Published in final edited form as:

Stat Med. 2017 February 20; 36(4): 671–686. doi:10.1002/sim.7152.

Targeted Use of Growth Mixture Modeling: A Learning Perspective

Booil Jo,
Stanford University

Robert L. Findling,
Johns Hopkins University

Chen-Pin Wang,
University of Texas Health Science Center

Trevor J. Hastie,
Stanford University

Eric A. Youngstrom,
University of North Carolina, Chapel Hill

L. Eugene Arnold,
Ohio State University

Mary A. Fristad, and
Ohio State University

Sarah McCue Horwitz
New York University

Abstract

From the statistical learning perspective, this paper shows a new direction for the use of growth mixture modeling (GMM), a method of identifying latent subpopulations that manifest heterogeneous outcome trajectories. In the proposed approach, we utilize the benefits of the conventional use of GMM for the purpose of generating potential candidate models based on empirical model fitting, which can be viewed as unsupervised learning. We then evaluate candidate GMM models on the basis of a direct measure of success; how well the trajectory types are predicted by clinically and demographically relevant baseline features, which can be viewed as supervised learning. We examine the proposed approach focusing on a particular utility of latent trajectory classes, as outcomes that can be used as valid prediction targets in clinical prognostic models. Our approach is illustrated using data from the Longitudinal Assessment of Manic Symptoms (LAMS) study.

Keywords

growth mixture modeling; unsupervised learning; latent trajectory class; early prediction; supervised learning; sensitivity; specificity

1 Introduction

The use of growth mixture modeling (GMM) [1–3] has been growing in various fields [4–15] as a flexible way of identifying latent subpopulations that manifest heterogeneous outcome trajectories. The main interest in GMM has been meaningful interpretation of longitudinal heterogeneity in the target population. Naturally, recovering true models became a central issue as different versions of trajectory class solutions will lead to different interpretations and potentially different policy and clinical implications.

As in any exploratory modeling involving latent classes, identification of trajectory classes can be affected by various factors such as sample size, parametric assumptions, model specification, and presence or absence of auxiliary variables such as predictors, concurrent outcomes, and distal outcomes of the trajectory classes [16–22]. As we try to identify trajectory classes considering the aforementioned components, complexities and variations in model specifications quickly increase along with computational difficulties. In the field of machine learning, this type of modeling strategy is categorized as unsupervised learning, which is generally considered a challenging task given the lack of direct measures of success [23–24].

In contrast, identifying models that perform well in terms of targeted utilities is a relatively straightforward task. In the field of machine learning, this type of modeling strategy is categorized as supervised learning, where the candidate models are evaluated in terms of direct measures of success such as prediction or classification accuracy. Shifting our focus from recovery of true models to specific utilities opens up new possibilities in terms of how we evaluate GMM models and how the GMM results can be used in clinical research and practice. Whereas recovering true models and interpreting them is important in improving our understanding of the population heterogeneity, being able to accurately predict or classify individual level outcomes is important in improving the quality in personalized treatment and intervention. This may seem like a subtle difference, although it makes considerable differences in terms of how candidate models are evaluated and utilized. We intend to examine these new possibilities focusing on a specific utility of GMM as a way of producing valid prediction targets. In this context, identifying models that capture individual heterogeneity without overfitting is the goal of model selection, which is consistent with the goal of model selection in supervised learning.

Specifically, we utilize the benefits of the conventional use of GMM for the purpose of generating potential candidate models based on empirical model fitting, which can be viewed as unsupervised learning. We then propose to evaluate candidate GMM models on the basis of a direct measure of success; how well they are predicted by clinically and demographically relevant baseline features (antecedent validators), which can be viewed as supervised learning. Establishing the validity of prediction targets is a challenging, but critical process to ensure that they are worthy of predicting and clinically meaningful. Assessing the validity of latent trajectory classes based on their relationships with other variables is not new [19, 22]. Latent trajectory classes may have various roles, for example, as an outcome, as a predictor of future outcomes, or as a key component in complex theoretical models. Embedding these features in GMM may support the validity of the

trajectory class solutions and lead to fuller interpretation of longitudinal heterogeneity in the target population. However, the main interest still has been in interpretation and therefore little attention has been given to the possibility of utilizing these features to directly evaluate the performance of GMM solutions at the individual level.

2 Motivating Example: The LAMS Study

The LAMS study [6, 25, 26] was designed to investigate phenomenology, development of bipolar disorder and related conditions, and to establish predictors of functional outcomes in children with elevated manic symptoms. Children aged 6 to 12 years at screening and their parents were recruited from nine outpatient clinics associated with four university affiliated LAMS sites. A primary outcome in the LAMS study is symptoms of mania as measured by the Parent General Behavior Inventory 10-item Mania Form (PGBI-10M) [27]. Parents completed the PGBI-10M at five assessments during the two-year period (at baseline, 6, 12, 18, and 24 months). Among 707 children enrolled in the study, we included 682 cases in our analyses, excluding ineligible cases and four cases with missing outcome information at all assessment points. Table 1 shows sample statistics of the PGBI-10M outcome and a small set of baseline variables used in our analyses.

Whether there are heterogeneous trajectory types of manic symptoms is a primary question in the LAMS study [6]. As in most GMM applications, identifying true trajectory classes is expected to be challenging as it can be affected by various factors such as sample size, model specification, and auxiliary information. Further, various practical questions arise surrounding the manic symptom trajectory class membership, such as whether we can predict the trajectory type early on (e.g., using baseline characteristics), whether the trajectory type is associated with relevant concurrent outcomes (e.g., depression, bipolar diagnosis), and whether we can predict distal outcomes (e.g., future delinquent behaviors) using the PGBI-10M trajectory type. As we try to identify true trajectory classes and also answer these questions, complexities and variations in model specifications quickly increase along with computational difficulties.

