**Longitudinal data, trajectories and time series: how to analyze them? An example of sleep data**

Ophélie Coiffier1, Ina Jandric1, Sébastien Bailly1,2

1 Univ. Grenoble Alpes, INSERM U1300, CHU Grenoble Alpes, HP2 Laboratory, Grenoble, France

2 Pole Thorax et Vaisseaux, Laboratoire EFCR (Explorations Fonctionnelles Cardiovasculaire et Respiratoire), CHU Grenoble Alpes, Grenoble, France

**Corresponding author:** Sébastien Bailly,Laboratoire EFCR, CHU de Grenoble, CS10217, 28043, Grenoble, France.E-mail: sbailly@chu-grenoble.fr

**Word count:** 3714 words

**Conflict of interest**

**Funding**

**Authors contribution**

**Abbreviations**

ACF, autocorrelation function; AIC, Akaike information criterion; ANOVA, analysis of variance; ARIMA, autoregression integrated moving average; BIC, Bayesian information criterion; CPAP, continuous positive airway pressure; ESS, Epworth sleepiness scale; GBTM, group-based trajectory modelling; GMM, growth mixture modelling; LCA, latent class analysis; LTA, latent transition analysis

**Keywords**

Longitudinal data; statistical approaches; description; classification; modeling; forecasting

# Abstract (299 words)

**Background:** Repeated measurement data, known as longitudinal data are common in the study of sleep apnea to assess changes in weight, symptoms of insomnia or daytime sleepiness, continuous positive airway pressure (CPAP) parameters such as adherence, leaks, apnea-hypopnea index. These data, collected at multiple points in lifetime, make it possible to examine how differences between individuals and within-person changes in outcome are associated with a longitudinal factor. However, working with longitudinal data presents some challenges, such as the choice of the statistical method to analyze them. An inadequate choice of method can add bias to parameter estimates and outcome predictions.

**Research question:** To provide a guide to the selection and application of statistical techniques for the analysis of longitudinal data.

**Study design and methods:** Eleven main methods used to analyze longitudinal data were described and applied to simulated variables of CPAP adherence and Epworth Sleepiness Scale score. The simulated data correspond to a dataset of 50 patients and 1,000 measurements with continuous, categorical and discrete outcomes.

**Results:** Firstly, the type and source of the data are two important elements in choosing an appropriate statistical method.Secondly, the study question and objectives must be well defined. Finally, the statistical method needs to be verified before it is applied, in particular with regard to its assumptions, limitations, objectives and specific type of data. With increasing application difficulties, there are ANOVA and χ² methods for comparison; LCA, K-means and LTA methods for classification; GMM, mixed model, GBTM, ARIMA and cross-correlation, joint model and Hidden Markov model for modeling, forecasting.

**Interpretation:** Selecting the best statistical method for analyzing longitudinal data requires careful consideration of data type, data source, study objectives and methodological criteria. This study provides a detailed overview of commonly used statistical methods, through a comprehensive step-by-step approach to longitudinal data analysis.

# Introduction

Traditionally, data is measured at specific points in time, making it impossible to analyze changes and evolutions over lifetime1,2. One approach to overcoming this limitation is to repeatedly measure the same variable of interest in a consistent population over a given period. This repeated measurement data is known as longitudinal data1 or trajectories3 with data from multiple time points in life separated by specific intervals, as shown in the study of weight associated with OSA risk2, ethnic disparities with insomnia symptoms4 or daytime sleepiness post-ischemic stroke and transient ischemic attack5. Population-based longitudinal studies are necessary to examine how between-person differences and within-person changes in outcome are associated with longitudinal factors (e.g., body mass2,6, insomnia symptoms4 or daytime sleepiness5). Numerous positioning technologies and remote sensors enable the analysis of a vast data set3. Thanks to this, patient monitoring and predictive analytics are progressively more efficient.

Another category of longitudinal data is the time series, which consist of 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.

However, working with longitudinal data presents some challenges, such as missing or extreme values, correlations between repeated measurements and variations in measurement 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, an inadequate choice of statistical method could add bias in parameter estimates and outcome predictions1.

Consequently, advanced models and methods have been developed, resulting in a need for a guide in selecting and applying statistical techniques for longitudinal data analysis.

# 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 method. Data can be broadly classified into numerical and categorical types (Figure 1). Numerical data can further be divided into discrete data – representing distinct and separate values, such as the number of hospitalization or of arousals per night - and continuous data, which consist of ordered measurements with equal intervals. Continuous data can be categorized into ratio and interval data. Ratio data have a true zero, allowing for operations such as addition, subtraction, multiplication, and division (e.g. height, weight, or blood pressure measurements). In contrast, interval data lack a true zero, meaning that multiplication and division are not meaningful (e.g., temperature).

