# A Primer on Latent Class Analysis

**Ernest H. Law, PharmD; Rachel Harrington**, Department of Pharmacy Systems, Outcomes and Policy, College of Pharmacy, University of Illinois at Chicago, Chicago, Illinois, USA

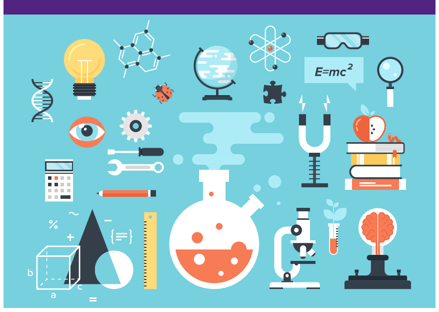
*Ernest H. Law, PharmD*

KEY POINTS . . .

Latent class analysis (LCA) is a modeling approach that identifies individuals that share common characteristics, allowing distinct “clusters” to be isolated.

LCA uses categorical indicators representing manifestations of a latent (or unobserved) variable (e.g., patterns of clinical signs and symptoms indicating different disease phenotypes).

LCA may be a useful tool in health outcomes research to better characterize the unobservable heterogeneity that exists within a population.




*The following article is the sixth in a series highlighting local student chapter activities and research talents. In this piece, we review latent class analysis.*

## Introduction to Latent Class Analyses

In outcomes research, it can be useful to represent underlying constructs as a model within which distinct subgroups, clusters, or categories of individuals exist. For example, a researcher may wish to determine the association between specific clinical factors and health outcomes. Traditional approaches typically model clinical factors as independent predictors of the outcome. However, this overlooks the fact that some factors (e.g., symptoms, hospitalization duration), do not exist in isolation, but rather share variance as a constellation of observed variables for a common latent (unobserved) variable. Latent class analysis (LCA) is one method that recognizes and leverages these relationships between observed variables to "cluster" together individuals for exploratory or explanatory investigations. This article will provide a brief introduction to LCA, including its important features and considerations, and potential applications in the field of health outcomes research.

## What is Latent Class Analysis and How Does it Work?

LCA is a person-centered approach that defines mutually exclusive and exhaustive subgroups of individuals within a population based on common characteristics. It is one of several frameworks that map individual items onto an underlying latent construct or variable (Table). Specifically, LCA uses observations of categorical dependent variables (also known as indicators) for every individual to define (categorical) class constructs. These classes may represent a multitude of underlying constructs, such as preferences, disease burden, symptom profiles, or genetic phenotypes. For comparison, factor analysis (or covariance structure analysis) is an analogous framework for mapping items onto latent variables on a continuous (e.g., normal) distribution. Since the latent classes cannot be observed directly, they must instead be inferred through relevant indicator variables measured from multiple observed items (or individuals). The Figure illustrates a hypothetical latent variable with three indicators (A, B, and C). It is important to note the direction of the arrows that point from the unobservable latent variable (and associated measurement error) to the indicator, signifying that observable indicators are caused by the unobservable latent variable. In other words, the observable indicators measure the latent variable, but do not in themselves cause the latent variable.

In traditional LCA models, two sets of parameters are estimated: class membership probabilities and item-response probabilities [1]. The class membership probability (or latent class prevalence) is the likelihood that an individual was properly classified, enabling each individual to be categorized into the best-fitting class. Class membership is estimated simultaneously with the overall model and is mutually exclusive and exhaustive (i.e., all class membership probabilities should sum to 1). Item-response probabilities form the basis for interpreting latent classes by indicating the probability of the present indicator, which is conditional for class membership. Comparing item-response probabilities between classes allows one to assess the distinctness of each identified class.

Latent class models are estimated by iteratively adding potential classes to determine which model is best fit to the data. Judging “best fit” requires consideration of several criteria, including information criteria and model parsimony. Information criteria (e.g., Akaike and/or Bayesian information criterion) reflect how well the model predicts the data, with smaller values indicating better model fit. After comparing candidate models based on model criteria, parsimony (i.e., interpretability) should be considered when making the final selection. However, the “optimal” model is not always clear and the investigator should be transparent in reporting the decision criteria and process.