Evaluating and utilizing numerous trajectory class solutions is a challenging issue when our goal is interpretation of the population heterogeneity. However, when the goal is simpler and specific, such as predicting or classifying individual outcomes, how we select and use GMM solutions also becomes a simpler problem. One of the key interests in LAMS is in establishing prognostic models that accurately predict the type of manic symptom course early on. Ultimately, these models are intended for the use in clinical practice to improve personalized treatment based on early prediction. Separating out a large proportion of children who would remain non-problematic is especially critical as it is the first step towards efficient clinical practice. If we can separate them out early on, clinicians will be able to treat individuals in the elevated risk trajectory classes with confidence, having ruled out those in the low risk trajectory class who may experience iatrogenic adverse effects from antimanic treatments. In this context, we can narrow the goal of GMM to formulating valid prediction targets with a specific contrast of trajectory classes (low risk vs. the rest). With this simple classification structure, validating the formulated prediction targets in terms of their relationship with relevant baseline, concurrent, or distal measures becomes a feasible

task. In particular, we will focus on assessing the quality of prediction by clinically and demographically relevant baseline characteristics (antecedent validators) given our interest in using trajectory types as outcomes.

3 Unsupervised Learning with GMM

Growth mixture modeling (GMM) is a flexible method of identifying latent subpopulations that manifest heterogeneous trajectory patterns [1–3]. For example, according to our previous investigation [6], qualitatively different patterns of manic symptom expression are apparent in LAMS over the two-year course of observation. In this situation, standard mixed effects modeling is unlikely to capture the longitudinal heterogeneity sufficiently well, necessitating consideration of multiple trajectory classes. GMM has been gaining popularity especially with accessible latent variable modeling software such as Mplus [28].

The resulting trajectory types from GMM may provide insights and useful summary information that cannot be directly obtained from observed data. Despite its potential utilities, GMM has been mostly used as an exploratory tool to generate theories, at least partly due to uncertainties surrounding model evaluation and selection. Identification of trajectory classes can be affected by various factors such as model specification and auxiliary information. As we explore various combinations of these components, we can easily end up with abundance of candidate models. With such variations and possible complexities, identifying and using latent classes is generally considered a challenging task. In machine learning, this type of exploratory modeling strategy is categorized as unsupervised learning given the lack of direct measures of success [23–24].

The same exploratory nature also makes GMM a convenient and effective tool that facilitates discovery of latent trajectory types. Using clinical thresholds would be a more conventional way of classification, although the strategy can be arbitrary and inefficient when classifying individuals based on repeated measures, in particular with substantial missing data. In LAMS, about half of the study sample have missing manic symptom data at one or more assessment points. GMM utilizes all available data and empirical model fitting, and therefore considerable gain in reliability and efficiency is expected. This is not a trivial advantage as it is directly related to better prediction quality. Statistically identified trajectory classes from GMM can readily serve as classification categories, which is another convenient feature we will utilize in formulating prediction targets.

Here we briefly describe the GMM procedures we used for the purpose of discovery of latent trajectory class solutions (unsupervised learning). To focus on illustrating the proposed approach using the LAMS data, we limited the range of possible models that will be validated in terms of the prediction quality (supervised learning). We used a simple quadratic growth specification with restricted variance/covariance structures. However, various alternative model specifications are possible, which will lead to a much wider array of GMM solutions. With our limited setting, we obtained 43 models that reached normal convergence using the LAMS data.

3.1 GMM without Covariates

We first conducted GMM without adjusting for any baseline covariates. The outcome Y (PGBI-10M) for individual i ($i = 1, 2, \ldots, N$) at time point t ($t = 1, 2, \ldots, T$) conditioned on trajectory class $C_i = j$ is expressed as

$$Y_{it}|C_i=j=\eta_{1ij}+\eta_{2ij}\,S_t+\eta_{3ij}\,S_t^2+\varepsilon_{ijt}, \quad (3.1)$$

$$\eta_{1ij}=\eta_{1j}+\zeta_{1ij}, \quad (3.2)$$

$$\eta_{2ij}=\eta_{2j}+\zeta_{2ij}, \quad (3.3)$$

$$\eta_{3ij}=\eta_{3j}+\zeta_{3ij}, \quad (3.4)$$

where there are J possible trajectory classes ($j = 1, 2, \ldots, J$). In line with LAMS, a quadratic growth model was chosen to capture potentially nonlinear patterns across five assessments (baseline, 6, 12, 18, and 24 month). This model includes three mean growth parameters: the initial status (η_{1j}), linear growth (η_{2j}), and quadratic growth (η_{3j}) for trajectory class j. The time score S_t reflects linear and S_t^2 quadratic growth. The measurement errors $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}, \cdots, \varepsilon_{ijT})$ are assumed to be normally distributed with $\boldsymbol{\varepsilon}_{ij} \sim MN(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$, where the associated variances are allowed to vary over time. The random effects (ζ_{1ij}, ζ_{2ij}, ζ_{3ij}) associated with growth parameters are also assumed to be normally distributed as $MN(0, \boldsymbol{\Sigma}_{\zeta})$. To maintain identifiability in models with larger numbers of classes, we imposed restrictions that $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is diagonal and that $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ and $\boldsymbol{\Sigma}_{\zeta}$ do not vary across trajectory classes. We used four variations of $\boldsymbol{\Sigma}_{\zeta}$: models allowing for all three random effects, models allowing for random linear slope/intercept ($Var(\zeta_{3i}) = 0$), models allowing for random intercept only ($Var(\zeta_{2i}) = Var(\zeta_{3i}) = 0$), and models with no random effects ($Var(\zeta_{1i}) = Var(\zeta_{2i}) = Var(\zeta_{3i}) = 0$).

The probability of subject i belonging to a certain trajectory class j ($\boldsymbol{\pi}_{ij} = Pr(C_i = j)$) is expressed in terms of a multinomial logit model,

$$logit(\pi_{ij})=\beta_{0j} \quad (3.5)$$

for $j = 1, 2, \ldots, (J - 1)$, where $logit(\boldsymbol{\pi}_{ij}) = log(\boldsymbol{\pi}_{ij}/\boldsymbol{\pi}_{iJ})$.

