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 (Figure 1). 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 (Figure 2). 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 (Figure 3). The methods also have a complexity gradation (Figure 4). Description of the data using ANOVA or χ² analyzing repeated measurements were the easier statistical models. They implement mostly categorical variables from medical data, questionnaires or interview. Then, the classifications methods as LCA method on repeated measurements or the study of longitudinal data with K-means and LTA models were less complex. Finally, methods used for modeling or forecasting were the most complex statistical approaches as GMM, mixed or GBTM methods using longitudinal data; ARIMA and Cross-correlation methods using time series or Joint and Hidden Markov models using mixed data. 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. Mixed data were both continuous and categorical data.

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.

First, description methods have made it possible to compare the characteristics of a population.

1. ANOVA model3

*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 method

*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, assuming that 2 nominal variables are conditionally independent in each stratum and 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.

Secondly, classification methods were used to divide the population into patients’ clusters to summarize certain information. The various clustering methods used different possible metrics such as Euclidean, Ward, Manhattan, Jaccard, Correlation-based or dynamic time warping distances.

1. LCA method4–10

*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 Supplementary Material (SM)). 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 (Figure 5). At the 2nd time point, it was more likely to be populated by the adherent patient and the 2nd cluster had as much as chance to include almost adherent and adherent patients. At the 3rd, 4th and 5th time points, the 1st cluster had a greater probability to be constituted by 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 method5,11–13

*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 (Figure 6). 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. LTA model5,9,14

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

*Advantages* - Model changes over time and identifies 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. All patients were included with 500 time points. Parameters and validation of the model were described in SM. 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 (Figure 9). 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.

Thirdly, modeling and forecasting methods were used to compare and study data trajectories, and to simulate the next trajectory.

1. GBTM model9,15,16

*Goal* - Analyze cluster trajectories to study and forecast 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 (Figure 8).

1. Mixed model15,17–19

*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 while time and baseline ESS score were categorical variables. All patients and all time points were included. A random intercept on patient was added.

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. GMM model5,9,13,20–23

*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* – All patients were included with 5 time points. There is a random intercept and slope and a mixture parameter on the time variable.

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). The first cluster overall tended to decrease over time while the second cluster overall tended to increase over the 5 time points (Figure 7).

1. ARIMA model & Cross-correlation method24–31

*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, QQpLot and Box-Ljung test were studied (in SM). For CPAP adherence, the final model was ARIMA(0, 0, 0) while for ESS score, it was ARIMA(1, 1, 0)(1, 0, 0).

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 (Figure 10); 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. Joint model17,32

*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) (Figure 9).

1. Hidden Markov model33–39

*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 analysis, CPAP adherence with 3 states and one hidden categorical variable with a known number of hidden states, e.g. 2 states: Adherent vs. Non-adherent. All time points but only one patient was included. We can repeat the model for each patient. Parameters and results were detailed in SM. The initial state probabilities model was 1.0 for the 2nd state. The transition probability matrix was around 0.5 for the initial state of non-adherence to the two arrived states, around 0.6 for the state of adherence to the non-adherent group and 0.4 for remaining in the adherent state (Table 1). 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, PCA, MCA, HAC for classification or BKMR for modeling, forecasting. Prediction methods are more complex than methods detailed above, but they can predict the likely trajectory of data.
* 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.

3. Understanding one-way ANOVA using conceptual figures [Internet]. [cited 2025 Feb 6];Available from: https://ekja.org/journal/view.php?doi=10.4097/kjae.2017.70.1.22

4. Cheng W-J, Finnsson E, Arnardóttir E, Ágústsson JS, Sands SA, Hang L-W. Relationship between Symptom Profiles and Endotypes among Patients with Obstructive Sleep Apnea: A Latent Class Analysis. *Annals ATS* 2023;20(9):1337–1344.

5. Hofmans J, Wille B, Schreurs B. Person-centered methods in vocational research. *Journal of Vocational Behavior* 2020;118:103398.

6. Mazzotti DR, Urbanowicz R, Jankowska M. Social risk factors and cardiovascular risk in obstructive sleep apnea: a systematic assessment of clinical predictors in community health centers [Internet]. In: Biocomputing 2025. WORLD SCIENTIFIC; 2024 [cited 2025 Feb 6]. p. 314–329.Available from: https://www.worldscientific.com/doi/10.1142/9789819807024\_0023

7. Sinha P, Calfee CS, Delucchi KL. Practitioner’s Guide to Latent Class Analysis: Methodological Considerations and Common Pitfalls. *Critical Care Medicine* 2021;49(1):e63.

