Introducing the fit-criteria assessment plot - A visualisation tool to assist class enumeration in group-based trajectory modelling

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2017, Vol. 26(5) 2424–2436

DOI: 10.1177/0962280215598665
journals.sagepub.com/home/smm
SAGE

Introducing the fit-criteria assessment plot – A visualisation tool to assist class enumeration in group-based trajectory modelling

Sven L Klijn,[1] Matty P Weijenberg,[2] Paul Lemmens,[3] Piet A van den Brandt[4] and Valéria Lima Passos[1]

Abstract

Background and objective: Group-based trajectory modelling is a model-based clustering technique applied for the identification of latent patterns of temporal changes. Despite its manifold applications in clinical and health sciences, potential problems of the model selection procedure are often overlooked. The choice of the number of latent trajectories (class-enumeration), for instance, is to a large degree based on statistical criteria that are not fail-safe. Moreover, the process as a whole is not transparent. To facilitate class enumeration, we introduce a graphical summary display of several fit and model adequacy criteria, the fit-criteria assessment plot. **Methods**: An R-code that accepts universal data input is presented. The programme condenses relevant group-based trajectory modelling output information of model fit indices in automated graphical displays. Examples based on real and simulated data are provided to illustrate, assess and validate fit-criteria assessment plot's utility. **Results**: Fit-criteria assessment plot provides an overview of fit criteria on a single page, placing users in an informed position to make a decision. Fit-criteria assessment plot does not automatically select the most appropriate model but eases the model assessment procedure. **Conclusions**: Fit-criteria assessment plot is an exploratory, visualisation tool that can be employed to assist decisions in the initial and decisive phase of group-based trajectory modelling analysis. Considering group-based trajectory modelling's widespread resonance in medical and epidemiological sciences, a more comprehensive, easily interpretable and transparent display of the iterative process of class enumeration may foster group-based trajectory modelling's adequate use.

Keywords

Group-based trajectory model, model selection, class-enumeration, finite mixture model, model dimensionality, proc traj

1 Introduction

Group-based trajectory modelling (GBTM) is a semi-parametric, model-based clustering technique that is mostly applied for the identification of latent groups of individuals following a similar progression of an outcome over time.[1,2] GBTM analysis can be run with a SAS macro '*proc Traj'*, developed by Jones and Nagin[3,4] , which can be downloaded from http://www.andrew.cmu.edu/user/bjones/. A version in Stata is also available.[5]

GBTM's cradle was in criminology, where theory-driven research has given much attention to the origins and development of crime and delinquency over a lifelong period.[6–11] Recently, GBTM has gained impetus for analysis

[1]Department of Methodology and Statistics, Maastricht University, Maastricht, the Netherlands
[2]Department of Epidemiology, GROW School for Oncology and Developmental Biology, Maastricht, the Netherlands
[3]Department of Health Promotion, Maastricht, the Netherlands
[4]Department of Epidemiology, GROW School for Oncology and Developmental Biology, CAPHRI School for Public Health and Primary Care, Maastricht University, Maastricht, the Netherlands

Corresponding author:
Valéria Lima Passos, Department of Methodology and Statistics, Maastricht University, Peter Debyeplein, 1, 6229 HA Maastricht, the Netherlands.
Email: valeria.limapassos@maastrichtuniversity.nl

of longitudinal data in clinical and health sciences.[12,13] Its popularity is especially noteworthy in life-course epidemiology.[14–20] Medical and health scientists, especially epidemiologists, often require developmental approaches to unravel the role of potential determinants, such as early life, life time (e.g. genetic) and environmental exposures in relation to individual paths of health and disease.[14,17,21]

GBTM's expansion is driven by many of its appealing features. Most conspicuously, it is a practical procedure for recognition and visualisation of different patterns of temporal change in a characteristic. Contrary to ex ante classifications, the user of GBTM does not presume the existence of a priori defined classes. Subgroups of individuals sharing a similar developmental course emerge from the data. The distinct longitudinal paths are referred to as latent trajectories (also known as latent classes, groups or strata). GBTM is, hence, suited for the task of uncovering unobserved heterogeneity in a population, i.e. data variability that is not attributable to any known factor. GBTM has been used for multiple purposes in clinical sciences. Most of these analyses aim at identification and characterisation of patients showing regular and irregular patterns in outcomes in a longitudinal design.[22,23] Modelling medication adherence[24,25] or modelling heterogeneity of treatment effect/response over time[26,27] are some examples. The information obtained by a GBTM analysis has been used to improve patients' profiling,[28–30] risk stratification,[19,31,32] patients' management[33,34] and prognosis.[35–40]

Despite widespread enthusiasm and rising number of adherents, a closer inspection of the specialised literature reveals that controversies around GBTM abound.[41–49] Such state of affairs is partly attributable to the fact that empirical and statistical properties of GBTM are not yet fully known. The existing literature on factors affecting mixture models performance is scant.[50–56] As Erosheva properly notes, 'although the applied literature on GMM[1] and GBTM is rich, publications in the statistical literature are relatively rare'.[60]

Most debates around the merits and demerits of GBTM tend to unfold within specialised circles and are often limited to social/psychopathological and criminological journals. Practitioners from other fields of expertise, especially newcomers, remain oblivious to these disputes. As a result, while publications of empirical studies based on GBTM are burgeoning, so do the inadequacies in some of its applications. For example, often the probabilistic nature of subjects' assignment to their respective latent trajectories (the uncertainty of cluster membership) is disregarded in the process of validating the trajectories.[61] Another moot point, which is neither properly recognised nor dealt with, is how the number of latent classes, k, is determined, a procedure often referred to as class enumeration or choosing the dimensionality of the model.

Class enumeration is a decisive stage of GBTM, representing the first step in lengthy and iterative analytical process. Several formal criteria of model fit and adequacy are available, some of which are part of *proc traj*'s standard output. These criteria, however, are known to be fallible. Moreover, they are presented in a fragmented manner making it difficult for researchers to weigh the different sorts of information, especially when comparing several models.

To facilitate class enumeration in GBTM, this paper introduces a graphical summary display of several fit and model adequacy criteria, the fit-criteria assessment plot (F-CAP). F-CAP is a visualisation tool that assists researchers in decisions on the number of latent classes in the initial phase of a GBTM analysis. To demonstrate, assess and validate F-CAP's utility, examples are provided based on real data from criminological and health/epidemiological sciences studies as well as simulated data.