3.2 GMM with Covariates

We also conducted GMM with baseline covariates as predictors of the trajectory class membership and as predictors of the growth parameters. The outcome Y (PGBI-10M) for

individual i ($i = 1, 2, \ldots, N$) at time point t ($t = 1, 2, \ldots, T$) conditioned on trajectory class j is now expressed as

$$Y_{it}|C_i=j=\eta_{1ij}+\eta_{2ij}\,S_t+\eta_{3ij}\,S_t^2+\varepsilon_{ijt}, \quad (3.6)$$

$$\eta_{1ij}=\eta_{1j}+\boldsymbol{\lambda}_1^{'}\,\boldsymbol{X}_i+\zeta_{1ij}, \quad (3.7)$$

$$\eta_{2ij}=\eta_{2j}+\boldsymbol{\lambda}_2^{'}\,\boldsymbol{X}_i+\zeta_{2ij}. \quad (3.8)$$

$$\eta_{3ij}=\eta_{3j}+\boldsymbol{\lambda}_3^{'}\,\boldsymbol{X}_i+\zeta_{3ij}, \quad (3.9)$$

where the relationship between the three growth factors and the vector of covariates $\boldsymbol{X}$ is captured by the vectors of regression coefficients $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, and $\boldsymbol{\lambda}_3$. These regression coefficients, in principle, can vary across trajectory classes ($j = 1, 2, \ldots, J$). We imposed the equality restrictions on these parameters to avoid serious convergence problems and to maintain identifiability in models with larger numbers of classes. However, in principle, one may choose to consider both sets of models with and without these restrictions.

The probability of subject i belonging to a certain trajectory class ($\boldsymbol{\pi}_{ij} = Pr(C_i = j)$) depends on the influence of covariates, and this association can vary by trajectory class. The multinomial logit model of $\boldsymbol{\pi}_{ij}$ conditioned on covariates subsumed in vector $\boldsymbol{X}_i$ is described as

$$logit(\pi_{ij}|\boldsymbol{X}_i)=\beta_{0j}+\boldsymbol{\beta}_{1j}^{'}\,\boldsymbol{X}_i, \quad (3.10)$$

for $j = 1, 2, \ldots, (J - 1)$, where $\boldsymbol{\beta}_{1j}$ is a vector of multinomial logit regression coefficients with dimension the same as the length of $\boldsymbol{X}_i$. In the LAMS application, we used the same set of baseline covariates (age, sex, Medicaid and CDRS-R) as the predictors of the growth parameters and the trajectory class membership.

Based on the model specifications described above, we conducted a series of GMM with varying numbers of classes. We calculated maximum likelihood (ML) estimates using the expectation maximization (EM) algorithm [29–32] implemented in the M*plus* program [28]. Within each type of $\boldsymbol{\Sigma}_\zeta$ restrictions, we increased the number of classes until the covariance matrices $\boldsymbol{\Sigma}_e$ and $\boldsymbol{\Sigma}_\zeta$ in any of the classes was not positive definite, any of the classes has less than ten individuals (using the most likely class membership), or the model could not be

identified. We used ample starting values to avoid potential convergence at local maxima (3000 for the initial and 300 for the final stage optimization in M*plus*).

From (3.6)–(3.10), the log-likelihood for the observed data $\{Y_i : i = 1, \cdots, N\}$ is

$$\sum_{i=1}^{N} log\{\sum_{j=1}^{J} \pi_{ij}(\boldsymbol{\beta}) f(Y_i|C_i=j, \boldsymbol{X}_i, \boldsymbol{\eta}_j, \boldsymbol{\lambda}, \sum_{\zeta}, \sum_{\varepsilon})\}, \tag{3.11}$$

where $\boldsymbol{\eta}_j = (\eta_{1j}, \eta_{2j}, \eta_{3j})$, $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3)$, $\sum_{j=1}^{J} \pi_{ij}(\beta)=1$ for $i = 1, \cdots, N$, and $\pi_{ij}(\boldsymbol{\beta}) = Pr(C_i = j|\boldsymbol{X}_i, \boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{11}, \cdots, \boldsymbol{\beta}_{0J}, \boldsymbol{\beta}_{1J})$ denotes the likelihood that Y_i arising from mixture component j given $\boldsymbol{X}_i$.

The log-likelihood of the complete-data $(Y_i, C_i : i = 1 \cdots, N)$ can be written as

$$LogL_C = \sum_{i=1}^{N} LogL_{Ci} = \sum_{i=1}^{N}\sum_{j=1}^{J} \{log[f(Y_i|C_i=j, \boldsymbol{X}_i, \boldsymbol{\eta}_j, \boldsymbol{\lambda}, \sum_{\zeta}, \sum_{\varepsilon})] + log[\pi_{ij}(\boldsymbol{\beta})]\}. \tag{3.12}$$

To maximize (3.11), the E step computes the expected values of the log-likelihood in (3.12) given observed data and the current parameter estimates $(\boldsymbol{\beta}^*, \boldsymbol{\eta}^*, \boldsymbol{\lambda}^*, \sum_{\zeta}^*, \sum_{\varepsilon}^*)$. Latent trajectory class C is considered as missing data in this step. That is, the E step computes

$$\sum_{i=1}^{N}\sum_{j=1}^{J} log[f(Y_i|C_i=j, \boldsymbol{X}_i, \boldsymbol{\eta}_j^*, \boldsymbol{\lambda}^*, \sum_{\zeta}^*, \sum_{\varepsilon}^*)]\, p_{ij}(\boldsymbol{\beta}^*, \boldsymbol{\eta}^*, \boldsymbol{\lambda}^*, \sum_{\zeta}^*, \sum_{\varepsilon}^*), \tag{3.13}$$

where $p_{ij}(\boldsymbol{\beta}^*, \boldsymbol{\eta}^*, \boldsymbol{\lambda}^*, \sum_{\zeta}^*, \sum_{\varepsilon}^*)$ is the posterior class probability of subject i belonging to class j conditioned on $(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}^*, \boldsymbol{\eta}_j^*, \boldsymbol{\lambda}^*, \sum_{\zeta}^*, \sum_{\varepsilon}^*)$ calculated as

$$p_{ij}(\boldsymbol{\beta}^*, \boldsymbol{\eta}^*, \boldsymbol{\lambda}^*, \sum_{\zeta}^*, \sum_{\varepsilon}^*) = \frac{\pi_{ij}(\boldsymbol{\beta}^*) f(Y_i|C_i=j, \boldsymbol{X}_i, \boldsymbol{\eta}_j^*, \boldsymbol{\lambda}^*, \sum_{\zeta}^*, \sum_{\varepsilon}^*)}{\sum_{j'=1}^{J} \pi_{ij'}(\boldsymbol{\beta}^*) f(Y_i|C_i=j', \boldsymbol{X}_i, \boldsymbol{\eta}_{j'}^*, \boldsymbol{\lambda}^*, \sum_{\zeta}^*, \sum_{\varepsilon}^*)}. \tag{3.14}$$

The M step computes the parameter estimates that maximize the quantity obtained from the E step. This procedure continues until it reaches the optimal status.