8. Weller BE, Bowen NK, Faubert SJ. Latent Class Analysis: A Guide to Best Practice. *Journal of Black Psychology* 2020;46(4):287–311.

9. Nguena Nguefack HL, Pagé MG, Katz J, et al. Trajectory Modelling Techniques Useful to Epidemiological Research: A Comparative Narrative Review of Approaches. *Clin Epidemiol* 2020;12:1205–1222.

10. P. Den Teuling NG, Heuvel ER van den, Aloia MS, Pauws SC. A latent-class heteroskedastic hurdle trajectory model: patterns of adherence in obstructive sleep apnea patients on CPAP therapy. *BMC Medical Research Methodology* 2021;21(1):269.

11. Mullin S, Zola J, Lee R, et al. Longitudinal K-means approaches to clustering and analyzing EHR opioid use trajectories for clinical subtypes. *Journal of Biomedical Informatics* 2021;122:103889.

12. Genolini C, Falissard B. KmL: k-means for longitudinal data. *Comput Stat* 2010;25(2):317–328.

13. Verboon P, Pat-El R. Clustering Longitudinal Data Using R: A Monte Carlo Study. *Methodology* 2022;18(2):144–163.

14. Chung H, Lanza ST, Loken E. Latent transition analysis: Inference and estimation. *Statistics in Medicine* 2008;27(11):1834–1854.

15. Charnigo R, Kryscio R, Bardo MT, Lynam D, Zimmerman RS. Joint Modeling of Longitudinal Data in Multiple Behavioral Change. *Eval Health Prof* 2011;34(2):181–200.

16. Mésidor M, Rousseau M-C, O’Loughlin J, Sylvestre M-P. Does group-based trajectory modeling estimate spurious trajectories? *BMC Medical Research Methodology* 2022;22(1):194.

17. Gasparini A, Abrams KR, Barrett JK, et al. Mixed-effects models for health care longitudinal data with an informative visiting process: A Monte Carlo simulation study. *Statistica Neerlandica* 2020;74(1):5–23.

18. Liu X. Chapter 1 - Introduction [Internet]. In: Liu X, editor. Methods and Applications of Longitudinal Data Analysis. Oxford: Academic Press; 2016 [cited 2025 Feb 6]. p. 1–18.Available from: https://www.sciencedirect.com/science/article/pii/B9780128013427000010

19. Statistical Learning Methods for Longitudinal High-dimensional Data - PMC [Internet]. [cited 2025 Feb 6];Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC4181610/

20. Rodrigues JF, Bailly S, Pepin J-L, Goeuriot L, Spadon G, Amer-Yahia S. CPAP Adherence Assessment via Gaussian Mixture Modeling of Telemonitored Apnea Therapy. *Applied Sciences* 2022;12(15):7618.

21. Hu Y, Stephenson K, Klare D. The dynamic relationship between daily caffeine intake and sleep duration in middle-aged and older adults. *Journal of Sleep Research* 2020;29(6):e12996.

22. Frontiers | Identifying longitudinal patterns of CPAP treatment in OSA using growth mixture modeling: Disease characteristics and psychological determinants [Internet]. [cited 2025 Feb 6];Available from: https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2022.1063461/full

23. Ram N, Grimm KJ. Growth Mixture Modeling: A Method for Identifying Differences in Longitudinal Change Among Unobserved Groups. *Int J Behav Dev* 2009;33(6):565–576.

24. Aloia MS, Goodwin MS, Velicer WF, et al. Time Series Analysis of Treatment Adherence Patterns in Individuals with Obstructive Sleep Apnea. *Annals of Behavioral Medicine* 2008;36(1):44–53.

25. Babbin SF, Velicer WF, Aloia MS, Kushida CA. Identifying Longitudinal Patterns for Individuals and Subgroups: An Example with Adherence to Treatment for Obstructive Sleep Apnea. *Multivariate Behavioral Research* [Internet] 2015 [cited 2025 Feb 6];Available from: https://www.tandfonline.com/doi/full/10.1080/00273171.2014.958211

26. Zhang H, Su K, Zhong X. Association between Meteorological Factors and Mumps and Models for Prediction in Chongqing, China. *International Journal of Environmental Research and Public Health* 2022;19(11):6625.

27. Tong S, Hu W. Climate variation and incidence of Ross river virus in Cairns, Australia: a time-series analysis. *Environmental Health Perspectives* 2001;109(12):1271.

