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

**Example of sleep data**

**Title for CHEST (50 characters), suggestions:**

1. **How to analyze longitudinal data? Sleep modeling**
2. **Analyzing longitudinal data - Methodology overview**
3. **Modeling framework for longitudinal data**
4. **Statistical methods for longitudinal data analysis**
5. **Longitudinal Data Analysis - Key Methodologies**

Authors

Affiliations

**Corresponding author**

**Word count:** 3727 words

**Conflict of interest**

**Funding**

**Authors contribution**

**Abbreviations**

ANOVA, analysis of variance; ARIMA, autoregression integrated moving average; 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; AIC, Akaike information criterion; BIC, Bayesian information criterion; ACF, autocorrelation function

**Keywords**

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

# Abstract (word count)

Background

Aims

Methods

Results

Conclusion

# Introduction

Traditionally, data is measured at a specific point in time, making it impossible to analyze changes and evolutions over time1. 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 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 dataset, namely trajectory data. 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 data source?

The type of data may depend on the data source, and may impact the choice of the 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 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 characteristics require different statistical approaches. For example:

* Medical records (history or follow-up visits) can provide ordinal, nominal, continuous and discrete data with a good 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, as responses are mandatory and the data volume depends on the time of 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 question 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 classification 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, GBTM or Mixed method 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 devices. 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 objective, 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, consisting of 50 patients with 1000 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 model3

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

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

*Limits* - 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; 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 0.09, indicating no statistically significant difference in CPAP adherence across time.

1. χ² method

*Objective* - Evaluate independence and 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 - Mc Nemar’s test for 2 or Mantel-Haenszel for more than 2 measurements; simple and computationally efficient model.

*Limits* - 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 0.16, 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 Euclidean4, Manhattan4, Cosine4, Correlation-based4 or dynamic time warping5–7.

1. LCA method8–14

*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.

*Limits* – 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 first time point, patients classified as almost adherent were more likely to belong to the Cluster 1, while adherent patients had a higher probability of belonging to the Cluster 2 (Figure 5). At the 2nd time point, Cluster 1 was mainly composed of adherent patients, while Cluster 2 included both almost adherent and adherent patients in similar proportions. At the 3rd, 4th and 5th time points, Cluster 1 had a greater probability of including adherent patients while the Cluster 2 was more likely to include almost adherent and, to a lesser extent, adherent patients (see table in SM).

1. K-means method9,15–17

*Objective* - Group patients’ trajectories into clusters based on their similarities.

*Advantages* – Does not require a priori assumptions and avoids the problems associated with model selection; capable of handling large datasets; can cluster trajectories that do not adhere to a polynomial trajectory.

*Limits* – Requires complete data; if missing values are present, imputation is necessary before statistical analysis, or a method that accommodates missing data must be applied; does not account for correlations between individuals; requires tests to determine initial parameters and the optimal number of clusters; may converge to a local rather than a global maximum, with no definitive way to verify it; evaluates the longitudinal trajectory of a single variable; clusters trajectories based on overall shape similarity, if 2 trajectories show a temporal shift, they may be assigned to separate clusters despite having similar patterns; lacks 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. Model parameters are detailed in SM. Based on the Calinski-Harabatz score, the optimal model consisted of 2 clusters. The distribution of individuals across clusters was relatively balanced, with 66% in Cluster 1 and 34% in Cluster 2 (Figure 6). The first cluster showed a fairly stable adherence around 4h, while the second cluster showed an increase in adherence after the third time point (from 4h to around 5h and 30 min).

1. LTA model9,13,18

*Objective* – Analyze the probability of transitioning from one cluster at a given point to another at the next.

*Advantages* - Model changes over time and identifies predictors of that change; enables comparison of different clusters to determine their characteristics and assess the contribution of different measures to each latent cluster.

*Limits* – Requires a large dataset, as the model has to estimate many parameters and typically uses a burn-in process; higher number of time points (>6) increases the complexity of the model; challenges in defining the optimal number of latent clusters and assigning appropriate labels; difficulties in incorporating covariates into the model.

*Example* - LTA model ~~implemented discrete or categorical outcome, thus~~ categorical CPAP adherence data was used. All patients were included with 500 time points. Model parameters and validation are described in SM. Based on Log-Likelihood, BIC or AIC, the optimal model identified 2 clusters. Conditional response probabilities for belonging to the 2nd cluster, were 0.22 for the [0h; 2h[ adherence group, 0.59 for the [2h; 4h[ group and 0.19 for the ≥4h group (Figure 9). In the 1st cluster, these probabilities were 0.03, 0.33, 0.63, respectively (see table in SM). The transition probabilities were: 0.29 probability of moving from the 1st to the 2nd cluster, 0.67 from the 2nd to the 1st cluster , 0.33 to remain in the 2nd cluster and 0.71 to remain in the 1st cluster.

