Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# A hybrid computational approach for mapping latent variable models in health research: development of an exploratory taxonomy

**Joffrey Marchi**

joffrey.marchi.cespa@gmail.com

French Armed Forces Center for Epidemiology and Public Health (CESPA)

**Bakridine Mmadi Mrenda**

French Armed Forces Center for Epidemiology and Public Health (CESPA)

**Jean-Marie Loreau**

French Armed Forces Center for Epidemiology and Public Health (CESPA)

**Hervé Chaudet**

French Armed Forces Center for Epidemiology and Public Health (CESPA)

**Mohammed Sedki**

Université Paris-Saclay, Gustave Roussy, Inserm, UVSQ, Center for Research in Epidemiology and Population Health – Unit 1018, Oncostat team, F-94805

**Véronique Letort**

Université Paris-Saclay, Centrale Supélec, MICS Lab, 91190 Gif-sur-Yvette, France; IHU PRISM National PRecISion Medicine Center in Oncology, Gustave Roussy

**Rodrigue S. Allodji**

French Armed Forces Center for Epidemiology and Public Health (CESPA)

**Gaëtan Texier**

French Armed Forces Center for Epidemiology and Public Health (CESPA)

**Research Article**

**Keywords:** Latent Variable Models, Models, Statistical, Artificial Intelligence, Natural Language Processing, Data Mining, Classification, Epidemiologic Methods

**Posted Date:** November 17th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11318410/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# A hybrid computational approach for mapping latent variable models in health research: development of an exploratory taxonomy


## Authors

Joffrey Marchi[1,2,5*], Bakridine Mmadi Mrenda[1,2], Jean-Marie Loreau[1,2], Hervé Chaudet[1,2,3], Mohammed Sedki[4], Véronique Letort[5], Rodrigue S. Allodji[1,6], Gaëtan Texier[1,2]

## Affiliations

[1]French Armed Forces Center for Epidemiology and Public Health (CESPA), Marseille, France.

[2]Aix Marseille University, SSA, APHM, RITMES, Marseille, France.

[3]University Hospital Institute (IHU) Méditerranée Infection, Marseille, France

[4]Université Paris-Saclay, Gustave Roussy, Inserm, UVSQ, Center for Research in Epidemiology and Population Health – Unit 1018, Oncostat team, F-94805, Villejuif, France.

[5]Université Paris-Saclay, Centrale Supélec, MICS Lab, 91190 Gif-sur-Yvette, France; IHU PRISM National PRecISion Medicine Center in Oncology, Gustave Roussy, Villejuif, France.

[6]Université Paris-Saclay, Gustave Roussy, Inserm, UVSQ, Center for Research in Epidemiology and Population Health – Unit 1018, EpiRad team, F-94805, Villejuif, France.

## Corresponding author

Joffrey Marchi,

Department of digital expertise in epidemiology, Center for Epidemiology and Public Health (CESPA), SSA, Camp de Sainte Marthe, 13568, Marseille, France.

E-mail: joffrey.marchi.cespa@gmail.com

**Keywords:** Latent Variable Models, Models, Statistical, Artificial Intelligence, Natural Language Processing, Data Mining, Classification, Epidemiologic Methods

## Abstract

**Background:** Latent variable models (LVMs) are fundamental in health research for capturing unobservable phenomena. However, their broader application is hindered by compartmentalised disciplines, inconsistent terminology, and a lack of methodological standardisation. This complexity often leads to methodological inertia, where researchers default to familiar rather than optimal models. This study aims to map LVMs using a computationally reproducible and expert-auditable approach to design a structured orientation tool. This framework helps researchers navigate existing models and narrow down candidate families based on data characteristics and hypotheses.

**Methods:** We applied a hybrid computational approach to map LVMs, beginning with a structured PubMed API search (from inception to 10 December 2025). First, LVM mentions were extracted from titles and abstracts using large language model (LLM) extraction based on mathematical criteria. Mentions were standardised through string processing, then grouped using natural language processing and clustering (TF-IDF and DBSCAN) to identify conceptually identical models and select representative exemplars. Subsequently, LLM-assisted data extraction from a targeted corpus, followed by rigorous expert validation, populated a characterisation grid detailing each exemplar's theoretical, computational, and software attributes. Finally, selected exemplars were classified into a three-level hierarchical taxonomy based on their directional architectures and methodological objectives.

**Results:** A total of 8,495 unique papers were included, yielding 9,555 extracted LVM mentions. After standardisation and filtering, 643 unique model mentions were clustered,

yielding 52 representative exemplar models. These exemplars were organised into the proposed exploratory taxonomy, which distinguishes three super-families. These comprise the reflective group (the vast majority, including classical latent class and growth curve models), the formative group (e.g., probabilistic principal component analysis), and the hybrid group (encompassing generative AI and bidirectional architectures). These super-families were further subdivided into nine meta-families relevant to health research. **Conclusions:** This study provides an exploratory taxonomy of LVMs in health research, offering a structured orientation tool. By integrating a hybrid computational approach with expert validation, this framework assists model selection across epidemiological, biostatistical, and data science applications. It provides a practical foundation to help researchers navigate disciplinary silos and identify modelling strategies suited to their specific data characteristics.

## Introduction

The accumulation of multidimensional health data, from electronic health records, omics sciences, or medical imaging, opens up perspectives for public health [1]. However, the high dimensionality and heterogeneity inherent in these databases pose fundamental challenges for statistical analyses [2, 3]. These challenges are further compounded as the observed variables often represent merely indirect manifestations of underlying biological, clinical, or environmental processes. In epidemiology and clinical research, the analysis of pathologies often requires identifying complex pathophysiological mechanisms, phenotypes, or environmental exposures that cannot be directly observed or measured [4]. Faced with the significant challenge represented by the reduction of this information into interpretable patterns, the use of advanced mathematical modelling frameworks has become highly relevant [5].

To meet this need, latent variable models provide an essential methodological framework. [6]. Historically rooted in psychometrics and the social sciences, these models, such as structural equation modelling, factor analysis, and latent class analysis, postulate the existence of unobserved variables whose effects manifest in observable data [6, 7]. Recently, the rapid development of artificial intelligence has expanded this analytical domain. Generative architectures derived from deep learning, such as variational autoencoders or probabilistic diffusion models, now fall within the latent variable paradigm [8]. These approaches offer great analytical flexibility in dimension reduction and the modelling of complex biomedical data [9].

Despite their potential, the scope of application of latent variable models remains limited by compartmentalised terminologies and disciplines [6]. Indeed, research has split into two cultures, contrasting the classical biostatistical approach, which prioritises inference and explainability, with the machine learning culture, which is focused on predictive performance [10]. In this heterogeneous context and in the absence of a methodological standardisation, it is difficult for a researcher to navigate in the multitude of existing models and evaluate their suitability for the specific characteristics of their data (longitudinal structures, non-linearity, variable types, heterogeneity, incompleteness) and their objectives. Faced with this complexity, researchers frequently yield to a form of methodological inertia, sometimes described as the “law of the instrument” which drives them to favour conventional models derived from their own disciplinary tradition rather than adopting the most appropriate approach for their research problem [10, 11].

Therefore, the combination of the “law of the instrument” and a lack of methodological standardisation frequently leads to the use of inadequate models to address the specific objectives of the study [11–13].

The objective of this study is therefore to map latent variable models using a computationally reproducible and expert-auditable approach, with the aim of developing a structured decision-support tool that helps researchers narrow down the range of candidate model families based on the characteristics of their data and their research hypotheses.

