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

Example of sleep data

# Introduction

Traditionally, data is measured at a specific point in time, making it impossible to analyze changes and evolutions over time1. One solution is to regularly measure the same variable of interest in a single, identical population over a given period. This repeated measurement data is known as longitudinal data. In models with longitudinal data, individuals are affected by a treatment or other risk factors over a number of time points separated by specific intervals.

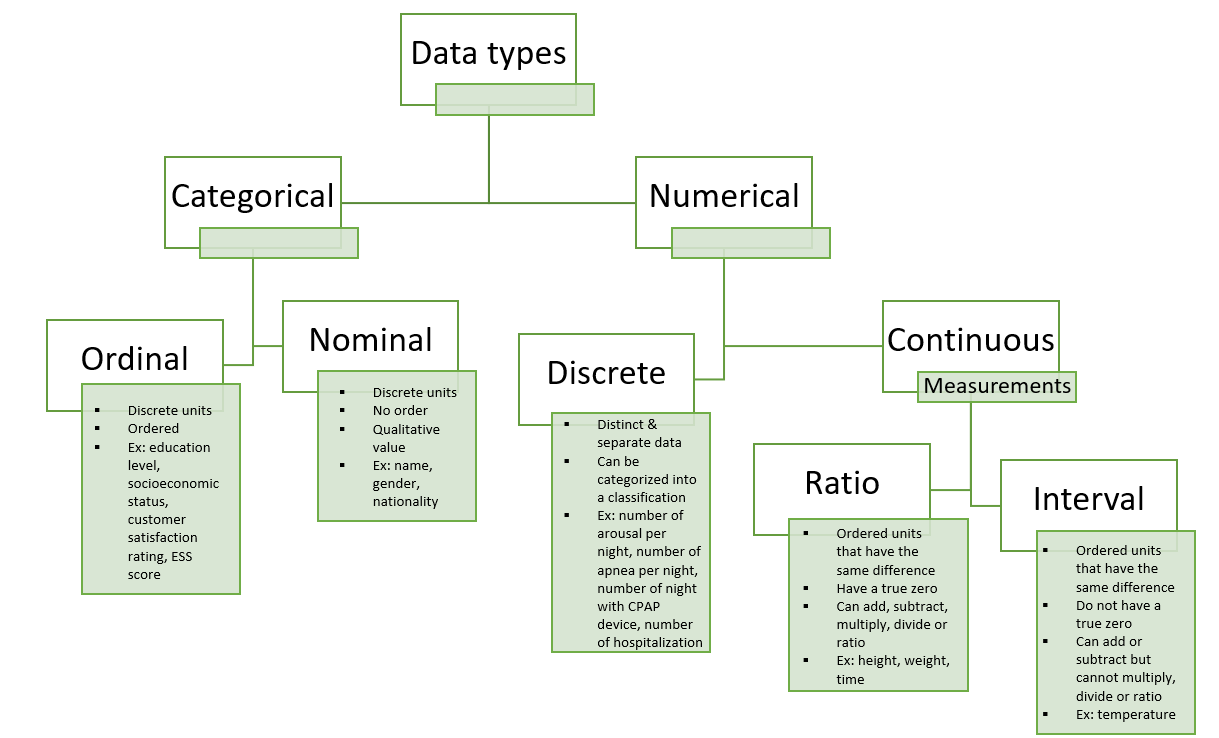
Similarly, with longitudinal data, the study of trajectories is emerging2. Numerous positioning technologies and remote sensors enable the analysis of a vast data set, namely trajectory data. Thanks to this data, patient follow-up and prediction, for example, is easier and progressively more efficient.

Another category of longitudinal data is the time series. A time series is a sequence of numerical values representing the evolution of a specific quantity over time. These developments of random variables can be analyzed to study their past evolution and future behavior.

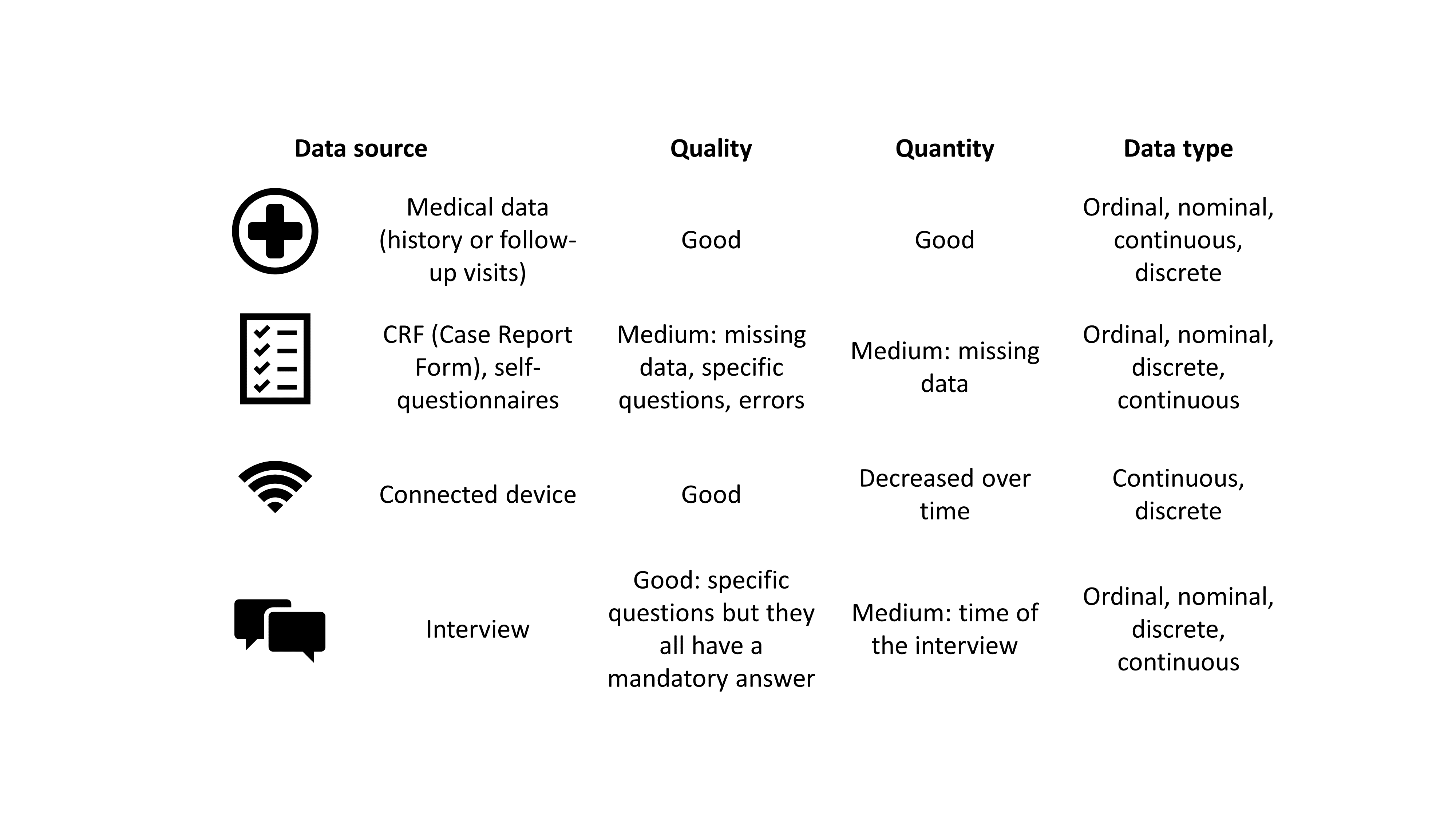
Longitudinal data has, however, some issues as missing or extreme values, correlations on variables due to repeated measures for the same observational unit or the data which are ordered by time in equal space or by unequal intervals1. The quality and quantity of data over time could influence the accuracy and the reliability of the statistical model used to analyze these data. Moreover, a bad or inappropriate use of statistical method could add bias in parameter estimates and outcome predictions.1