Thirdly, modeling and forecasting methods were applied to analyze and compare data trajectories, as well as to stimulate future trajectories.

1. GBTM model13,19,20

*Objective* – Analyze cluster trajectories in order to study and forecast variations over time.

*Advantages* - Simpler than the GMM method, as there are fewer parameters to estimate; faster computation with lower error rates; ability to handle missing values and correlated residuals; easier to interpret, especially visually, due to its reduced complexity.

*Limits* - Missing data must be Missing Completely At Random; clusters must be qualitatively distinct from the dependent variable; assumes no direct relationship between dependent and independent variables; strong assumptions on trajectory distributions must be respected; may overestimate the number of clusters and trajectories, particularly when individual trajectories follow a similar pattern and are distributed on a continuum around the mean trajectory; does not account for no intra-class variation.

*Example* - Continuous CPAP adherence data was analyzed, including 5 time points and all patients. Various curves and numbers of clusters were tested to determine the optimal model (detailed in SM). Based on multiple criteria – including BIC, loglikelihood, Average Posterior Probability and Proportion of assignment parameters - the best-fitting model had 2 clusters. The first cluster showed a decreased CPAP adherence unlike the second cluster (Figure 8).

1. Mixed model19,21–23

*Objective* – Estimate the relationship between the dependent variables and both the fixed and random effects of the independent variables.

*Advantages* – Allows for the simultaneous analysis of 2 or more dependent variables; ability to deal with missing values; enables the estimation of odds and rate ratios.

*Limits* - Interpretation of coefficients possible if random effects are controlled by the analyst; even if the difference between estimated trajectories of the dependent variable is statistically significant, this difference may not be clinically relevant; unobserved variables are assumed to be Missing At Random (MAR).

*Example* - ~~Any type of variable was accepted~~ but for this example, a continuous outcome was analyzed while time and baseline ESS score were treated as categorical variables. All patients and all time points were included, with a random intercept for each patient:

Model validation and results are detailed in SM. Findings indicated that CPAP adherence was negatively associated with certain time points, while baseline ESS score was not significantly associated with CPAP adherence.

1. GMM model9,13,17,24–27

*Objective* - Identify trajectory patterns and describe longitudinal changes within each unobserved group.

*Advantages* - Accommodates missing data and correlated residuals; captures both between- and within-individual differences over time; allows for qualitative trajectory changes across different groups.

*Limits* - Many parameters are estimated; interpreting results can be challenging; some parameters must be defined a priori; risk of identifying false clusters.

*Example* – All patients were included across 5 time points. A random intercept and slope were incorporated, along with a mixture parameter for the time variable:

Based on BIC, the optimal model consisted of 2 clusters (detailed in SM). The distribution of patients across clusters was relatively balanced (detailed in SM). Over the 5 time points, the first cluster demonstrated a decline in CPAP adherence, whereas the second cluster showed an increasing adherence trend (Figure 7).

1. ARIMA model & Cross-correlation method28–35

*Objective* - Analyze time series and evaluate the correlation between two time series that vary over time, coinciding or not across time intervals.

*Advantages* - Assumption of local stationarity only; robust results even if non-linear trends are present in the data or time series have different time scales; ability to identify correlations in complex systems with multiple linked signals.

*Limits* – Requires linear relationships between multiple signals; better performance with at least 100 observations.

*Example* - First, the ARIMA model was applied to transform numerical outcomes into time series. CPAP adherence and the ESS score were both transformed into time series, including all time points, though the analysis was conducted on a single patient. The model can be repeated for each patient. The frequency parameter was set to 7 to reflect a weekly scale. To validate the time series, autocorrelation, partial autocorrelation, QQpLot and Box-Ljung test were studied (in SM). The final models selected were: ARIMA(0, 0, 0) for CPAP adherence and ARIMA(1, 1, 0)(1, 0, 0) for ESS score.

Next, a cross-correlation function was performed to analyze the relationship between the two time series after detrending them using the ARIMA model.

Scatterplots revealed no correlation between ESS score and CPAP adherence, with or without lag, except for the 11th time points, as observed on the ACF plot (Figure 10). Interestingly, this suggests that a higher CPAP adherence value is likely to lead to a higher ESS score, approximately11 time points later.

These lags could be incorporated into a regression model to further study the association between CPAP adherence and ESS score at different time lags (details in SM).

1. Joint model21,36

*Objective* - Account for the joint behavior of the evolution of a quantitative longitudinal marker and the time-to-event occurrence, considering their joint density.