The method of data collection impacts both data type and data quality (Figure 2). Consequently, variations in data quality and quantity require different statistical approaches. For example:

* Medical records (history or follow-up visits) can provide ordinal, nominal, continuous and discrete data with a high quality and quantity.
* Case Report Form (CRF) or self-administered questionnaires provide ordinal, nominal, discrete and continuous data but their quality and quantity are moderate, often affected by missing data or reporting errors. Moreover, data collected through specific questions may limit the scope of the analysis.
* Interviews generate similar data types with high quality but moderate quantity, since responses are mandatory and the data volume depends on the length of the interview.
* Connected devices primarily collect continuous or discrete data with high quality, though the quantity tends to decline over time due to user compliance issues or device limitations.

Careful consideration of the study’s context, including its objectives, population included, data type and source, and research questions enhances model selection and ensures robust and meaningful analysis.

# Which statistical method to use?

Selecting the appropriate statistical method requires consideration of several key questions:

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

Depending on the research question, the statistical method may differ (Figure 3). The methods also vary in complexity (Figure 4), ranging from relatively simple descriptive analyses to more advanced modeling and forecasting approaches.

Descriptive methods, such as ANOVA or χ², were among the simplest statistical models. They are mostly used to analyze repeated measurements, particularly categorical variables derived from medical records, questionnaires or interviews. Then, the classifications methods, such as LCA on repeated measurements, K-means clustering and LTA for longitudinal data were less complex. Finally, methods used for modeling or forecasting were the most complex statistical approaches, such as GMM, mixed or GBTM methods using longitudinal data; ARIMA and Cross-correlation methods for time series or Joint and Hidden Markov models with mixed data.

In terms of data type and data source, longitudinal data methods incorporate both categorical and numerical data, derived from medical records, questionnaires, connected device or interviews. Time series methods mostly rely on numerical data from medical records or connected device. Mixed data approaches handle both continuous and categorical variables, allowing for diverse data sources.

*Is the chosen statistical method correct?*

Once the type and source of data, as well as the study objectives are clearly defined, it is necessary to verify that the selected statistical method is appropriate. Before conducting the analysis, the goal, Advantages and limitations of the chosen method should be carefully assessed to ensure its suitability. To illustrate the application of different statistical methods, an example of sleep data analysis was carried out. A dataset was simulated using R version 4.3.0, consisting of 300 patients with 90 time points, representing CPAP adherence and ESS score. The CPAP adherence followed a normal distribution (with negative values replaced by 0; μ = 4, σ = 1.5), while the ESS score was a discrete variable ranging from 0 to 24. For certain methods that only accept categorical variables, both variables were transformed into categorical data: non-adherent ([0h; 2h[) vs. almost adherent ([2h; 4h[) vs. adherent (≥4h) for CPAP adherence; with excessive sleepiness (≥10) vs. without (<10) for the ESS score.

First, descriptive methods were applied to compare the population characteristics, providing an initial overview of the data distribution and trends.

1. ANOVA model7

*Objective* - Assess whether there is a statistically significant interaction effect between 2 or 3 within-subjects factors in explaining a continuous outcome.

*Advantages* – Allows comparison between more than 2 groups.

*Limitations* - If the null hypothesis is rejected, it indicates that at least one group differs, but does not specify which group(s), although some post hoc tests are available to counter this limitation; assumes normal distribution, metric scale and equal variance across groups; susceptible to outliers, which can affect the model’s reliability.

*Example* – ANOVA was used to analyze continuous data on CPAP adherence over time, including all time points and all patients. The p-value was 0.88, indicating no statistically significant difference in CPAP adherence across time.

1. χ² method

*Objective* - Evaluate independence and assess differences between variables across a series of contingency tables; assess whether the proportions of the binary variable vary over time.

*Advantages* - Different models are available based on the number of measurements, as Mc Nemar’s test for 2 measurements or Mantel-Haenszel for more than 2 measurements; simple and computationally efficient model.

*Limitations* - All theoretical numbers must be greater than 5; assumes that all individuals transition between states (no dropouts); the sample must be random; does not account for covariates.

*Example* - The χ² Mantel-Haenszel method was applied, assuming that 2 nominal variables are conditionally independent in each stratum and that there is no 3-way interaction. This analysis included 4 time points and all patients, using a contingency table where CPAP adherence and ESS score were treated as categorical variables. The p-value of the χ² test was 0.28, suggesting no significant difference between groups.

Secondly, classification methods were used to cluster patients and summarize key information. Various clustering approaches employed different distance metrics including Euclidean8, Manhattan8, Cosine8, Correlation-based8 or dynamic time warping9–11 distances.

1. LCA method12–18

*Objective* - Identify unobserved (latent) clusters sharing common characteristics.

*Advantages* - Powerful tool for analyzing relationships between categorical variables, exploring and interpreting complex contingency tables and testing hypotheses on the structures of categorical latent variables; low classification error rate and a robust model; supports mixed data types, allowing for variables with different measurements scales; if continuous variables are involved, Latent Profile Analysis can be applied.

*Limitations* – Computationally intensive, limiting the number of variables based on available computing power; sensitive to outliers; the percentage of individuals in each cluster is unknown; requires making several a priori decisions, such as the number of clusters.