28. Boker SM, Rotondo JL, Xu M, King K. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods* 2002;7(3):338–355.

29. Yuan N, Fu Z, Zhang H, Piao L, Xoplaki E, Luterbacher J. Detrended Partial-Cross-Correlation Analysis: A New Method for Analyzing Correlations in Complex System. *Sci Rep* 2015;5(1):8143.

30. Shen C. Analysis of detrended time-lagged cross-correlation between two nonstationary time series. *Physics Letters A* 2015;379(7):680–687.

31. Zheng Y, Wang K, Zhang L, Wang L. Study on the relationship between the incidence of influenza and climate indicators and the prediction of influenza incidence. *Environ Sci Pollut Res* 2021;28(1):473–481.

32. Philipson P, Hickey GL, Crowther MJ, Kolamunnage-Dona R. Faster Monte Carlo estimation of joint models for time-to-event and multivariate longitudinal data. *Computational Statistics & Data Analysis* 2020;151:107010.

33. Das R, Muldoon M, Lunt M, McBeth J, Yimer BB, House T. Modelling and classifying joint trajectories of self-reported mood and pain in a large cohort study. *PLOS Digital Health* 2023;2(3):e0000204.

34. Midelet A, Bailly S, Tamisier R, et al. Hidden Markov model segmentation to demarcate trajectories of residual apnoea-hypopnoea index in CPAP-treated sleep apnoea patients to personalize follow-up and prevent treatment failure. *EPMA Journal* 2021;12(4):535–544.

35. Bartolucci F, Pandolfi S, Pennoni F. LMest: An R Package for Latent Markov Models for Longitudinal Categorical Data. *Journal of Statistical Software* 2017;81:1–38.

36. Efthimiou O, Welton N, Samara M, Leucht S, Salanti G, Package 4 on behalf of GW. Α Markov model for longitudinal studies with incomplete dichotomous outcomes. *Pharmaceutical Statistics* 2017;16(2):122–132.

37. Zhou J, Song X, Sun L. Continuous time hidden Markov model for longitudinal data. *Journal of Multivariate Analysis* 2020;179:104646.

38. Haan-Rietdijk S de, Kuppens P, Bergeman CS, Sheeber LB, Allen NB, Hamaker EL. On the Use of Mixed Markov Models for Intensive Longitudinal Data. *Multivariate Behavioral Research* 2017;52(6):747–767.

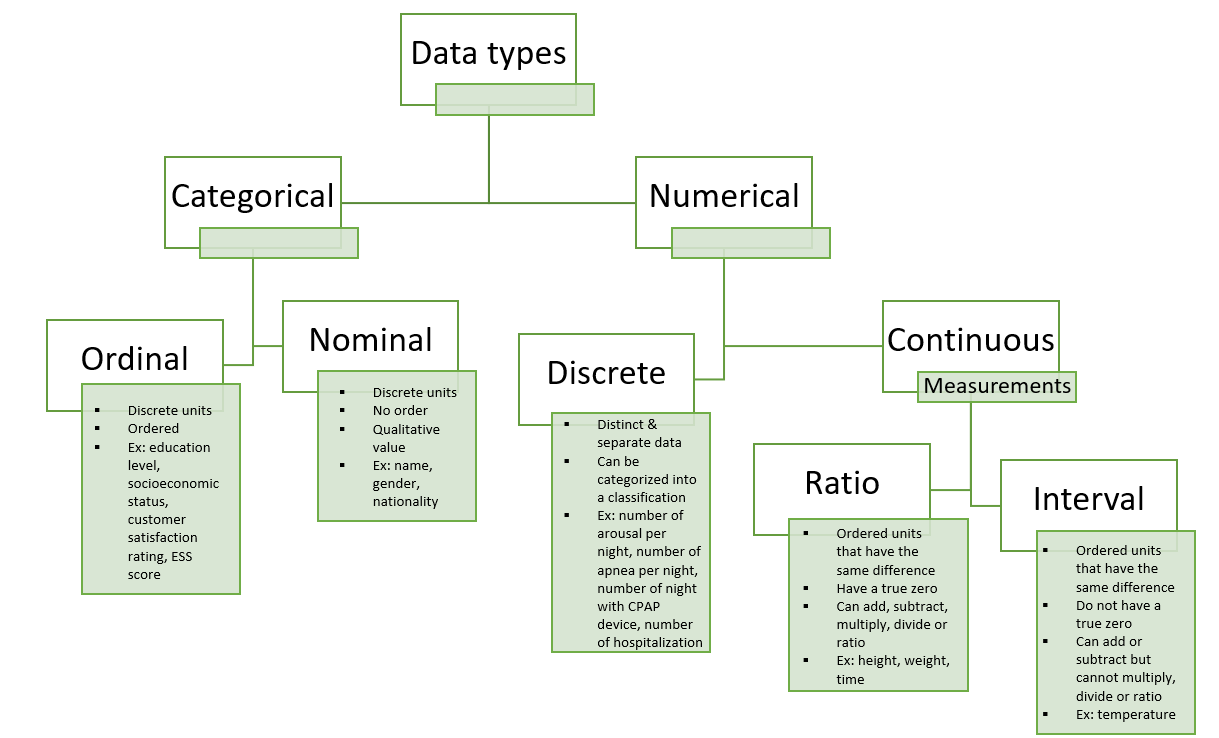
39. Pandolfi S, Bartolucci F, Pennoni F. A hidden Markov model for continuous longitudinal data with missing responses and dropout. *Biometrical Journal* 2023;65(5):2200016.

