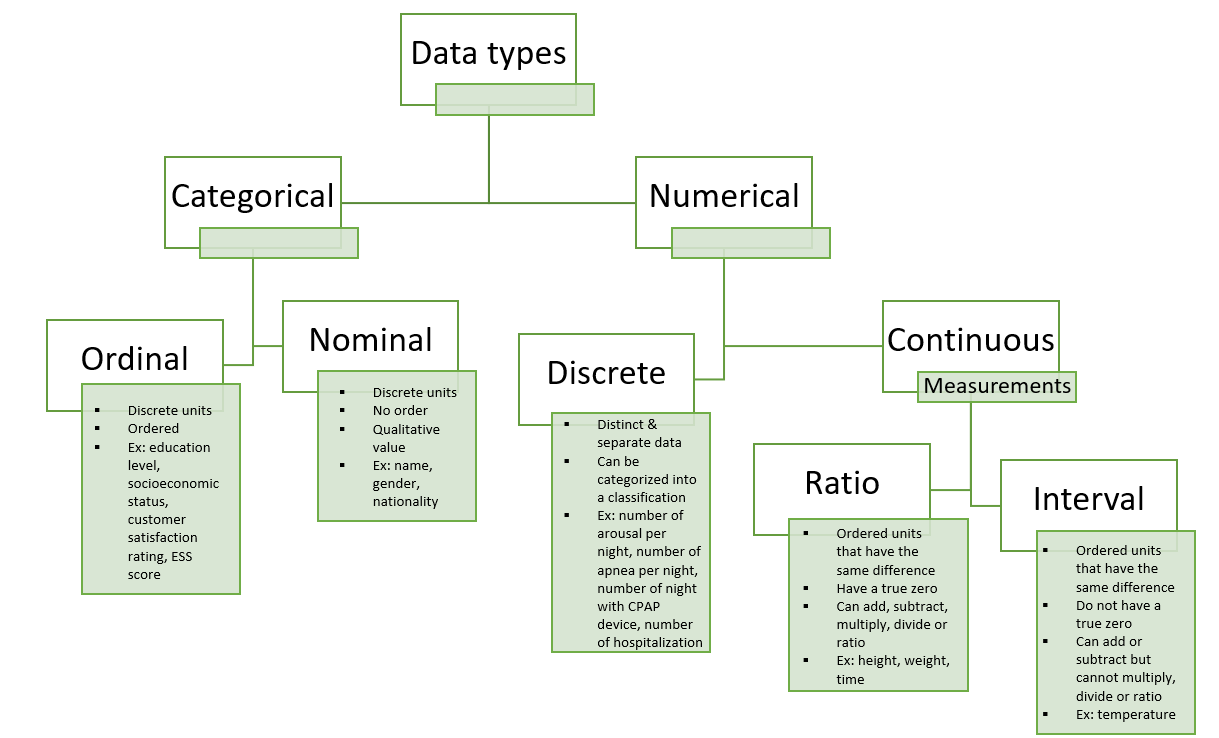
Longitudinal data, trajectories and telemonitoring: how to analyze them?