*Example* – LCA was performed using CPAP adherence as a categorical variable across 5 time points. The optimal number of clusters was determined to be 2, based on the smallest AIC and BIC (see Supplementary Material (SM)). At the 1st, 2nd and 3rd time points, patients classified as adherent were more likely to belong to the Cluster 1 as adherent patients had a higher probability of belonging to the Cluster 2 (Figure 5). At the 4th and 5th time points, Cluster 1 had a greater probability of including almost adherent patients while the Cluster 2 was more likely to include adherent (see table in SM).

1. K-means method13,19–21

*Objective* - 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.

*Limitations* - 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: 51% in the 1st cluster and 49% in the second cluster (Figure 6). Both groups remained relatively stable around 4 hours for the first 3 measuring points, then the 1st group decreased to around 3 hours at the 4th measuring point and finally increased to around 5 hours at the last measuring point, while the second group increased after the 3 measuring points from around 4 hours to around 5 hours and finally decreased at the last measuring point to around 4 hours.

1. LTA model13,17,22

*Objective* - 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.

*Limitations* - 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 categorical CPAP adherence was used. All patients were included with all 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 0.17 for the CPAP adherence [0h; 2h[, 0.65 for the [2h; 4h[ group and 0.19 for the ≥4h group (Figure in SM). To the 1st cluster, the conditional response probabilities were, respectively 0.06, 0.29, 0.65 (table in SM). The transition probability was 0.28 from the 1st cluster to the 2nd cluster, 0.62 from the 2nd cluster to the 1st cluster and 0.38 to stay in the 2nd cluster, 0.72 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 model17,23,24

*Objective* - 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.

*Limitations* - Missing data must be 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 linear curve and 2 clusters was the best model. The first group decreased their CPAP adherence (from more than 4h to less than 3.5h), while the second group increased it slightly (around 4h, Figure 7).

1. Mixed model1,23,25,26

*Objective* - 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.

*Limitations* - 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 Missing At Random (MAR).

*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 was not significantly associated with any of the time points or with the ESS baseline.

1. GMM model13,17,21,27–30

*Objective* - 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.

*Limitations* - 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 group tended to decrease over the first 4 time points, then to increase, while the second group tended to decrease over the first few time points, then to stabilize thereafter (Figure 8).

1. ARIMA model & Cross-correlation method31–38

*Objective* - 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.

*Limitations* - 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)(1, 0, 1) while for ESS score, it was ARIMA(0, 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 (seen on the ACF plot, in SM).

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 model25,39

*Objective* - 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.

*Limitations* - 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 = 0.47 > 0.05).

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

1. Hidden Markov model40–46

*Objective* - 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.

*Limitations* - 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: Non-adherent vs. 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 1st state. The transition probability matrix was around 0.32 for the initial state of non-adherence to the arrived states of adherence, 0.40 for the state of adherence to the non-adherent group, 0.60 for remaining in the adherent state and 0.68 for remaining in the non-adherent state (Table 1). The states prediction included 55.6% of time points in the 1st state and 44.4% in the 2nd state (detail in SM).

# Summary and perspectives

Selecting the best statistical method for analyzing longitudinal data requires careful consideration of data type, data source and study objectives. Statistical methods vary in complexity, depending on the data characteristics and the study design. Before conducting an analysis, it is essential to evaluate the Advantages and limitations of each method to ensure its suitability. Some methods are more suitable for comparisons (e.g., ANOVA, χ² tests), classification (e.g., LCA, K-means clustering, LTA), or modeling and forecasting (e.g., GBTM, mixed models, GMM, ARIMA, cross-correlation, joint models, and Hidden Markov Models).

This study did not consider all possible data limitations, yet proper data description and verification remain crucial first steps in longitudinal data analysis. Outliers and missing values are common in such datasets, and specific techniques exist for their imputation or removal47. Another limitation of the study is its focus on the most commonly used statistical methods. Many other techniques can be used for classification (e.g., Configural Frequency Analysis13, Latent growth Curve Model23, hierarchical methods48–50, such as Principal Component Analysis, Multiple Correspondence Analysis, Hierarchical Ascending Classification, as well as partitioning methods48–50 like X-means, DBSCAN, and K-medoids) or for modeling and forecasting (e.g.,BKMR51). There are more complex methods that can be applied to predict data trajectories, but were outside the scope of this study.

This work is not an exhaustive review of statistical methods, but provides a detailed overview of methods commonly used in the literature (illustrated with implementation examples in R and Python). A comprehensive, step-by-step approach for analyzing longitudinal data is proposed.

Longitudinal data are particularly relevant in healthcare studies, where they are often collected through follow-up visits, hospitalizations and, more recently, through connected devices (e.g., CPAP treatment for sleep apnea). However, the methodological framework outlined in this study is broadly applicable to other domains, including climat34,38, finance52 or insurance53,54.