## Material and methods

Our methodology comprised a sequential computational pipeline followed by a structured synthesis, encompassing seven distinct stages. First, data acquisition and cleaning aims to retrieve and standardise a targeted set of relevant scientific publications from the PubMed database. Second, LLM (Large Language Model) extraction was used to automatically extract model mentions based on strict mathematical and probabilistic criteria. Third, string processing standardises the extracted terminology and eliminates textual artifacts. Fourth, natural language processing and clustering groups conceptually identical models and removes orthographic variations through text mining and density-based clustering. Fifth, exemplar identification selects the most representative model for each established group. Following this computational extraction, a characterisation grid is developed to systematically profile the theoretical, computational, and software attributes of each exemplar. Finally, the latent variable models taxonomy classifies these selected exemplars into a three-level hierarchical framework based on their directional architectures and methodological objectives.

### Data acquisition and cleaning

The literature search was conducted in the PubMed database. The study period included all articles published from inception up to 10 December 2025. The request was executed via the application programming interface (API) using the pubmedR [14] package in R [15] (version 4.4.3). The search equation targeted articles titles using the following syntax:

(variable[Title] OR class[Title] OR patterns[Title] OR profile[Title] OR model[Title] OR modeling[Title] OR methods[Title] OR method[Title]) AND (latent[Title] OR hidden[Title]). The raw data were cleaned using a protocol adapted from Govindasamy et al. [16]. This preprocessing phase involved converting titles to lowercase, removing apostrophes, and cleaning the Digital Object Identifier (DOI) strings (by removing the 'doi.' prefix). Duplicates were removed based on standardised titles. Finally, all articles without a title or abstract were excluded for the next stage.

**LLM extraction**

Automated extraction of latent variable models from titles and abstracts was performed using the LLM gemma3:27b [17], deployed in R using the biorecap [18] package. The model was prompted to extract relevant methodological entities, specifically targeting statistical and machine learning concepts. The detection scope was strictly limited to any model that met the following criteria: it explicitly posited the existence of at least one unobserved variable; it linked this variable to observed data through a mathematical, probabilistic, or generative formulation; and it involved an estimation or inference procedure. Additionally, the model had to be identifiable from the title or abstract. The extraction instructions required the exclusion of all other statistical models; missing even one criterion led to immediate exclusion, returning only the full canonical name of the identified model(s) without any abbreviations. If multiple models were detected, they were to be separated by a forward slash ("/"). If no models were detected, or if they were not latent variable models, the required output was the string "no latent model". The complete prompt used for this extraction is available in the `prompts/03_prompt_inference.md` file within the project's GitHub repository (detailed in the Availability of data and materials section).

### String processing

The extracted data were processed in R in five distinct steps, the last three of which involved normalising the strings. The first step consisted of pivoting the data to isolate each extracted model by decomposing the model lists into individual rows. The second step involved purging the dataset of non-responses and textual artifacts generated by the LMM. Missing or superfluous values, such as “not detected,” “no latent model,” “na,” “nan,” “n.a.,” “n/a,” or “uncertain,” were removed from the final dataset, as were specific prefixes (e.g., “the latent variable model mentioned…”). Before the next step, suffixes and residual punctuation from strings were removed. The third step focused on linguistic standardization through conversion to Latin-ASCII format, removing accents via transliteration, converting to lowercase, and replacing hyphens with spaces. The fourth step was orthographic homogenisation, where regular expression substitution rules (regex) were applied for the systematic conversion of British to American spelling. The final step involved targeted lemmatization to enforce the singular form and semantically harmonize core concepts, using additional regular expressions to standardize common plurals and correct recurring terminology, such as converting “modeling” to “model.” The final list of standardised model names was deduplicated.

## Natural language processing and clustering

To group conceptually identical entities exhibiting residual orthographic variations, a text mining approach was implemented in Python (version 3.13.9). The normalised names were vectorised using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm (TfidfVectorizer) with scikit-learn [19] package. The analysis was configured to use character n-grams only from text inside word boundaries (*analyzer=“char_wb”*) with an n-gram range of 3 to 6 and a minimum document frequency of 1 (*min_df=1*). A k-nearest neighbours (NearestNeighbors) algorithm was configured with the cosine metric

and a stringent raduis threshold of 0.3 (*radius=0.3*) to construct a connectivity graph and merge the orthographic variants. Within each identified group, the shortest character string was retained as the canonical name. Thematic clustering was subsequently performed using the DBSCAN (density-based spatial clustering of applications with noise) algorithm on the reduced matrix (weighted by the size of each orthographic error group). The parameters were set to *eps=0.4* and *min_samples=4*, utilising the cosine distance metric. These thresholds were selected as pragmatic choices supported by local cluster stability, representing a compromise between the number of initial conceptual groups and the number of latent model mentions within each group. Following this stage, the coherence of the resulting clusters was rigorously evaluated to ensure their semantic and mathematical validity. This verification process, predicated on expert judgement, facilitated the refinement of the cluster structures. While the computational steps of this pipeline are strictly reproducible, this final cluster refinement relies on auditable, domain-specific expert judgement. Applying this same qualitative rationale, elements categorised as noise during the initial clustering phase were individually examined to allow for their reclassification into relevant groups. Finally, models classified as noise and deemed unclassifiable were permanently exclude.

**Exemplars identification**

The identification of exemplar models was conducted in R, predicated on the previously established groupings. The models were merged with the cluster nomenclature database and subsequently categorised according to their Bayesian or frequentist nature (identified via a regular expression targeting the root “baye”). The exemplar model for each cluster was determined by selecting the most frequently occurring model within its respective group. In the event of a tie, the tie-breaking procedure was based on the most recent publication date (sorted in descending order by year, and subsequently by month).

## Characterisation grid for latent variable models

To construct the characterisation grid, a targeted documentary corpus was assembled for each exemplar to ensure data extraction. Where available, this corpus systematically included: a seminal methodological paper, a pedagogical or tutorial review, a benchmark or comparative study, and the reference literature or official documentation for the canonical software implementations (R packages and Python modules). The key attributes of each exemplar were then synthesised across three user-centric categories. The theoretical category covers the models' main goals, the distributional assumptions they rely on, and how they are estimated and evaluated. The computational and data-handling category looks at practical constraints, such as data types, how they handle missing values, their ability to scale to high-dimensional data, and their overall algorithmic efficiency. Lastly, the functional and software category identifies common domains of use and the specific R packages or Python modules needed to run them. Data extraction from these specific corpora involved using NotebookLM, followed by a validation phase to ensure the accuracy and traceability of all retrieved information. Due to the extensive number of models identified across the three super-families (see next paragraph), and the high granularity of the extracted characteristics, the exhaustive characterisation grid is provided as Supplementary Material 1. A curated selection of representative models, condensed for clarity, is presented in the main manuscript to illustrate the proposed taxonomy.

## Latent variable models taxonomy

All exemplars were classified according to a three-level taxonomy. The first taxonomic level describes the predominant statistical or generative relationship between latent and observed variables. In reflective-type models, observed variables are modelled conditionally on a latent structure. In formative-type approaches, the latent representation

is constructed or estimated primarily from combinations of observed variables. The hybrid category includes approaches involving both encoding and generative components, or for which the relationship cannot be adequately represented by a single direction. These labels are used as pragmatic organisational descriptors and should not be interpreted as establishing causal relationships. Their meaning is broader than their traditional use in psychometric measurement models and does not imply that the included models are intrinsically causal. At the second level, models are grouped into meta-families based on the nature of the object of study and the theoretical objective. The third level comprises families, or umbrella families. Umbrella families refer to the “parent” models that encompass both direct extensions and specific variations of a single central concept. Families, on the other hand, represent a grouping of models based on algorithmic or disciplinary relationships without necessarily constituting variations of a parent model.