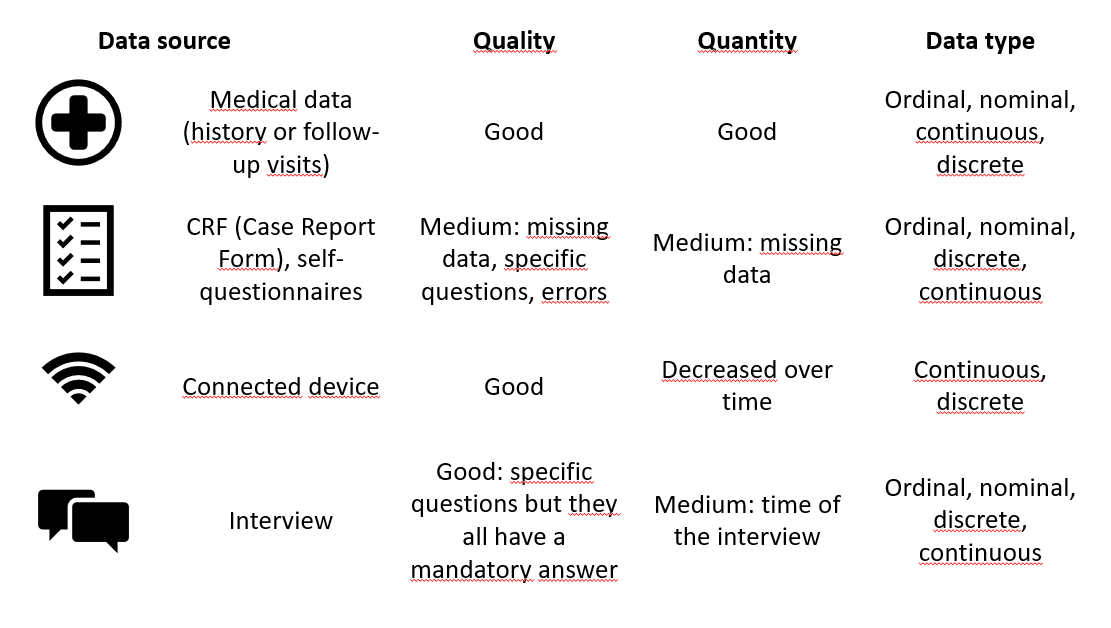
Example of sleep data

# Introduction

* Longitudinal data definition
* Trajectories definition
* Time series definition
* Repeated measurements
* Issues: missing values, extreme values, bad use of statistical method, quality and quantity of data over time
* Solutions? A guide to the use and selection of statistical methods in these contexts

# What is the source of the data?

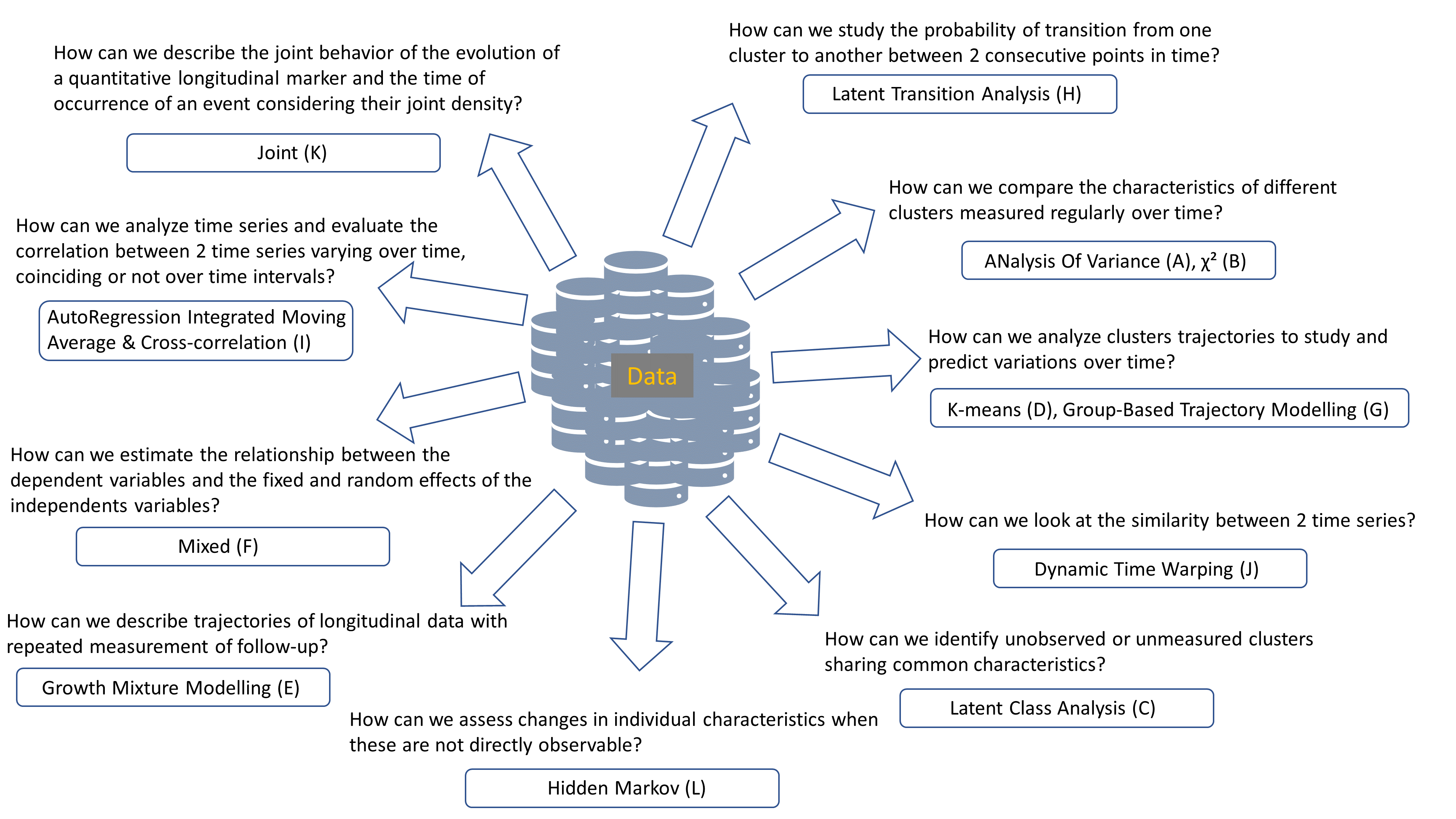
*  The type of data may depend on the data source, and may have an impact on the choice of statistical analysis method.
* Different data source = different quality and quantity of the data = different statistical approaches to analyzing them