Consequently, advanced models and methods are created. A guide is needed to the use and selection of these 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. Data can separate into two different groups, the numerical one and the categorical data. Numerical data is discrete, i.e. distinct and separate data which can be categorized into a classification as the number of hospitalization or number of arousals per night; or continuous. In the continuous data (e.g. measurements), that can be described by ordered units that have the same difference. Ratio data have a true zero and can add, subtract, multiply, divide or ratio while interval data have not a true zero and cannot multiply, divide or ratio.

The manner of having these data influence the type of the data and the quality, quantity of them. Consequently, different quality and quantity of the data implicate different statistical approaches to analyzing them. For example, medical data (history or follow-up visits) can provide ordinal, nominal, continuous and discrete data with a good quality and quantity. In the same wat, the Case Report Form (CRF) or self-questionnaires can provide ordinal, nominal, discrete and continuous data but the quality and quantity are medium. Missing data, errors can be including in the data set and the data source is based on specific questions that limits the analyze. Interview data source provide the same data type with a good quality but medium quantity of data. Indeed, there are specific questions but they all have a mandatory answer and the quantity depend on the time of interview. Another example is data collected by connected device, i.e. only continuous or discrete data. The quality of the data is good but the quantity of them decreased over time.



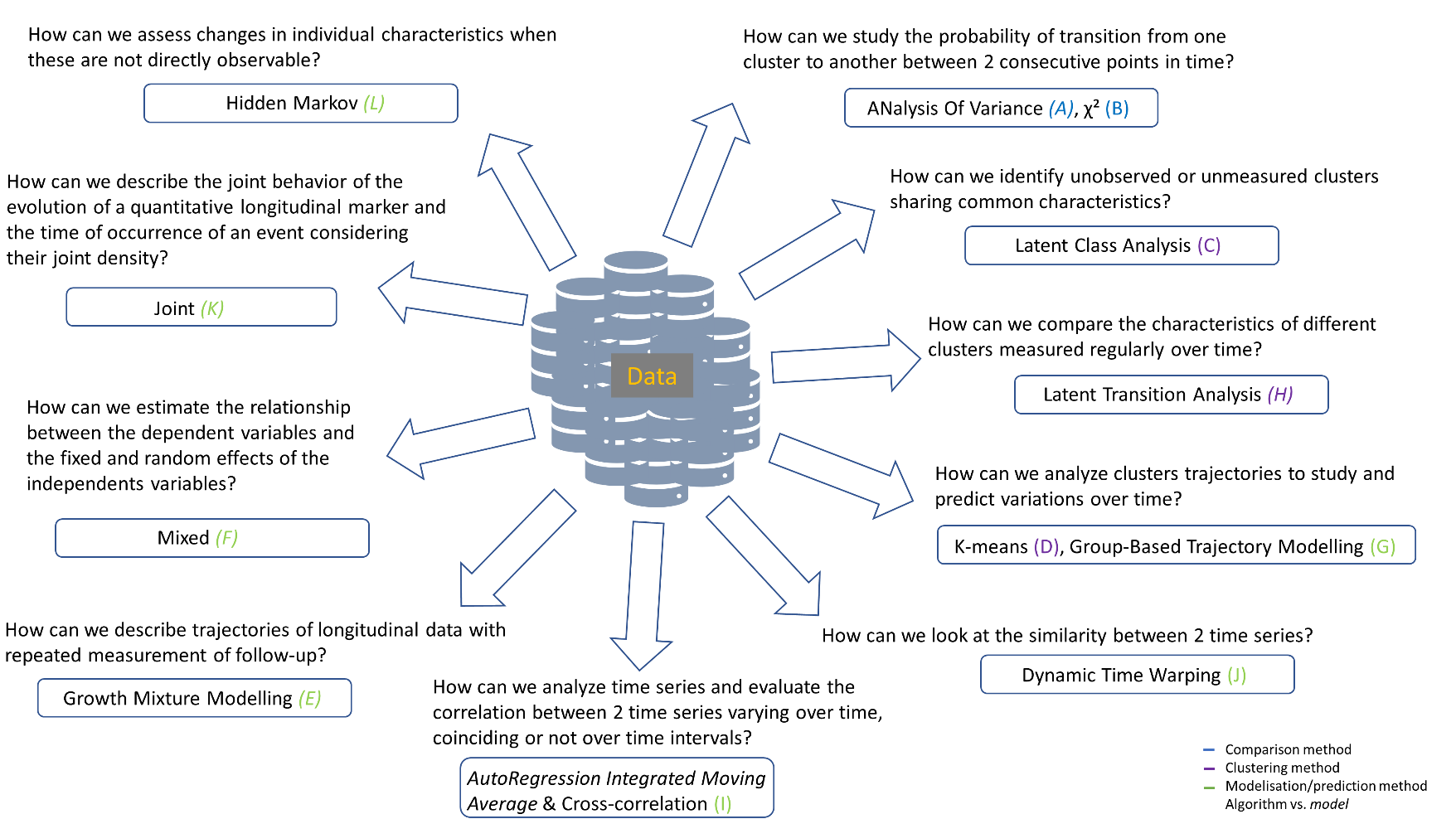
Some questions have to be asked in front of these data analyses. Is it best to have good data quality, good data quantity, or both? How can we improve data quality or quantity? What is the best source of data in terms of accuracy and reliability? How can we analyze it?