**Methodological validation**

To ensure the robustness of our hybrid computational pipeline, two distinct methodological validations were performed: a sensitivity analysis of the primary search strategy and a manual evaluation of the LLM extraction accuracy. First, to evaluate the potential selection bias introduced by restricting our initial literature search exclusively to article titles, we conducted a targeted sensitivity analysis on the 52 exemplar models identified at the end of the pipeline. For each exemplar, a broad query was executed via the PubMed API to retrieve records mentioning the model name in either the title or the abstract (e.g., "[MN]"[Title/Abstract] AND ("1800"[Date - Publication] : "2025/12/10"[Date - Publication], where MN was the Model Name). When a query returned more than 10,000 records, the maximum retrieval limit imposed by the *retstart* parameter of the NCBI Entrez Programming Utilities API, data extraction was restricted to the 10,000 most recent records [20]. Within each retrieved dataset, we reapplied the exact Boolean logic of our

primary search equation to the text of both titles and abstracts. This allowed us to quantify the proportion of articles where the primary query criteria were met exclusively in the title versus exclusively in the abstract, deliberately excluding articles meeting the criteria in both fields to strictly isolate single-field occurrences. Second, to rigorously assess the reliability of the automated extraction process, a manual validation was conducted on a simple random sample of 100 articles drawn from the corpus of 8,495 unique records. One investigator acted as the human reference standard. The title and abstract of each selected article were manually reviewed to either extract all explicitly mentioned latent variable models or confirm their absence. These human-extracted results were subsequently compared against the LLM's outputs. Performance was evaluated by calculating the percentage of agreement with 95% confidence intervals (CIs) across two levels: (i) perfect agreement, defined as an exact match between the LLM output and the human reference, and (ii) partial accuracy, evaluating whether the LLM correctly identified at least one true model present in the text (assessed on the subset of articles where the LLM detected at least one model).

## Results

The initial data acquisition phase yielded a total of 8,850 articles. Following the exclusion of 355 records, comprising 21 duplicates and 334 articles lacking abstracts, a corpus of 8,495 unique articles was retained for large language model extraction (Figure 1). From this corpus, the extraction process generated an initial pool of 9,555 model mentions. Following the LLM extraction and string processing, a total of 958 model mentions were marked as non latent models, resulting in their direct removal. Furthermore, this string processing (linguistic standardisation, orthographic homogenisation, and targeted lemmatisation) facilitated the identification of duplicate entries, leading to the removal of 7,954 duplicate model mentions. In total, 643 unique latent model mentions were retained

for natural language processing and clustering. After the clustering process, a total of 43 initial conceptual groups were identified. From these, 1 residual group was permanently excluded as noise (n=7 unique latent model mentions). The remaining groups yielded 52 eligible conceptual clusters, which expanded to 53 groups after the separation of frequentist and Bayesian models. The exemplars identification process driven by the selection frequency of the model relative to the total number of articles within each respective group identified 52 exemplars, which were included in the final synthesis (Figure 1).

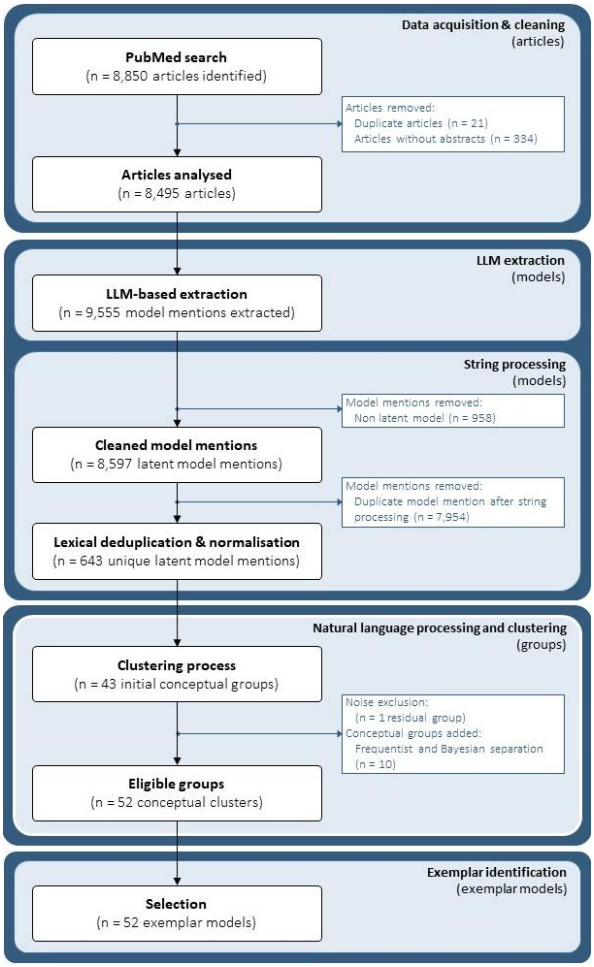



**Figure 1:** Flowchart of the hybrid (human and computational) pipeline for latent variable model identification and exemplar selection. The diagram details the sequential attrition of identified models through three main phases: LLM-based extraction and string processing, natural language processing and clustering (via TF-IDF and DBSCAN), and the final frequency and expert -based identification yielding 52 representative exemplars.

Mapping latent variable models results in a hierarchical taxonomy structured according to decreasing levels of abstraction. The first level of classification defines three super-families, categorising all models according to the orientation of the statistical or generative relationship between latent constructs and observed indicators. This initial level is further subdivided into nine distinct meta-families based on theoretical objectives and analytical architectures.

The results highlight a predominance of the reflective super-family, which represents 78.8% (n=41/52) of the identified exemplars (Figure 2). This super-family encompasses the vast majority of classical biostatistical and epidemiological approaches. It is distributed across five meta-families including three umbrella families, that address core methodological needs, ranging from general structural modelling and latent categorisation (e.g., latent class models) to temporal dynamics and probabilistic mixtures.

In contrast, the remaining models are distributed between two less prevalent directional structures. The hybrid super-family accounts for 15.4% (n=8/52) of the models, including one umbrella family. This group is particularly notable as it bridges classical statistical theory with modern artificial intelligence, encompassing bidirectional architectures as well as the “Deep Learning and Generative Models” meta-family (such as variational autoencoders and latent diffusion models). Finally, the formative super-family constitutes the remaining 5.8% (n=3/52) of the exemplars, primarily grouping approaches dedicated to variance extraction, component synthesis, and geometric partitioning (Figure 2).