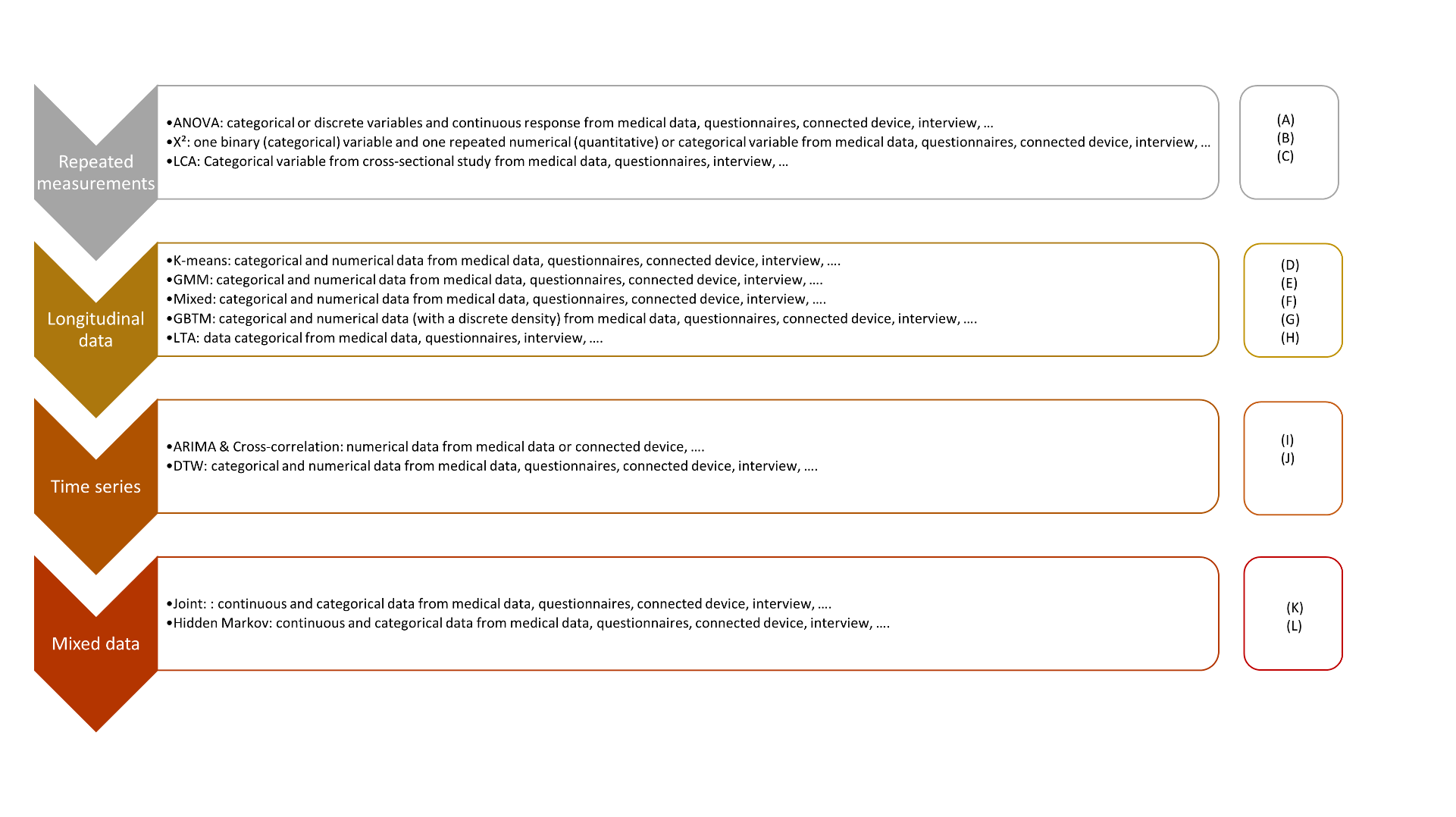


* In the event of poor quality, solutions can be implemented, such as imputing missing values or recoding variables according to errors.
* Questions to ask: is it best to have good data quality, good data quantity, or both? How can we improve data quality or quantity? What is the best source of data in terms of accuracy and reliability? How can we analyze it? 🡺 Solutions depend on the study (aims, population, type and source of data, study questions)🡺 Has an impact on the statistical method but helps in model selection.

# Which statistical method to use?

To choose the right statistical method, we need to think about and check some points:

1.  What is the purpose of the study and what are its objectives?
2. What is the source and type of data?

Complexity gradation:

1. Is the chosen statistical method correct?

Brief description of the methods:

1. ANOVA

*Goal:* Assess whether there is a statistically significant interaction effect between 2 and 3 within-subjects’ factors to explain a continuous outcome.

*Advantages:* Differences between more than 2 groups.

*Limits:* If the null hypothesis of the test is rejected, the means of the groups may differ, and at least one group may show a difference, but the different group(s) is (are) not known, however some post hoc tests are available to counter this limitation; the data must be normally distributed and have a metric scale level, the variance must be equal; the model is influenced by outliers.

1. χ²

*Goal:* Evaluate independence, the difference between variables on a series of contingency tables; assess whether the proportions of the binary variable vary over time?

*Advantages:* According to the number of measurements, different models exist as Mc Nemar for 2 measurements or Mantel-Haenszel for more than 2 measurements; simple and fast model.

*Limits:* All theoretical numbers must be greater than 5; all individuals must move from one state to another (no dropouts); the sample must be random; no covariate.

1. LCA

*Goal:* Identify unobserved, unmeasured clusters sharing common characteristics.