In the event of poor quality, solutions can be implemented, such as imputing missing values or recoding variables according to errors. More generally, solutions depend on the study context: the aims, the population included, the type and source of the data, the study questions and the answers have 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:

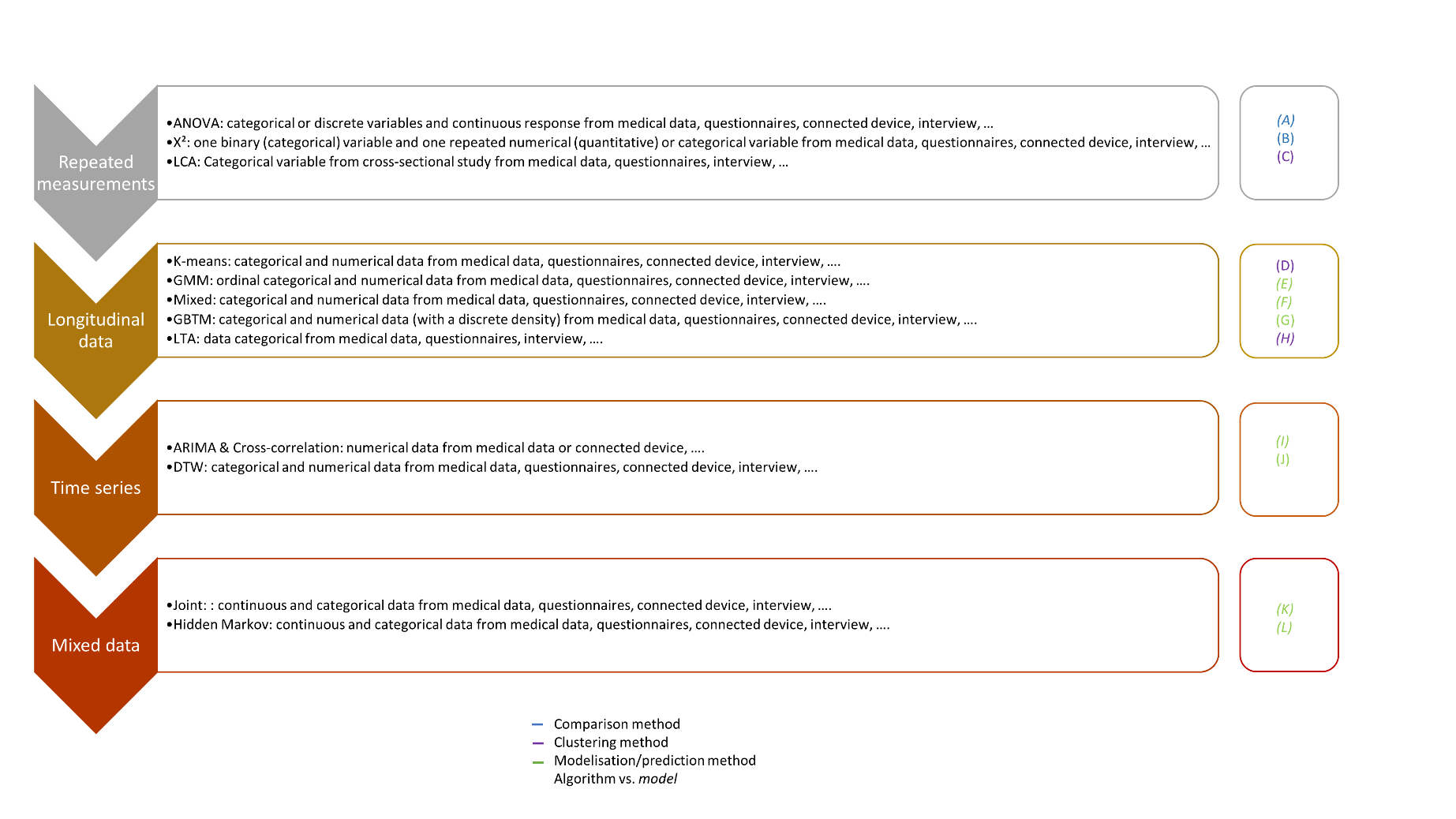
1. What is the purpose of the study and what are its objectives?

Depending on the question of the study, the statistical method differed. The creation and study of trajectories clusters will be performed by Latent Transition Analysis (LTA), Analysis Of Variance (ANOVA) or χ², K-means, Latent Class Analysis (LCA) and Group-Based Trajectory Modelling (GBTM). The study of times series will be performed by Dynamic Time Warping (DTW), AutoRegression Integrated Moving Average (ARIMA) and Cross-correlation or Hidden Markov; the description of trajectories with repeated measurements will be performed by the Growth Mixture Modelling (GMM), the study of longitudinal data will be performed by Mixed model and the analysis of a longitudinal data and survival data will be performed by a joint model.

1. What is the source and type of data?

Once the question is well defined and the method is known, the corresponding between type and source of the data and the statistical method have to be checked.

The methods have a complexity gradation. ANOVA, χ² and LCA analyzing repeated measurements were the easier statistical models. They implement mostly categorical variables from medical data, questionnaires or interview. Then, the study of longitudinal data with K-means, GMM, Mixed, GBTM and LTA models were less complex then the study of time series with ARIMA and Cross-correlation or DTW. Longitudinal data methods implemented categorical and numerical data from medical data, questionnaires, connected device, interview while time series methods mostly implemented numerical data from medical data or connected device. The more complex model implemented mixed data, i.e. continuous and categorical data using joint model or Hidden Markov model.