We monitored Bayesian information criteria (BIC; [33]) and conducted bootstrapped likelihood ratio test (BLRT; [34]). Using BIC or BLRT has been examined as a preferred method of model selection in GMM [21, 35]. Within each type of $\boldsymbol{\Sigma}_{\zeta}$ restrictions, best fitting

models were identified based on BIC and BLRT. In this study, we report BIC and BLRT results, although we do not use them for model evaluation and selection.

3.3 Formulating Practical Prediction Targets

In this paper, we focus on a specific utility of GMM as a way of constructing prediction targets based on empirical model fitting. Ultimately, these GMM-based prediction targets are intended for the use as reliable and valid outcomes in clinical prognostic models. Classifying patients is common in clinical practice and research, often implemented by applying fixed clinical thresholds. This is necessary and practical even when the outcome is dimensional as most treatment, prevention, and intervention decisions are made in a categorical manner (e.g., surgery or not, prescription or not). In this context, GMM has a great potential as a tool for patient classification. To improve the practicality of GMM-based classification, we propose to formulate prediction targets with a specific contrast of trajectory classes. This is an important consideration as building prognostic models with good predictive power and validating them is likely more challenging when aiming to classify patients into multiple categories. Further, using prediction targets with fewer categories makes the prognostic models easier to understand and use in clinical practice.

In LAMS, separating out low risk children is of great clinical importance. We visually inspected all candidate GMM models to examine if this clinical intention can be aligned with the results of empirical clustering. All GMM models, with and without covariates, consistently exhibited trajectory classes that start with the PGBI-10M score of around 12 or lower and then gradually decrease. From the clinical perspective, these classes are clearly the least problematic. This interpretation is also supported by a previously suggested clinical threshold that sets PGBI-10M ≥ 12 as elevated in manic symptoms [25–27]. One way to categorize trajectory classes would be to use a fixed threshold (e.g., PGBI-10M < 12), which will provide validation results more in line with those from the conventional method of using observed measures. In this study, we instead categorize only the bottom class with the lowest mean trajectory as low risk. This approach allows the level of PGBI-10M to vary across different GMM models, and therefore will inform us how the prediction quality varies depending on how conservatively individuals are classified as low risk. Identifying the bottom trajectory class was straightforward in the LAMS example as all GMM models had a trajectory class that has the lowest estimated mean PGBI-10M across all five assessments. In different applications, different rules may need to be formulated depending on the clinical purpose of classification.

Let us designate the J^{th} class as the lowest risk trajectory class in each GMM model with J classes. From (3.14), the posterior probability of belonging to the low risk class (class J) for person i is $\hat{p}_{iJ}$, which denotes p_{iJ} evaluated at the maximum likelihood estimates of model parameters ($\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}_{\zeta}$, $\boldsymbol{\Sigma}_{\varepsilon}$). The posterior probability of belonging to the elevated risk category is simply calculated as $1-\hat{p}_{iJ}=\sum_{j=1}^{J-1}\hat{p}_{ij}$.

Then, in each GMM model, we can define the risk status of individual i as

$$L_i = \begin{cases} 1\,(\text{low risk}) & \text{if } \hat{p}_{iJ} > 0.5 \\ 0\,(\text{elevated risk}) & \text{otherwise.} \end{cases} \quad (3.15)$$

It is a common practice to ignore uncertainties when classifying individuals based on observed measures. Similarly, when using GMM, we can classify individuals into the most likely category based on their posterior class probabilities as is done in (3.15). Another way of utilizing posterior class probabilities is to create multiple versions of L_i via multiple independent pseudoclass draws [37–38], which makes it possible to reflect uncertainties in classification. That is, in each GMM model, using the posterior class probability of belonging to the low risk trajectory class ($\hat{p}_{iJ}$), the risk status of individual i in the r^{th} pseudoclass draw is defined as

$$L_i^{(r)} = \begin{cases} 1\,(\text{low risk}) & \text{if } \; 0 \le u^{(r)} \le \hat{p}_{iJ} \\ 0\,(\text{elevated risk}) & \text{if } \; \hat{p}_{iJ} < u^{(r)} \le 1, \end{cases} \quad (3.16)$$

where $u^{(r)}$ denotes the realization of drawing a random sample from a uniform (0, 1) distribution for $r = 1, 2, 3, \ldots, R$, and R is the total number of psudoclass draws. We used 20 draws in the LAMS application, resulting in 20 versions of L_i for person i based on each GMM model. For each GMM model, the quality of each version of the prediction target (i.e., $L_1^{(r)}, \cdots, L_N^{(r)}$) is assessed, and the overall quality of the prediction is obtained by averaging the results across multiple versions of pseudoclass draws.

4 Supervised Learning for Validation

The validity of GMM-based prediction targets is supported by the use of empirically derived trajectory classes and clinical insights. However, this initial validation process, embedded in formulation of prediction targets, may not necessarily narrow the range of prediction targets. Additional validation efforts will further support the validity of formulated prediction targets and help narrow our choice. One popular way of model selection in GMM is to assess the model fit, which significantly reduces the number of candidate models. For example, 11 models are selected by BIC and/or BLRT in the LAMS application. However, it is unclear whether this approach is ideal when we intend to use trajectory classes as outcomes to be predicted at the individual level. Further, the formulated prediction targets (e.g., 2 types in LAMS: low and elevated risk trajectory types) are not necessarily the same as the GMM-generated trajectory classes (e.g., 6 classes). In this study, instead of screening GMM models based on fit measures, we directly evaluate formulated prediction targets in the prediction framework focusing on prediction by relevant baseline measures. A wide array of GMM-based prediction targets can be evaluated in terms of prediction accuracy, which is in line with how models are evaluated and selected in supervised learning problems. What is unique about our approach is that the goal is to select prediction targets, not to select predictors.

