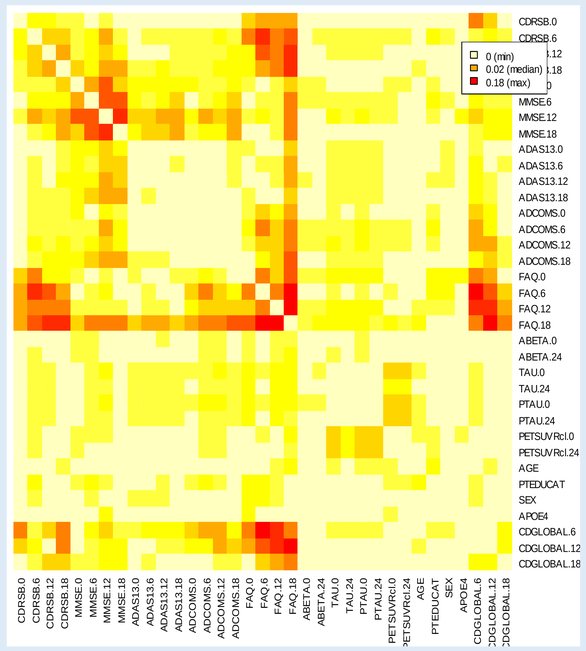
https://github.com/ronhandels/synthetic-correlated-data

Result part 1a – Figure below: the synthetic distribution and mean at each time point.



Histogram/grey = original data
Line solid black = original data density
Line dashed black = beta distribution parameters density rescaled to original scale
Line solid blue = simulated data density

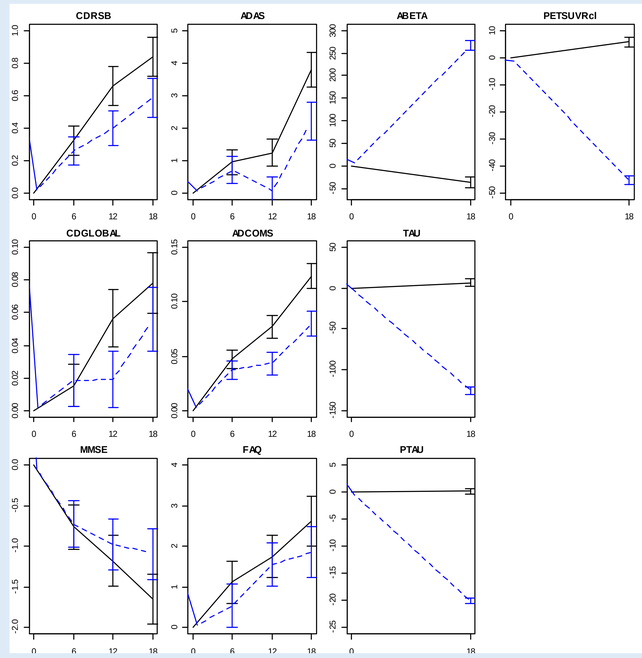
Result part 1b – Figure below: The absolute difference in pairwise correlations between original and synthetic data median was 0.02 (95th percentile=0.11, max=0.18).



Method part 2 – treatment effect

12. Of the simulated data keep half as control arm, and half as intervention arm, and estimate change from baseline.
13. Multiply intervention change from baseline with self-defined hypothetical relative treatment effect.

Result part 2 – Figure below: synthetic data of a hypothetical AD treatment trial (mean and 95% confidence interval over time of the original data (black) and simulated data (blue)).



Discussion

- We judge the synthetic data moderately to strongly similar to the original data.
- Limitation:
  - Correlations on normalized scale are assumed identical to correlations on original scale.
  - No simulation of missing data or drop-out.
  - Synthetic data are only as good as the underlying models generating them.
  - Not compared to alternatives (e.g., R package synthpop).
- Results were successfully used as benchmark scenario in the IPECAD cross-comparison of decision-analytic models for Alzheimer's disease (www.ipecad.org/workshop).

Acknowledgment

Contact: ron.handels@maastrichtuniversity.nl

ADNI: 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

Maastricht University · Karolinska Institutet · Lund University