2 The model

GBTM, also known as latent class growth model is a form of finite mixture model that identifies homogenous subgroups of a population sharing similar patterns of change of an outcome over time. Extensive details about the model can be found elsewhere[2,3] and a short introductory tutorial how to use the *proc traj* is presented by Andruff et al.[62] The general model is given by

$$\boldsymbol{P}(\boldsymbol{Y_i}) = \sum_{j}^{k} \pi_j \boldsymbol{P^j}(\boldsymbol{Y_i}) \tag{1}$$

where $P(Y_i)$ is the unconditional probability of an outcome (Y) of individual i. Group membership is denoted by j, while k is the total number of latent groups or trajectories to be identified from the data. π_j represents the estimated proportion of the sample belonging to group j and $P^j(Y_i)$ is the probability of observing the outcome of individual i, conditional upon membership of group j. Model parameters estimates are maximum likelihood based. Conditional on group membership the repeated measures of a subject are assumed to be independent. Moreover, between-subjects variability within every latent trajectory is assumed to be zero (no random

coefficients within a trajectory, contrary to GMM). Both assumptions are rigid but reduce the complexity of the model considerably.

The link between time and the outcome variable is determined by the type of data: for continuous outcomes, the link function is the censored normal; for count data, the zero-inflated Poisson and for dichotomous outcomes, the binary logit distribution. For a continuous outcome, the equation determining the shape and level of the trajectory j of the latent variable y_{it}^{*} is given by

$$y_{it}^{*} = \beta_0^j + \beta_1^j Time_{it} + \beta_2^j Time_{it}^2 + \beta_3^j Time_{it}^3 + \varepsilon_{it} \tag{2}$$

where the regression coefficients determine the trajectories' starting levels (intercepts) and temporal shape (here cubic polynomial regression coefficients, but *proc traj* can estimate coefficients up to the fifth order). Each trajectory will have its own intercept and polynomial parameters that are generally expected to be quantitatively and qualitatively distinct from those of the other trajectories. For the error term, the usual assumption holds $\varepsilon_{it} \sim N(0, \sigma)$.

For the zero-inflated Poisson link function, the model assumes that

$$ln(\lambda_{it}^j) = \beta_0^j + \beta_1^j Time_{it} + \beta_2^j Time_{it}^2 + \beta_3^j Time_{it}^3 \tag{3}$$

where $ln(\lambda_{it}^j)$ is the natural logarithm of expected numbers of occurrences of an event for subject i at time t, conditional on trajectory j. As for the logit model, also conditional on j, if y_{it} is a binary response of subject i at time t, the probability of the outcome of interest is given by

$$p_t^j(y_{it}) = \frac{e^{\beta_0^j + \beta_1^j Time_{it} + \beta_2^j Time_{it}^2 + \beta_3^j Time_{it}^3}}{1 + e^{\beta_0^j + \beta_1^j Time_{it} + \beta_2^j Time_{it}^2 + \beta_3^j Time_{it}^3}} \tag{4}$$

3 Model selection

The existence of goodness-of-fit and model adequacy indices, used to guide the choice of the number of latent classes, is a prominent advantage of model-based clustering compared to its heuristic counterparts, such as k-means clustering. However, these criteria are not foolproof, having their own share of unresolved issues. Their functioning has been shown to be influenced by sample size, e.g. a large N may lead to over-extraction of latent trajectories, by the magnitude of underlying data heterogeneity (within and between cluster variability), by the degree of clusters' separation, the number of time points and by violations of model assumptions, to cite a few.[41,51,55,63,64]

To make matters worse, the criteria may not coincide in their respective selection of k. These shortcomings are reinforced by the lengthy and iterative model fitting procedure. Though some information-based indices are default elements of a *proc traj* output, others need to be computed by hand, for every fitted model before reaching a decision. At the end, the user is left with bits and pieces of indices that need to be properly organised, compared and evaluated with respect to the theoretical meaningfulness of the selected model before proceeding to the next phase of the analysis, which is usually the model validation.

On a whole, although *proc traj* is easy to use, the model specification itself is, at its best, cumbersome and, at its worst, obscure. In practice, GBTM-based research articles either omit fit indices or only present the Bayesian information criterion (BIC) for a limited number of models. Other indices are seldom reported. The reader is left unable to evaluate the appropriateness of the selected model and has to give the author the benefit of the doubt. Considering GBTM's widespread use in clinical sciences, a more comprehensive, easily interpretable and transparent display of the iterative process of class enumeration is welcome.

3.1 Overview of common criteria for class enumeration

Several criteria are available for the selection of k. The first three are Akaike's information criterion (AIC), the likelihood (L) of the model, and the frequently used, and often favoured, BIC

$$BIC = \log(L) - 0.5\, p \log(\mathrm{N}) \tag{5}$$

where L is the model's (maximised) likelihood, N is the sample size and p is the number of parameters in the model.

Both the AIC (the same expression as in equation (5), without N) and BIC penalise for the amount of parameters used in the model, though the BIC does so more severely by additionally taking the sample size into account.

Nagin also recommends four other criteria of model adequacy based on the posterior probability of assignment.[2] These are the average posterior probability of assignment (APPA), odds of correct classification (OCC), mismatch between estimated and assigned group probabilities, and standard deviation of group membership probabilities (SD-GMP). A final criterion is the percentage of individuals estimated to be assigned to the smallest group, thus the smallest π_j. For this, a cut-off point of 1% is often applied. This threshold, however, may be adjusted to the sample size, with smaller samples requiring a larger minimum percentage as to warrant sufficient number of members in the smallest latent class.

The APPA for a group is the mean of the PPA of individuals assigned to this group. A threshold of 0.70 is considered an acceptable limit.[2]

The OCC compares the odds of correctly classifying subjects into group j based on the maximum probability classification rule (APPA for class j), correcting for the OCC based on random assignment. Per group (j), OCC is defined as

$$\boldsymbol{OCC_j} = \frac{\boldsymbol{APPA_j}/(1 - \boldsymbol{APPA_j})}{\hat{\pi}_j/(1 - \hat{\pi}_j)} \tag{6}$$

Note that in the denominator, the odds of the probability of assignment to j take the estimated population proportion $\hat{\pi}_j$ as the probability of correct classification. If this is, for instance, 0.25, then chance would already classify 25% of the subjects to j correctly. Higher OCCs indicate a better fitting model and an OCC above 5.0 for all groups shows good assignment accuracy.[2] For models with only one group, the equation fails and the OCC is infinite. In these cases, the OCC is arbitrarily set to 999.0, to convey a good fit.

The mismatch is the difference between the estimated group probability $\hat{\pi}_j$ and the real proportion of the sample assigned to group ($P_j = N_j / N$). In cases of a perfectly fitting model, the mismatch ($\hat{\pi}_j - P_j$) score is zero.

The SD-GMP calculates the standard deviation of assignment probability per group, based on the individuals assigned to that group. This measure can be expected to have lower values, when models have more groups. Lower SD-GMP scores indicate a better fitting model.

3.2 Analysis strategy

Nagin recommends a two-step procedure for model selection.[2] Firstly, the number of latent trajectories, k, is selected based on fit indices. Subsequently, the order of the polynomials describing the level and shape of the latent trajectories is determined.

Whether the stipulation of k is done first, holding the order of polynomials constant, or whether both class enumeration and the tests of the regression coefficients are conducted simultaneously can be left to the researchers' own consideration.