# Tables

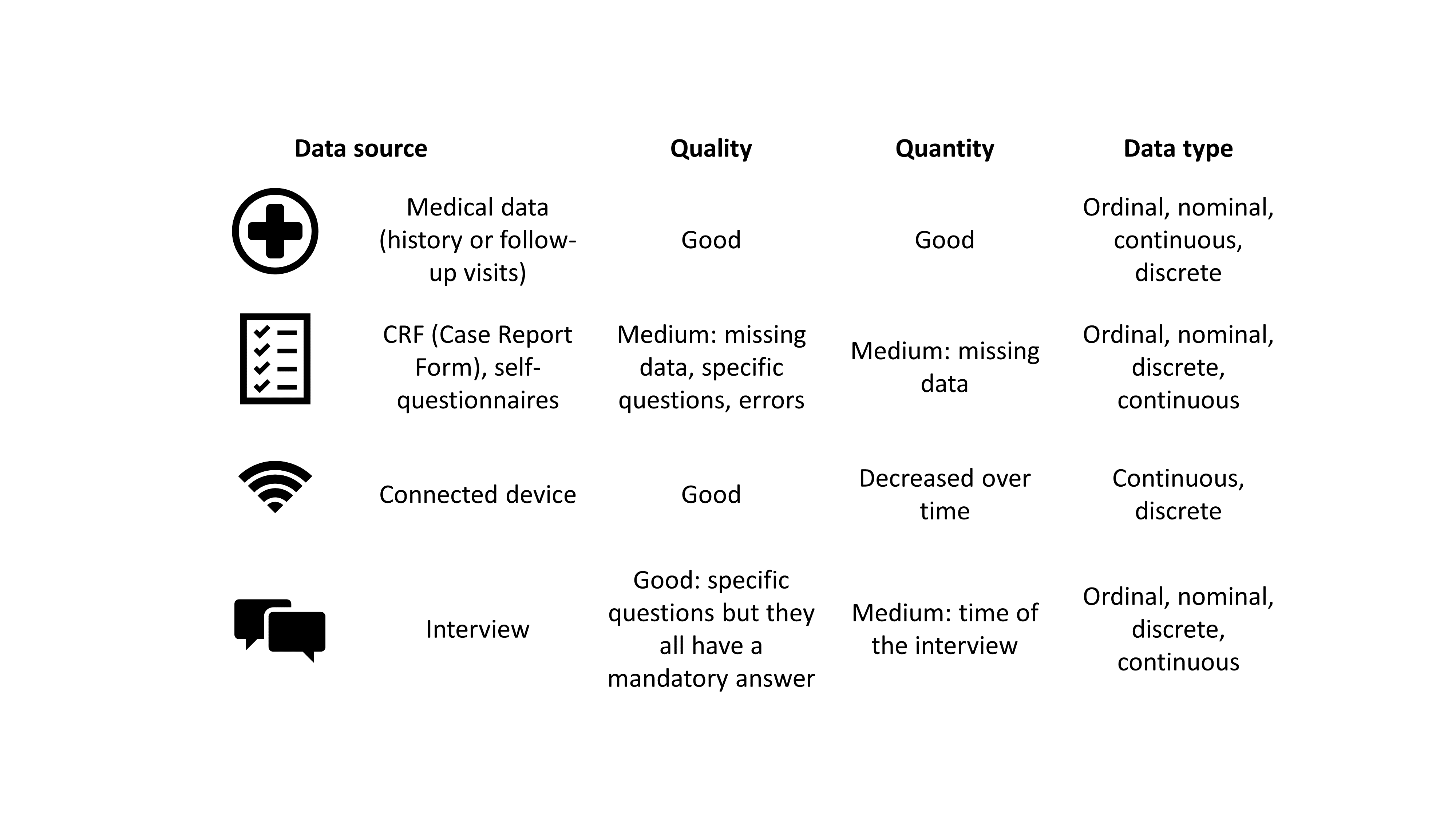
**Table 1: Matrix of transition probability from number of hours to CPAP use to hidden state of CPAP adherence using Hidden Markov model.**