4.1 Prediction by Antecedent Validators

As a way of validating prediction targets with the emphasis in clinical relevance, we will evaluate their relationship with other relevant measures. In particular, given the intended purpose of trajectory types as outcomes, we will focus on assessing the quality of prediction by relevant baseline variables. The idea is that a good prediction target should be well predicted by demographically and clinically relevant baseline characteristics, known as antecedent validators.

In line with LAMS, we will focus on a simple binary target (low vs. elevated risk) defined in (3.15). With this simplified classification structure, a prediction model with a binary outcome L_i can be expressed as

$$logit(\pi_{L_i}(\boldsymbol{W}_i))=\gamma_0+\boldsymbol{\gamma}_1^{'}\boldsymbol{W}_i, \quad (4.1)$$

where $\boldsymbol{\pi}_{L_i}Pr(L_i = 1|\boldsymbol{W}_i)$ denotes the probability of subject i belonging to the low risk type varying as a function of a vector of baseline covariates $\boldsymbol{W}_i$. The relationship between the risk status L and covariates is captured by a vector of logit coefficients $\boldsymbol{\gamma}_1$. The baseline covariates included here are basic demographic characteristics (sex, age, health insurance as proxy for family SES, smoking status) and clinically relevant measures (depression measured by CDRS-R, bipolar diagnosis, baseline PGBI-10M). In our LAMS application, the covariates included in GMM ($\boldsymbol{X}_i$) and the covariates used in the validation process ($\boldsymbol{W}_i$) overlap, but are not identical. The baseline bipolar diagnosis, which is considered the most clinically relevant antecedent validator, was not used in formation of trajectory classes using GMM, but used in the validation step. We consistently used this fixed set of covariates as predictors during the validation process.

Based on (4.1), the predicted risk status of individual i is defined as

$$\hat{L}_i(\boldsymbol{W}_i)=\begin{cases} 1\,(\text{low risk}) & \text{if } \hat{\pi}_{L_i}(\boldsymbol{W}_i)>0.5 \\ 0\,(\text{elevated risk}) & \text{otherwise.} \end{cases} \quad (4.2)$$

4.2 Evaluation of Prediction Quality

Using the prediction model in (4.1), we evaluate the quality of prediction on the basis of common measures of classification performance. They are sensitivity (S), specificity (P) and accuracy (A) defined as

$$\mathrm{S}=TP/(TP+FN), \quad (4.3)$$

$$\mathrm{P}=TN/(TN+FP), \quad (4.4)$$

$$\mathrm{A}=(TP+TN)/(TP+FN+TN+FP), \quad (4.5)$$

where true positives (*TP*) are individuals who belongs to the low risk group ($L_i = 1$) and correctly classified as low risk by the prediction model ($\hat{L}_i(\boldsymbol{W}_i) = 1$), true negatives (*TN*) are those who are not in the low risk group ($L_i = 0$) and correctly predicted as not belonging to the low risk group ($\hat{L}_i(\boldsymbol{W}_i) = 0$), false positives (*FP*) are those who are not in the low risk group ($L_i = 0$) and incorrectly predicted as low risk ($\hat{L}_i(\boldsymbol{W}_i) = 1$), and false negatives (*FN*) are individuals who belongs to the low risk group ($L_i = 1$) and incorrectly classified as not belonging to the low risk group by the prediction model ($\hat{L}_i(\boldsymbol{W}_i) = 0$).

In our validation framework using antecedent validators, evaluating the prediction performance itself is already a process of internal validation. This process can be enhanced by taking into account variability in prediction quality across different samples. Paying much attention to the prediction capability of learning methods on new data is a signature feature of supervised learning. However, comparable independent data sets are often rare as is the case in LAMS. Given that, internally examining the prediction performance across different portions of the data at hand has been an important component in supervised learning. For this purpose, a simple, but effective, validation technique known as K-fold cross-validation is widely used [23, 24, 39, 40].

4.2.1 K-Fold Cross-Validation—In this method, we randomly divide the total sample into K equal size subsamples ($k = 1, 2, \ldots, K$). Of the K subsamples, we set aside one subsample (k^{th} fold) to be used as a validation sample. With the rest of the subsamples (training data), we build a prediction model. The validation sample (k^{th} fold) is then used to estimate the expected prediction quality when the model is applied to a data set that is not used to formulate the prediction model. This process is repeated K times and then the results are averaged over K results. Specifically, the three measures of prediction performance, sensitivity, specificity and accuracy will be calculated K times and then averaged over K subsamples. That is,

$$CV_S=(\sum_{k=1}^{K} S_k)/K, \quad (4.6)$$

$$CV_P=(\sum_{k=1}^{K} P_k)/K, \quad (4.7)$$

$$CV_A = (\sum_{k=1}^{K} A_k)/K, \quad (4.8)$$

where S_k, P_k and A_k are sensitivity, specificity and accuracy in the k^{th} fold.

We will also use another popular way of model selection based on the one-standard-error rule, where we select the most parsimonious model with CV values that are within one standard error range of those of the best performing model [23–24]. The standard errors of CV_S, CV_P and CV_A are calculated as

$$SE_S = \sqrt{Var(S_1, S_2, \ldots, S_K)}/\sqrt{K}, \quad (4.9)$$

$$SE_P = \sqrt{Var(P_1, P_1, \ldots, P_K)}/\sqrt{K}, \quad (4.10)$$

$$SE_A = \sqrt{Var(A_1, A_1, \ldots, A_K)}/\sqrt{K}. \quad (4.11)$$

The calculation of CV_S, CV_P and CV_A and their associated standard errors are straightforward when the class membership has one version as in (3.15). Alternatively, to capture the uncertainty in classification, one could also derive CV_S, CV_P and CV_A based on multiple independent pseudoclass draws as in (3.16). That is, for each iteration of pseudoclass draw (say the r^{th} iteration), we first obtain $L_i^{(r)}$ by drawing a random sample from the binomial distribution with probability equal to the posterior probability p_{iJ} of belonging to the low risk class. Then, based on the prediction model in (4.1), $CV_S^{(r)}$, $CV_P^{(r)}$ and $CV_A^{(r)}$ are calculated from each r^{th} pseudoclass draw. Based on the pseudoclass theory, CV_S, CV_P and CV_A can be estimated by averaging the estimate of each model validation index over multiple independent pseudoclass draws. That is,