From recent research articles that used GBTM (http://www.pubfacts.com/search/Group-based + trajectory + modeling), the following analysis strategy emerged as a common operational procedure:

1. Model Specification:
   - First, Settle for a starting polynomial order (the same for all ks) and maximum number of trajectories k. These calls are partly arbitrary, partly theory driven, depending also on the nature of the data (a large number of time points allows for a higher order of polynomials to be explored, whereas large sample sizes give the researcher the freedom to play around with a larger number of clusters). Keeping the order of polynomials constant, run several models for different 'ks';
   - Second, among all models, select the one yielding the best BIC, with corroboration by the supportive roles of APPA (above 0.7), and the 1% rule for $\hat{\pi}_j$. The indices themselves are seldom provided in publications;
   - Third, keep the selected k constant, conduct significance tests of polynomials terms keeping a watch on the BIC;
2. Model validation:
   Lastly, in the overwhelmingly vast majority of papers, a classify–analyse[ii] strategy follows, either for latent classes profiling (latent classes' comparisons based, e.g. on ANCOVA/ANOVA), or for predictive validity in terms of the association between trajectories and a distal outcome (e.g. logistic regression

for dichotomous outcome). At this stage, a prominent feature of GBTM, which is the uncertainty of cluster membership, is taken out of the equation. A closer discussion on model validation is beyond the scope of the present work.

4 Introducing the F-CAP

Coherent to this analysis of plan, F-CAP is introduced as an assisting and exploratory tool for the selection of number of latent classes, *k* (associated R-code, Excel template for the F-CAP as well as the instructions manual can be obtained from the corresponding author or found in the supplementary material (available at: http://smm.sagepub.com/)). Though our analyses were conducted with the SAS *proc traj,* the R-code accepts universal data input with model fit indices.

4.1 Data sources

The first data (The Montreal Study) were used to introduce GBTM in the seminal work of Nagin and Jones. The second data stem from a cohort study conducted in the Netherlands. For the simulation, the outcome variable and random effects (when applicable) were generated from a multivariate normal distribution for each latent cluster separately and then merged. *Proc traj* based on the censored normal distribution was run for all examples.

4.2 The F-CAP

The F-CAP combines eight goodness-of-fit and model-adequacy criteria in compact graphs for several user-stipulated varying *ks*.

The basic idea behind the F-CAP is that the user can see how indices change by increasing the number of latent trajectories. The visual display of their behaviour in such condensed form provides an insightful perspective. Different criteria can be assessed side by side to reach a well-informed decision. Moreover, it fosters transparency in the process both for the user and the critical appraiser. The downside of the F-CAP is that a visual inspection of the indices' behaviour provides little information about whether the undergone changes, by varying *k*, are substantive or not. For example, does the magnitude of the BIC change favour one model over the other adjacent? The use of the Bayes factor has been recommended for models' comparisons. This is, however, no constituent element of the F-CAP. Moreover, by using model selection indices to decide among competing models, one is disregarding the impact, often disproportionate, that a few cases may have on model ranking.[55]

5 Results

Each F-CAP is composed of four sub-plots (b to e in Figures 1 and 2). The AIC, BIC and L of the models are displayed in sub-plot (b). The next sub-plots display the APPA, mismatch and SD (c), OCC (d) and at last π_j(e). For some criteria (APPA, OCC, mismatch and SD), two lines are displayed. The first one is the mean value of all groups for that criterion. The second value is denoted by either a plus or a minus sign. This indicates that the given value is that of the group with, respectively, the highest or lowest score for that criterion. The complementary information of the worst-case scenario is a reassurance that for all selected classes, model fit and adequacy indices are kept within the acceptable boundaries. In the F-CAPS for the real data, the vertical green bar indicates the final selection of *k*.

Due to space constraints, all F-CAP subplots have been compressed and are displayed for the Montreal longitudinal and Leefstijl en Gezondheid Onderzoek – Lifestyle and Health Research (LEGO) studies. For the simulated data, only the information-based criteria graph is given. Excel files of all original F-CAPs can be found in the supplementary material (available at: http://smm.sagepub.com/).

5.1 F-CAP – Montreal longitudinal study

In the Montreal longitudinal study,[1] behavioural attributes of 1000 boys were annually rated by their teachers at ages 6 and 10–15 years. Three externalising behavioural problems were quantified (physical aggression, opposition and hyperactivity).