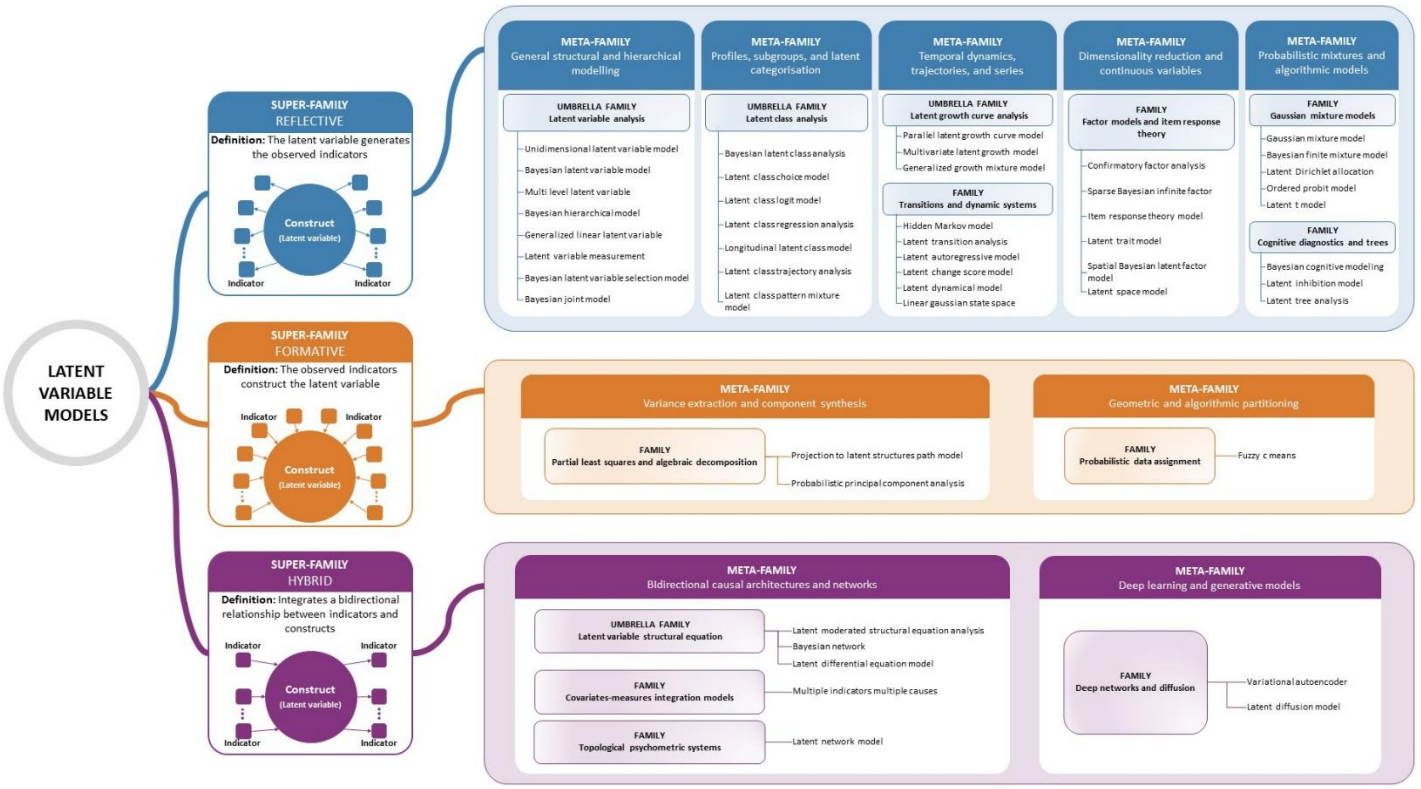



**Figure 2:** Hierarchical taxonomy of latent variable models in health research. Models are classified across three descending levels of abstraction: super-families (reflective, formative, and hybrid) determining the directionality between constructs and indicators, methodological meta-families, and representative model families or exemplars.

Tables 1, 2, and 3 summarise the key methodological and computational characteristics of representative exemplars for the reflective, formative, and hybrid super-families, respectively. The complete, fully detailed characterisation grid, encompassing all identified models and their attributes, is available in Supplementary Material 1.

The objectives of the reflective super-family focus on explaining, understanding, and predicting events by identifying hypothetical constructs measured by observed indicators. This encompasses identifying unobserved heterogeneity, modelling longitudinal trajectories, and grouping individuals into distinct subgroups, unified by the theoretical assumption of conditional independence among observed variables. Input data types include continuous, categorical, longitudinal, spatial, and textual variables (Table 1). Estimation methods typically rely on maximum likelihood approaches and Bayesian

methods, such as Markov Chain Monte Carlo (MCMC) or Gibbs sampling. Computationally, these models handle missing data through robust strategies, notably using Expectation-Maximization (EM) algorithms. The assessment of model fit is primarily based on information criteria, such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the Integrated Completed Likelihood (ICL). Latent class analysis (LCA), latent transition analysis (LTA), latent Dirichlet allocation (LDA), and latent growth curve models are representative exemplars of this super-family (Table 1). Finally, their application is supported by extensive software implementations in R (e.g., lavaan, lme4, brms) and Python (e.g., Bambi, PyMC) (Supplementary Material 1).

**Table 1:** Selected representative models of the reflective super-family and their methodological specifications

| Family / model name | Description & Objectives | Data types | Assumptions include | Commun computational framework |
|---|---|---|---|---|
| Latent variable analysis / Bayesian hierarchical model | Multi-level framework estimating prior distributions from data to enable partial pooling, improve power, and control bias | **In:** Continuous, categorical, binomial, survival<br>**Out:** Posterior distributions, credible intervals | Group-level parameters normally distributed; zero covariance between grouping factors | **Est:** MCMC, Gibbs, HMC, NUTS<br>**Miss:** Impute missing predictor values<br>**Fit:** WAIC, LOO, DIC, Bayes factors |
| Latent class analysis / Latent class analysis | Finite mixture model identifying unobserved subgroups (classes) based on categorical responses to investigate confounding | **In:** Multivariate categorical, binary, ordinal, continuous<br>**Out:** Latent class sizes, posterior probabilities | Local conditional independence; mutually exclusive classes | **Est:** ML, EM, MCMC<br>**Miss:** FIML, listwise deletion<br>**Fit:** BIC, SABIC, Entropy, VLMR-LRT |
| Latent growth curve analysis / Latent growth curve analysis | SEM approach capturing developmental trajectories and inter-individual variability in intra-individual change over time | **In:** Repeated measures, longitudinal panel<br>**Out:** Intercept/slope means, factor covariances | Change is systematically related to time; normally distributed measures | **Est:** ML, FIML, Bayesian<br>**Miss:** FIML, MI<br>**Fit:** $\chi 2$, RMSEA, CFI, BIC |
| Transitions and Dynamic Systems / Hidden Markov model | Doubly stochastic process characterising time series data driven by an unobserved Markov chain to forecast future states | **In:** Observation sequences, continuous signals<br>**Out:** Probability sequence, optimal state sequence | Markov property; conditional independence given current state | **Est:** ML, EM, MCMC, Viterbi<br>**Miss:** Modified EM algorithm<br>**Fit:** AIC, BIC, log-likelihood, MPE |
| Transitions and Dynamic Systems / Latent transition analysis | Longitudinal extension of LCA to model interrelations of latent variables and explore transitions across profiles over time | **In:** Categorical, longitudinal, ordinal<br>**Out:** Item probabilities, transition probabilities | Conditional independence; Markov property (time t influenced by t−1) | **Est:** ML, EM, Bayesian<br>**Miss:** Missing at random (MAR)<br>**Fit:** AIC, BIC, BLRT, Entropy |
| Factor Models and Item Response Theory / Confirmatory factor analysis | Statistical tool testing a priori hypotheses about relations between observed indicators and latent constructs to assess construct validity | **In:** Test scores, covariance matrices, ordinal, continuous<br>**Out:** Parameter estimates, factor loadings, correlations | Continuous multivariate normal distribution (or normal latent for categorical) | **Est:** ML, robust ML, WLS, ULS<br>**Miss:** FIML, MI, EM<br>**Fit:** $\chi 2$, CFI, RMSEA, SRMR |
| Factor Models and Item Response Theory / Item response theory model | Latent framework using logistic functions to specify relationships between a latent trait and categorical item responses for adaptive testing | **In:** Dichotomous, polytomous, multiple-choice<br>**Out:** Item parameters, latent trait scores, information functions | Unidimensionality; local conditional independence | **Est:** Marginal ML, EM, MCMC<br>**Miss:** EM, Plausible value (MI)<br>**Fit:** AIC, BIC, $G^2$, RMSEA |
| Factor Models and Item Response Theory / Spatial Bayesian latent factor model | Bayesian framework capturing spatial dependencies conditioning on latent factors to identify consistent activation regions | **In:** Spatial point pattern data, census variables<br>**Out:** Latent factors, posterior predictive probabilities | Normal factors/residuals; spatial dependence via CAR structures | **Est:** MCMC, HMC, Gibbs<br>**Miss:** Data augmentation or replicate datasets<br>**Fit:** WAIC, CRPS, Gelman-Rubin |
| Gaussian Mixture Models / Gaussian mixture model | Parametric density estimation via weighted Gaussian components for model-based clustering and discriminant analysis | **In:** Continuous measurements, features, time series<br>**Out:** Mixture weights, mean vectors, covariance matrices | i.i.d observations; ellipsoidal clusters determined by covariance matrix | **Est:** EM, MAP, CEM<br>**Miss:** Conditional expectations via EM<br>**Fit:** BIC, ICL, ARI (only if reference values are available) |
| Gaussian Mixture Models / Latent Dirichlet allocation | Generative probabilistic finite mixture model for dimensionality reduction and finding short descriptions of discrete collections (topics) | **In:** Text corpora, bag-of-words vectors<br>**Out:** Latent topics, document-topic assignments | Bag-of-words assumption; exchangeable documents; independent topics | **Est:** Variational methods, Gibbs, MCMC<br>**Miss:** [Not specified]<br>**Fit:** Perplexity, topic coherence, F1 scores |

