Latent Class Analysis in health research

Many health conditions are influenced by biological, psychological and social factors that interact to determine individuals’ prognoses and likely treatment responses.1–5 Many initiatives have been undertaken to address this complexity, by identifying homogenous subgroups of patients who have similar outcomes.3–5 Traditionally, subgroups have been identified by finding patient characteristics associated with the outcome of disease. However, such characteristics may not work the same way in all patients. Interest in a better understanding of individual differences between patients has prompted the use of person-oriented techniques to subgrouping that do not assume the relationship between variables to be the same for all patients. One such technique is Latent Class Analysis, which is based on people’s different scoring patterns across variables, rather than being driven by associations with an outcome. Such approaches are not being supervised by the outcome and therefore they are also referred to as unsupervised techniques. Latent Class Analysis is a very flexible technique that is available in several software packages.6 Based on established knowledge,7–11 the intention of the present paper was to provide readers with a non-technical introduction to the general principles of Latent Class Analysis and to the interpretation of the results of such analyses.

What is Latent Class Analysis and why would you use it?

Latent Class Analysis aims to identify subgroups of people who share common characteristics in such a way that people within the subgroups have a similar scoring pattern on the measured variables, while the difference in scoring patterns between the subgroups are as distinctly different as possible. Latent Class Analysis uses a mixture of distributions to identify the most likely model describing the heterogeneity of data as a finite number of classes (subgroups); this is known as finite mixture models.1 The model defines a categorical latent variable in which each level represents a subgroup. Because Latent Class Analysis is a probabilistic approach looking for the most likely model, subgroup membership is not fixed and all individuals are assigned a probability of belonging to each subgroup.7,8 A latent variable is not directly measured, but identified from some measured variables. The latent variable often represents a complex construct that cannot be directly measured (such as happiness or social behaviours). In essence, the purpose of any Latent Class Analysis is to define the latent variable in order to: identify a number of classes (in this paper called subgroups) that describe the underlying scoring patterns in the data; estimate the prevalence of the subgroups; and estimate each individual’s probability of belonging to each subgroup.

Latent Class Analysis is relevant when it is suspected that people do not only differ by levels on a continuum (eg, severity), but that different underlying distributions among the variables characterise the heterogeneity of the group. This means that different scoring patterns are expected with, for example, some people scoring high on one parameter and low on another, while others have the opposite scoring pattern. When the measured variables, often referred to as observed variables, manifest variables or indicator variables, include few items and simple response options, the possible scoring patterns and their frequency in data can be relatively easily described. However, the picture becomes more complex as more information about the patients is added.

For example, a study by Lacey et al investigated patterns of pain location based on information about the presence of pain in each of 16 body sites.12 With two options (pain or no pain) possible at 16 body sites, there were 2^{16} = 65,536 possible pain patterns. Latent Class Analysis helped to identify the pain patterns that provided the best balance between considering all individuals to belong to the same group and considering all existing patterns to be a relevant subgroup on their own. Everybody being in one subgroup would have meant that all patients’ scores originated from one underlying distribution with scores from 0 to 16 (from no pain at any site to pain in all sites). The assumption of one underlying distribution implied that it was not important in which body site the pain was located, and having pain in, for example, the neck and the shoulder would have ‘counted’ the same as having pain in the neck and the knee. Latent Class Analysis had the potential to identify distinct pain patterns in the data that best explained the scoring patterns (latent data structure) in the data.

General considerations when performing a Latent Class Analysis

Observed variables

A merit of Latent Class Analysis is its ability to handle different types of observed variables.3 Sometimes, analyses with continuous observed variables are referred to as ‘Latent profile analysis’ and those with categorical observed variables as ‘Latent class analysis’. However, in practice, the distinction is not that rigid and the two can be combined.