Table: Types of latent variable models [9]

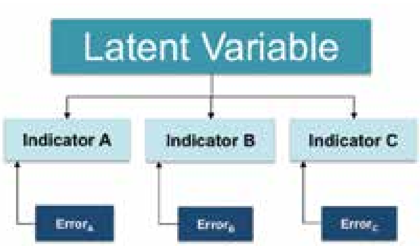
| | | Latent Variable | |
|---|---|---|---|
| | | Continuous | Categorical |
| Indicator | Continuous | Factor Analysis | Latent Profile Analysis |
| | Categorical | Latent Trait Analysis* | Latent Class Analysis |

## What are the Major Considerations when Conducting an LCA?

There are several issues that should be considered when using LCA. First, by using categorical data, strict assumptions about the distributions of indicators are not required. However, the assumption of local independence needs to be considered. Local independence assumes indicator variables are not correlated and are only related to each other through the latent variable (i.e., there are no arrows connecting indicators A, B, or C in the figure). It is important to note that the local independence assumption pertains to the observed variables within a latent class (i.e., the "local" in local independence). Indeed, observed variables (or indicators) may exhibit varying degrees of dependency outside of the latent classes. Researchers may wish to develop a conceptual framework to represent relationships between variables and to identify potential assumption violations.

Second, subgroup formation in LCA is typically determined by baseline data only, and is therefore not necessarily dependent on any one outcome or treatment. Therefore, subgroups detected by LCA can later be studied against a range of treatments and outcomes. However, subgroups identified this way may lack clinical relevance when used out of context. Subgroups should be tested to determine if latent class membership is predictive of clinical meaningful outcomes.

Figure: Conceptual representation of latent variable and observable indicators




Finally, LCA provides a potential solution to concerns about dimensionality in variable-centered regression analysis (e.g., when exploring interactions in a multivariable logistic regression). When the number of higher-order interactions (i.e., >2 variable interactions) measures statistical problems such as collinearity and reduced statistical power may result. Person-centered statistical approaches such as LCA can be used to mimic higher-order interaction terms, offering a simple summary of complex relationships.

## Example: Using LCA with Clinical Outcomes Research

Dumas et al. sought to identify distinct profiles among pediatric patients with severe bronchiolitis [2]. The authors identified four clinical profiles among a cohort of hospitalized children in the US and Finland using LCA. The investigators used 18 clinical variables (indicators) to form the clinical profiles (latent classes), which included history of wheezing, presenting symptoms, hospital length, and viral etiology. Profile membership and indicator-response probabilities were reported, and the authors discussed clinical interpretation of these classes. For example, children in "Profile C" were considered "the most severely ill group" [2]. They were more likely to have moderate-to-severe reactions and to have longer hospital stays. For predictive validity, the authors used the results of the LCA to evaluate the association between profile membership and antibiotic and asthma medication use during emergency department and inpatient stay.

## Conclusion

LCA is a method that can be applied to complex datasets to organize observed variables that represent unobservable clinical, sociodemographic, economic, and behavioral constructs into two or more potentially meaningful, homogenous subgroups. This increasingly popular approach has been applied to a variety of fields in outcomes research, including pharmacoepidemiology [3], analysis of clinical trial data [4], and treatment and health preferences studies [5-8]. Readers interested in learning more about LCA are encouraged to access the references cited throughout this primer for more comprehensive discussions and examples of application.

## Acknowledgements

We are indebted to Drs. Simon Pickard and Patrick Zueger for their assistance in reviewing the article prior to submission.

### References

[1] Kongsted A, Nielsen AM. An introduction to latent class analysis in health research. J Physiotherapy. 2016. [2] Dumas O, Mansbach JM, Jartti T, et al. A clustering approach to identify severe bronchiolitis profiles in children. Thorax. 2016;71:712-718. [3] Feuillet F, Bellanger L, Hardouin J-B, et al. On comparison of clustering methods for pharmacoepidemiological data. J Biopharm Statistics. 2015;25:843-856. [4] Leoutsakos JMS, Muthen BO, Breitner J, Lyketsos CG. Effects of non-steroidal anti-inflammatory drug treatments on cognitive decline vary by phase of pre-clinical Alzheimer disease: findings from the randomized controlled Alzheimer's Disease Anti-inflammatory Prevention Trial. Int J Geriatric Psych. 2012;27:364-374. [5] Mühlbacher AC, Bethge S. Reduce mortality risk above all else: a discrete-choice experiment in acute coronary syndrome patients. Pharmacoeconomics. 2015;33:71-81. [6] Kenzik KM, Martin MY, Fouad MN, Pisu M. Health-related quality of life in lung cancer survivors: Latent class and latent transition analysis. Cancer. 2015;121:1520-1528. [7] Hauber AB, González JM, Groothuis-Oudshoorn CG, et al. Statistical methods for the analysis of discrete choice experiments: a report of the ISPOR Conjoint Analysis Good Research Practices Task Force. Value Health. 2016;19(4):300-315. [8] Goossens LM, Utens CM, Smeenk FW,et al. Should I stay or should I go home? a latent class analysis of a discrete choice experiment on hospital-at-home. Value Health. 2014;17:588-596. [9] Collins LM, Lanza ST. Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences: John Wiley & Sons; 2013.