|  |  |  |
| --- | --- | --- |
|  | **Non-adherent** | **Adherent** |
| **Non-adherent** | 0.53 | 0.47 |
| **Adherent** | 0.56 | 0.44 |

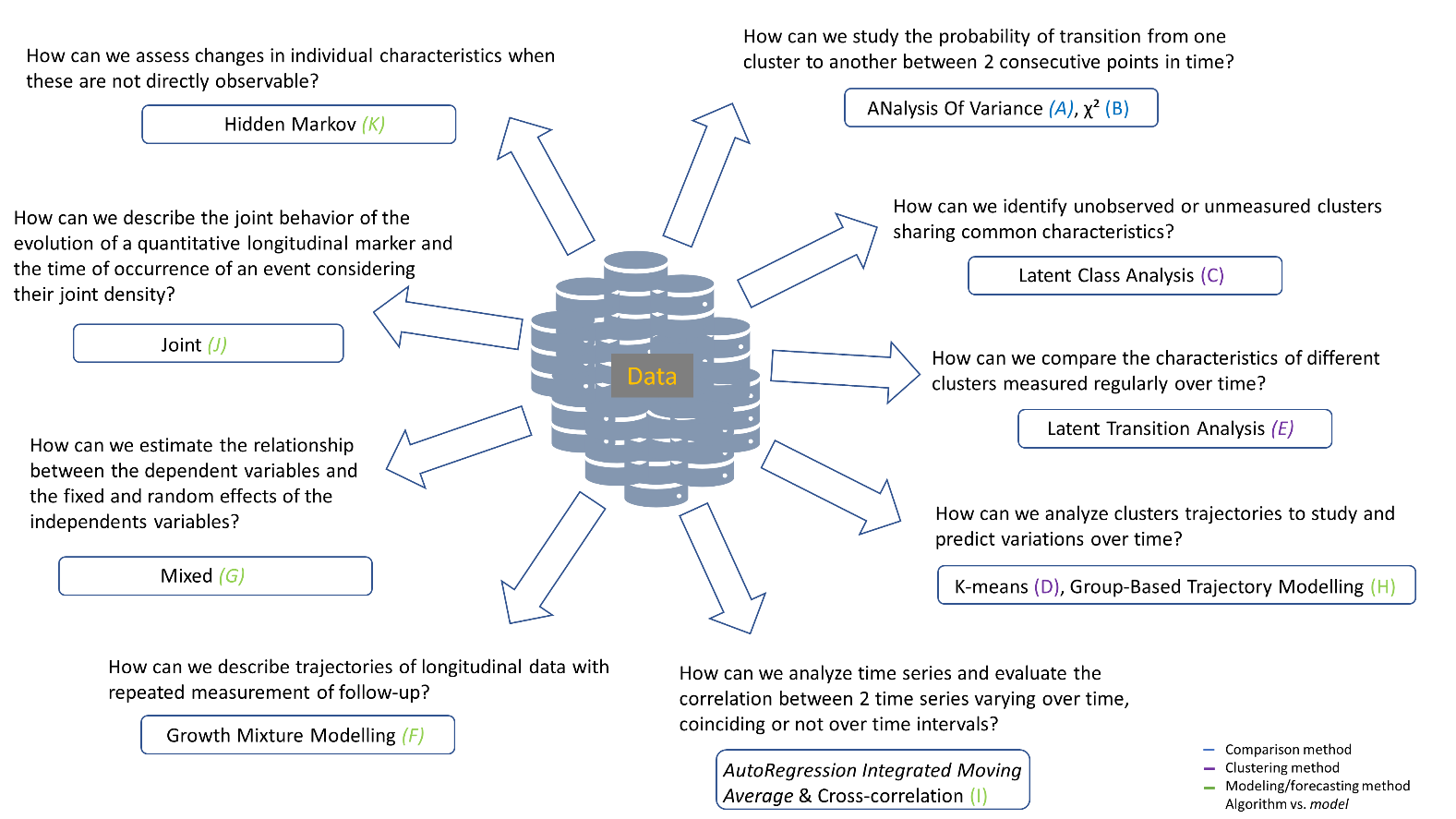