***Table structure:*** ***Est:*** *Estimation method(s);* ***Miss:*** *Missing data handling;* ***Fit:*** *Model fit & Evaluation metrics.*
***Abbreviations:*** ***AIC:*** *Akaike Information Criterion;* ***ARI:*** *Adjusted Rand Index;* ***BIC:*** *Bayesian Information Criterion;* ***BLRT:*** *Bootstrapped Likelihood Ratio Test;* ***CAR:*** *Conditional Autoregressive structure;* ***CEM:*** *Classification EM;* ***CFI:*** *Comparative Fit Index;* ***CRPS:*** *Continuous Ranked Probability Score;* ***DIC:*** *Deviance Information Criterion;* ***EM:*** *Expectation-Maximisation;* ***FIML:*** *Full-Information Maximum Likelihood;* ***HMC:*** *Hamiltonian Monte Carlo;* ***ICL:*** *Integrated Classification Likelihood;* ***i.i.d.:*** *Independently and identically distributed;* ***LCA:*** *Latent Class Analysis;* ***LOO:*** *Leave-One-Out;* ***MAP:*** *Maximum A Posteriori;* ***MAR:*** *Missing At Random;* ***MCMC:*** *Markov Chain Monte Carlo;* ***MI:*** *Multiple Imputation;* ***ML:*** *Maximum Likelihood;* ***MPE:*** *Minimum Prediction Error;* ***NUTS:*** *No-U-Turn Sampler;* ***RMSEA:*** *Root Mean Square Error of Approximation;* ***SABIC:*** *Sample-Size Adjusted BIC;* ***SEM:*** *Structural Equation Modelling;* ***SRMR:*** *Standardised Root Mean Square Residual;* ***ULS:*** *Unweighted Least Squares;* ***VLMR-LRT:*** *Vuong-Lo-Mendell-Rubin Likelihood Ratio Test;* ***WAIC:*** *Watanabe-Akaike Information Criterion;* ***WLS:*** *Weighted Least Squares;* ***$\chi 2$:*** *Chi-square*

The objectives of the formative super-family focus on maximizing explained variance (projection to latent structures path modelling), reducing dimensionality (probabilistic principal component analysis [PPCA]), and the soft clustering of data (fuzzy c-means clustering) (Table 2). These exemplars estimate latent variables as linear combinations of observed indicators. Input data types include continuous and nominal variables, high-dimensional vectors, and observations in Euclidean space. Parameter estimation is performed using iterative algorithms, such as ordinary least squares, expectation-maximisation (EM), or Picard iteration. Computationally, these models handle missing data through specific strategies, ranging from pairwise deletion and mean replacement to iterative EM algorithms. The assessment of model fit is based on measures such as the coefficient of determination ($R^2$), the cross-validation score ($Q^2$), or partition coefficients (e.g., fuzzy partition coefficient [Fc], standard partition coefficient [PC], or modified partition coefficient [MPC]) (Table 2). Finally, their application is supported by dedicated software implementations, primarily through specific R packages (e.g., semPLS, pcaMethods, fclust) (Supplementary Material 1).

**Table 2:** Selected representative models of the formative super-family and their methodological specifications

| Family / model name | Description & Objectives | Data types | Assumptions include | Commun computational framework |
|---|---|---|---|---|
| Partial least squares and algebraic decomposition / Projection to latent structures path model | Variance-based structural equation modelling estimating latent variables as linear combinations to ²maximise explained variance of endogenous variables | **In:** Continuous, nominal, ordinal, interval, ratio, categorical, binary **Out:** Latent scores, outer weights/loadings, path coefficients, total effects | Systematic portion of OLS regressions equals conditional expectation; zero-mean residuals uncorrelated with latent variables | **Est:** OLS, PLS algorithm (NIPALS) **Miss:** Pairwise deletion, mean replacement **Fit:** R2, f2, Q2, q2, GoF index, composite reliability, AVE, Cronbach's α |
| Partial least squares and algebraic decomposition / Probabilistic principal component analysis | Gaussian latent variable model decomposing high-dimensional data into a low-dimensional subspace and noise via ML estimation to compress data | **In:** High-dimensional vectors, continuous, incomplete datasets **Out:** Principal components, latent variables, noise variance, conditional latent mean | Normally distributed latent variables and noise; isotropic (spherical) noise covariance | **Est:** ML, EM **Miss:** Iterative EM algorithm **Fit:** Log-likelihood, Bayesian model selection, cross-validation score, RMS error |
| Probabilistic data assignment / Fuzzy c means | Soft clustering algorithm generating fuzzy partitions by minimising a generalised least-squares function to find dense regions via soft assignment | **In:** Numerical data, observations in Euclidean space, object data **Out:** Fuzzy partitions, cluster centres, membership degrees | Memberships between 0 and 1 summing to unity; linearly separable structures; spherical clusters | **Est:** Picard iteration, Alternating optimisation (AO) **Miss:** [Not specified] **Fit:** Partition coefficient, partition entropy, MPC, Xie-Beni index, Fukuyama-Sugeno, DBI |

***Table structure: Est:*** *Estimation method(s);* ***Miss:*** *Missing data handling;* ***Fit:*** *Model fit & Evaluation metrics*
***Abbreviations: AO***: *Alternating Optimisation*; ***AVE***: *Average Variance Extracted*; ***DBI***: *Davies-Bouldin Index*; ***EM***: *Expectation-Maximisation*;; ***f2***: *Cohen's effect size*; ***GoF***: *Goodness-of-Fit*; ***ML***: *Maximum Likelihood*; ***MPC***: *Modified Partition Coefficient*; ***NIPALS***: *Non-linear Iterative Partial Least Squares*; ***OLS***: *Ordinary Least Squares*; ***PLS***: *Partial Least Squares*; ***Q2***: *Stone-Geisser's Q2*; ***q2***: *Predictive relevance*; ***R2***: *Coefficient of determination*; ***RMS***: *Root Mean Square*.

The objectives of the hybrid super-family focus on combining the principles of factor analysis, path analysis, and probabilistic learning to model complex structural relationships, handle measurement errors, and optimise the performance of complex generative models. Input data types are highly diverse, ranging from continuous variables and time series to unstructured data, such as images or videos Table 3). Estimation methods combine classical statistical algorithms, like maximum likelihood, with deep learning or graph-based algorithms, such as stochastic gradient variational Bayes. Computationally, this super-family demonstrates broad missing data handling capabilities, spanning from classical full-information maximum likelihood (FIML) and expectation-maximization (EM) algorithms to advanced deep learning techniques like inpainting for masked regions. Evaluation relies on structural fit indices (e.g., the root mean square error of approximation [RMSEA] and the comparative fit index [CFI]), network scores, or the Kullback-Leibler divergence. Bayesian networks, MIMIC (Multiple Indicators and Multiple

Causes) models, as well as variational autoencoders and latent diffusion models, are representative exemplars of this super-family (Table 3). Finally, their application relies on a dual software ecosystem, bridging classical structural and network packages in R (e.g., lavaan, bnlearn, lvnet) with modern probabilistic and generative frameworks in Python (e.g., pomegranate, Pythae) (Supplementary Material 1).