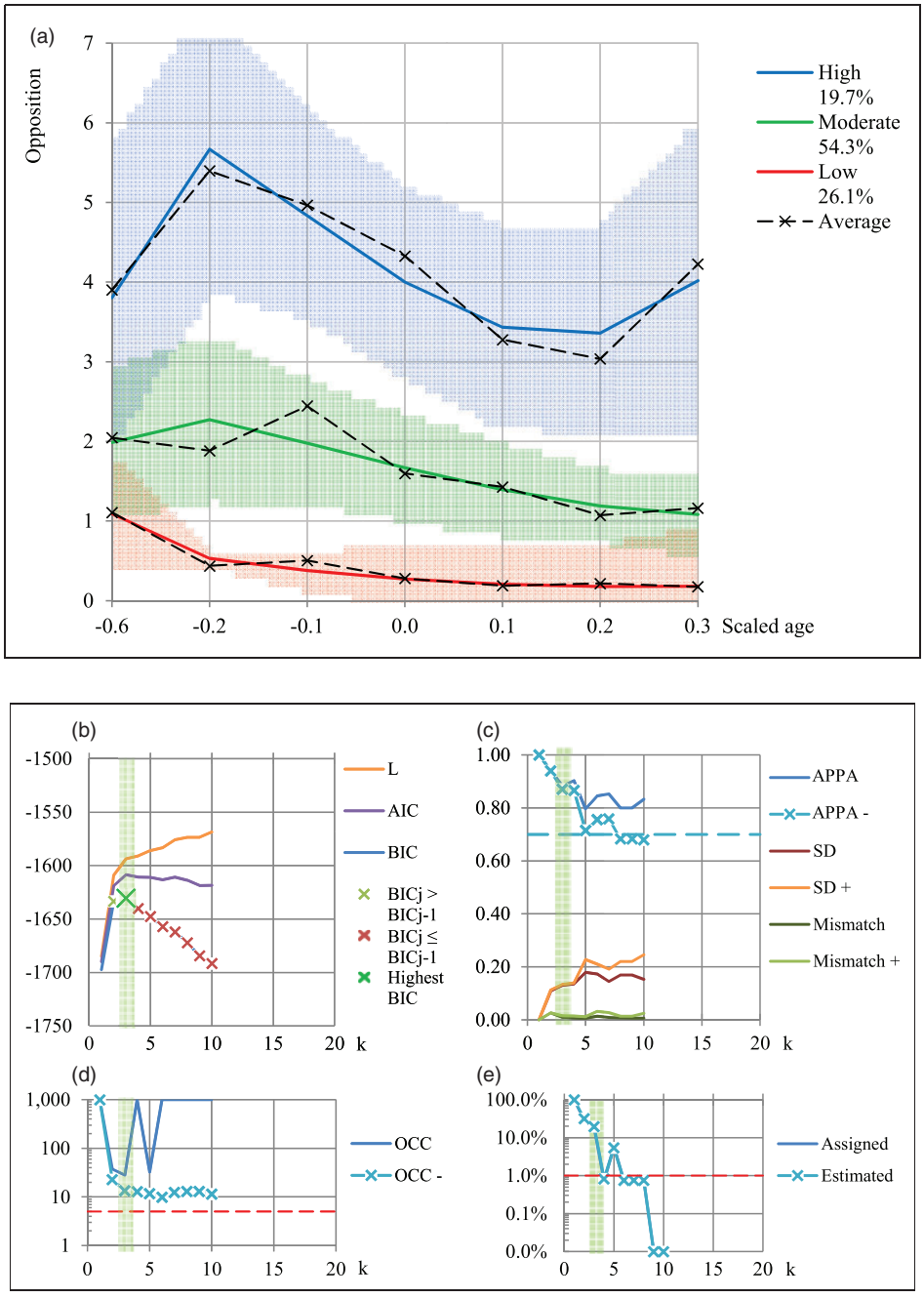

Figure 1. (a) Average oppositions scores of the final trajectories: estimated values (coloured lines) with corresponding 95% confidence intervals (shaded areas) versus observed values (interrupted lines). Estimated groups' proportions are given in %. (b) to (e) AIC, BIC and L (b); APPA, mismatch and SD (c); OCC (d) and percentage of individuals estimated to be assigned to the smallest group (e). Ten models of opposition scores were fit (x-axis). Shaded vertical bar indicates the selected k (here $k = 3$).

5.1.1 Dataset

For the present purpose, only the data on the opposition scores (scale range 0–10) of 138 subjects were analysed with GBTM.

The best fit model had $k = 3$, with cubic trajectories, as selected by the authors. The plot with individual trajectories and their 95% confidence band is presented below (Figure 1(a)). The coloured lines represent the average change of the outcome longitudinally, as estimated by the model, whereas the dotted lines were the

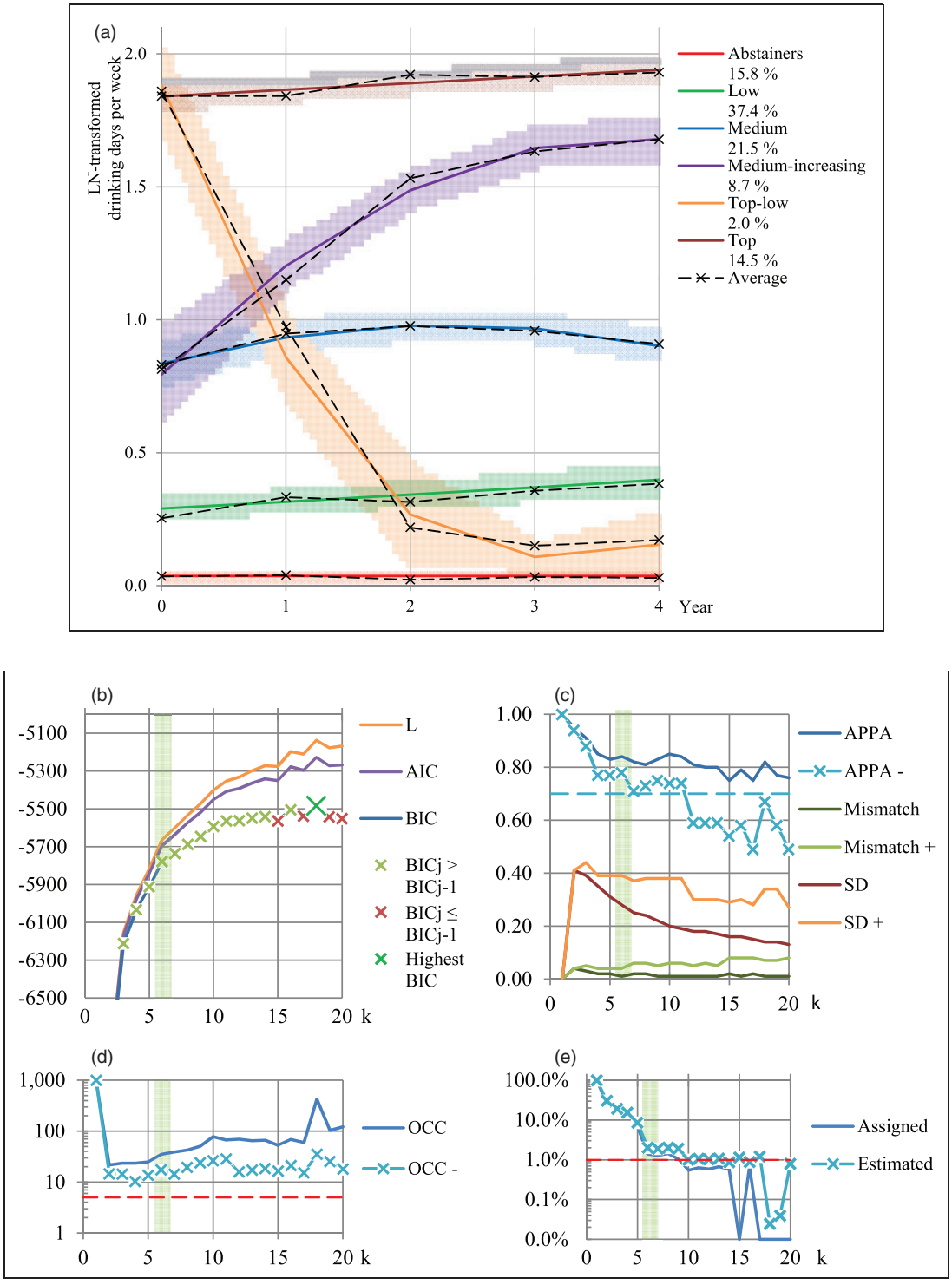

Figure 2. (a) Average frequencies of drinking wine per week of the final trajectories: estimated values (coloured lines) with corresponding 95% confidence intervals (shaded areas) versus observed values (interrupted lines). Estimated groups' proportions are given in %. (b) to (e) AIC, BIC and L (b); APPA, mismatch and SD (c); OCC (d) and percentage of individuals estimated to be assigned to the smallest group (e). Twenty models for frequency of drinking wine per week were fit (x-axis). Shaded vertical bar indicates the selected k (here $k = 6$).

observed values. The 95% confidence limits of the estimated trajectories are given by the shaded areas. Estimated latent classes' proportions (in %) are also displayed.

5.1.2 F-CAP results

For the display of the F-CAPs, models with $k = 1$–10 were fitted. The results presented in Figure 1(b) to (e) exemplify a relatively unambiguous case scenario of class enumeration. Note the elbow behaviour of the BIC curve when $k = 3$ (green shaded column), after which not only BIC but also other criteria deteriorate. The choice of k was thus relatively straightforward.