Complexity gradation:

1. Is the chosen statistical method correct?

The type and source of the data, the study questions are well defined and has allowed to choose the right statistical method. Before the analyze, the goal, advantages and limits of the chosen method has to be checked to fill well. An example of sleep data analysis was carried out for each statistical method described below. A total of 50 patients with 1000 time points were simulated for the representation of CPAP adherence and ESS score. CPAP adherence was normally distributed (with negative values replaced by 0; μ = 4, σ = 1.5) and ESS score was a discrete variable ranging from 0 to 24. For some methods, only categorical variables were accepted. Consequently, we transformed both data into categorical variables: non-adherent ([0h; 2h[) vs. almost adherent ([2h; 4h[) vs. adherent (≥4h) to CPAP and with (≥10) vs. without (<10) ESS.

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.

*Example* - Continuous data are used to compare CPAP adherence over time. All time points and all patients were included in the analysis. The p-value of the ANOVA test was 0.09. We cannot therefore say that adherence to CPAP differs according to time.

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.

*Example* - The χ² Mantel-Haenszel method were performed, admitting that 2 nominal variables are conditionally independent in each stratum assuming that there is no 3-way interaction. For this analysis, 4 time points and all patients were included to calculate a contingency table. CPAP adherence and ESS score were defined as categorical variables. The p-value of the χ² test was 0.16. Therefore, the odd ratio did not draw away from 1, i.e. the difference between groups was not really evident.

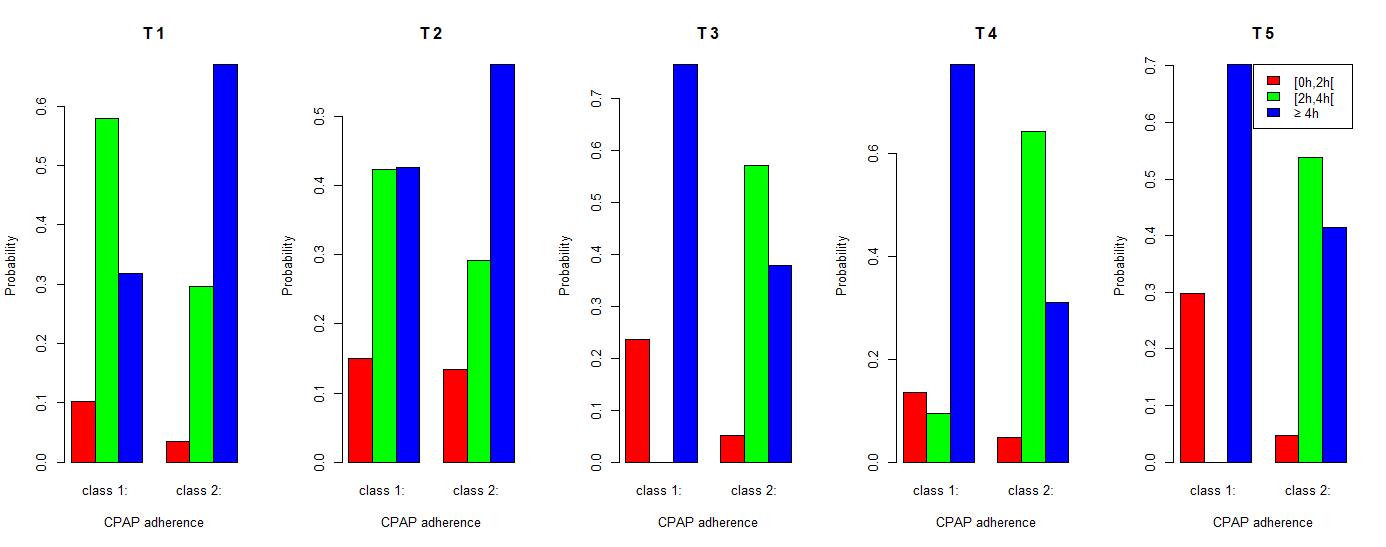
1. LCA

*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.

*Example* - Categorical variable for CPAP adherence and 5 time points were selected to perform LCA method. First, the optimal number of clusters was 2, according to the smallest AIC and BIC (in SM), then LCA method was able to be implemented.



Consequently, at the first time point, patients almost adherent had a greater probability to belong to the 1st cluster while adherent patients had greater probability to belong to the 2nd cluster. At the 2nd time point, it was the same probability for the 2nd cluster but the 1st cluster had more chance to include almost adherent and adherent patients at the same probability. At the 3rd, 4th and 5th time points, the 1st cluster had a greater probability to be constituted by adherent patients and had no or a small probability to include almost adherent patients and a small probability to include non-adherent patients while the 2nd cluster had a greater probability to include almost adherent or (with smaller probability) adherent patients (table in SM).

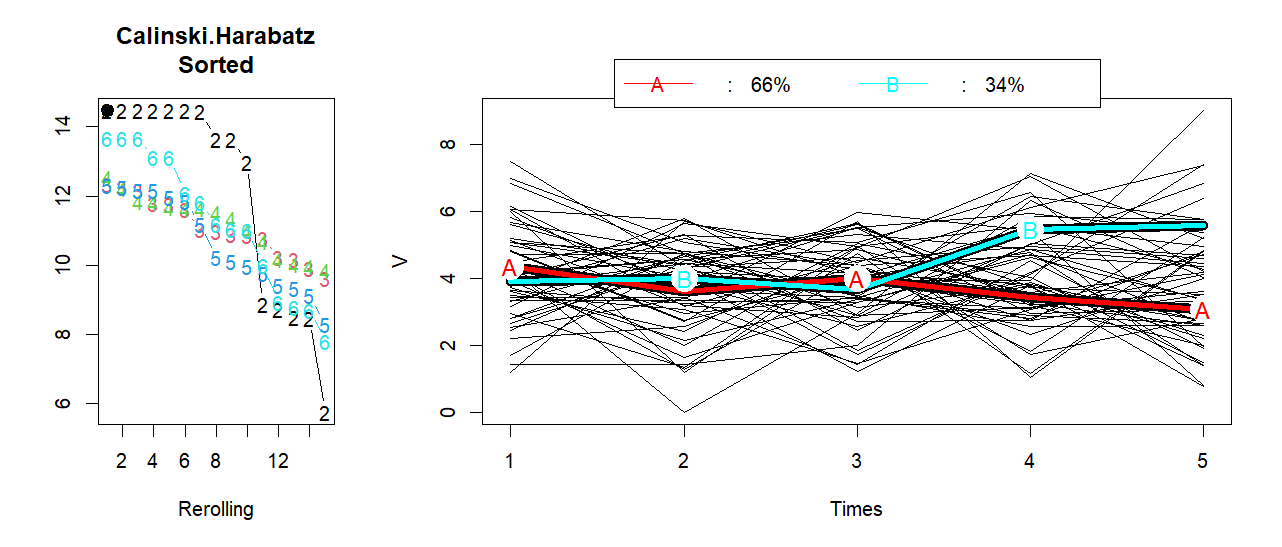
1. K-means

*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 and these patients are to remain included, 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.

