

INTRODUCTION



Background and rationale

*Partly based on:
Jos Twisk & Trynke Hoekstra*

*Classifying developmental trajectories over time should be done with great caution:
A comparison between methods
Journal of Clinical Epidemiology 65:1078-1087
2012*

Epidemiologists and medical researchers are often interested in analysing longitudinal data. One of the main advantages of longitudinal data is the possibility to study individual development over time [1]. Studying these individual developmental trajectories helps to better understand how risk factors for diseases naturally develop and aids the understanding and unravelling of aetiology of diseases, which is important for early detection and prevention. When analysing longitudinal data, the fact that repeated observations of each individual in the dataset are not independent (i.e. correlated) should be taken into account [1]. Standard regression analyses are not able to fully incorporate these correlated measures and therefore, more sophisticated statistical techniques are needed to analyse these data accordingly. Several techniques to do this are available, such as mixed models [2, 21, 22] and generalised estimating equations [1, 23]. In particular mixed models are now common in the epidemiological literature and many clinically applied applications exist [24]. Besides taking into account the dependent observations within the data, mixed models can, to some extent, also account for individual variability [2]. This possibility, however, is limited and for example does not allow for the revelation of subgroups of individuals with different trajectories of risk factors and consequently, potentially differential risks of disease. In particular in the fields of psychology and criminology, acknowledgement of heterogeneity in developmental trajectories has led to new theories of multiple developmental pathways of, for example, binge drinking [13–15, 25] and criminal- or misbehaviour [16, 26, 27] and has improved the understanding of the aetiology of these behaviours. In this context, Odgers et al. [3] were one of the first researchers to empirically or statistically (i.e. determined by the data) derive distinct trajectories of criminal behaviours, comparable to those proposed in theories from developmental psychology [28]. Odgers et al. [3] were able to statistically reveal distinct trajectories of conduct disorder, as predicted by the theories of Moffitt [28], hereby providing an external validation of this longstanding theory of criminal behaviour. More examples followed in the fields of criminology and psychology in particular [15], but the so-called group-based approach (as opposed to the more common classification-based approach where individuals are classified into predefined subgroups [29]) was and is less often applied in epidemiology [29]. However, it is clear that in order to answer similar research questions in this field such an approach is more often required. The knowledge of distinct trajectories including their determinants and consequences can allow for the identification of high-risk individuals, who

may need supplementary treatment or preventive measures in addition to the standard care or prevention. Moreover, such an approach can also allow the subsequent identification of low-risk individuals who might need less treatment or preventive measures.

There are several statistical techniques available to study distinct developmental trajectories, the most commonly used and the most flexible techniques available are probably those based on structural equation models [30–32], i.e. *latent class models*. These techniques have previously shown their merit in other fields of research [33], and have relatively recently been introduced in epidemiology. However, the techniques are fairly complex to understand and apply. This poses challenges for a confident familiarity with the understanding of underlying assumptions, the possible extensions of the models and knowledge of using the available software packages. These issues will be highlighted in this thesis and aim to guide epidemiologists in making correct, well-thought through decisions before, during and after the latent class modelling process.

Outline of this thesis

Several latent class models will be presented in this thesis in the context of epidemiological research. Within the flexible framework of the Mplus statistical software [34], several of these models will be described in detail, in order to aid researchers in epidemiology in the selection and application of them. Difficult mathematical formulas are kept to a minimum throughout the thesis and the general approach will be to explain the techniques by taking a range of research questions as focus points; each chapter describes an example (or examples) of the application of a latent class model to answer a particular research question. However, readers are assumed to have knowledge of (cross-sectional as well as longitudinal) regression modelling.

First, chapter 2 will provide an overview of the most common latent class models, including important methodological considerations and chapter 2 describes the datasets used in this thesis. Chapter 4 then demonstrates in detail the application of a latent class growth mixture model, which is one of the most flexible longitudinal models available. Chapter 5, 6 and 7 describe more longitudinal examples of latent class models in more applied settings. Chapter 8 and 9 describe two examples of the use of latent class models in cross-sectional studies. With cross-sectional data, we are not so much interested in heterogeneity in development, but more in the unobserved subgroups with distinct characteristics, symptom profiles or behaviours. Finally, chapter 10 reviews the main findings of the work presented in this thesis in the context of the current literature, discusses several unresolved (clinical and methodological) issues and presents suggestions for future research opportunities.