5.2 F-CAP – The LEGO study

LEGO was a prospective cohort study conducted in the Netherlands analysing the relation between alcohol and cardiovascular disease. It started in 1996 and ran for five years. Several studies have already made use of the LEGO data before[65–68] but the analysis with GBTM is unprecedented.

5.2.1 Dataset

For the current purpose, only the outcome frequency of drinking wine per week (averaged and log transformed) is considered. Alcohol consumption of 2382 individuals (1211 men; 1171 women) with a mean age of 55.5 years was obtained.

Given the large sample size, k was varied between 1 and 20. The model was selected for which $k = 6$, with cubic polynomial. The final trajectories plot is displayed in Figure 2a.

5.2.2 F-CAP results

Contrary to the previous case, the BIC (2b) keeps improving with larger k, showing an asymptotic behaviour common for large sample sizes. Such steady improvement is a well-known problem of this criterion.[60] OCC (2d) behaves similarly with no clear elbow-bend discernible, whereas APPA (2c) and the groups' proportion (2e) would favour a $k < 7$. The final call fell for $k = 6$. This decision was a trade-off between parsimony, interpretability of the model, the distinctiveness of the trajectories, and, not least, the recommended limits for APPA, OCC and π_j as indicated by the horizontal lines on the figures above.

5.3 F-CAP – Simulated data

For the simulation, six case scenarios were considered, depicting varying degrees of latent class separation. The model's assumption of uncorrelated repeated measures was violated by introducing random effects of different magnitudes. The true number of latent trajectories ($k = 4$), the sample size ($N = 600$), the mixture proportions ($p = 0.25$ each class) and the polynomial fixed effects capturing the average trajectories trends (stable, linearly increasing, quadratic and cubic) were kept constant. Error variances and random effects variances were changed but kept homogeneous among latent classes.

Results of the simulation are presented in Figures 3 and 4. The partial F-CAPs (BIC) are accompanied by the spaghetti plot of a random sample together with the estimated trajectories plot for the model, whose k yielded the best BIC. In case model parameters were not estimable for the selected k, a smaller one was used for the trajectories plot.

Figure 3(a) and (b) displays cases of well-separated classes. Class recovery was unequivocal in the data without random effects. Within-class variability was due to (independent) errors only with either small or large error variances (Figure 3(a) and (b), respectively). Class recovery was less clear in the case depicted in Figure 3(c) (random intercept added). An inspection of the trajectory plot suggests over-extraction. The overall picture of temporal pattern, compared to the previous case is not changed. Additional trajectories seemed to have been formed by partitioning of those identified with the smaller k. Despite the plateauing of the BIC function in Figure 3(c) (and Figure 4(a)), an elbow bend is discernible in the vicinities of $k = 4$. This stands in contrast to the cases in Figure 4(b) and (c) (with random intercept and random slope) where the BIC curves are smoother. These exemplify a more challenging class recovery. As visible in the spaghetti plots, class separation is less evident due to larger between-subjects variability within classes. The trajectories plot in Figure 4(a) resembles that of Figure 3(c), with close, parallel trajectories, indicating that over-extraction may be at play.

The difference between Figure 4(b) and (c) is the presence of a strong correlation between random effects (Figure 4(c)). Although too many classes were also recovered in these examples, the distinct trajectories' shapes

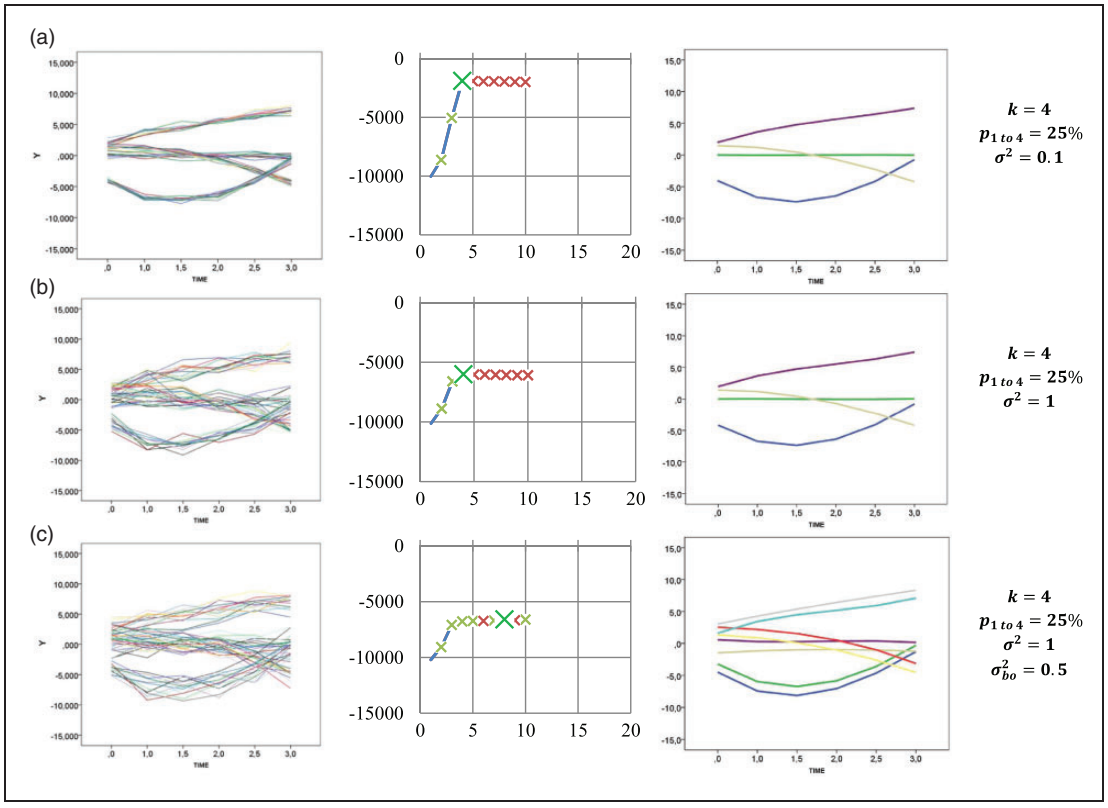

Figure 3. (a) to (c) Spaghetti plot of a random sample of simulated data (left column), F-CAP of information-based criteria for models with up to 10 latent classes (middle column) and estimated trajectories plot the *k* with the best BIC (right column). Parameters for data generation are specified.

beg the question whether these classes can be interpreted in a meaningful way. In clinical and epidemiological settings, the link of extreme onset values (large or small intercepts) to steeper trajectories' slopes warrants a closer look when validating the model.