# 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. Qian Y, Perret JL, Hamilton GS, et al. Early-to-Midlife Body Mass Index Trajectories and Obstructive Sleep Apnoea Risk 10 Years Later. *Respirology* [Internet] [cited 2025 Feb 14];n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/resp.70002

3. 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.

4. Singh R, Atha R, Lenker KP, et al. Racial/ethnic disparities in the trajectories of insomnia symptoms from childhood to young adulthood. *Sleep* 2024;47(5):zsae021.

5. Baillieul S, Tamisier R, Gévaudan B, et al. Trajectories of self-reported daytime sleepiness post-ischemic stroke and transient ischemic attack: A propensity score matching study versus non-stroke patients. *Eur Stroke J* 2024;9(2):451–459.

6. Liu Y, Barnet JH, Hagen EW, Peppard PE, Mignot E, Reither EN. Objectively measured daytime sleepiness predicts weight change among adults: Findings from the Wisconsin Sleep Cohort Study. *Sleep Health* 2024;10(3):327–334.

7. 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

8. Kumar V, Chhabra JK, Kumar D. Performance Evaluation of Distance Metrics in the Clustering Algorithms. *INFOCOMP Journal of Computer Science* 2014;13(1):38–52.

9. Bottaz-Bosson G, Hamon A, Pépin J-L, Bailly S, Samson A. Continuous positive airway pressure adherence trajectories in sleep apnea: Clustering with summed discrete Fréchet and dynamic time warping dissimilarities. *Statistics in Medicine* 2021;40(24):5373–5396.

10. Yuan Y, Chen Y-PP, Ni S, et al. Development and application of a modified dynamic time warping algorithm (DTW-S) to analyses of primate brain expression time series. *BMC Bioinformatics* 2011;12(1):347.

11. Tormene P, Giorgino T, Quaglini S, Stefanelli M. Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. *Artificial Intelligence in Medicine* 2009;45(1):11–34.

12. 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.

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

14. 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

15. 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.

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

17. 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.

18. 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.

19. 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.

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

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

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

23. 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.

24. 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.

25. 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.

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

27. 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.

28. 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.

29. 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

30. 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.

31. 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.

32. 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

33. 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.

34. 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.

35. 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.

36. 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.

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

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

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

45. 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.

46. 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.

47. Bottaz-Bosson G, Midelet A, Mendelson M, et al. Remote Monitoring of Positive Airway Pressure Data: Challenges, Pitfalls, and Strategies to Consider for Optimal Data Science Applications. *CHEST* 2023;163(5):1279–1291.

48. Kaya M-F, Schoop M. Analytical Comparison of Clustering Techniques for the Recognition of Communication Patterns. *Group Decis Negot* 2022;31(3):555–589.

49. Rujasiri P, Chomtee B. Comparison of Clustering Techniques for Cluster Analysis. *Agriculture and Natural Resources* 2009;43(2):378–388.

50. Rodriguez MZ, Comin CH, Casanova D, et al. Clustering algorithms: A comparative approach. *PLOS ONE* 2019;14(1):e0210236.

51. Devick KL, Bobb JF, Mazumdar M, et al. Bayesian kernel machine regression–causal mediation analysis. *Stat Med* 2022;41(5):860–876.

52. Yay Donderici E, Forbes SP, Zhang NJ, et al. Cost-effectiveness of blood-based colorectal cancer screening – a simulation model incorporating real-world longitudinal adherence. *Expert Review of Pharmacoeconomics & Outcomes Research* 0(0):1–7.

53. Massari S, Bauleo L, Gariazzo C, et al. Cancer mortality and sectors of employment: a cohort study in Italy. *BMC Public Health* 2025;25(1):458.