*Example* - For this analysis, numerical CPAP adherence was used including all patients and 5 time points. Parameters of the model were detailed in SM. According to the Calinski-Harabatz score, the model with 2 clusters was the best model. Moreover, there was a pretty equal number of individuals in each cluster: 66% in the 1st cluster and 34% in the second cluster.



The first cluster was fairly stable around 4h while the 2nd cluster increased after the 3rd time point from around 4h to around 5h30.

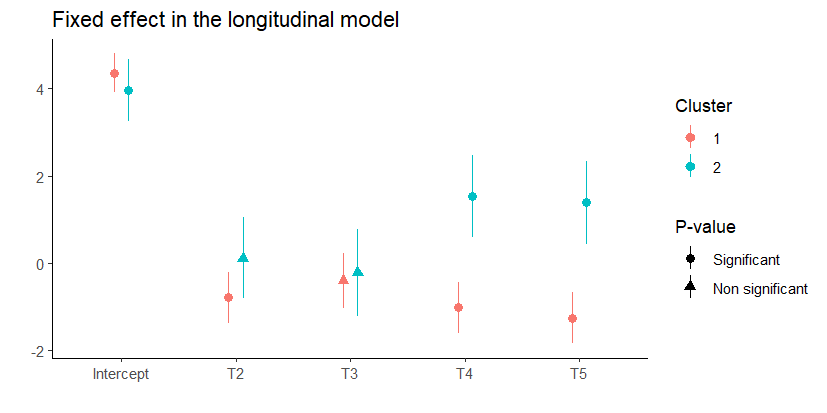
1. GMM

*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.

*Example* - GMM method implemented data in long format. The example used included 5 time points with all patients. There is a random intercept and slope and a mixture parameter on the time variable. The equation was:

According to the BIC criteria, the model with 2 clusters was the best model (detailed in SM). Moreover, the distribution of patients in the clusters was fairly equally distributed (detailed in SM). 

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).

*Example* - Any type of variable was accepted but for this example, continuous outcome was used (in long format data) while Time and ESS\_baseline were categorical variables. All patients and all time points were included. A random intercept on patient were added. The equation was:

Validation of the model and results were detailed in SM. According to the results, CPAP adherence were negatively associated with some time points and ESS baseline was not significantly associated with CPAP adherence.

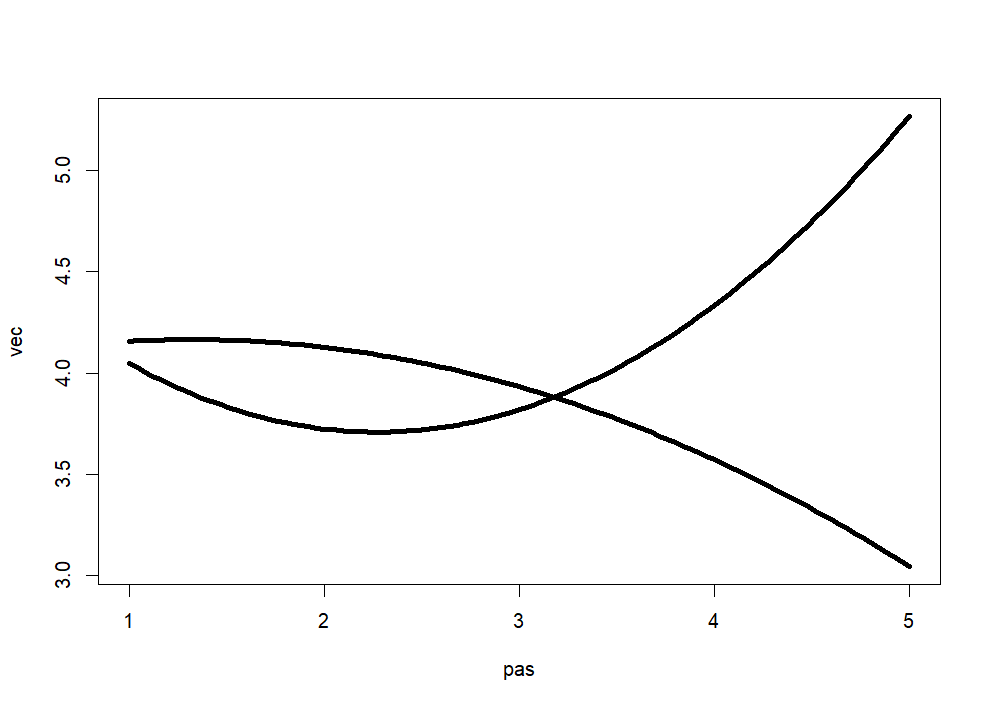
1. GBTM

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

*Advantages* - Simpler than the GMM method, 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.

*Example* - Continuous CPAP adherence was used including 5 time points and all patients. First, we performed GBTM method with different curve and number of clusters to find the best model (detailed in SM). According to some criteria (BIC, loglikelihood, Average Posterior Probability and Proportion of assignment parameters), the model with 2 clusters was the best model. The first cluster had a decreased CPAP adherence unlike the second cluster.