6 Discussion

The current state of knowledge regarding class enumeration is yet equivocal. The consensus is, and remains that several indices should be used, keeping track simultaneously of practicalities and of the theoretical plausibility of the model. Model-based clustering is, therefore, not to be handled lightly. As Ialongo properly put it, analysing data with latent trajectories model is a daunting endeavour, as 'a scientifically strong case for the existence of trajectories needs to be built'.[69] GBTM requires from researchers to go the extra length of providing a sound motivation for its application and validating the identified clusters.

A newcomer to GBTM should refrain from following a mechanistic, tool-box analysis plan for model selection. Recipe-like model-selection procedures preclude users from recognising underlying problems or difficulties they may encounter, perpetuating poor understanding or sloppy applications. The lengthy, theory substantiated model selection can be neither circumvented, nor shortened by focusing on statistical criteria only. Encompassing graphical displays like F-CAP are meant to assist in the course of action, not to replace it. It should be noted that, after settling for the number of latent trajectories, users still have to determine the order of the trajectories' polynomials in the standard operating procedure, i.e. stepwise elimination of non-significant polynomial higher orders, one model at a time. Moreover, the criteria displayed in the F-CAP are not exhaustive. Others, like entropy-based statistics and the bootstrapping likelihood ratio (BLR) tests have been used. However, simulation has shown that the information-based BIC, despite shortcomings, fares comparatively well in recovering the true number of latent classes. Because BIC is computationally less intensive than the BLR test, it is recommended for this initial stage of model exploration.[63]

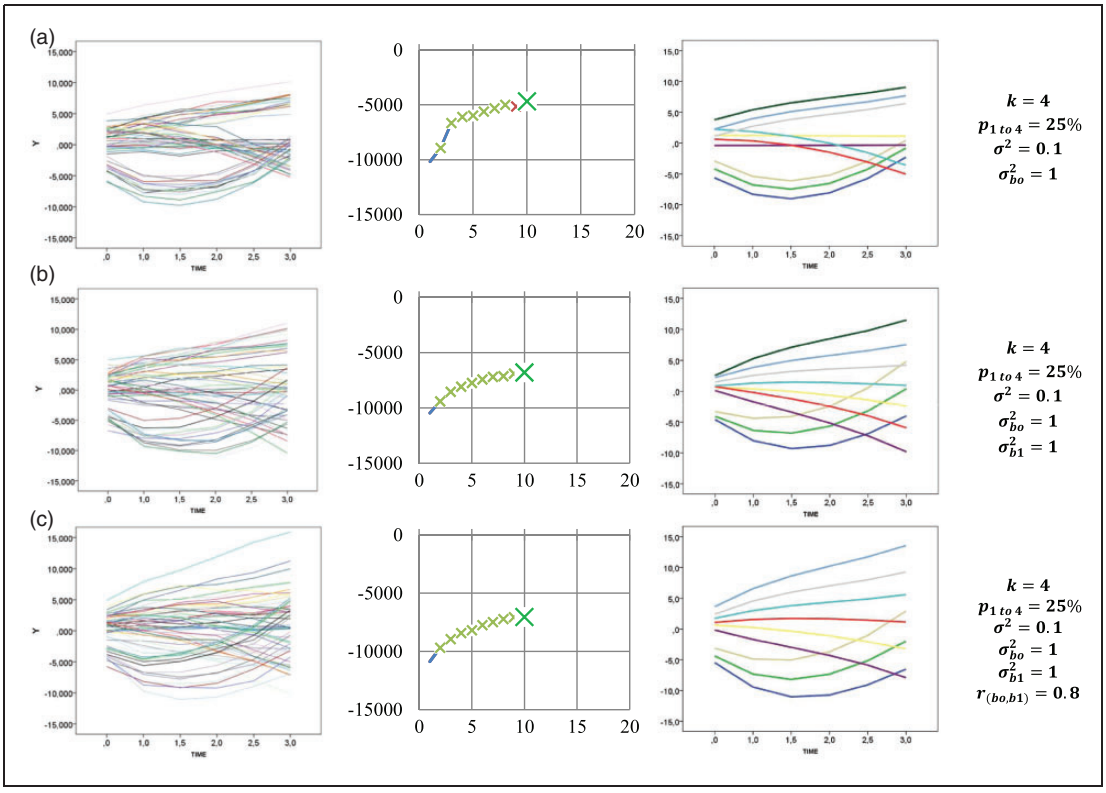

Figure 4. (a) to (c) Spaghetti plot of a random sample of simulated data (left column), F-CAP of information-based criteria for models with up to 10 latent classes (middle column) and estimated trajectories plot for the *k* with the best BIC (right column). Parameters for data generation are specified.

Amidst disputes, the plea for a cautionary use has been put forward by both enthusiasts and critics of finite mixture models.[69,70] Given GBTM's momentum, a cautionary tone may be an insufficient strategy to halt the expansion of a plug-and-chug approach for model selection. More than only caution, practitioners need enabling tools. In this sense, the F-CAP may play an eye-opening role, allowing them to recognise more easily what is going on with the data. As an automated visual graph, F-CAP provides an overview of fit criteria on a single page. It is important to emphasise that the F-CAPs do not automatically select the most appropriate model but considerably ease the model assessment procedure. For instance, in the Montreal data, the final decision was rather straightforward and all criteria reinforced each other. For the LEGO data the overall picture was less clear-cut. The indices were in disagreement and information based criteria were inconclusive. Both situations were mirrored in the illustrative simulations. Though tentative, the findings confirmed that the behaviours of fit indices in the F-CAPs were different between the plain versus ambiguous class-recovery scenarios. In the latter case, over-extraction emerged as a problem. Theoretical input and subsequent class validation is therefore of paramount importance, an issue not addressed here. In this respect, it should be underlined that, despite validation, latent trajectories should not be reified, i.e. considered to be real entities. Warnings notwithstanding, the 'fallacy of reification' is widespread among GBTM users. Hence, a new reminder is warranted. The observed latent groups are statistical, and to a certain extent also heuristic classifications, and as such inherently moot. To counteract the reification fallacy, the developmental trajectories have been lately referred to as 'latent strata' or alternatively 'distinctive features in the data' (Nagin DS 2015, personal communication).

Lastly, the available R-code was written to run unadjusted models with *proc traj*. The procedure, however, allows for adjustment to time-fixed and varying covariates and to model joint trajectories.

The topic of model selection in GBTM is far from being settled. Class enumeration is not the only aspect not yet fully grasped. In this sense, a cautionary attitude is never misplaced, nor is creating means that enhance the researchers' awareness and promote a more reflective involvement in the modelling process. Currently, such GBTM tools are yet scarce, but some auxiliary spin-offs have been put forward. Elsensohn et al.,[71] for instance, suggested graphical methods to assess the residuals distributional assumption for GBTM, which, if

violated, may have a substantial impact on the final model. Shah et al.[72] came up with an entropy discrimination index to assess the quality of group assignments. While we await a final resolution regarding GBTM's disputes (which may never come), such aiding tools are bound to foster its more judicious application.