It is a basic assumption of Latent Class Analysis that the observed variables are not highly correlated within the identified subgroups, which is known as local independence. Ignoring this assumption would bias the statistical fit measures and generally result in models with a larger number of subgroups.10 Ignoring local independence may mean that variables that represent the same information get weighted too heavily in the Latent Class Analysis,7 and violating the assumption of local independence has been demonstrated to bias the subgroup solution in the special case of diagnostic tests with binary outcomes.10 It is believed that the practical implications of ignoring local dependent variables have not been investigated in other situations. It is unclear whether including observed variables that do not actually carry additional information might harm the model or not.10,11

Latent Class Analysis often includes a procedure for handling missing values and does not require complete data. However, a large proportion of missing data, and data that are not missing at

http://dx.doi.org/10.1016/j.jphys.2016.05.018
1836-9553 © 2016 Australian Physiotherapy Association. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
random, seems to reduce the accuracy of the model fit indices (discussed further below).^10^  

### Sample size

The risk of overfitting the model to the data will increase with a higher number of subgroups relative to the sample size, and an insufficient sample size can provide difficulties with model convergence. However, there is no simple way to estimate the required sample size.\(^13\) According to Finch and Bronk ‘it would appear that LCA (Latent Class Analysis) requires samples well into the hundreds, with most simulation studies suggesting 500 as a worthy goal in practice’.\(^14^\) A large number of informative and complete variables can, to some degree, compensate for a small sample size.\(^11^\) However, small samples limit the potential for identifying small but meaningful subgroups.

### Deciding the number of subgroups

A central element of Latent Class Analysis is deciding the number of subgroups. When performing Latent Class Analysis, a number of analyses are conducted, starting from a model with one subgroup and adding more subgroups until the ‘optimal’ solution is identified. The decision about the number of subgroups is commonly based on statistical measures of model fit, number of patients in the subgroups, the certainty of individuals’ membership to the subgroups and the clinical characteristics of the identified subgroups.

The Akaike information criterion and Bayesian information criterion statistics take model complexity into account and therefore attempt to balance the fit of the latent class model to the data, while favouring the most parsimonious model.\(^8^\) A decrease in Akaike information criterion or Bayesian information criterion indicates that a model with more subgroups (eg, a five-subgroup model) has a better trade-off between model fit and model complexity than the four-subgroup model. In contrast, the Likelihood Ratio simply examines whether the fit of a larger model is significantly better than that of a smaller one, without regard to model parsimony. These statistics guide the choice of models based on the probability of observing the data, given the subgrouping model; they do not tell whether any of the models are actually good at describing the health condition. For that reason, and because different model fit statistics often result in different ‘preferred’ model solutions, the models also have to be compared on other parameters.

The certainty of patients’ subgroup membership is quantified by the posterior probabilities, which sum to 1 across subgroups for each individual. People who have a pattern that fits very well with a certain group have a high probability of belonging to that group (close to 1) and a very low probability of belonging to other subgroups. In contrast, people with characteristics that do not fit very well with any group have low posterior probabilities and might have the same probability of belonging to more than one group. A commonly used measure for presenting the overall certainty of subgroup classification is ‘entropy’, which is a rescaling of the posterior probability.\(^8^\)

In the example of the pain pattern subgroups introduced above, the Bayesian information criterion decreased with each added subgroup up to the seven subgroups explored, and the Likelihood Ratio indicated that adding a subgroup made the fit significantly better at each step (Table 1). However, the change in the fit measures with more than five subgroups was considered minor, and patients’ average posterior probabilities were larger with four than with five subgroups, making the four-subgroup solution the preferred choice.