# Figures legend

**Figure 1: Data types**. *CPAP, continuous positive airway pressure; ESS, Epworth sleepiness scale*

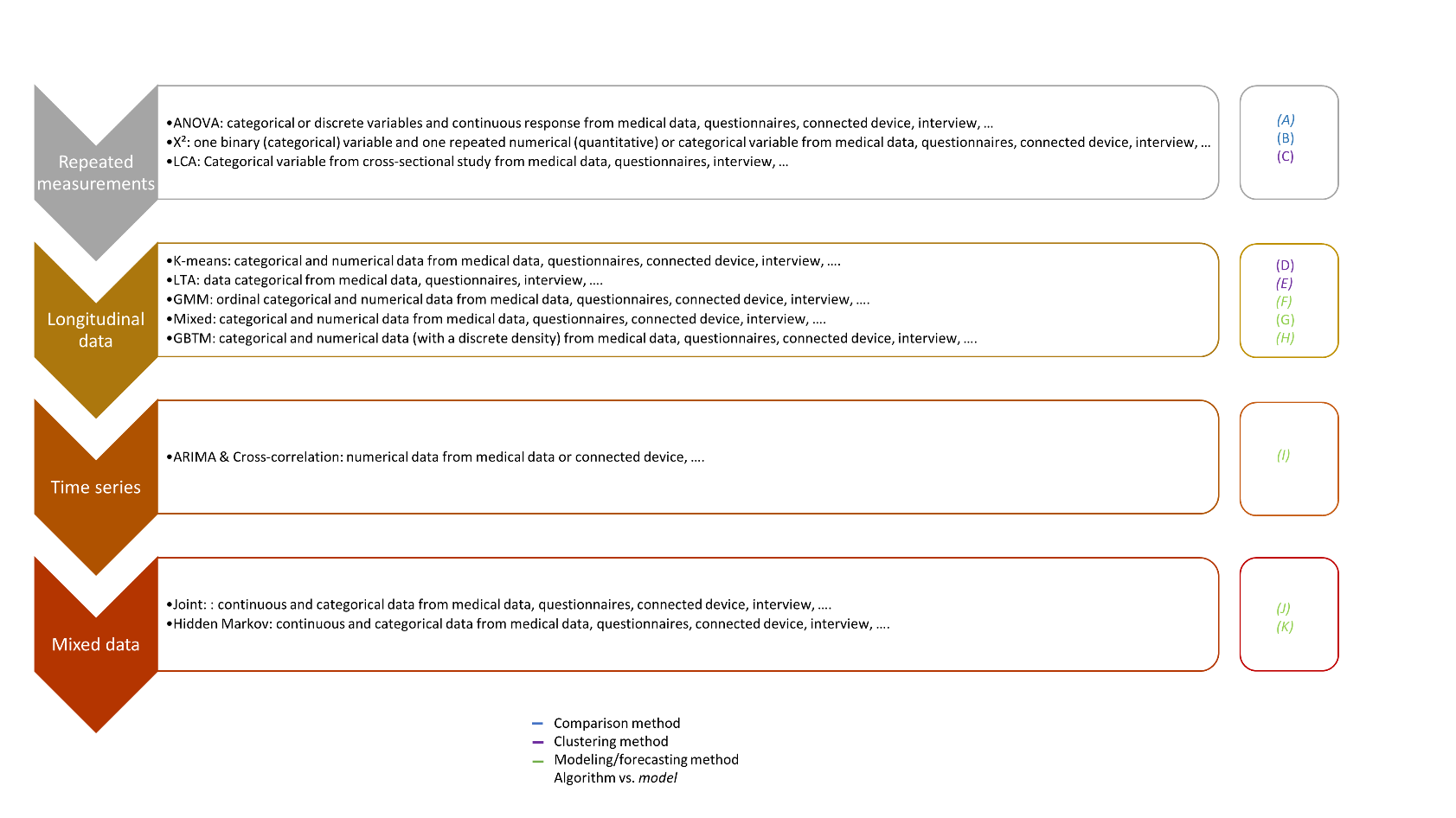
**Figure 2: Data sources**.