Acknowledgements

The authors are greatly indebted to Prof. Daniel Nagin and the reviewers for their constructive suggestions to improve the manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

i. GMM stands for growth mixture model, another finite mixture approach, the importance of which is equally increasing.[57–59]
ii. Classify–analyze strategy: The widespread approach of first establishing the presence of latent classes via a finite mixture procedure, classifying individuals into these classes, based on the maximum posterior probability, and treating class membership as known in a subsequent analysis (i.e. deterministic instead of probabilistic).

References

1. Nagin DS and Tremblay RE. Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency. *Child Dev* 1999; **70**: 1181–1196.
2. Nagin DS. *Group-based modeling of development*. Cambridge, MA: Harvard University Press, 2005.
3. Jones BL, Nagin DS and Roeder K. A SAS procedure based on mixture models for estimating developmental trajectories. *Sociol Methods Res* 2001; **29**: 374–393.
4. Jones BL and Nagin DS. Advances in group-based trajectory modeling and an SAS procedure for estimating them. *Sociol Methods Res* 2007; **35**: 542–571.
5. Jones BL and Nagin DS. A note on a Stata plugin for estimating group-based trajectory models. *Sociol Methods Res* 2013; **42**: 608–613.
6. Osgood DW. Making sense of crime and the life course. *Ann Am Acad Pol Soc Sci* 2005; **602**: 196–211.
7. Piquero AR. Taking stock of developmental trajectories of criminal activity over the life course. In: AM Liberman (ed.) Taking stock of developmental trajectories of criminal activity over the life course. *The long view of crime: A synthesis of longitudinal research*, pp.23–78, 2008.
8. Saunders JM. Understanding random effects in group-based trajectory modeling: an application of Moffitt's developmental taxonomy. *J Drug Issues* 2010; **40**: 195–220.
9. Nagin DS and Odgers CL. Group-based trajectory modeling (nearly) two decades later. *J Quant Criminol* 2010; **26**: 445–453.
10. Bosick SJ. Crime and the transition to adulthood: A person-centered approach. *Crime Delinq [Internet] 2012*. Available from: http://cad.sagepub.com/content/early/2012/11/09/0011128712461598.full.pdf+html.
11. Laub JH, Nagin DS and Sampson RJ. Trajectories of change in criminal offending: good marriages and the desistance process. *Am Sociol Rev* 1998; **63**: 225–238.
12. Nagin DS and Odgers CL. Group-based trajectory modeling in clinical research. *Annu Rev Clin Psychol* 2010; **6**: 109–138.
13. Nagin DS. Group-based trajectory modeling: an overview. *Ann Nutr Metab* 2014; **65**: 205–210.
14. Liu S, Jones RN and Glymour MM. Implications of lifecourse epidemiology for research on determinants of adult disease. *Public Health Rev* 2010; **32**: 489–511.
15. Tilling K, Howe LD and Ben-Shlomo Y. Commentary: Methods for analysing life course influences on health—untangling complex exposures. *Int J Epidemiol* 2011; **40**: 250–252.
16. Sandman C, Cordova C, Davis E, et al. Patterns of fetal heart rate response at ~30 weeks gestation predict size at birth. *J Dev Orig Health Dis* 2011; **2**: 212–217.
17. Ng J, Barrett LM, Wong A, et al. The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities. *Genome Biol* 2012; **13**: 246.

18. Tu Y-K, Tilling K, Sterne JA, et al. A critical evaluation of statistical approaches to examining the role of growth trajectories in the developmental origins of health and disease. *Int J Epidemiol* 2013; **43**: 1664–1665.
19. Zheng H, Tumin D and Qian Z. Obesity and mortality risk: new findings from body mass index trajectories. *Am J Epidemiol* 2013; **178**: 1591–1599.
20. Ziyab AH, Karmaus W, Kurukulaaratchy RJ, et al. Developmental trajectories of body mass index from infancy to 18 years of age: prenatal determinants and health consequences. *J Epidemiol Community Health* 2014; **68**: 934–941.
21. Carter MA, Dubois L, Tremblay MS, et al. Trajectories of childhood weight gain: the relative importance of local environment versus individual social and early life factors. *PLoS One* 2012; **7**: e47065.
22. Van Ryzin MJ, Chatham M, Kryzer E, et al. Identifying atypical cortisol patterns in young children: the benefits of group-based trajectory modeling. *Psychoneuroendocrinology* 2009; **34**: 50–61.
23. Ferro MA, Camfield CS, Levin SD, et al. Trajectories of health-related quality of life in children with epilepsy: a cohort study. *Epilepsia* 2013; **54**: 1889–1897.
24. Maddox TM, Ross C, Tavel HM, et al. Blood pressure trajectories and associations with treatment intensification, medication adherence, and outcomes among newly diagnosed coronary artery disease patients. *Circ Cardiovasc Qual Outcomes* 2010; **3**: 347–357.
25. Franklin JM, Shrank WH, Pakes J, et al. Group-based trajectory models: a new approach to classifying and predicting long-term medication adherence. *Med Care* 2013; **51**: 789–796.
26. Willke RJ, Zheng Z, Subedi P, et al. From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC Med Res Methodol* 2012; **12**: 185.
27. Levine SZ and Leucht S. Treatment response heterogeneity in the predominant negative symptoms of schizophrenia: analysis of amisulpride vs placebo in three clinical trials. *Schizophr Res* 2014; **156**: 107–114.
28. Hsu H-C and Jones BL. Multiple trajectories of successful aging of older and younger cohorts. *Gerontologist* 2012; **52**: 843–856.
29. Chen CF, Lee WC, Yang HI, et al. Changes in serum levels of HBV DNA and alanine aminotransferase determine risk for hepatocellular carcinoma. *Gastroenterology* 2011; **141**: 1240–1248e2.
30. Zaslavsky O, Cochrane BB, Herting JR, et al. Application of person-centered analytic methodology in longitudinal research: Exemplars from the women's health initiative clinical trial data. *Res Nurs Health* 2014; **37**: 53–64.
31. M'Bailara K, Cosnefroy O, Vieta E, et al. Group-based trajectory modeling: a novel approach to examining symptom trajectories in acute bipolar episodes. *J Affect Disord* 2013; **145**: 36–41.
32. Huang DY, Lanza HI, Wright-Volel K, et al. Developmental trajectories of childhood obesity and risk behaviors in adolescence. *J Adolesc* 2013; **36**: 139–148.
33. Østbye T, Malhotra R and Landerman LR. Body mass trajectories through adulthood: results from the National Longitudinal Survey of Youth 1979 Cohort (1981–2006). *Int J Epidemiol* 2011; **40**: 240–250.
34. Soicher JE, Mayo NE, Gauvin L, et al. Trajectories of endurance activity following pulmonary rehabilitation in COPD patients. *Eur Respir J* 2012; **39**: 272–278.
35. Von Ehrenstein OS, Mikolajczyk RT and Zhang J. Timing and trajectories of fetal growth related to cognitive development in childhood. *Am J Epidemiol* 2009; **170**: 1388–1395.
36. Berger RP, Bazaco MC, Wagner AK, et al. Trajectory analysis of serum biomarker concentrations facilitates outcome prediction after pediatric traumatic and hypoxemic brain injury. *Dev Neurosci* 2011; **32**: 396–405.
37. Smith ORF, Kupper N, De Jonge P, et al. Distinct trajectories of fatigue in chronic heart failure and their association with prognosis. *Eur J Heart Fail* 2010; **12**: 841–848.
38. Smith ORF, Kupper N, Denollet J, et al. Vital exhaustion and cardiovascular prognosis in myocardial infarction and heart failure: predictive power of different trajectories. *Psychol Med* 2011; **41**: 731–738.
39. Niyonkuru C, Wagner AK, Ozawa H, et al. Group-based trajectory analysis applications for prognostic biomarker model development in severe TBI: a practical example. *J Neurotrauma* 2013; **30**: 938–945.
40. Schmitz N, Gariépy G, Smith KJ, et al. Trajectories of self-rated health in people with diabetes: associations with functioning in a prospective community sample. *PLoS One* 2013; **8**: e83088.
41. Bauer DJ and Curran PJ. Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychol Methods* 2003; **8**: 338–363.
42. Sampson RJ, Laub JH and Eggleston EP. On the robustness and validity of groups. *J Quant Criminol* 2004; **20**: 37–42.
43. Nagin DS and Tremblay RE. Further reflections on modeling and analyzing developmental trajectories: a response to Maughan and Raudenbush. *Ann Am Acad Pol Soc Sci* 2005; **602**: 145–154.
44. Nagin DS and Tremblay RE. From seduction to passion: a response to Sampson and Laub. *Criminology* 2005; **43**: 915.
45. Raudenbush SW. How do we study "what happens next"? *Ann Am Acad Pol Soc Sci* 2005; **602**: 131–144.
46. Sampson RJ and Laub JH. Seductions of method: rejoinder to Nagin and Tremblay's developments trajectory groups: fact or fiction. *Criminology* 2005; **43**: 905.
47. Bauer DJ and Reyes HLM. Modeling variability in individual development: differences of degree or kind? *Child Dev Perspect* 2010; **4**: 114–122.
48. Skardhamar T. Distinguishing facts and artifacts in group-based modeling. *Criminology* 2010; **48**: 295–320.