$$CV_S^{PCD} = \{\sum_{r=1}^{R} CV_S^{(r)}\}/R, \quad (4.12)$$

$$CV_P^{PCD} = \{\sum_{r=1}^{R} CV_P^{(r)}\}/R, \quad (4.13)$$

$$CV_A^{PCD}=\{\sum_{r=1}^{R} CV_A^{(r)}\}/R, \quad (4.14)$$

where $r = 1, 2, \ldots, R$, and the associated standard error is estimated by the squared root of the variances averaged across pseudoclass draws as shown below:

$$SE_S^{PCD}=\sqrt{\{\sum_{r=1}^{R}(SE_S^{(r)})^2\}/R+Var(CV_S^{(1)},CV_S^{(2)},\ldots,CV_S^{(R)})}, \quad (4.15)$$

$$SE_P^{PCD}=\sqrt{\{\sum_{r=1}^{R}(SE_P^{(r)})^2\}/R+Var(CV_P^{(1)},CV_P^{(2)},\ldots,CV_P^{(R)})}, \quad (4.16)$$

$$SE_A^{PCD}=\sqrt{\{\sum_{r=1}^{R}(SE_A^{(r)})^2\}/R+Var(CV_A^{(1)},CV_A^{(2)},\ldots,CV_A^{(R)})}, \quad (4.17)$$

where $SE_S^{(r)}$, $SE_P^{(r)}$, and $SE_A^{(r)}$ correspond to the calculation of (4.9)–(4.11) for the r^{th} pseudoclass draw.

5 Application to LAMS

In the proposed GMM approach, prediction targets are first generated by categorizing individuals based on their posterior class probabilities estimated in each GMM model as described in (3.15) and (3.16). In the completely separate next step, we validate formulated prediction targets in the prediction framework in (4.1). We included basic demographic characteristics (sex, age, smoking status, health insurance as proxy for family SES) and clinically relevant measures (bipolar diagnosis, depression measured by CDRS-R, baseline PGBI-10M) as antecedent validators. Based on these predictors, the quality of prediction was assessed in terms of sensitivity and specificity. In this validation step, we used 10-fold cross-validation as another enhancement of internal validation, taking into account potential variation in prediction quality across different samples.

Before examining the prediction quality using GMM-based prediction targets, we examined the prediction quality when prediction targets are generated using observed PGBI-10M at each assessment with a fixed clinical threshold, which is a more conventional method of classification. We classified individuals as low and elevated risk using the threshold of 12, which has been proposed in previous studies [25–27]. The resulting four prediction targets (low vs. elevated risk at 6, 12, 18, and 24 month) were evaluated in the same method used to

evaluate GMM-based prediction targets. Table 2 shows the prediction quality when prediction targets are formulated using this conventional method. The level of sensitivity is high for all four targets, although the level of specificity is very low especially when predicting targets farther away from the baseline.

In the proposed approach, we first conduct GMM to extract heterogeneous trajectory classes. Based on model specifications in (3.1) to (3.10), 36 GMM models with two or more classes reached normal convergence (20 without, 16 with covariates), which are summarized in Table 3. Based on the simple binary classification structure described in (3.15), 36 binary prediction targets were generated based on 36 GMM models. In this study, we monitored BIC and BLRT as a way of assessing the model fit. However, instead of screening GMM models based on the model fit, we directly evaluate formulated prediction targets in the prediction framework.

We focused on a simple classification setting, where separating out a considerable proportion of children who would maintain non-problematic levels of manic symptoms is of great clinical importance. All candidate GMM models consistently exhibited a trajectory class that starts with the estimated PGBI-10M of around 12 or lower and then gradually decreases. From the clinical perspective, this bottom trajectory class is clearly the least problematic. This interpretation is also supported by a previously suggested clinical threshold that sets PGBI-10M ≥ 12 as elevated in manic symptoms [25–27]. However, in models with no random effects with seven or more classes, the low risk trajectory split into smaller subclasses, resulting in 2–3 classes that could be potentially categorized as low risk. In these models, we categorized only the class with the lowest mean trajectory as low risk. In comparison to other models, these models will inform us how the prediction quality varies depending on how conservatively individuals are classified as low risk. Some examples of GMM results are shown in Figure 1, where the solid line at the bottom in each model is the trajectory class we categorize as low risk. The rest of the classes are combined and categorized as elevated risk.

Once we designate a low risk trajectory class, prediction targets can be generated by classifying each individual based on his or her posterior class probability of belonging to that low risk trajectory class, as described in (3.15) and (3.16). As this classification happens under each GMM model, some individuals, in particular those in the grey zone, may be classified as low risk under one model, but not under another model. Examining differences in classification for these cases may lead to better understanding of which prediction target is more aligned with our clinical intention, and therefore lead to a better informed choice.

Table 4 shows an example of potential disagreement in classification across different GMM models. In this example, 5- and 6-class random intercept models with covariates (XC6i and XC5i) are compared. In terms of their posterior probabilities of belonging to the low risk class ($\hat{p}_{ij}$), the two models are highly correlated ($r = 0.997$). Based on the classification in (3.15), 15 individuals (2.2%) out of 682 were classified differently between the two models. We examined the observed data of these 15 cases based on 1) whether PGBI-10M ≥ 12 at any point and 2) whether the baseline CDRS-R (mean=34.73, SD=10.73) is above the mean. Cases 1 to 11 are more likely to be categorized as elevated risk based on XC6i and as low

risk based on XC5i. Cases 1–3 have PGBI-10M ≥ 12 only once and the deviation is not serious. However, their baseline depression level is considerably higher than average. Cases 4–8 have PGBI-10M ≥ 12 only once, although the deviation is quite large. Given this observation, a safe choice would be to classify these cases (1–8) as elevated risk. Cases 9–11 are more questionable, especially with missing data, which is reflected in the estimated class probabilities close to 0.5 in both models. Cases 12 to 15 are more likely to be categorized as low risk based on XC6i, which seems reasonable, although 13 and 15 are somewhat questionable given moderately high PGBI-10M scores and missing assessments. Overall, XC6i tends to classify uncertain cases as elevated risk, whereas XC5i as low risk. At the same time, the agreement between the two models is very high, assuring the possibility of stable classification.