1. LTA

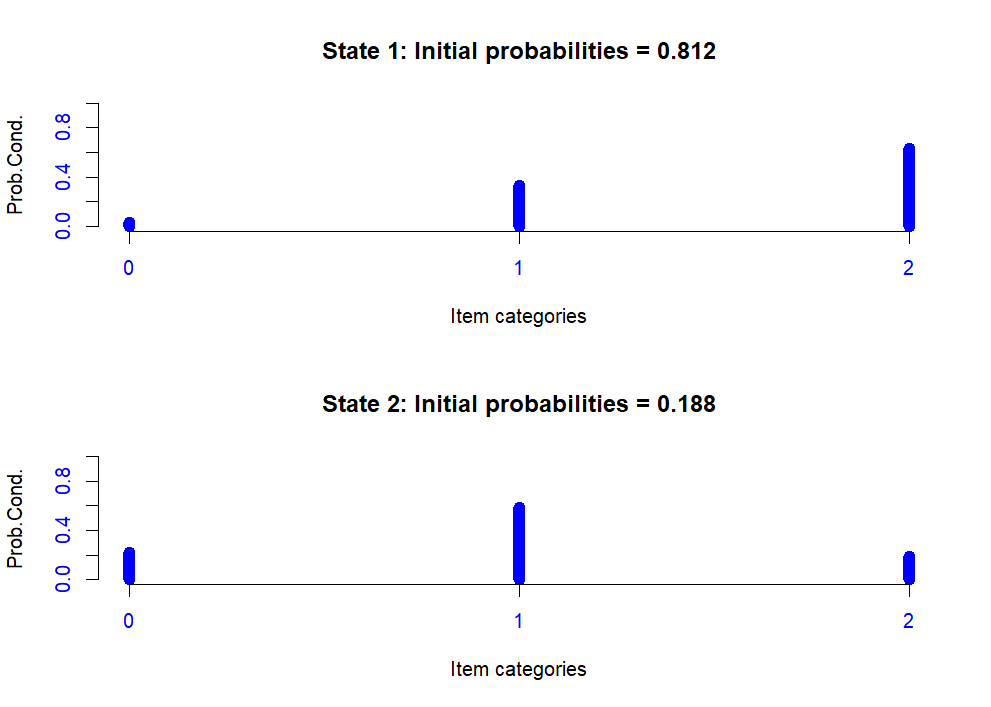
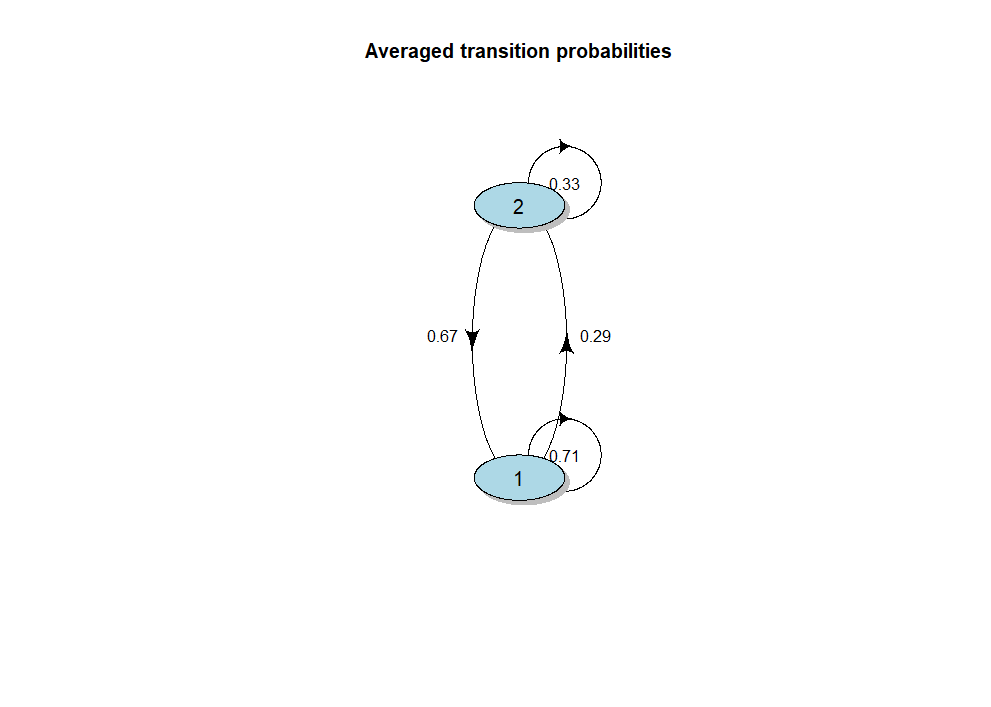
*Goal* - Study the probability of transition from one cluster at one time to another at the next.

*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.

*Example* - LTA model implemented discrete or categorical outcome, so categorical CPAP adherence was used (in long format). All patients were included with 500 time points. Parameters and validation of the model were described in SM. The first step is to define the optimal number of clusters. According to the Loglikelihood and the BIC or AIC criteria, the best model created 2 clusters.

The conditional response probabilities to belong to the 2nd cluster, were 0.22 for the CPAP adherence [0h; 2h[, 0.59 for the CPAP adherence [2h; 4h[ and 0.19 for the CPAP adherence ≥4h. To the 1st cluster, the conditional response probabilities were, respectively 0.03, 0.33, 0.63 (table in SM). The transition probability was 0.29 from the 1st cluster to the 2nd cluster, 0.67 from the 2nd cluster to the 1st cluster and 0.33 to stay in the 2nd cluster, 0.71 to stay in the 1st cluster.