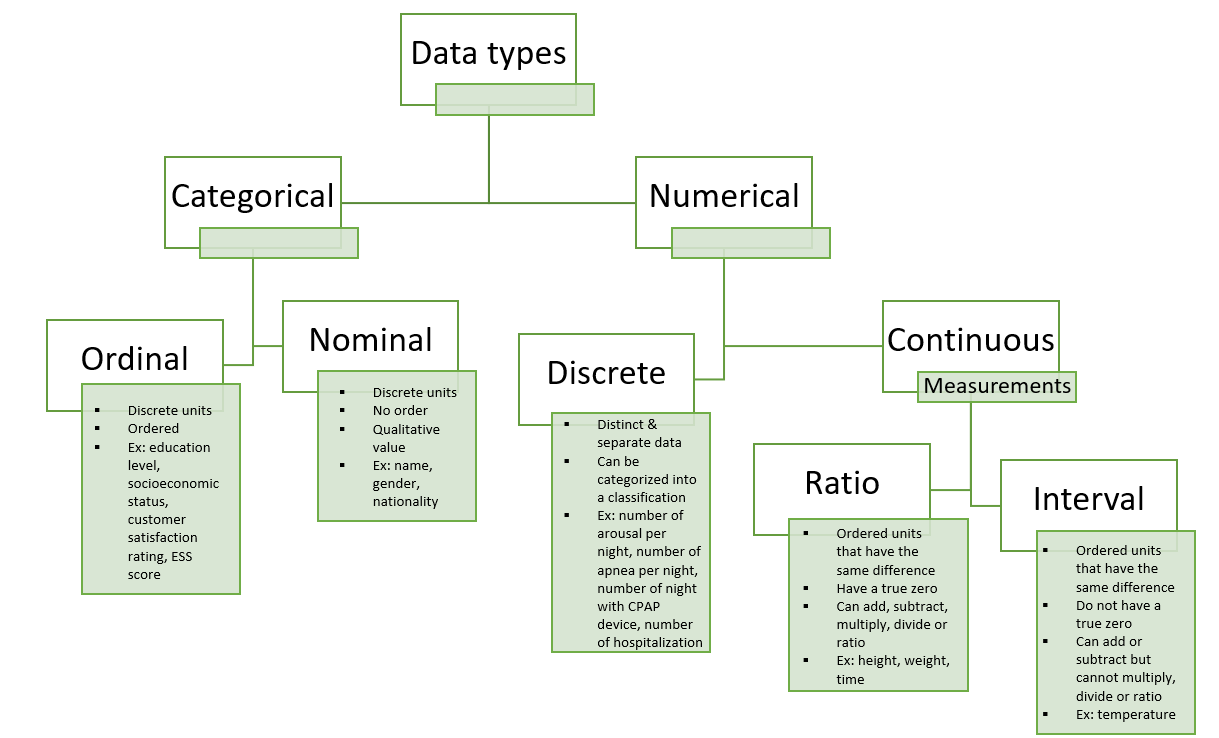
54. Serván-Mori E, Pineda-Antúnez C, Cerecero-García D, et al. Health system financing fragmentation and maternal mortality transition in Mexico, 2000–2022. *Int J Equity Health* 2025;24:32.

# Tables

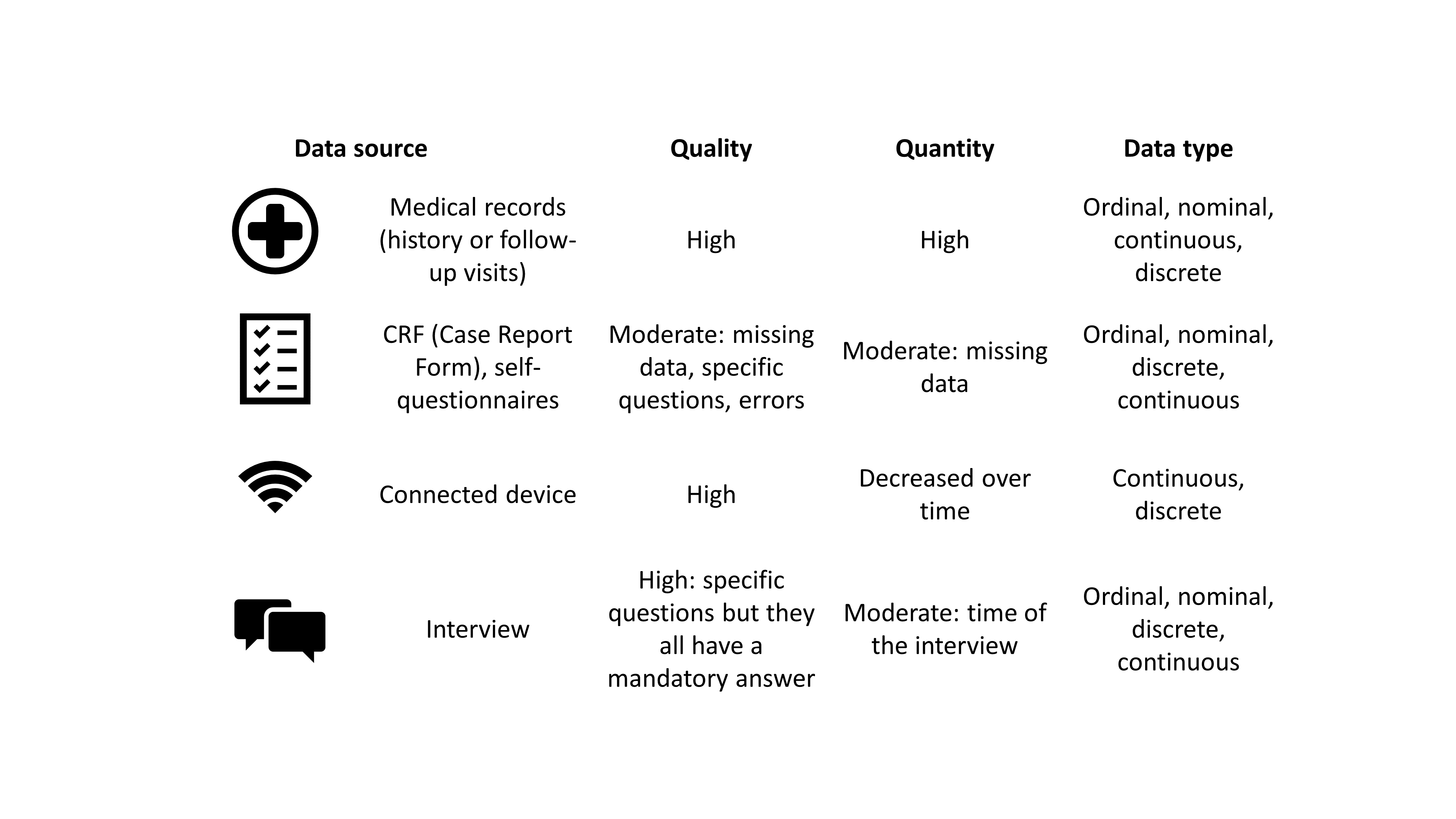
**Table 1: Matrix of transition probability from number of hours to CPAP use to hidden state of CPAP adherence using Hidden Markov model.**