In the proposed approach, prediction targets are generated by categorizing patients based on their estimated posterior class probabilities. In the completely separate next step, we validate formulated prediction targets in the prediction framework as described in (4.1). We first evaluated GMM-based prediction targets without taking into account uncertainties in classification as defined in (3.15). Figure 2 shows the assessed prediction quality based on this approach. As sensitivity and specificity are equally important in the LAMS context, we intend to select GMM-based prediction targets that not only lead to the highest overall accuracy, but also the highest sensitivity and specificity. We also intend to improve prediction accuracy by using GMM-based prediction targets instead of using observed measures (see Table 2). Given these considerations, we used 0.7 as the lower limit for the sensitivity and specificity estimates of the GMM-based prediction targets. We screened all candidate targets based on the rule that $CV_S - SE_S \geq 0.7$ and that $CV_P - SE_P \geq 0.7$, considering possible variations in prediction quality. Among 36 GMM models considered, 15 models that satisfied this rule are shown in Figure 2. The best sensitivity was achieved using the prediction target based on XC5i ($CV_S = 0.83$), and the best specificity based on XC6i ($CV_P = 0.90$). The next best prediction targets are based on random intercept models without covariates with 5–7 classes. The quality of prediction based on these models is close to that based on XC6i and XC5i.

We repeated our cross-validation taking into account uncertainties in classification. We used pseudoclass draws based on the posterior class probability of belonging to the low risk trajectory class ($\hat{p}_{ij}$), as defined in (3.16), which basically creates multiple versions of prediction target based on each GMM model. Using each version of the target (i.e., each pseudoclass draw), the quality of prediction can be assessed, and then the results are averaged across multiple versions. We used 20 pseudoclass draws. We again screened all candidate targets based on the rule that $CV_S - SE_S \geq 0.7$ and that $CV_P - SE_P \geq 0.7$. Figure 3 shows the quality of prediction when uncertainties in classification is taken into account. Among 36 GMM models considered, 10 models satisfied the rule this time. Overall, the CV results show somewhat lower sensitivity and specificity compared to when uncertainties are not taken into account (Figure 2). However, the main story remains the same. The best sensitivity was achieved using the prediction target based on XC5i ($CV_S = 0.77$), and the best specificity based on XC6i ($CV_P = 0.87$). The next best prediction targets are based on random intercept models without covariates with 5–7 classes. The quality of prediction based on these models is again quite close to that based on XC6i and XC5i.

Table 5 provides some details of the results shown in Figure 3 that take into account uncertainties in classification. The two best performing prediction targets based on XC6i and XC5i are basically tied in terms of prediction quality, although XC6i is a better choice if we prefer classifying individuals somewhat more conservatively. As discussed with Table 4, a more informed choice can be made by carefully examining individuals who are differently classified by different GMM models. If selecting a more parsimonious model is of interest, we can also apply the one-standard-error rule, where we pick the most parsimonious model with CV values that are within one standard error range of those of the best performing model. For sensitivity, the minimum in the model XC5i is 0.737, and therefore most models with fewer parameters in Table 4 are comparable to XC5i. For specificity, the minimum in XC6i is 0.855, and therefore four simpler models (XC5i, C6i, C7i, C5i) are comparable to XC6i. Together, five best performing models are comparable in terms of prediction quality, and C5i would be the most parsimonious choice with only 25 parameters. The five best performing models consistently categorized 39–40% of individuals as low risk. Using information on model fit will considerably narrow the selection, although it may prematurely exclude potentially useful models. Among the five best performing models, XC6i and C7i were selected by BLRT, C6i was selected by BIC, but XC5i and C5i were not selected either by BLRT or by BIC.

6 Conclusions

GMM has been increasingly used in various research fields for the purpose of meaningful interpretation of longitudinal heterogeneity in the target population and inference about how these trajectory classes are related to other variables. Expanding such common use of GMM, this study showed that GMM can also serve as a useful tool in the context of individual level prediction. In particular, we focused on a specific utility of GMM as a way of constructing reliable and valid prediction targets. In the proposed GMM approach, prediction targets are first generated by categorizing individuals based on their posterior class probabilities estimated in each GMM model. In the completely separate next step, formulated prediction targets are validated in terms of how well they are predicted by relevant baseline covariates (antecendent validators). In this approach, a large array of prediction targets generated by GMM (unsupervised learning) are validated in the prediction framework (supervised learning).

A small fixed set of antecedent validators was used in the proposed approach to swiftly validate GMM-based prediction targets and to narrow the choice among them. In our LAMS application, it turned out that the level of prediction accuracy based on this limited selection of variables is quite high (both sensitivity and specificity over 0.75) even after taking into account uncertainties in classification. We find the results quite promising as our validation process can be considered a precursor to developing prognostic models fully considering all possible baseline predictors. For this next step of developing prognostic models, a large number of baseline covariates, including the ones we used as antecedent validators, can be considered as potential predictors. The LAMS study indeed collected rich data in multiple domains, providing a valuable opportunity to develop prognostic models with a practical level of accuracy. We leave this next step, which naturally involves feature selection, as a topic for future investigation.

Whether covariates should be included in extracting trajectory classes is a debatable issue as it may help better classify individuals, but may also lead to misspecified models. Further, it is not clear whether we should use covariates redundantly, both as predictors in GMM and as predictors in the validation step. Neither seemed to be an issue in our LAMS application, where prediction targets formulated based on both GMM with and without covariates (XC6i, XC5i, C6i, C7i, C5i) performed comparably well in the validation process. These GMM models are also highly correlated ($r > 0.98$) in terms of the posterior class probability of belonging to the low risk trajectory class ($\hat{p}_{Li}$), which leaves little room for possible impacts of model misspecification involving covariates. In fact, both the GMM and the validation process can utilize various types of auxiliary information, not only from the baseline, but also from concurrent and distal measures. Further investigation, both theoretical and empirical, seems necessary as there is little research in this regard in the context of individual level prediction.

