Summary

Epidemiologists and medical researchers are often interested in analysing longitudinal data, because it makes possible the study of individual development of an outcome over time. Studying these developmental trajectories helps to understand how risk factors for diseases develop and aids the unravelling of aetiology of diseases, which is important for (early) detection and prevention.

There are many ways to analyse longitudinal data, of which the most common techniques in epidemiology include generalised estimating equations (GEE) [1] and mixed models [2]. Both techniques take into account the repeated observations (correlated data) and only to some extent are able to handle individual variability in development. The possibility to do so is limited, however, and both GEE and mixed models do not allow for the revelation of subgroups of individuals with varying trajectories of risk factors and consequently, potentially different risks of health- or disease outcomes. The acknowledgement (and identification) of distinct developmental trajectories is relatively common in the fields of psychology and criminology and has led to new theories of multiple developmental pathways of binge drinking [13–15] and criminal behaviour [16], amongst others.

To identify these distinct trajectories, a range of techniques are available, of which *latent class models* are probably the most flexible. The merit [4] of these models for psychological- and criminological research has been demonstrated previously, but they have only recently been introduced to epidemiological researchers.

Latent class models do not assume that individuals in the sample come from one underlying population, but rather may originate from multiple, underlying (or latent) subpopulations. The main aim of the latent class models is to (statistically) identify the number and characteristics of these subpopulations (or classes). Each class has its own growth parameters (intercept, slope) in the case of longitudinal data or item characteristics in the case of cross-sectional data. Because latent class models are clustering techniques, the optimal model will be a model in which homogeneity *within*- and heterogeneity *between* classes is maximised.

To overcome the relative uncommonness with latent class models in epidemiology, latent class models are the focus of this thesis. The overall aim is to guide epidemiologists in making correct, well-thought through decisions before, during and after the latent class modelling process.

Chapter 2 of this thesis provides an overview of the most common latent class models, namely latent class growth models (LCGM), latent class growth mixture models (LCGMM) for longitudinal data and latent class analysis for cross-sectional data. Model assumptions (within-class conditional normality, overall goodness-of-fit, missing data are missing at random, independent observations between study participants and local independence) are discussed and the model building process is explained. Important issues in the model building process include 1) criteria for determining of the optimal number of classes (e.g. model fit indices [17], sample size issues [18, 19], posterior probabilities [4], model parsimony); 2) assigning individuals to classes; 3) the choice for a conditional- versus an unconditional model [5] and 4) studying predictors and consequences of latent class membership [20].
In chapter 4, a detailed example of a LCGMM is provided using body mass index data from the Amsterdam Growth and Health Longitudinal Study [8].
Chapter 5, 6 and 7 describe several applied examples of latent class models in longitudinal datasets too. Besides the Amsterdam Growth and Health Longitudinal Study (chapter 6), the Spokane Heart Study [9, 10] (chapter 5) and data from the Diabetes Care West-Friesland [11] (chapter 7) are analysed. A variation of topics, ranging from measures of cardiovascular disease-risk, to mental health, diabetes and colorectal cancer symptoms is studied. Chapters 8 and 9 concern applications of cross-sectional latent class models using data from the LINH-study [12]. Finally, chapter 10 reviews the main findings of previous chapters and highlights some research opportunities especially relevant to the field of epidemiology.

A final note: The recognition and subsequent analysis of heterogeneity within a study population can be a valuable insight, and can instigate new research ideas, provided that the models are conducted with caution. Therefore, the continuing interaction between clinicians, epidemiologists and biostatisticians is encouraged; awareness of the potential of latent class models for epidemiology, including the correct use of them will be beneficial to both the evolvement of the techniques as well as the epidemiological research field.