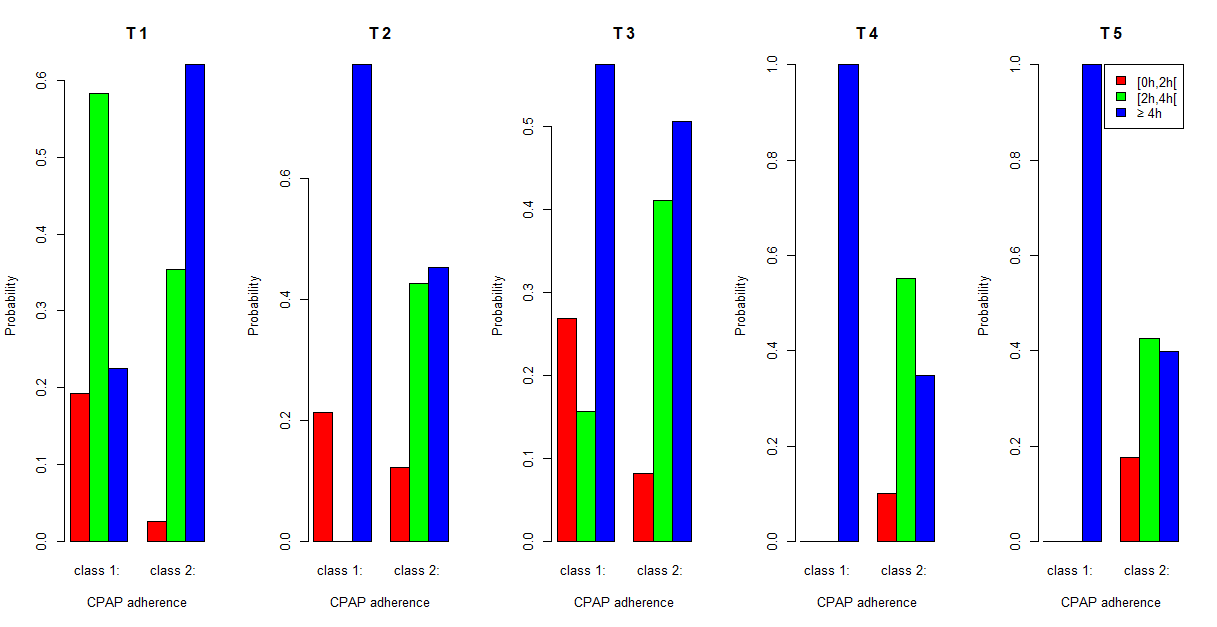
**Figure 3: Statistical methods according to objectives and question of the study**.



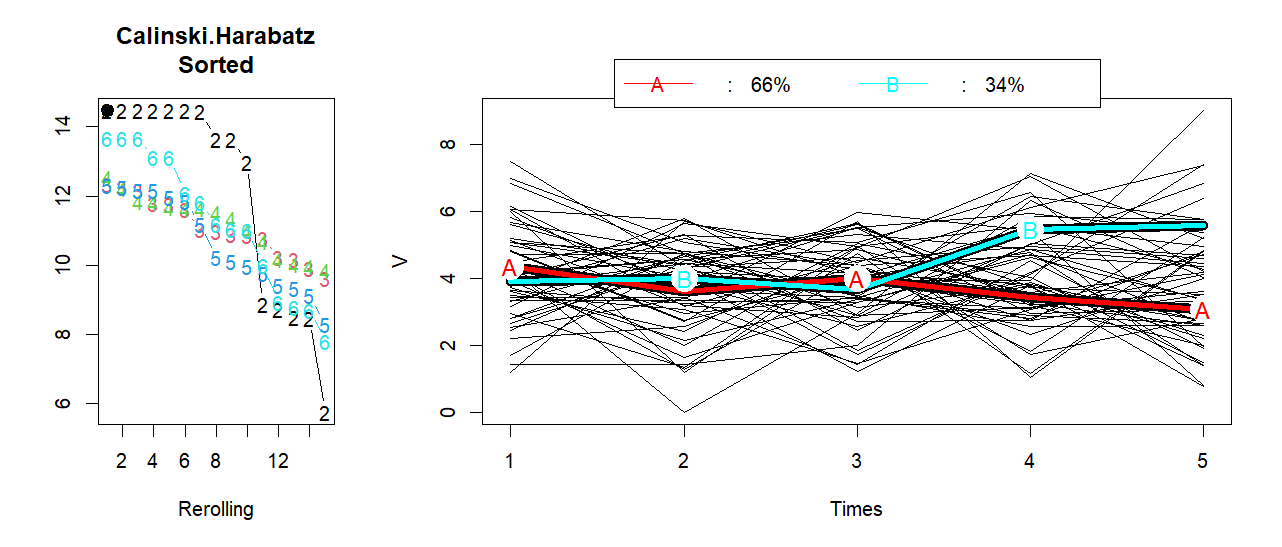
**Figure 4: Statistical methods classified according to complexity, data format and objectives**. *ANOVA, analysis of variance; ARIMA, autoregression integrated moving average; DTW, dynamic time warping; GBTM, group-based trajectory modelling; GMM, growth mixture modelling; LCA, latent class analysis; LTA, latent transition analysis.*