1. ARIMA & Cross-correlation

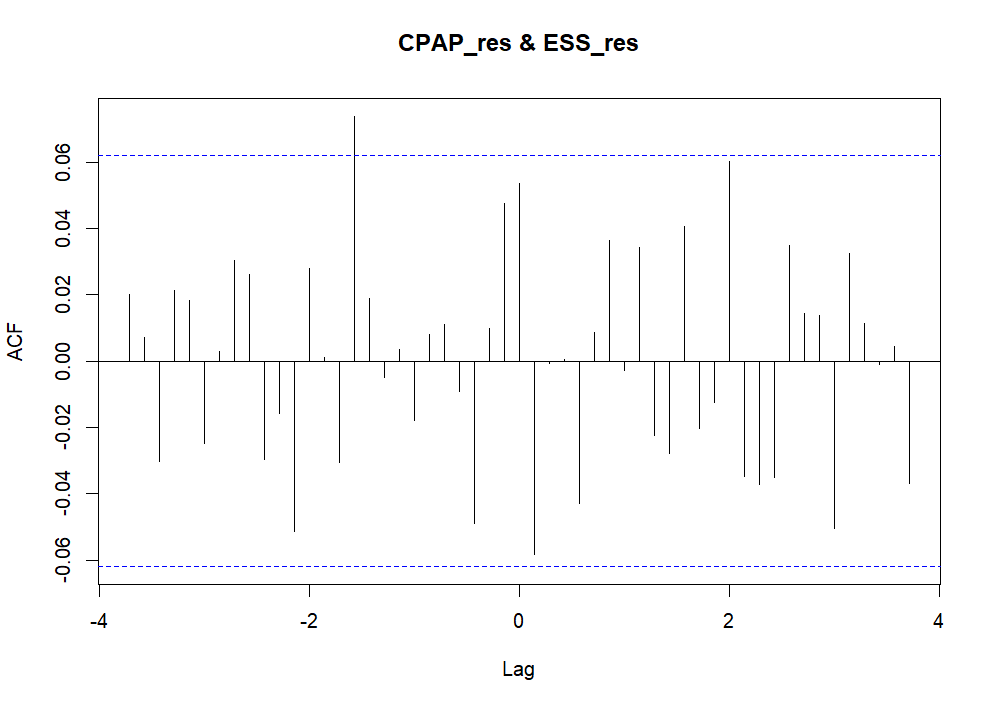
*Goal* - Analyze time series and evaluate the correlation between two time series varying over time, coinciding or not over time intervals.

*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; better with at least 100 observations.

*Example* - First, the ARIMA model used numerical outcome for time series. We compared the CPAP adherence and the ESS score, transformed into time series. All time points were used but only one patient was included. We can repeat the model for each patient. The frequency used was 7, for week scale. To validate the time series, autocorrelation, partial autocorrelation and Box-Ljung test were studied. For CPAP adherence, the final model was ARIMA(0, 0, 0) while for ESS score, it was ARIMA(1, 1, 0)(1, 0, 0). For time series validation, QQ Plot and ACF in SM.

Then, a cross-correlation function was performed to compare the correlation between these two time series (the variables detrend by the ARIMA model).



There was no correlation between ESS score and CPAP adherence with or without lag according to scatterplots; except for the 11th time points (seen on the ACF plot). Interestingly, this meant that an above value of CPAP is likely to lead to an above value of ESS score, about 11 time points later.

These lags could be implemented to a regression to study the association of the ESS score and the CPAP adherence at different lags (in SM).

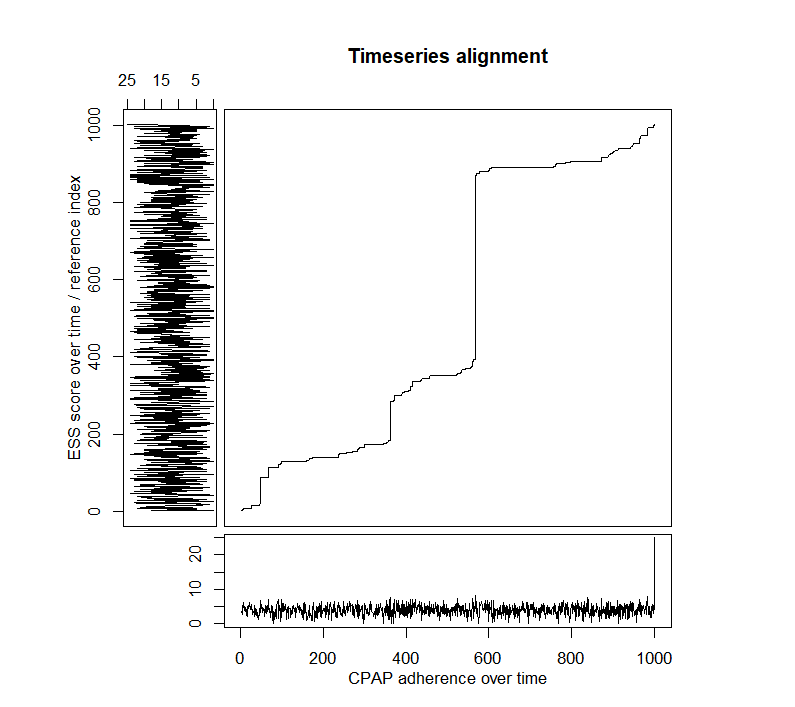
1. DTW

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

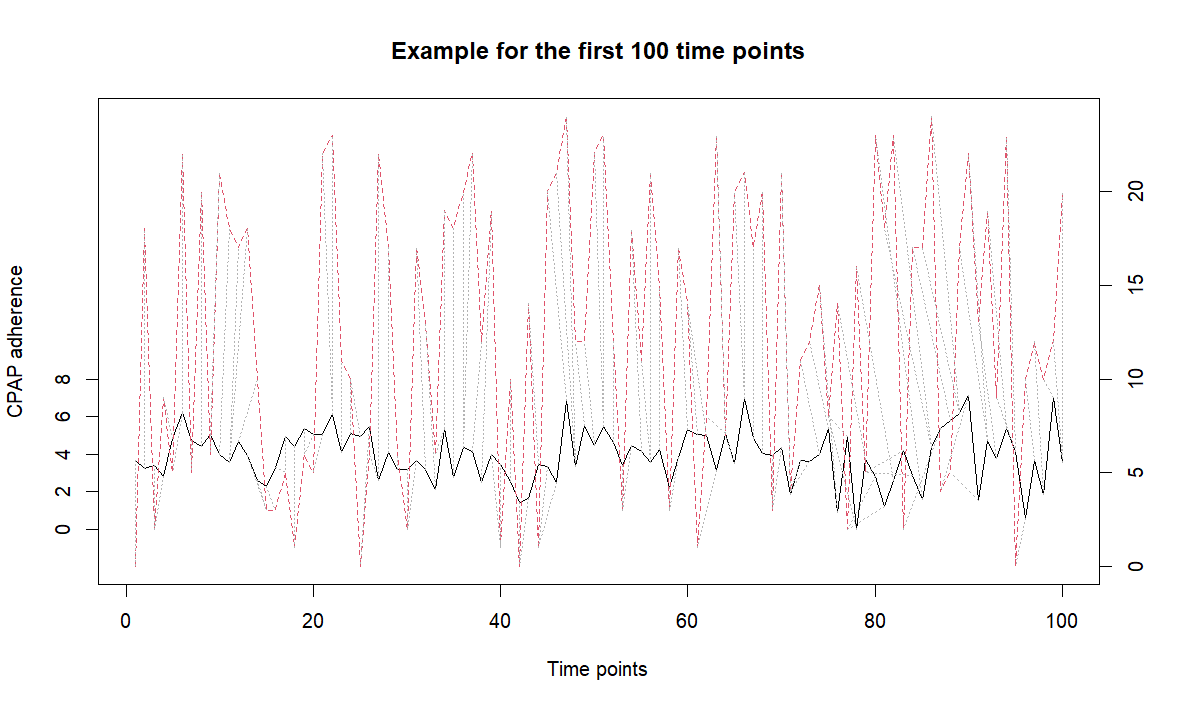
*Advantages* - Deal with incomplete time series with complete references; deal with trajectories with a different number of time points; quickly and efficiently calculate the time lag between two time series; estimate variations in time lag amplitude and direction.