*Advantages* – Unbiased estimates of the regression coefficients; 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; accommodates patients lost to follow-up by incorporating them into the survival model.

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

*Example* – A linear mixed-effects model was first developed with sex of the patient as a covariate. This was a categorical variable, randomly assigned as Male or Female. The joint modeling process was separated into three steps: 1) the linear mixed-effect model, 2) the Cox model and 3) the joint model.

The analysis included all patients and seven time points.

First, the mixed model examined continuous CPAP adherence with a random intercept and slope per patient. Parameters of the model are detailed in SM.

No variables were found to be statistically significant.

Next, the Cox model was applied using the categorical ESS score: ESS score < 10 was coded as event occurrence (death) and ESS score ≥ 10 was coded as no event (survival). The sex variable was added as a covariate and the model was clustered by patient:

Model validation and results are detailed in SM. However, according to the results, the sex did not significantly influence the survival curve (p-value = 0.64 > 0.05).

Finally, a joint model combining the mixed-effects and survival model was implemented. However, no significant associations were found. While the model validation showed poor fir for CPAP adherence, the other parameters were well-fitted (detailed in SM).

For example, the prediction of the cumulative risk for the patient 49 increased (risk of having ESS score < 10) from the 4th time point (cumulative risk around 0.00) to the 5th time point (cumulative risk around 0.58) (Figure 9).

1. Hidden Markov model37–43

*Objective* - Assess changes in individual characteristics when these are not directly observable.

*Advantages* - Flexible and robust approach for Bayesian estimation; more appropriate for small sample sizes; 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); algorithms are computationlly efficient; possible predictions of future states.

*Limits* - The number of classes must be well chosen to prevent overestimation or missing occasional clusters; no standardizedcriteria for selecting 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 low dimensionality for the model to function properly; Bayesian estimation assumes that the distribution of model parameters is known a priori; label change imposes an order restriction on parameters for different states; patients lost to follow-up cannot be reassigned to other states.

*Example* - This method requires at least one observed categorical variable. In our analysis, CPAP adherence was categorized into 3 states and paired with one hidden categorical variable with a predefined two-state classification: Adherent vs. Non-adherent. The analysis included all time points, but was performed on a single patient (the model can be repeated for each patient). Model parameters and results are detailed in SM. The initial state probability was 1.0 for the second state. The transition probability matrix indicated a probability of around 50% of transitioning from the initial non-adherent state to either of the two outcome states, around 60% probability of switching from the adherent to the non-adherent group and 40% probability of 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).

# 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 strengths 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 removal44. 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 Analysis9, Latent growth Curve Model19, hierarchical methods45–47, such as Principal Component Analysis, Multiple Correspondence Analysis, Hierarchical Ascending Classification, as well as partitioning methods45–47 like X-means, DBSCAN, and K-medoids) or for modeling and forecasting (e.g.,BKMR48). 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 climate31,35, finance49 or insurance50,51.

# 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. 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.

5. 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.

6. 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.

7. 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.

8. 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.

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

10. 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

11. 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.

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

13. 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.

14. 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.

15. 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.

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

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

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

19. 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.

20. 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.

21. 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.

22. 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

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

24. 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.

25. 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.

26. 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

27. 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.

28. 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.

29. 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

30. 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.

31. 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.

32. 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.

33. 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.

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

35. 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.

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

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

42. 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.

43. 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.

44. 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.

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

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

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

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

49. 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.

50. 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.