*Advantages:* Powerful tool for analyzing the structure of relationships between categorical variables, for exploring and interpreting complex contingency tables, for testing hypotheses on the structures of categorical latent variables; few classification errors; robust model; possibility of using mixed data, including different scales, for the variables defining the clusters; if continuous variables are involved, possibility of using profiles and therefore the LPA method.

*Limits:* Costly, so number of variables limited by computer power; sensitive to outliers; percentage of individuals in clusters unknown; many a priori decisions to be made.

1. K-means

*Goal:* To group patients’ trajectories into clusters based on their similarities.

*Advantages:* No need for a priori assumptions and avoids the problems associated with model selection; ability to analyze a large data set; can group trajectories that do not follow a polynomial trajectory.

*Limits:* Complete data are required; if missing data are observed, imputations must be performed before statistical analysis, or the chosen method must take data imputations into account; correlation between individuals is not taken into account; tests to find the initial parameters and the optimal number of clusters must be performed; no way of knowing whether it's a global maximum or one of the local maxima when the algorithm converges to the maximum; assessed the longitudinal trajectory of only one variable; the algorithm agglomerates trajectories with a similar overall shape, but if 2 trajectories are transferred in time, they could be in 2 distinct clusters; no tests to check the algorithm's goodness of fit.

1. GMM

*Goal:* Identify trajectory patterns and describe longitudinal changes for each unobserved group identified.

*Advantages:* Deal with missing data and correlated residuals; identify differences between and within individuals over time; trajectory may change qualitatively over time according to different groups.

*Limits:* Many parameters are estimated; complexity of interpreting results; some parameters need to be defined a priori; possibility of identifying false clusters.