|  |  |  |
| --- | --- | --- |
|  | **Non-adherent** | **Adherent** |
| **Non-adherent** | 0.68 | 0.32 |
| **Adherent** | 0.40 | 0.60 |

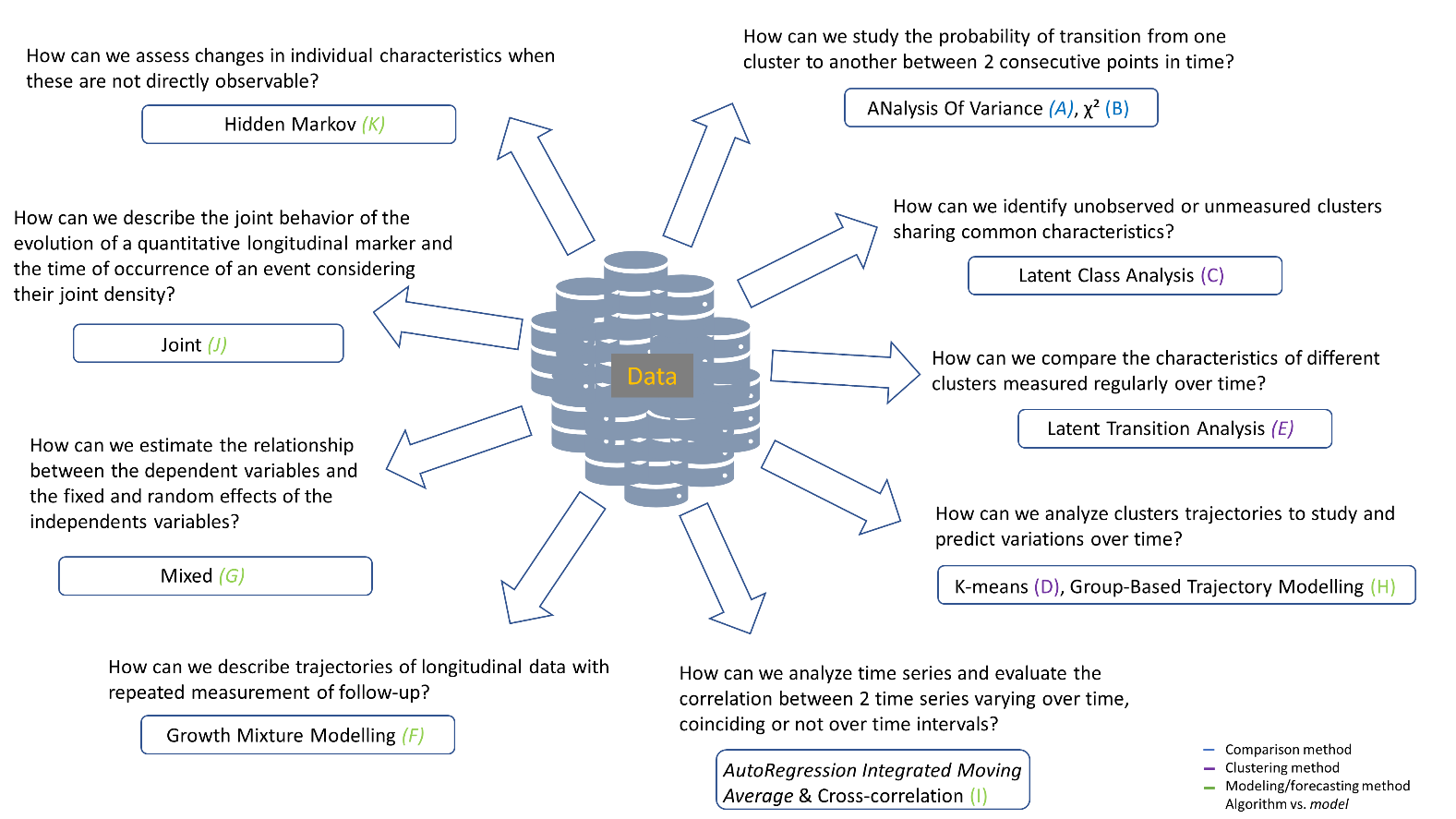