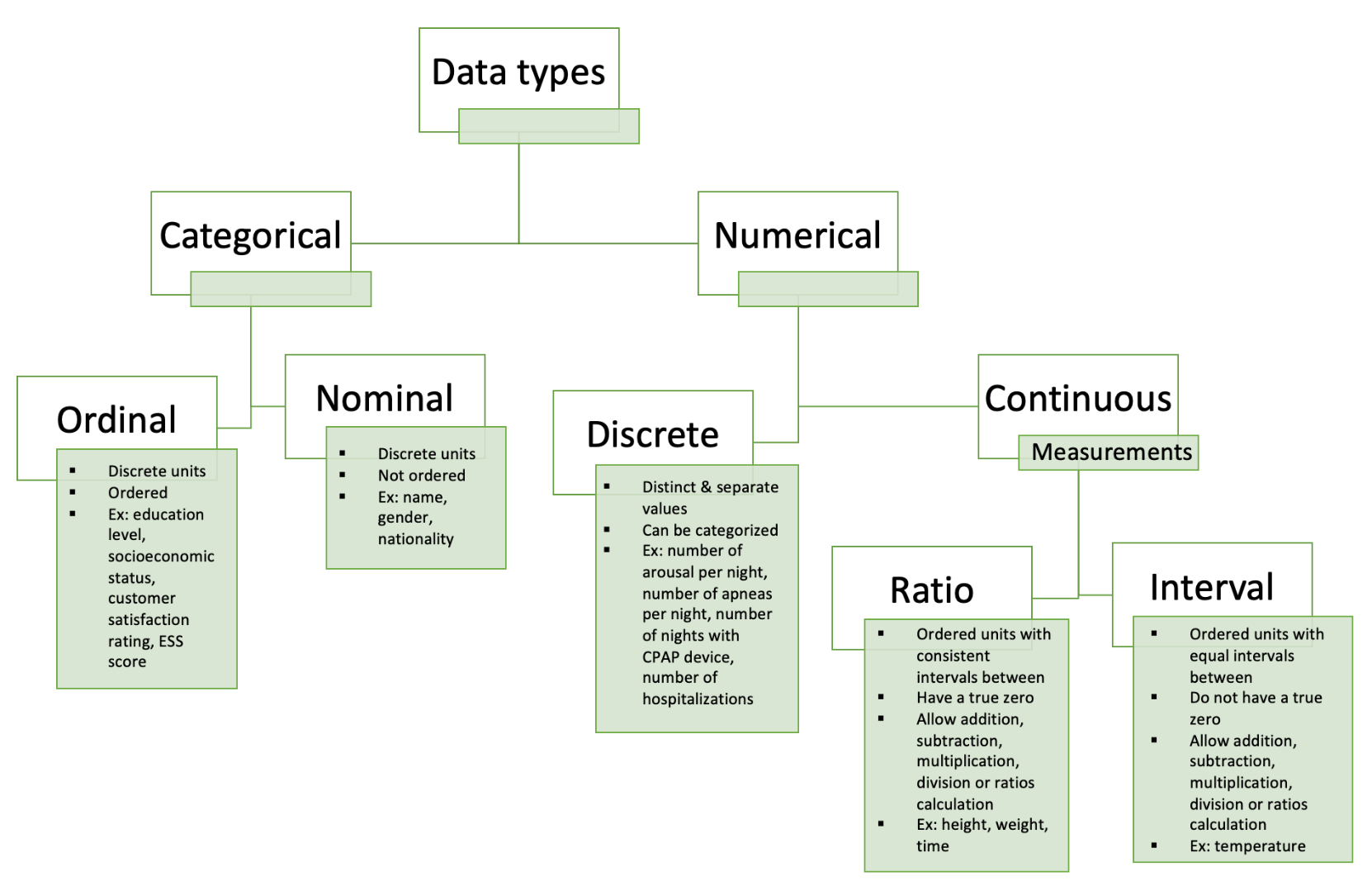
51. Serván-Mori E, Pineda-Antúnez C, Cerecero-García D, et al. Health system financing fragmentation and maternal mortality transition in Mexico, 2000–2022. *Int J Equity Health* 2025;24:32.