1. Mixed

*Goal:* Estimate the relationship between the dependent variables and the fixed and random effects of the independent variables.

*Advantages:* Ability to simultaneously analyze 2, 3 or more dependent variables; ability to deal with missing values; estimation of the odd ratios and the rate ratios.

*Limits:* Interpretation of coefficients possible if random effects are controlled by the analyst; even if differences are statistically significant between estimated trajectories of the dependent variable, these may be non-different in terms of clinical relevance; unobserved variables are assumed to be MAR (Missing At Random).

1. GBTM

*Goal:* Analyze cluster trajectories to study and predict variations over time.

*Advantages:* Simpler than the GMM model, as there are fewer parameters to estimate; faster, with fewer errors; ability to handle missing data and correlated residuals; easier to interpret, especially visually, as less complex.

*Limits:* Missing data must be MCAR (Missing Completely At Random); clusters must be qualitatively different from the dependent variable; dependent and independent variables must have no direct relationship; strong assumptions on trajectory distributions must be respected; possibility of overestimating the number of clusters and the number of trajectories when individual trajectories have the same profile and are distributed on a continuum around the mean trajectory; no intra-class variation.

1. LTA

*Goal:* Study the probability of transition from one cluster at one time to another at the next.

*Advantages:* Model change over time and identify predictors of that change; compare different clusters to determine their characteristics and assess the contribution of different measures for each latent cluster.

*Limits:* Need for a large data set as the model has to estimate many parameters and generally uses the burn-in process; a large number of time points (>6) increase the complexity of the model; problems in defining the optimal number of latent clusters and assigning them a label; problems in including covariates.

1. ARIMA & Cross-correlation

*Goal:* Analyze time series and evaluate the correlation between two time series varying over time, coinciding or not over time intervals.

*Advantages:* Assumption of local stationarity only; robust results even if non-linear trends are mixed in the data or if the time scale is different between time series; ability to define correlations when multiple signals are linked, when the system is complex.

*Limits:* Multiple signals must have linear relationships.

1. DTW

*Goal:* Study the similarity between two time series.

*Advantages:* Deal with incomplete time series with complete references; deal with trajectories with a different number of time points; quickly and efficiently calculate the time lag between two time series; estimate variations in time lag amplitude and direction.

*Limits:* The two-time series must cover the same time window; there must be no time loop (monoticity); constraints on the local slope of the trajectory; computationally expensive; requires a specific averaging process to create cluster centroids; assumes that the training and validation time series are perfectly known.

1. Joint

*Goal:* Account for the joint behavior of the evolution of a quantitative longitudinal marker and the time of occurrence of an event considering their joint density.

*Advantages:* The estimated regression coefficients are unbiased; the association between two outcomes can be estimated; additional random effects can be added; the functional form of the time effect can be generalized using fractional polynomials or splines; patients lost to follow-up can be added to the survival model.

*Limits:* For some Monte-Carlo methods (e.g. Quasi Monte-Carlo), MC error estimation is not possible.

1. Hidden Markov

*Goal:* Assess changes in individual characteristics when these are not directly observable.

*Advantages:* For Bayesian estimates, the model has a very flexible and robust approach; the model is more appropriate for small samples; Bayesian multiple imputation can handle missing data (MAR) without loss of information or introduction of bias; possibility of obtaining the latent clusters of the final model (LMM); efficient algorithms.

*Limits:* The number of classes must be well chosen, as the model could be overestimated or fail to find occasional clusters ; no criteria or model selection to choose the optimal number of latent clusters ; the estimation process cannot be generalized for non-homogeneous transitions ; the number of hidden states must be sufficiently small and/or the covariates must have small dimension for the model to function properly; Bayesian estimation assumes that the distribution of model parameters must be known a priori; label change imposes an order restriction on the parameters for the different states; the status lost of follow-up cannot be exchanged with other states.

Other methods: configural frequency analysis, latent growth curve model, BKMR

# Concrete examples of sleep data studies

Detail in each case: the types and sources of data, the statistical methods used and the reasons why these methods were chosen.

* E-Meuse: telemonitoring (BP), questionnaires (patients’ characteristics)
* Agir A’Dom, MARS: connected devices (CPAP parameters), questionnaires (patients’ characteristics, ESS score)

# Conclusion