# Figures legend

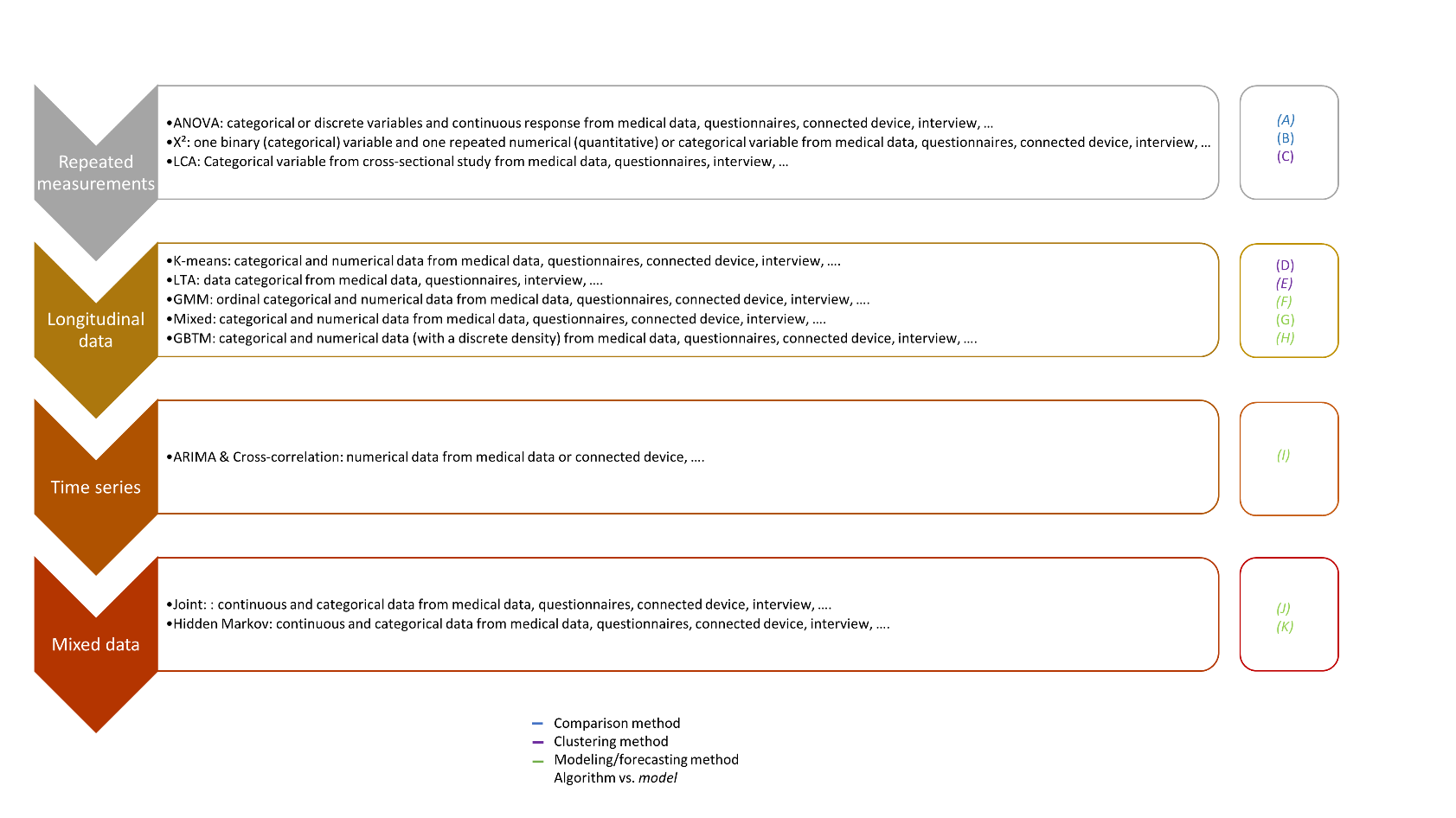
**Figure 1: Data types**. *CPAP, continuous positive airway pressure; ESS, Epworth sleepiness scale*