# Tables

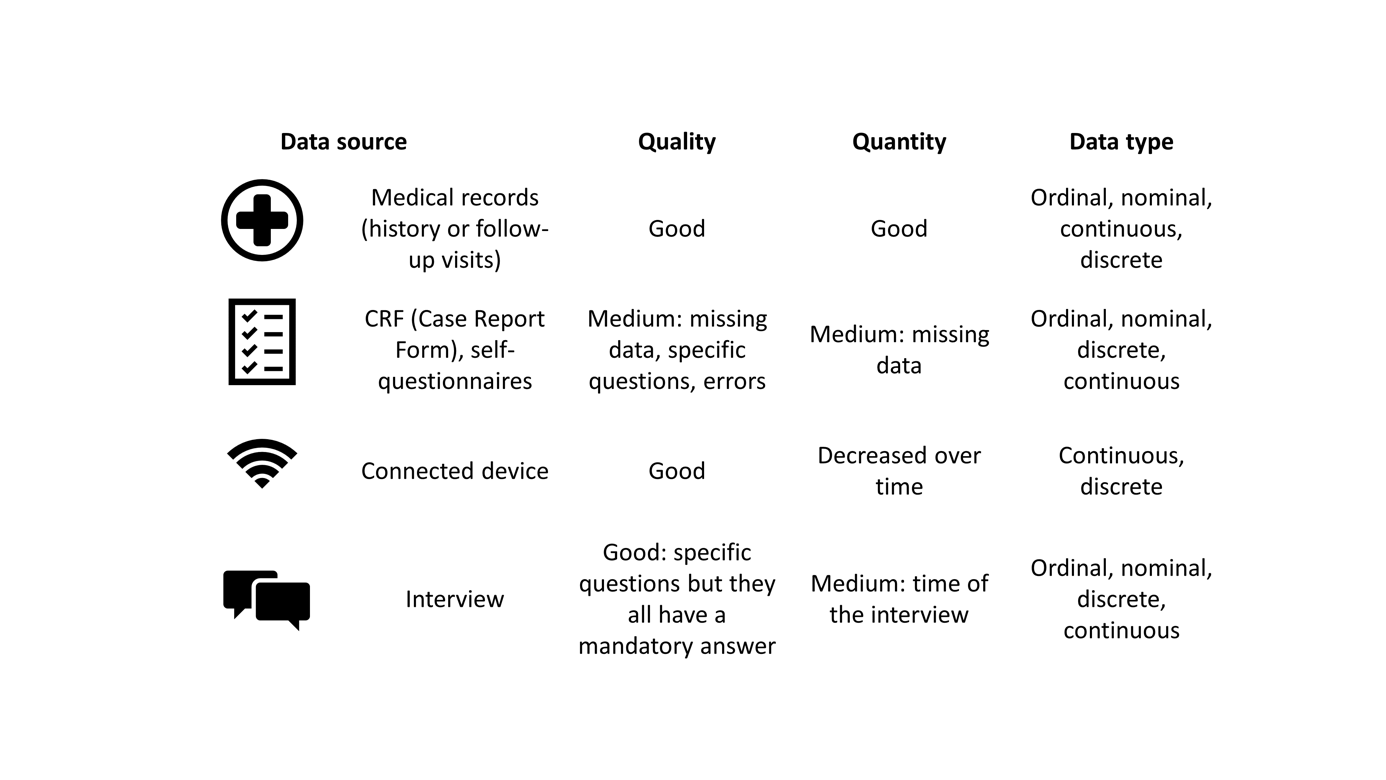
**Table 1: Transition probability matrix** **mapping CPAP usage duration (in hours) to latent adherence states, as estimated by the Hidden Markov Model**