49. Brame R, Paternoster R and Piquero AR. Thoughts on the analysis of group-based developmental trajectories in criminology. *Justice Q* 2012; **29**: 469–490.
50. Eggleston EP, Laub JH and Sampson RJ. Methodological sensitivities to latent class analysis of long-term criminal trajectories. *J Quant Criminol* 2004; **20**: 1–26.
51. Brame R, Nagin DS and Wasserman L. Exploring some analytical characteristics of finite mixture models. *J Quant Criminol* 2006; **22**: 31–59.
52. Loughran T and Nagin DS. Finite sample effects in group-based trajectory models. *Sociol Methods Res* 2006; **35**: 250–278.
53. Sterba SK and Bauer DJ. Statistically evaluating person-oriented principles revisited. *Dev Psychopathol* 2010; **22**: 287–294.
54. Morin AJ, Maïano C, Nagengast B, et al. General growth mixture analysis of adolescents' developmental trajectories of anxiety: the impact of untested invariance assumptions on substantive interpretations. *Struct Equ Model Multidisc J* 2011; **18**: 613–648.
55. Sterba SK, Baldasaro RE and Bauer DJ. Factors affecting the adequacy and preferability of semiparametric groups-based approximations of continuous growth trajectories. *Multivariate Behav Res* 2012; **47**: 590–634.
56. Gilthorpe M, Dahly D, Tu Y-K, et al. Challenges in modelling the random structure correctly in growth mixture models and the impact this has on model mixtures. *J Dev Orig Health Dis* 2014; **5**: 197–205.
57. Muthén B. The potential of growth mixture modelling. *Infant Child Dev* 2006; **15**: 623–625.
58. Ram N and Grimm KJ. Methods and measures: growth mixture modeling: a method for identifying differences in longitudinal change among unobserved groups. *Int J Behav Dev* 2009; **33**: 565–576.
59. Pickles A and Croudace T. Latent mixture models for multivariate and longitudinal outcomes. *Stat Methods Med Res* 2009; **19**: 271–289.
60. Erosheva EA, Matsueda RL and Telesca D. Breaking bad: two decades of life-course data analysis in criminology, developmental psychology, and beyond. *Ann Rev Stat Appl* 2014; **1**: 301–332.
61. McIntosh CN. Pitfalls in subgroup analysis based on growth mixture models: a commentary on van Leeuwen et al. (2012). *Qual Life Res* 2013; **22**: 2625–2629.
62. Andruff H, Carraro N, Thompson A, et al. Latent class growth modelling: a tutorial. *Tutor Quant Methods Psychol* 2009; **5**: 11–24.
63. Nylund KL, Asparouhov T and Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct Equ Model Multidisc J* 2007; **14**: 535–569.
64. Tofighi D and Enders CK. Identifying the correct number of classes in growth mixture models. In: Hancock GR and Samuelsen KM (eds) *Advances in latent variable mixture models*. Greenwich, CT: Information Age Publishing, Inc, 2008, pp.317–341.
65. Friesema I, Veenstra MY, Zwietering PJ, et al. Measurement of lifetime alcohol intake: utility of a self-administered questionnaire. *Am J Epidemiol* 2004; **159**: 809–817.
66. Friesema IH, Zwietering PJ, Veenstra MY, et al. The effect of alcohol intake on cardiovascular disease and mortality disappeared after taking lifetime drinking and covariates into account. *Alcohol Clin Exp Res* 2008; **32**: 645–651.
67. Veenstra M, Friesema I, Zwietering P, et al. Lower prevalence of heart disease but higher mortality risk during follow-up was found among nonrespondents to a cohort study. *J Clin Epidemiol* 2006; **59**: 412–420.
68. Veenstra MY, Lemmens PH, Friesema IH, et al. Coping style mediates impact of stress on alcohol use: a prospective population-based study. *Addiction* 2007; **102**: 1890–1898.
69. Ialongo N. Steps substantive researchers can take to build a scientifically strong case for the existence of trajectory groups. *Dev Psychopathol* 2010; **22**: 273–275.
70. Twisk J and Hoekstra T. Classifying developmental trajectories over time should be done with great caution: a comparison between methods. *J Clin Epidemiol* 2012; **65**: 1078–1087.
71. Elsensohn M-H, Klich A, Ecochard R, et al. A graphical method to assess distribution assumption in group-based trajectory models. *Stat Methods Med Res* [Internet] 2013. Available from: http://smm.sagepub.com/content/early/2013/02/14/0962280213475643.full.pdf + html.
72. Shah NH, Hipwell AE, Stepp SD, et al. Measures of discrimination for latent group-based trajectory models. *J Appl Stat* 2015; **42**: 1–11.