**Figure 2: Data sources**.

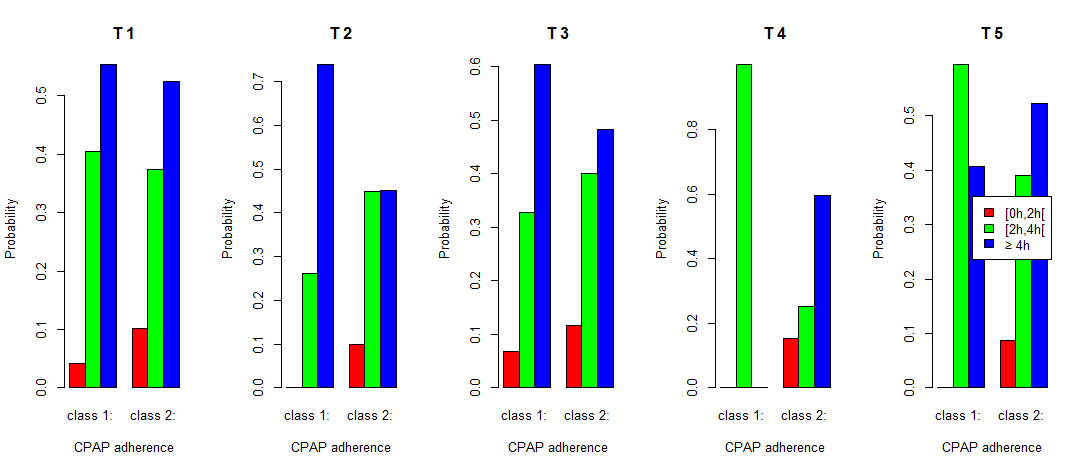


**Figure 3: Statistical methods according to objectives and question of the study**.

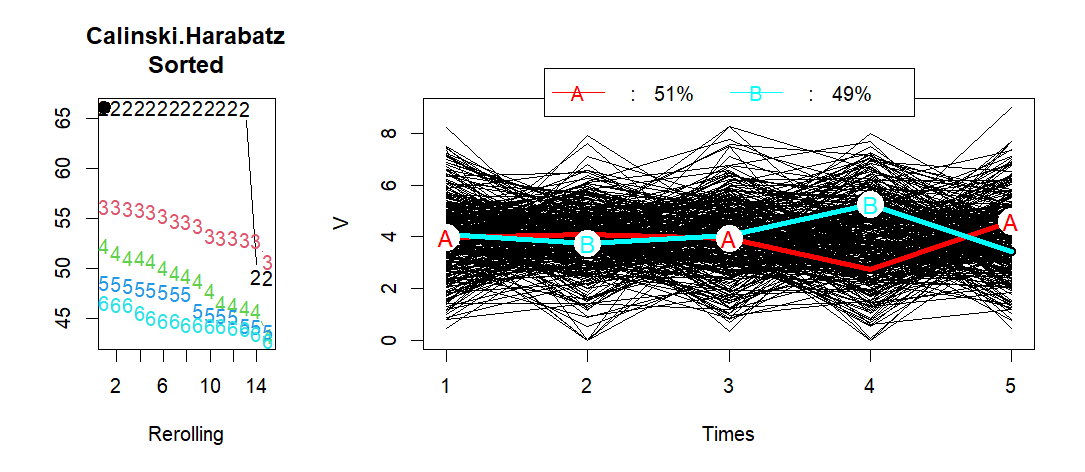


**Figure 4: Statistical methods classified according to complexity, data format and objectives**. *ANOVA, analysis of variance; ARIMA, autoregression integrated moving average; DTW, dynamic time warping; GBTM, group-based trajectory modelling; GMM, growth mixture modelling; LCA, latent class analysis;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

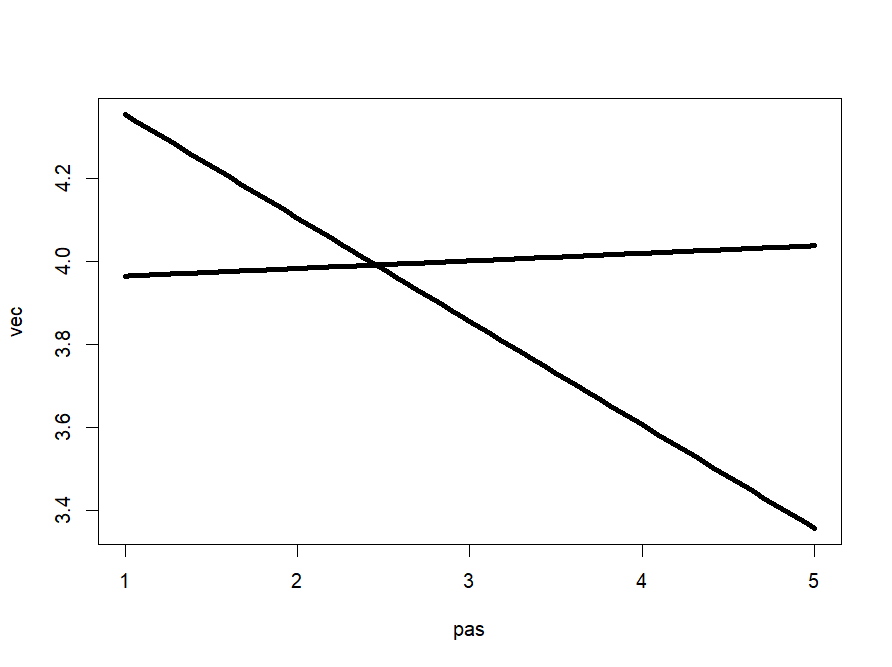
**Figure 5: Probability to belong to clusters using LCA method**. *CPAP, continuous positive airway* *pressure*



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



**Figure 7: Trajectories of the two clusters using GBTM model.**



**Figure 8: Trajectories of the clusters over 5 time points using GMM model.**