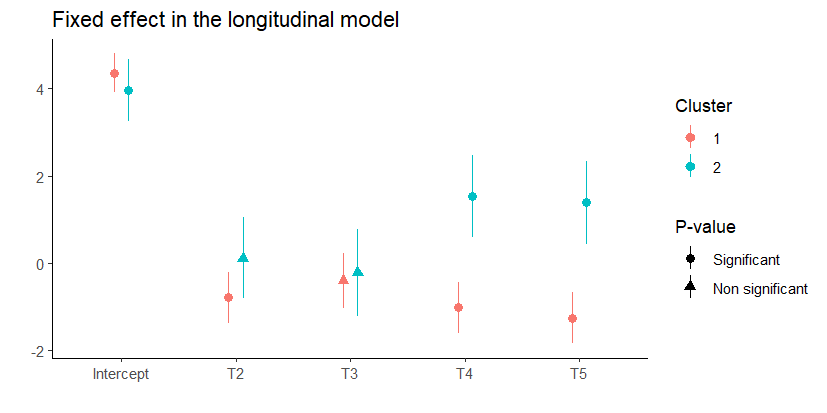
**Figure 5: Probability to belong to clusters using LCA method**. *CPAP, continuous positive airway* *pressure*



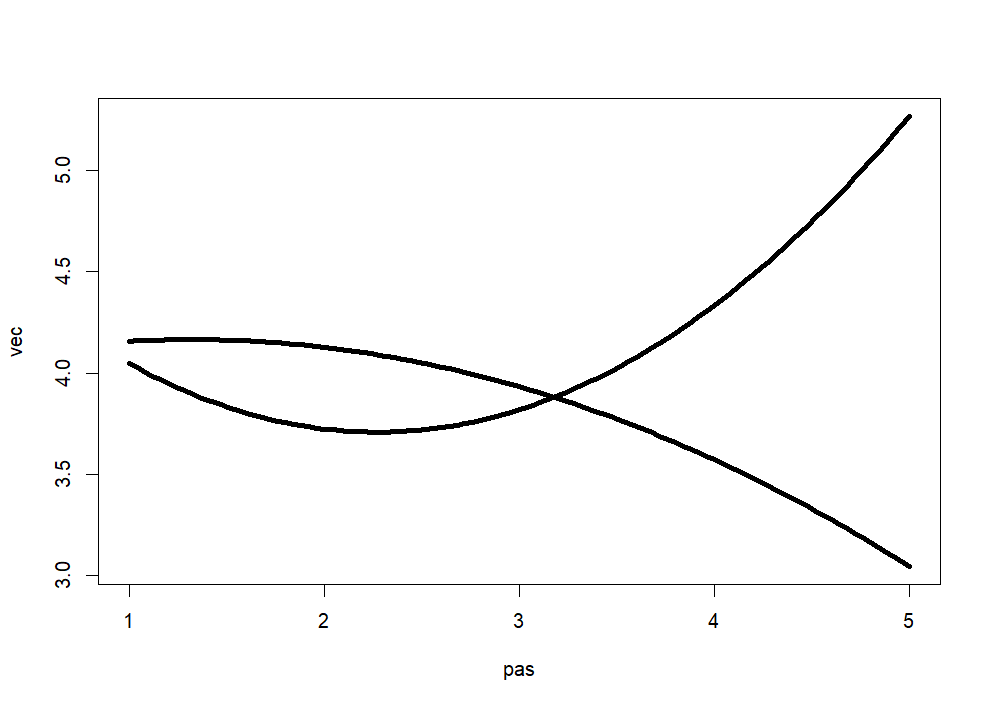
**Figure 6: Trajectories of clusters using K-means.**



**Figure 7: Trajectories of the clusters over 5 time points using GMM model.**



**Figure 8: Trajectories of the two clusters using GBTM model.**



**Figure 9: Prediction of the cumulative risk of sleepiness according to CPAP adherence for one patient, using joint model.**