Face validity of the subgroup solutions cannot be objectively measured, but it is an important consideration in model selection because statistical measures do not tell if the identified subgroups are informative.\(^15^\) One parameter that assists in choosing an apparently informative model is the extent to which the models identify qualitative subgroup differences rather than only quantitative subgroup differences.\(^16^\) Qualitative differences imply that the

---

### Table 1

Model fit indices of the Latent Class Analysis models of 16 pain sites, as reported by Lacey et al.\(^12^\)

<table>
<thead>
<tr>
<th>Subgroups</th>
<th>Log-likelihood</th>
<th>% reduction in Log-likelihood from the previous model</th>
<th>BIC</th>
<th>Likelihood ratio test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 subgroup</td>
<td>-98,807</td>
<td>-</td>
<td>197,764</td>
<td>-</td>
</tr>
<tr>
<td>2 subgroups</td>
<td>-85,254</td>
<td>14</td>
<td>170,819</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>3 subgroups</td>
<td>-81,974</td>
<td>4</td>
<td>164,419</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>4 subgroups</td>
<td>-79,633</td>
<td>3</td>
<td>159,898</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>5 subgroups</td>
<td>-78,181</td>
<td>2</td>
<td>157,155</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>6 subgroups</td>
<td>-77,499</td>
<td>1</td>
<td>155,950</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>7 subgroups</td>
<td>-76,978</td>
<td>1</td>
<td>155,068</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>


The term ‘cluster’ is used in the original paper instead of ‘subgroup’. BIC = Bayesian information criterion.

### Figure 1

Hypothetical example of profile plots of three subgroups based on measures of pain, disability and depression.  
(A) Illustrates three subgroups that differ in severity but share a common pattern.  
(B) Illustrates qualitative differences between subgroups 1 and 2. The subgroups are similar for pain, but subgroup 2 is characterised by high disability and subgroup 1 by high depression.
Subgroups are characterised by different parameters, whereas purely quantitative differences between the subgroups simply represent different levels of severity across parameters (Figure 1).

Interpretation of latent class models

Two parameters are important for the interpretation of the results from a Latent Class Analysis: whether patients can be uniquely allocated to one subgroup (the posterior probability) and whether the subgroups have distinct clinical profiles.

Table 2 presents the characteristics of the subgroups identified by Lacey et al. The average posterior probabilities ranged from 0.86 to 0.95, indicating that although not all participants’ pain patterns were perfectly represented by the subgroups, there was a high certainty of the classification. The average posterior probability does not tell for how many patients the classification was uncertain. This is sometimes reported as the proportion of patients with a posterior probability below a certain value such as 0.70.[17]

How distinct the subgroups are is judged from the conditional probabilities or conditional means of scores on the variables that informed the subgroup formation, as well as from the differences of patient characteristics and outcomes that were not part of the Latent Class Analysis. ‘Conditional’ refers to the within subgroup probability or mean; the reported value is dependent on the specific subgroup. For example, the conditional probabilities reported in Table 2 indicate that in subgroup four there is a high probability of pain at many body sites, but it does not show the number of sites at which individuals in that group reported pain. Often, the conditional probabilities are visualised in a profile plot (Figure 2). To support that truly different subgroups are identified and to understand how the identified subgroups are different, it is also useful to compare the subgroups on parameters that were not part of the Latent Class Analysis.[18]

Discussion and summary

Latent Class Analysis and other person-centred techniques are relevant when there is a desire for an individualised approach, whilst at the same time a need to understand something general about a patient population. Latent Class Analysis explores a number of subgroup solutions that are compared on statistical and clinical parameters to choose the ‘optimal’ model. However, there is often not one optimal model. The preferred model from a Latent

---

**Table 2**

Characteristics of four pain-site subgroups identified by Latent Class Analysis, presented as probabilities of reporting pain at each of 16 pain sites and the average posterior probability in four Latent Class Analysis-derived subgroups, as reported by Lacey et al.[12]