**Table 3:** Selected representative models of the reflective super-family and their methodological specifications

| Family / model name | Description & Objectives | Data types | Assumptions include | Commun computational framework |
|---|---|---|---|---|
| Latent variable structural equation / Latent variable structural equation | Hybrid of factor and path analysis providing a parsimonious summary of interrelationships to test complex, multidimensional constructs | **In:** Continuous, ordinal, categorical, binary, non-normal data<br>**Out:** Factor loadings, path coefficients, error variances, standard errors, modification indices | Multivariate distribution is normally distributed (for ML); techniques do not imply causality; underlying continuous variables are normal | **Est:** ML, LS, unweighted LS, GLS, ADF, cat-WLS, cat-LS, cat-DWLS<br>**Miss:** Listwise/pairwise deletion, ML, MI, FIML<br>**Fit:** χ2, GFI, CFI, RMSEA, SRMR, TLI, AIC, BIC, BL89, RNI, Gamma Hat, Mc |
| Latent variable structural equation / Latent moderated structural equation analysis | SEM approach estimating latent interactions via finite normal mixtures and EM algorithm to handle non-normality, yielding unbiased, efficient estimators and precise hypothesis testing | **In:** Continuous, observed binary, raw indicator data<br>**Out:** ML estimators, parameter estimates, standard errors, likelihood ratio test statistics, posterior probabilities | Exogenous indicator vector is multivariate normal; errors/disturbances are multivariate normal, independent, zero mean | **Est:** ML, EM, MLR<br>**Miss:** [Not specified]<br>**Fit:** Likelihood ratio test statistic |
| Latent variable structural equation / Bayesian network | Probabilistic graphical model representing variables and conditional dependencies via a directed acyclic graph (DAG) to model uncertainty | **In:** Discrete, continuous, mixed variables, time-series<br>**Out:** DAG, partially directed graph, adjacency matrix, posterior probabilities | Markov property (variable depends only on parents); conditional independence; discrete BNs assume multinomial; Gaussian BNs assume multivariate normal | **Est:** MLE, EM, MCMC, Constraint-based (PC, GS, IAMB, MMPC), Score-based (GES, hill-climbing)<br>**Miss:** EM algorithm, virtual evidence (dummy nodes)<br>**Fit:** BIC, AIC, BDeu, BGe, log-likelihood ratio, Pearson's χ2, Fisher's Z, SHD, ROC, AUC |
| Latent variable structural equation / Latent Differential Equation Model | SEM technique fitting differential equations to repeated observations to model self-regulating processes, calculate group trajectories, test dynamic hypotheses, and isolate random fluctuations | **In:** Repeated observations, multiple indicators, multi-wave panel data<br>**Out:** Factor loadings, differential equation parameters, expected group trajectories, factor scores | Latent constructs evolve continuously with linear derivative relationships; i.i.d zero-mean errors | **Est:** MLE<br>**Miss:** MLE under MAR assumption<br>**Fit:** -2LogL, χ2, RMSEA, GFI, BIC |
| Covariates-measures integration models / Multiple indicators multiple causes | Causal SEM where a single latent variable is linearly determined by observable causes and measured by multiple endogenous indicators | **In:** Exogenous causes, endogenous indicators, dummy variables, background variables<br>**Out:** ML estimators, asymptotic variance-covariance matrix, parameter estimates | Disturbances are mutually independent with expectation zero; distributions are multivariate normal; successive observations independent | **Est:** MLE, robust weighted least squares<br>**Miss:** [Not specified]<br>**Fit:** Likelihood ratio statistic, χ2 goodness-of-fit test |
| Topological psychometric systems / Latent network model | Generalisation of SEM modelling the covariance structure of latent variables as a Gaussian Graphical Model to obtain conditional independence | **In:** Multivariate Gaussian/normal data, cross-sectional data<br>**Out:** Conditional independencies, parameter estimates, network structures | The response vector is centered and follows a multivariate Gaussian density (multivariate normal data) | **Est:** MLE, step-wise model search, penalised MLE (LASSO)<br>**Miss:** Full-information maximum likelihood (FIML)<br>**Fit:** χ2, RMSEA, CFI, AIC, BIC, extended BIC (EBIC) |
| Deep networks and diffusion / Variational autoencoder | Deep architecture jointly learning latent-variable and inference models via neural networks to approximate intractable posterior distributions | **In:** Continuous, discrete, binary, images, video, text, sound<br>**Out:** Continuous latent variables, generated artificial data, latent representation | i.i.d. datasets; prior and likelihood come from parametric families with PDFs differentiable almost everywhere | **Est:** ML, MAP, SGVB estimator, EM<br>**Miss:** Inpainting, imputation of missing words<br>**Fit:** ELBO, marginal likelihood, KL divergence, MSE, FID, IS |
| Deep networks and diffusion / Latent diffusion model | Generative model training diffusion processes in the latent space of pretrained autoencoders to optimise limited computational resources | **In:** Text prompts, bounding boxes, semantic layouts, low-res/masked images<br>**Out:** High-resolution images | Model learns the reverse process of a fixed Markov Chain; denoised variable normal; representational space perceptually equivalent | **Est:** Denoising score-matching, reweighted variational bound<br>**Miss:** Inpainting (filling masked regions)<br>**Fit:** FID, Inception Score (IS), Precision, Recall, PSNR, LPIPS |

***Table structure: Est:*** *Estimation method(s);* ***Miss:*** *Missing data handling;* ***Fit:*** *Model fit & Evaluation metrics*

***Abbreviations: -2LogL***: *Minus Two Log-Likelihood*; ***ADF***: *Asymptotic Distribution-Free*; ***AIC***: *Akaike Information Criterion*; ***AUC***: *Area Under the Curve*; ***BDeu***: *Bayesian Dirichlet equivalent uniform*; ***BGe***: *Bayesian Gaussian equivalent*; ***BIC***: *Bayesian Information Criterion*; ***BL89***: *Bollen 1989 Index*; ***BNs***: *Bayesian Networks*; ***cat-WLS/LS/DWLS***: *Categorical Weighted Least Squares / Least Squares / Diagonally Weighted Least Squares*; ***CFI***: *Comparative Fit Index*; ***DAG***: *Directed Acyclic Graph*; ***EBIC***: *Extended BIC*; ***ELBO***: *Evidence Lower Bound*; ***EM***: *Expectation-Maximisation*; ***FID***: *Fréchet Inception Distance*; ***FIML***: *Full-Information Maximum Likelihood*; ***GES***: *Greedy Equivalence Search*; ***GFI***: *Goodness-of-Fit Index*; ***GLS***: *Generalised Least Squares*; ***GS***: *Grow-Shrink*; ***IAMB***: *Incremental Association Markov Blanket*; ***i.i.d.***: *Independently and identically distributed*; ***IS***: *Inception Score*; ***KL***: *Kullback-Leibler divergence*; ***LASSO***: *Least Absolute Shrinkage and Selection Operator*; ***LPIPS***: *Learned Perceptual Image Patch Similarity*; ***LS***: *Least Squares*; ***MAP***: *Maximum A Posteriori*; ***MAR***: *Missing At Random*; ***Mc***: *McDonald's Noncentrality Index*; ***MCMC***: *Markov Chain Monte Carlo*; ***MI***: *Multiple Imputation*; ***ML***: *Maximum Likelihood*; ***MLE***: *Maximum Likelihood Estimation*; ***MLR***: *Maximum Likelihood Robust*; ***MMPC***: *Max-Min Parents and Children*; ***MSE***: *Mean Squared Error*; ***PC***: *Peter-Clark algorithm*; ***PDFs***: *Probability Density Functions*; ***PSNR***: *Peak Signal-to-Noise Ratio*; ***RMSEA***: *Root Mean Square Error of Approximation*; ***RNI***: *Relative Noncentrality Index*; ***ROC***: *Receiver Operating Characteristic*; ***SEM***: *Structural Equation Modelling*; ***SGVB***: *Stochastic Gradient Variational Bayes*; ***SHD***: *Structural Hamming Distance*; ***SRMR***: *Standardised Root Mean Square Residual*; ***TLI***: *Tucker-Lewis Index*; ***$\chi^2$***: *Chi-square*