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

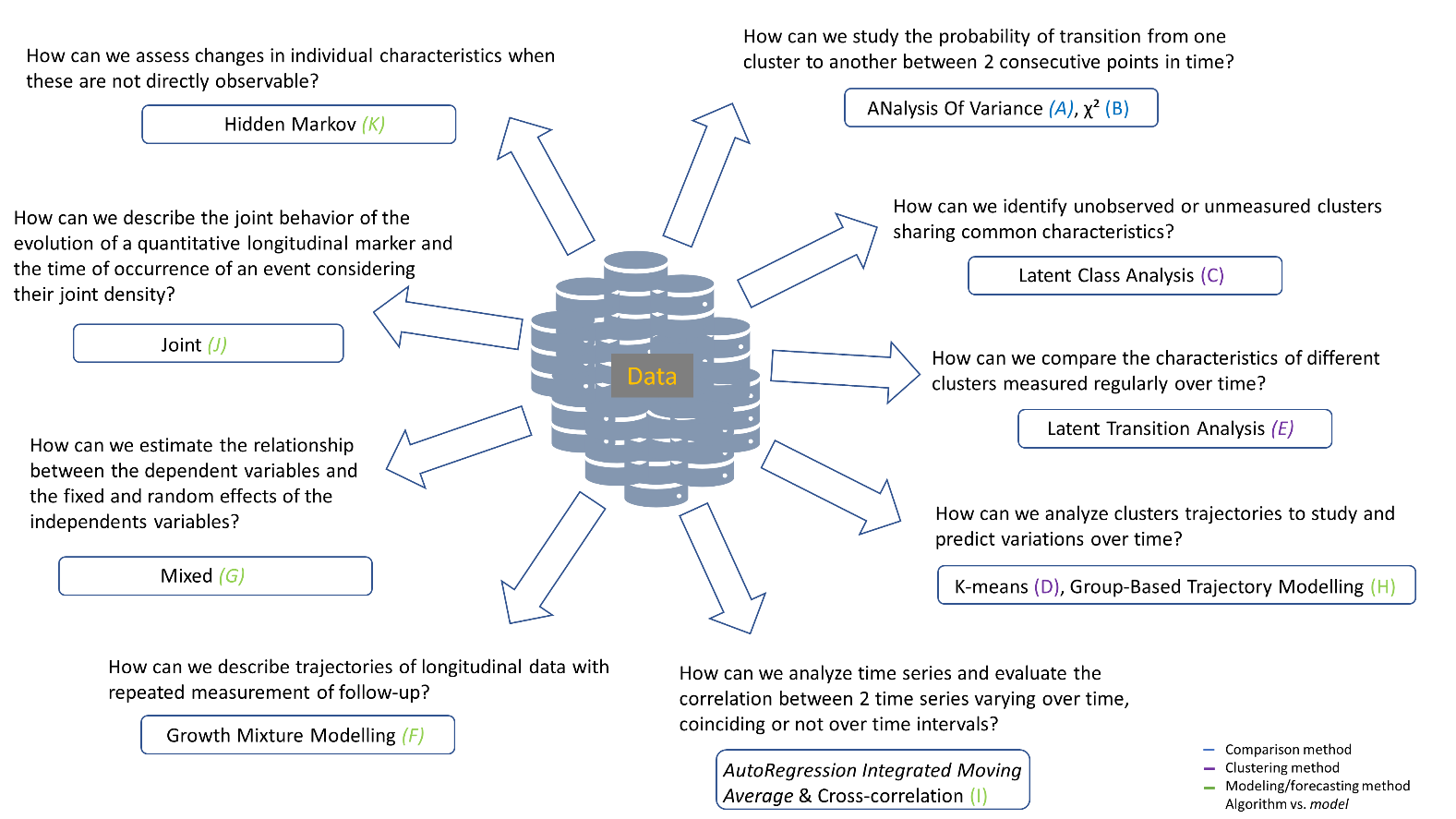
# Figures legend

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

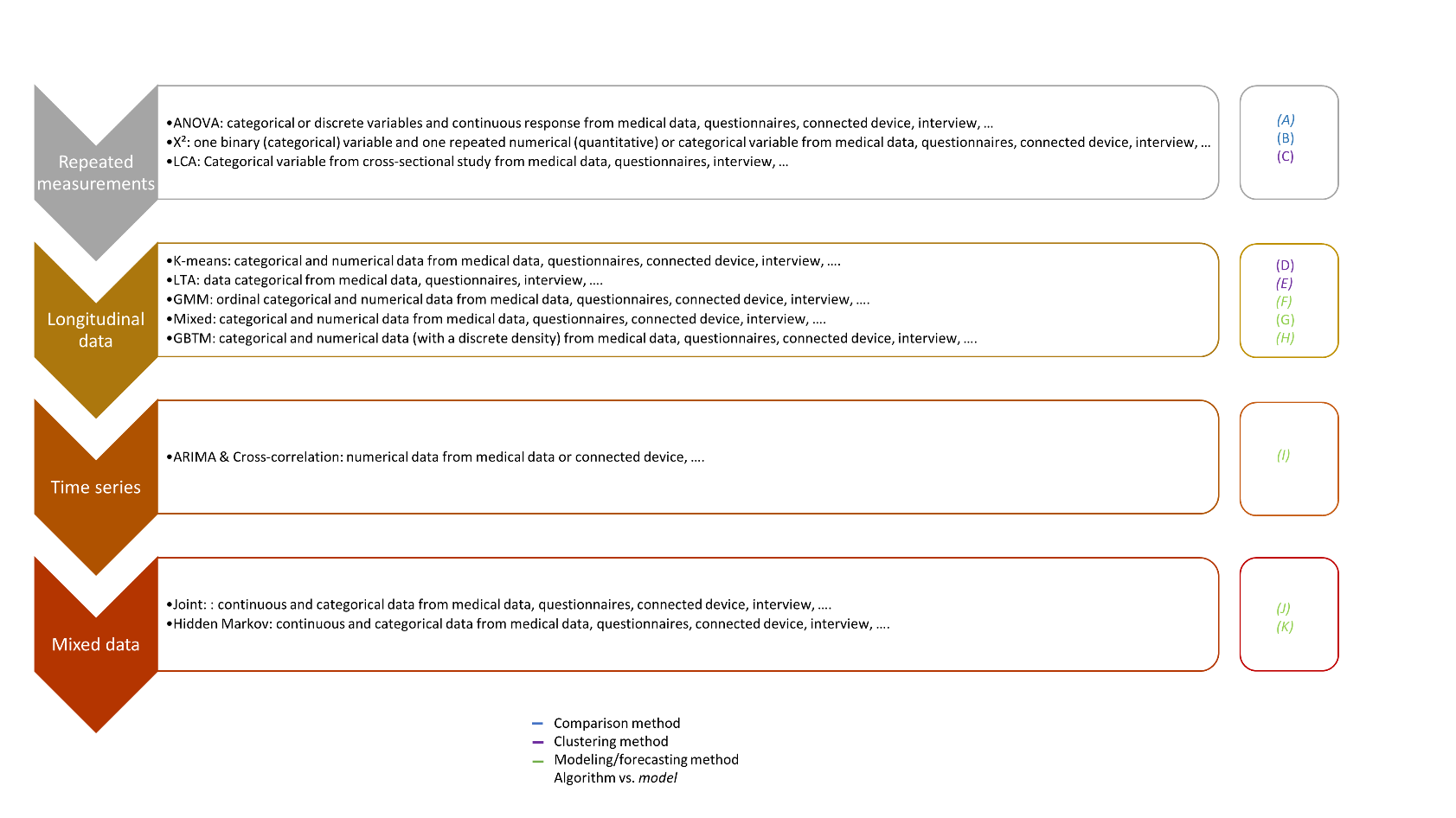
**Figure 2: Data sources**



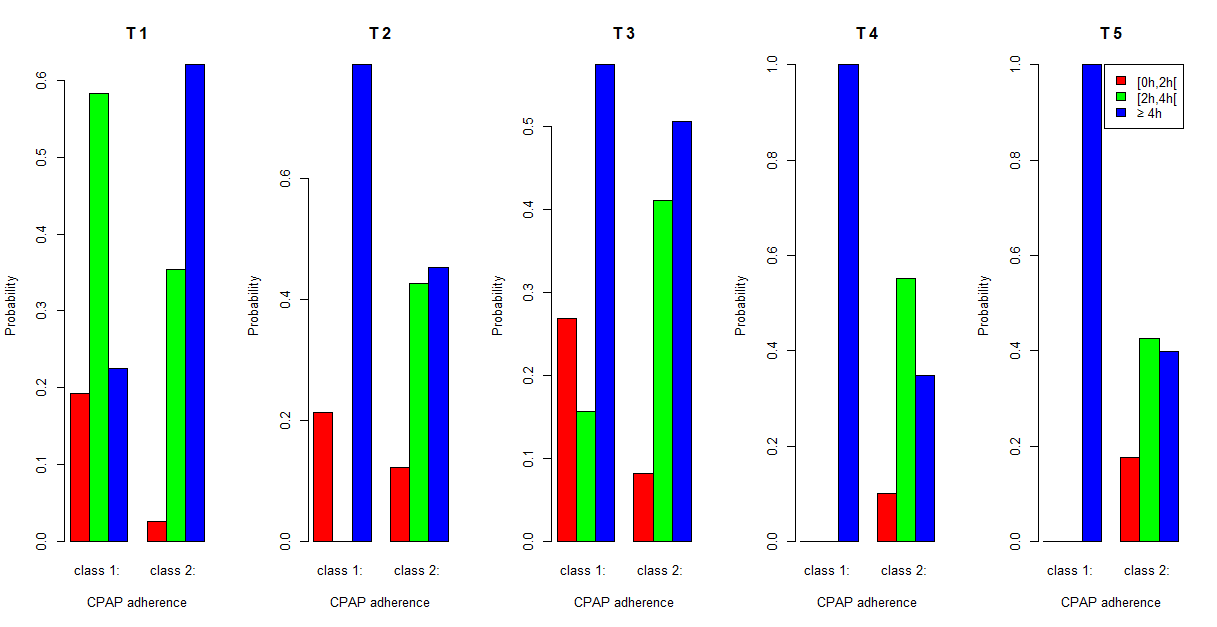
**Figure 3: Statistical methods according to study objectives and research questions**



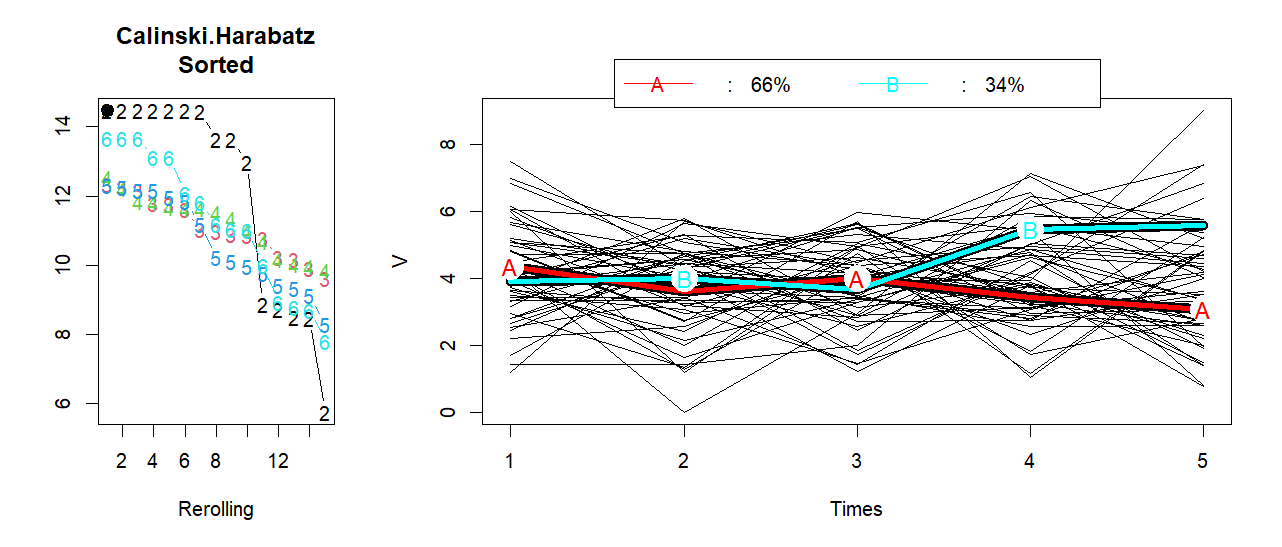
**Figure 4: Classification of statistical methods based on complexity, data format and objectives**. *ANOVA, analysis of variance; ARIMA, autoregression integrated moving average; DTW, dynamic time warping; GBTM, group-based trajectory modelling; GMM, growth mixture modelling; LCA, latent class analysis;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



