Checklist for Initial Data Analysis (IDA) – longitudinal studies

Lusa L, Proust-Lima C, Schmidt CO, Lee KJ, le Cessie S, Baillie M, Lawrence F, Huebner M. Initial data analysis for longitudinal studies to build a solid foundation for reproducible analysis. PLoS ONE 2024, to appear. https://stratosida.github.io/

Topic	Item	Features
IDA screening domain: Participation profile		
Time frame	P1	Provide number of time points and intervals at which measurements are taken, using the time metric that best reflects the time of inclusion in the study (typically time from enrollment, or calendar time in studies that involve long enrollment times). Highlight the differences between the time of first measurements and follow-up times.
Time metric	P2	Describe the time metric and corresponding time points specified in the analysis strategy, if different from the time metric described in P1.
Participants	P3	Provide the number of participants who attended the assessment by time metric(s).
Extensions: Participation Profile		
Other time metrics	PE1	Use different time metric(s) to describe the time frame of the study, if applicable and appropriate, e.g. calendar time or measurement occasion.
Data collection	PE2	Describe aspects of the data collection process that can have an impact on the data, if applicable. For example, describe if baseline and longitudinal measurements are different, possible changes in variable measurements though time, etc.
IDA screening domain: Missing data		
Non-enrollment	M1	Describe the non-enrolled, i.e., the participants that were selected but did not enter the study (and the reasons, if available), if applicable.
Drop-out	M2	Describe the participants who dropped out from the study during the follow-up (loss to follow-up and other possible reasons: death, withdrawal, missing by design, if applicable).
Intermittent visit missingness	M3	Describe the participants that have missing data for some of the measurements (intermittent, occasional omission, but do not drop out out of the study).
Variable (item) missingness	M4	Provide the number and proportion of missing values for each variable at each time point as appropriate for fixed or time-varying variables. Describe missingness stratifying the summaries by variables that might influence the frequency of missing values, if relevant (for example: structural variables or levels of measurement).
Patterns	M5	Describe patterns of missing values across variables at each time point and across time points.
Extensions: Missing data		
Non-enrollment	ME1	Compare the characteristics of the participants that entered the study with those of the non-enrolled or with the characteristics of the target population, if applicable and data are available.
Probability of drop-out	ME2	Estimate the probability of drop-out after inclusion, taking appropriately into account the reasons for drop-out.
Dropout effect on outcome	ME3	Visualize mean profiles of a continuous outcome by time metric stratified by time to drop-out.
Predictors of missingness	ME4	Explore whether there are predictors of missingness by comparing complete vs incomplete cases or investigate predictors of time to dropout, as appropriate; this can assist in understanding of the missing data mechanism.
IDA screening domain: Univariate descriptions		
Description of the variables at baseline	U1	Summarize the outcome variable and the explanatory variables with numerical and graphical summaries at baseline.

Description of the time-varying variables at later points	U2	Summarize the outcome variable and the time-varying explanatory variables also at later time points. This might require discretization of time intervals and/or the use of different time metrics.
IDA screening domain: Multivariate descriptions		
Association at baseline	V1	Visualize the association between each explanatory variable with the structural variables at baseline.
Correlation at baseline	V2	Quantify association with pairwise correlation coefficients between all explanatory variables in a matrix or heatmap at baseline.
Interactions at baseline, if applicable	V3	Evaluate bivariate distributions of the variables specified in the analysis strategy with an interaction term; include appropriate graphical displays.
Extensions: Multivariate descriptions		
Stratification	VE1	Compute summary statistics and describe variation between strata defined based on level of measurement, e.g. centers, providers, locations, or by structural variables or other variables described as stratification variables in the analysis strategy (at baseline, other time points/time intervals can be also included).
Associations and correlations at time-points beyond baseline	VE2	Associations and correlations between explanatory variables at time points later than baseline to explore their possible change across time; his could be useful for the identification of auxiliary variables. Selection might bias the results.
Sampling design	VE3	If relevant, identify the stratifying variables and/or the clusters used in the sampling design; explore the distribution of the number of clusters (by stratification variables).
IDA screening domain: Longitudinal aspects		
Profiles	L1	Summarize changes and variability of variables within subjects, e.g. profile plots (spaghetti-plots) for groups of individuals.
Trends	L2	Describe numerically or graphically longitudinal(average) trends of the outcome variable.
Correlation and variability	L3	Estimate the strength of the within-participant correlation of the outcome variable between time points and its variability across time points.
Trends of time-varying explanatory variables	L4	Describe numerically or graphically the longitudinal trends of the time-varying variables.
Extensions: Longitudinal aspects		
Cohort/Period effects	LE1	If appropriate, summarize possible cohorts or period effects (for example, age birth cohorts or period cohorts defined by the calendar time/wave of measurement) on the outcome, and on the explanatory variables, to assess if the variation of the outcome can occur because of these effects.