Regardless of their super-family affiliation, these modelling frameworks encompass a very broad spectrum of data types, ranging from continuous, categorical, and ordinal variables to longitudinal and unstructured data (such as texts or images). While certain exemplars are highly specialised (e.g., for strictly continuous or dichotomous data), others offer generalist approaches for the joint processing of mixed data types. The handling of missing data is generally addressed natively by the estimation methods, frequently utilising the EM algorithm, Full Information Maximum Likelihood (FIML), or multiple imputation. The applications of these exemplars are highly interdisciplinary, being widely used across psychology, the social sciences, medicine, biology, and economics (Supplementary Material 1).

## Validation of the computational pipeline

### Search strategy sensitivity

The sensitivity analysis of the search strategy across the 52 exemplar models identified a total of 15,767 instances where the primary search criteria were met in either the title or the abstract (Supplementary Material 2 - table S2). Within this total, 3,996 occurrences (25.3%) were identified exclusively in the article titles, while 11,771 (74.7%) were found exclusively in the abstracts (Supplementary Material 2 - table S2). Consequently, the title-restricted search captured approximately one quarter of the total mentions for these models.

### LLM extraction reliability

In the manual validation of the automated extraction process conducted on a random subset of 100 articles, agreement between the LLM output and the human reference standard was 88.0% (95% CI: 79.6% - 93.4%) (Supplementary Material 2 - table S3). Furthermore, in the subset of cases where the LLM extracted at least one methodological

entity (n=90), the algorithm correctly identified at least one true latent variable model in 91.1% of the instances (95% CI: 82.7% - 95.8%) (Supplementary Material 2 - table S3).

## Discussion

This study maps latent variable models through a computationally reproducible and expert-auditable approach, providing a structured orientation tool for researchers. To handle the considerable volume of publications extracted from PubMed, the development of a hybrid pipeline integrating an API, natural language processing, and large-scale language model inference, facilitated the analysis of an extensive dataset. This approach enabled the structuring of fragmented concepts into a user-centred taxonomy, collating the primary objectives, underlying assumptions, supported data types, and associated estimation methods for each identified model within a characterisation grid.

Manual systematic reviews currently face significant scalability challenges, given the exponential growth of the scientific literature [21], as traditional data extraction requires, on average, more than a year of intensive human labour [22]. The integration of the PubMed API, automated inference via a locally deployed large language model, and data processing in R and Python made it possible to circumvent this conceptual obstacle. This hybrid pipeline enabled the processing of a volume of data that would be prohibitively time-consuming and resource-intensive by a conventional manual approach [23], thus demonstrating the feasibility of automation for future methodological mapping. Manual validation supported the viability of this approach, with the local LLM achieving an 88.0% agreement rate with human expert extraction (Supplementary Material 2 - Table S3), suggesting that automated high-throughput screening can maintain acceptable analytical reliability.

The abundance and heterogeneity of isolated models are a striking result of this study. This hyperproliferation is thought to be largely due to terminological inconsistency and the absence of a formal ontology linking statistics and data science. Many mathematically analogous models are frequently rediscovered and renamed according to their fields of application (epidemiology, psychometrics, machine learning, etc.) [12]. The complexity of the cleaning process implemented in this study requiring semantic standardization, filtering, and validation by experts further illustrates the extent of this conceptual noise. The domain suffers from a lack of taxonomic consensus, which this study attempts to address.

Beyond semantic confusion, the findings highlight the mathematical and epistemological diversity of latent variable models, which are structured around the duality of frequentist and Bayesian paradigms [24]. Historical models, such as Latent Class Analysis (LCA), generally have their equivalents in the Bayesian paradigm. Concurrently, new horizons are illustrated by the emergence of novel models such as Variational Autoencoders (VAEs) or Latent Diffusion Models (LDMs) that rely on a Bayesian theoretical framework while being trained by frequentist-inspired empirical estimation methods [8, 25]. This contemporary hybridization suggests that the historical opposition between these two schools of thought is gradually fading in favour of a growing algorithmic synergy.

The proliferation of latent variable models' places researchers with a complex dilemma regarding the choice of the most appropriate method for their objectives. Furthermore, they frequently succumb to methodological inertia, or the "law of the instrument," and tend to favour classical models from their own disciplinary tradition rather than adopting the algorithm best suited to their specific problem [10, 11]. A primary contribution of this review is the provision of an orientation tool designed to help researchers navigate disciplinary

522 silos. To meet this challenge, we recommended that model selection should be guided by
523 four distinct axes: (i) the direction of the statistical or generative relationship, (ii) the
524 intended aims, (iii) the nature of the input data, and (iv) the availability of software
525 implementations. For example, a researcher aiming to identify unobserved patient
526 subgroups from categorical data (axes ii and iii) can use this matrix to immediately rule
527 out formative approaches and focus on the reflective super-family, specifically exploring
528 the latent class model umbrella. However, this framework has not been prospectively
529 validated and should not be interpreted as yielding an unequivocally ‘optimal’ model. It is
530 designed to explicitly structure hypotheses and reduce the field of candidate options; the
531 final model selection invariably remains contingent upon domain-specific knowledge and
532 rigorous statistical expertise.

533 **Limitations**

534 The search equation represents a methodological limitation of this study, as it exclusively
535 targeted article titles via the PubMed API to define the analysed corpus. Our sensitivity
536 analysis quantified this restriction, revealing that 74.7% of single-field model mentions
537 occurred exclusively in the abstracts (Supplementary Material 2 - Table S2). This confirms
538 a substantial reduction in the volume of retrievable literature, meaning this approach is
539 inherently unsuitable if the objective is an exhaustive systematic review [26]. However,
540 the primary goal of this study was exploratory methodological mapping, which does not
541 require exhaustive literature retrieval. For the automated pipeline to successfully extract
542 a method, a single article mentioning the model is sufficient. Notably, despite the reduced
543 volume of articles, all 52 exemplar models included in the final taxonomy were detected
544 at least once through the title-only search . In this context, restricting the search to titles
545 represents a pragmatic compromise: it significantly reduces API query times and

subsequent computational burden while retaining sufficient sensitivity to map the relevant methodologies.

Furthermore, the automation approach proposed in this study presents a limitation related to the decision to execute the LLM locally. Commercial cloud-hosted LLMs demonstrate superior performance in terms of information extraction [27]. However, accepting this performance trade-off resolves critical privacy compliance issues (Health Insurance Portability and Accountability Act, General Data Protection Regulation) [28]. Additionally, commercial models, operating as black boxes, undergo opaque updates that compromise the long-term reproducibility of results [29]. Deploying a local model allows for a fixed computing environment, thus ensuring the reproducibility of analyses. However, this approach remains susceptible to the potential loss of subtle algorithmic and mathematical nuances. Although our validation demonstrated a high agreement rate for entity extraction, this inherent risk of nuance loss underscores the necessity of expert validation at the end of the pipeline.

## Conclusions

This study demonstrates that the synergy between data science (automation, LLMs) and epidemiological rigour offers a promising avenue for mapping and structuring rapidly proliferating research methodologies. The proposed exploratory taxonomy of latent variable models provides a practical orientation tool designed to reduce methodological ambiguity and help researchers navigate disciplinary silos. Ultimately, this work establishes an evolving, computationally assisted framework capable of processing large volumes of literature, providing a foundation for the structured mapping of other complex analytical methods in the future.