<table>
<thead>
<tr>
<th>Site</th>
<th>Probability of reporting pain</th>
<th>Subgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Head</td>
<td>0.018</td>
<td>0.109</td>
</tr>
<tr>
<td>Neck</td>
<td>0.001</td>
<td>0.072</td>
</tr>
<tr>
<td>Chest</td>
<td>0.003</td>
<td>0.125</td>
</tr>
<tr>
<td>Abdomen</td>
<td>0.022</td>
<td>0.120</td>
</tr>
<tr>
<td>Shoulder</td>
<td>0.021</td>
<td>0.439</td>
</tr>
<tr>
<td>Elbow</td>
<td>0.001</td>
<td>0.203</td>
</tr>
<tr>
<td>Forearm</td>
<td>0.000</td>
<td>0.191</td>
</tr>
<tr>
<td>Hand</td>
<td>0.030</td>
<td>0.317</td>
</tr>
<tr>
<td>Spine</td>
<td>0.062</td>
<td>0.222</td>
</tr>
<tr>
<td>Upper back</td>
<td>0.000</td>
<td>0.173</td>
</tr>
<tr>
<td>Lower back</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Buttock</td>
<td>0.031</td>
<td>0.162</td>
</tr>
<tr>
<td>Knee</td>
<td>0.087</td>
<td>0.561</td>
</tr>
<tr>
<td>Calf</td>
<td>0.004</td>
<td>0.321</td>
</tr>
<tr>
<td>Foot</td>
<td>0.038</td>
<td>0.324</td>
</tr>
<tr>
<td>Spine</td>
<td>0.028</td>
<td>0.270</td>
</tr>
<tr>
<td>Average posterior probability</td>
<td>0.95</td>
<td>0.86</td>
</tr>
</tbody>
</table>


---

**Figure 2.** A profile plot illustrating four subgroups identified from pain charts of participants in Lacey et al.[12]. People were assigned to the subgroup for which they had the highest posterior probability. The profile plot illustrates the proportion assigned to each subgroup that indicated pain at each body site.

Subgroup 1: characterised by a low probability of pain in any of the listed sites

Subgroup 2: characterised by a high probability of pain in the knees and a low probability of spine pain

Subgroup 3: characterised by a high probability of pain in the spine and buttocks

Subgroup 4: characterised by high probability of pain at many sites.

Class Analysis is one that represents the data well, has a reasonable distribution of patients across subgroups, has a high certainty of classification, and for which the subgroups have clear clinical characteristics. Even when a convincing subgroup model is identified, it may not necessarily improve clinical practice. This would be determined from validation studies testing if the subgroups can be identified in other samples and if the subgroups have different prognoses or respond differently to treatment.

It should be recognised that Latent Class Analysis does not provide a firm answer to how many subgroups exist within a condition. The result illustrates how many subgroups provide a good balance between recognising nuanced between-patient differences and understanding general patterns in relation to the chosen measured variables. For instance, a Latent Class Analysis based on pain and disability is likely to result in a model with fewer subgroups than would be the case if psychological variables were used in addition. With the available user-friendly software packages, it has become quite straightforward to run a Latent Class Analysis and choose the model with the best statistical fit. It is, however, important for the reader of such analyses to be aware that a large number of decisions have been made concerning the measured variables and the model selection, and that these have not always been straightforward.

Latent Class Analysis and similar techniques have become popular in health research and have been applied, for example: in subgrouping people with musculoskeletal pain based on their pain distributions,12,19 to identify pain phenotypes of knee osteoarthritis,20 and in identifying distinct trajectories of chronic obstructive pulmonary disease.21 However, it is believed that the potential clinical impact of Latent Class Analysis–based models has still not been tested in any condition. Therefore, the true value of these techniques in health research still warrants investigation.

**Ethics approval:** Nil

**Competing interests:** The authors declare to have no competing interests. No funding was received in direct relation to this work.

**Source(s) of Support:** Nil

**Acknowledgements:** AK and AMN’s positions are financially supported by the Foundation for Chiropractic Research and Post Graduate Education (Denmark); additionally AMN is supported by the Faculty of Health Sciences, University of Southern Denmark.

**Provenance:** Invited. Peer reviewed.

Alice Kongsted and Anne Molgaard Nielsen

Department of Sports Science and Clinical Biomechanics, University of Southern Denmark and Nordic Institute of Chiropractic and Clinical Biomechanics, Odense, Denmark

**References**