*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; the template variable must be numerical.

*Example* - In our example, 2 numerical variables were used. All time points but only one patient was included in the model. We can repeat the model for each patient. The aim of the method was to determine the best match with the canonical recursion formula.



There were lags between the ESS score and CPAP adherence, e.g. for a given time point CPAP adherence was fixed but it corresponded to an evolution of the ESS score over time (e.g. for the 600th CPAP adherence point or the 400th CPAP adherence point). The interpretation of these shifts could be that CPAP adherence at the 400th time point, for example, would predict all ESS scores between the 400th time point and the 850th time point. Sometimes, it was the other way around, i.e., ESS score predicted CPAP adherence (e.g. for the 852th time point of the ESS score). Some of these changes can be seen in the graph below (only the first 100 time points have been displayed, for ease of reading).



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.

*Example* - One covariate was added to the linear mixed-effect model, the sex of the patient. This variable was a categorical random sample between Male and Female values. The joint model was separated into three steps: 1) the linear mixed-effect model, 2) the Cox model and 3) the joint model. All patients and 7 time points were included in these analyses.

First, the mixed model was performed using continuous CPAP adherence and a random intercept and slope on patient. Parameters of the model were detailed in SM.

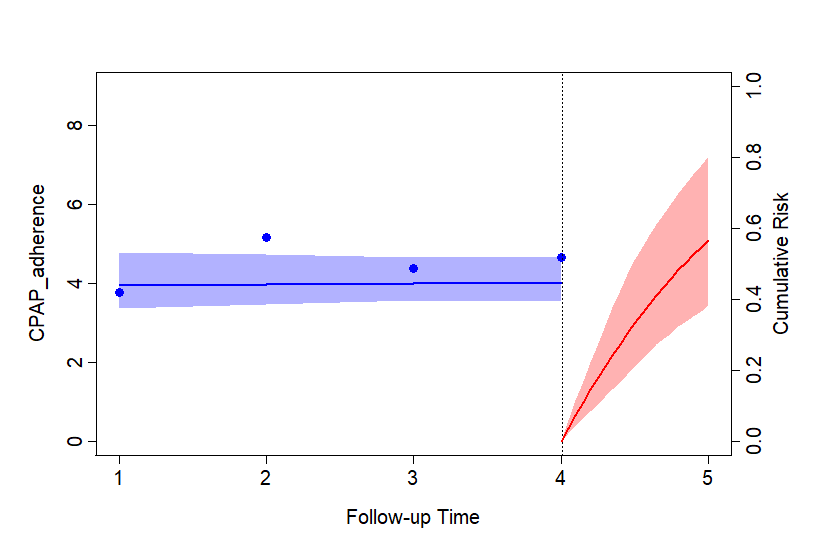
No variable was significant.

Next, the Cox model was run using the categorical ESS score (ESS score < 10 corresponded to the value of death and ESS score ≥ 10 corresponded to the value of life). The sex variable was added as a covariate and the model was clustered by patient.

Validation of the model and results were detailed in SM. However, according to the results, the sex did not significantly influence the survival curve (p-value = 0.64 > 0.05).

Finally, the joint model did not highlight significant result. The verification of the model showed not good observations for the CPAP adherence but good observations for the other parameters (in SM).

For example, the prediction of the cumulative risk for the patient 49 increased (risk to have ESS score < 10) from the 4th time points (cumulative risk around 0.00) to the 5th time points (cumulative risk around 0.58).



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; possible predictions.

*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.

*Example* - This method needed one known categorical variable, e.g. in our analyses, CPAP adherence with 3 states and one hidden categorical variable with a known number of hidden states, e.g. in our analyses, 2 states: Adherent vs. Non-adherent. The independence of observation variable from all covariates were added to the model. All time points but only one patient was included. We can repeat the model for each patient. Parameters and results were detailed in SM.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **[0h;2h [** | **[2h;4h [** | **≥ 4h** |
| **Non-adherent** | 0.56 | 0.44 | 0.00 |
| **Adherent** | 0.23 | 1.31 | -10.60 |

The matrix of transition probability was:

The matrix of emission probability was:

|  |  |  |
| --- | --- | --- |
|  | **Non-adherent** | **Adherent** |
| **Non-adherent** | 0.00 | 1.00 |
| **Adherent** | 0.53 | 0.47 |

The states prediction included 59.1% of time points in the 1st state and 40.9% in the 2nd state (detail in SM).

# Discussion/Conclusion

* Data type and source affect the choice of model.
* Study objectives and design influence method choice.
* Different methods exist, each with its own limitations, advantages and objectives, which we need to consider when choosing a model.
* If there are missing data, there are specific methods and steps for imputing or deleting it. [ref article Bottaz-Bosson, 2023, CHEST, <https://doi.org/10.1016/j.chest.2022.11.034>]
* Data description is the first important step in data analysis. This is followed by considering any missing values, if necessary, and choosing the method. Finally, interpret the results.
* Other statistical methods: configural frequency analysis, latent growth curve model, BKMR.
* In addition to sleep data, other applications are possible (climat, finance, ….)

# References

1. Liu X. Introduction [Internet]. In: Methods and Applications of Longitudinal Data Analysis. Elsevier; 2016 [cited 2024 Dec 20]. p. 1–18.Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780128013427000010

2. Oueslati W, Tahri S, Limam H, Akaichi J. A systematic review on moving objects’ trajectory data and trajectory data warehouse modeling. *Computer Science Review* 2023;47:100516.