## List of abbreviations

- AIC Akaike Information Criterion
- API Application Programming Interface
- ASCII American Standard Code for Information Interchange
- BIC Bayesian Information Criterion
- CFI Comparative Fit Index
- DBSCAN Density-Based Spatial Clustering of Applications with Noise
- DOI Digital Object Identifier
- EM Expectation-Maximization
- Fc Fuzzy partition coefficient
- FIML Full-Information Maximum Likelihood
- ICL Integrated Completed Likelihood
- LCA Latent Class Analysis
- LDA Latent Dirichlet Allocation
- LDMs Latent Diffusion Models
- LLM Large Language Model
- LMM Mixed Linear Model
- LTA Latent Transition Analysis
- LVMs Latent Variable Models
- MCMC Markov Chain Monte Carlo
- MIMIC Multiple Indicators and Multiple Causes
- MPC Modified Partition Coefficient
- PC Standard Partition Coefficient

- PPCA Probabilistic Principal Component Analysis
- $Q^2$ Cross-validation score
- $R^2$ Coefficient of determination regex Regular expression
- RMSEA Root Mean Square Error of Approximation
- TF-IDF Term Frequency-Inverse Document Frequency VAEs Variational Autoencoders

## Supplementary Information

**Supplementary Material 1:** Complete characterisation grid of the latent variable model taxonomy. (Format: .xlsx). A table merging the reflective, formative, and hybrid super-families. This file details the methodological objectives, short descriptions, input/output data types, statistical assumptions, estimation methods, missing data handling strategies, and evaluation metrics for all 52 identified models.

**Supplementary Material 2:** Methodological validation tables. (Format: .docx). This file provides the detailed results used to validate the computational pipeline and search strategy. It includes Table S2, which presents the comparison of search query sensitivity between article titles and abstracts for identifying the 52 exemplar models, and Table S3, which details the manual validation metrics of the automated LLM extraction process compared to a human reference standard.

**Acknowledgements**

The authors wish to thank Leïla Chassery, for their expertise in ergonomics, which contributed to refining the usability of the proposed orientation tool.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Availability of data and materials

The datasets generated and/or analysed during the current study, as well as the R and Python code used for text mining and LLM extraction, are available in the GitHub repository JoffreyCespa/Taxonomy-LVMs: https://github.com/JoffreyCespa/Taxonomy-LVMs.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

JM, HC, RA and GT took part in conceiving the initial plans of the study. JM, RA and GT developed the search strategy. JM and HC developed the R code for the automated extraction of literature via the PubMed API and the local implementation of the large language model (LLM) extraction. JM performed the research and analysed the data. JM, RA and GT interpreted the results. JM drafted the manuscript with input from GT. RA and GT supervised the project. BM, HC, JML, MS and VL critically revised the manuscript. All authors read and approved the final manuscript.

# REFERENCES


1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44–56. https://doi.org/10.1038/s41591-018-0300-7.

2. Fan J, Han F, Liu H. Challenges of Big Data Analysis. Natl Sci Rev. 2014;1:293–314. https://doi.org/10.1093/nsr/nwt032.

3. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19:1236–46. https://doi.org/10.1093/bib/bbx044.

4. Wild CP. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev. 2005;14:1847–50. https://doi.org/10.1158/1055-9965.EPI-05-0456.

5. Blei DM. Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models. Annual Review of Statistics and Its Application. 2014;1:203–32. https://doi.org/10.1146/annurev-statistics-022513-115657.

6. Skrondal A, Rabe-Hesketh S. Latent Variable Modelling: A Survey. Scandinavian Journal of Statistics. 2007;34:712–45. https://doi.org/10.1111/j.1467-9469.2007.00573.x.

7. Saporta G. Les classes latentes. In: Droesbeke J-J, Lejeune M, Saporta G, editors. Modèles statistiques pour données qualitatives. Editions Technip; 2005. p. 71–82.

8. Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv.org. 2013. https://arxiv.org/abs/1312.6114v11.

9. Wei R, Mahmood A. Recent Advances in Variational Autoencoders With Representation Learning for Biomedical Informatics: A Survey. IEEE Access. 2021;9:4939–56. https://doi.org/10.1109/access.2020.3048309.

10. Breiman L. Statistical Modeling: The Two Cultures. Statistical Science. 2001;16:199–215.

11. Kaplan A. The Conduct of Inquiry: Methodology for Behavioural Science. New York: Routledge; 2017. https://doi.org/10.4324/9781315131467.

12. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11:e048008. https://doi.org/10.1136/bmjopen-2020-048008.

13. Mrenda BM, Pellegrin L, Marchi J, Chaudet H, Texier G. Scientific workflow for state sequence analysis: a simple and reproducible tool. Discov Appl Sci. 2026;8:552. https://doi.org/10.1007/s42452-026-08591-4.

14. Aria M. pubmedR: Gathering Metadata About Publications, Grants, Clinical Trials from 'PubMed' Database [Internet]. 2026. Available from: https://CRAN.R-project.org/package=pubmedR

15. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2025.

16. Govindasamy P, Noor AM, Isa NJM, Olmos A, Green K. A Practical Guide to Conducting Systematic Literature Review Using R Software. Evaluation Studies in Social Sciences. 2020;1:35–43. https://doi.org/10.37134/esss.vol1.1.4.2020.

17. Gemma Team, Kamath A, Ferret J, Pathak S, Vieillard N, Merhej R, et al. Gemma 3 Technical Report. arXiv [Preprint]. 2025. Available from: https://doi.org/10.48550/arXiv.2503.19786

18. Turner S. biorecap: Retrieve and summarize bioRxiv and medRxiv preprints with a local LLM using ollama [Internet]. 2025. Available from: https://github.com/stephenturner/biorecap

19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–30.

20. Sayers E. A General Introduction to the E-utilities. In: Entrez® Programming Utilities Help [Internet]. National Center for Biotechnology Information (US); 2022 [cited 2026 Oct 1]. Available from: https://www.ncbi.nlm.nih.gov/sites/books/NBK25497/

21. Bornmann L, Mutz R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology. 2015;66:2215–22. https://doi.org/10.1002/asi.23329.

22. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017;7:e012545. https://doi.org/10.1136/bmjopen-2016-012545.

23. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open. 2023;13:e072254. https://doi.org/10.1136/bmjopen-2023-072254.

24. Efron B. Bayes' Theorem in the 21st Century. Science. 2013;340:1177–8. https://doi.org/10.1126/science.1236536.

25. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2020. p. 6840–51.

26. Villa M, Le Pera M, Cassina T, Bottega M. Reporting quality of abstracts from randomised controlled trials published in leading critical care nursing journals: a

methodological quality review. BMJ Open. 2023;13:e070639. https://doi.org/10.1136/bmjopen-2022-070639.

27. Yu A, Weile J, Courtot M. Benchmarking LLM-based Information Extraction Tools for Medical Documents. medRxiv [Preprint]. 2026. Available from: https://doi.org/10.64898/2026.01.19.26344287

28. Wada A, Nishizawa M, Yamamoto A, Akashi T, Hagiwara A, Irie R, et al. Bridging the performance gap: systematic optimization of local LLMs for Japanese medical PHI extraction. Sci Rep. 2026;16:5910. https://doi.org/10.1038/s41598-026-36904-5.

29. Gong EJ, Bang CS, Shin YS. Applications of Large Language Models in Medical Research: From Systematic Reviews to Clinical Studies. Bioengineering. 2026;13:365. https://doi.org/10.3390/bioengineering13030365.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SummplementaryMaterial1.xlsx
- SupplementaryMaterial2.docx