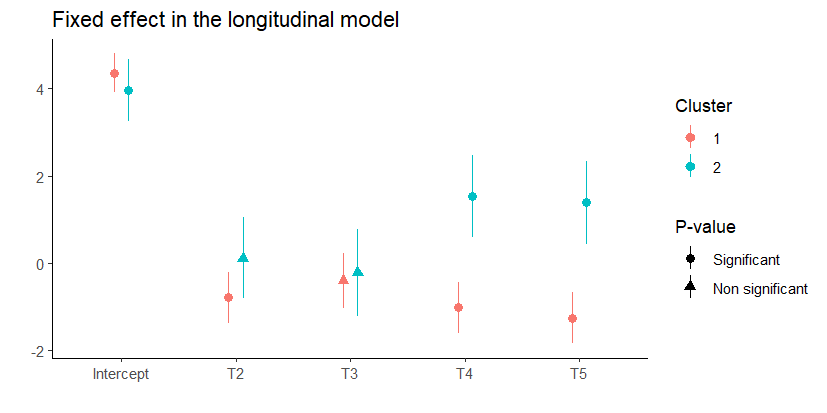
**Figure 5: Probability of belonging to clusters using LCA method**. *CPAP, continuous positive airway* *pressure*



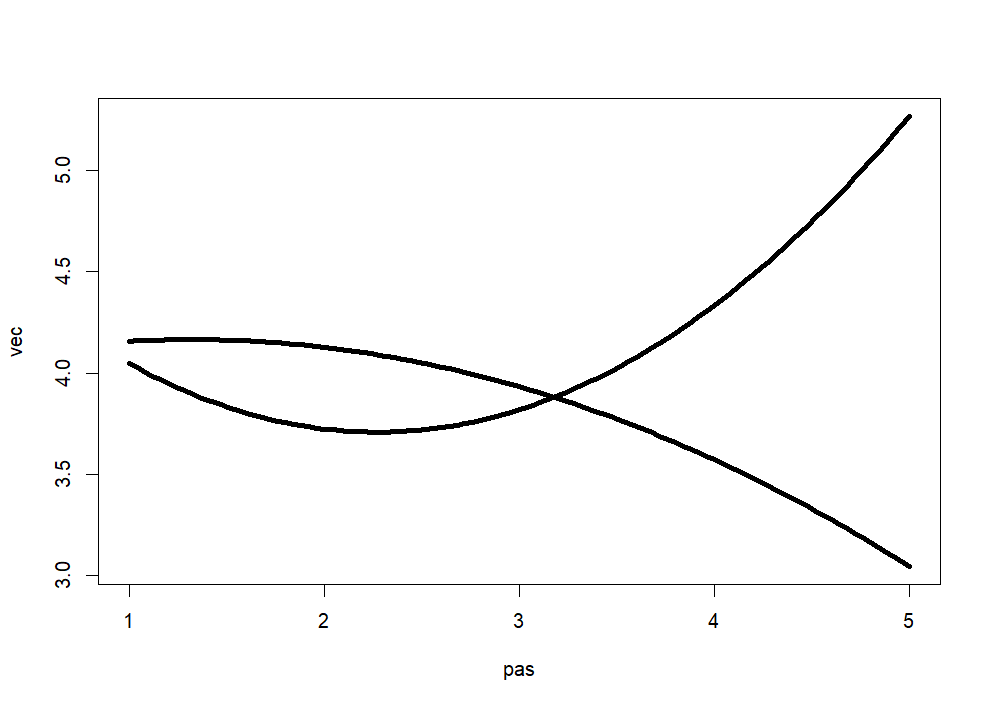
**Figure 6: Trajectories of clusters using K-means**



**Figure 7: Trajectories of 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**