To focus on demonstrating our new approach of using GMM, we limited the range of models to be examined. We used a simple quadratic growth specification with restricted variance structures, although various alternative model specifications are possible. For example, our investigation did not include models with higher order polynomial growth specifications or class-varying variances. Considering various model specifications will lead to a much wider pool of GMM models. According to our investigation with the LAMS data, once prediction targets are formulated based on GMM, it seems feasible to process a large number of them in the validation process. However, having numerous candidate models can still overwhelm the model fitting process, which calls for some guiding principles. Future research is warranted to examine the benefits and limitations of different GMM model specifications under various application settings.

The proposed method of utilizing and evaluating GMM solutions may take several different directions for further development and refinement. In the LAMS application, we constructed prediction targets with the purpose of separating out children who would remain non-problematic over the two-year course. However, prediction targets can be formulated in many different ways depending on the specific purpose of classification. For example, our next interest is in identifying children who would develop the most problematic trajectory type. The best GMM models in the current application may or may not turn out to be the best with this shifted target. Another issue that deserves further attention is how to best utilize model-based classification and fixed thresholds. In this study, we used model-based classification, where we categorize the class with the lowest mean trajectory as low risk and the rest as elevated risk. This approach allows the level of outcome in the low risk category to vary across different GMM models, allowing us to observe how the prediction quality varies depending on how conservatively individuals are classified as low risk. An alternative way would be to use fixed thresholds (e.g., estimated PGBI-10M < 12 in LAMS), relying less on classification provided by GMM. In this method, the external thresholds can be directly applied to estimated individual trajectories, and therefore prediction targets can be formulated not only using multi-class, but also using single-class trajectory solutions. Further research is needed to examine similarities and differences between these alternative methods in various situations.

Acknowledgments

This study was supported by the National Institute on Drug Abuse (DA031698) and National Institute of Mental Health (Case Western Reserve University: R01 MH073967-06A1, Cincinnati Childrens Hospital Medical Center: MH073816-06A1, The Ohio State University: MH073801-06A1, University of Pittsburgh: MH073953-06A1). We appreciate helpful feedback from the LAMS (Longitudinal Assessment of Manic Symptoms) project team and the PSMG (Prevention Science Methodology Group). We also thank Helena Kraemer, Linda and Bengt Muthén, and Karen Bandeen-Roche for their valuable input.

References

1. Muthén, BO. Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In: Collins, LM.; Sayer, A., editors. New Methods for the Analysis of Change. Washington, DC: APA; 2001. p. 291-322.
2. Muthén BO, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics. 1999; 55:463–469. [PubMed: 11318201]
3. Muthén, BO. Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In: Kaplan, D., editor. Handbook of quantitative methodology for the social sciences. Newbury Park, CA: Sage; 2004. p. 345-368.
4. Dunn KM, Jordan K, Croft PR. Characterizing the course of low back pain: A latent class analysis. American Journal of Epidemiology. 2006; 163:754–761. [PubMed: 16495468]
5. Dunn LB, Cooper BA, Neuhaus J, West C, Paul S, Aouizerat B, Abrams G, Edrington J, Hamolsky D, Miaskowski C. Identification of distinct depressive symptom trajectories in women following surgery for breast cancer. Health Psychology. 2011; 30:683–692. [PubMed: 21728421]
6. Findling RL, Jo B, Frazier TW, Youngstromd EA, Demeter CA, Fristad MA, Birmaher B, Kowatch RA, Arnold E, Axelson DA, Ryan N, Hauser JC, Brace DJ, Marsh LE, Gill MK, Depew J, Rowles BM, Horwitz SM. The 24-month course of manic symptoms in children. Bipolar Disorders. 2013; 15:669–679. [PubMed: 23799945]
7. Gueorguieva R, Mallinckrodt C, Krystal JH. Trajectories of depression severity in clinical trials of duloxetine insights into antidepressant and placebo responses. Archives of General Psychiatry. 2011; 68:1227–1237. [PubMed: 22147842]
8. Jo B, Wang C-P, Ialongo NS. Using latent outcome trajectory classes in causal inference. Statistics and Its Interface. 2009; 2:403–412. [PubMed: 20445809]
9. Kellam SG, Brown CH, Poduska JM, Ialongo NS, Wang W, Toyinbo P, Petras H, Ford C, Windham A, Wilcox HC. Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. Drug and Alcohol Dependence. 2008; 95:S5–S28. [PubMed: 18343607]
10. Muthén BO, Brown CH. Estimating drug effects in the presence of placebo response: Causal inference using growth mixture modeling. Statistics in Medicine. 2009; 28:3363–3385. [PubMed: 19731223]
11. Muthén BO, Brown CH, Masyn K, Jo B, Khoo ST, Yang CC, Wang C-P, Kellam SG, Carlin JB, Liao J. General growth mixture modeling for randomized preventive interventions. Biostatistics. 2002; 3:459–475. [PubMed: 12933592]
12. Rodriguez D, Audrain-McGovern J. Team sport participation and smoking: Analysis with general growth mixture modeling. Journal of Pediatric Psychology. 2004; 29:299–308. [PubMed: 15148352]
13. Stulz N, Thase ME, Klein DN, Manber R, Crits-Christoph P. Differential effects of treatments for chronic depression: a latent growth model reanalysis. Journal of Consulting and Clinical Psychology. 2010; 78:409–419. [PubMed: 20515216]
14. van Lier PAC, Muthén BO, van der Sar RM, Crijnen AAM. Preventing disruptive behavior in elementary schoolchildren: Impact of a universal classroom-based intervention. Journal of Consulting and Clinical Psychology. 2004; 72:467–478. [PubMed: 15279530]

15. Wang C-P, Jo B, Brown CH. Causal inference in longitudinal comparative effectiveness studies with repeated measures of a continuous intermediate variable. Statistics in Medicine. 2014; 33:3509–3527. [PubMed: 24577715]
16. Bauer DJ, Curran PJ. Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. Psychological Methods. 2003; 8:338–363. [PubMed: 14596495]
17. Jo, B. Growth mixture modeling and causal inference. In: Hancock, GR.; Harring, JR., editors. Advances in Longitudinal Methods in the Social and Behavioral Sciences. Greenwich, CT: Information Age; 2012. p. 193-214.
18. Masyn, KE. Latent class analysis and finite mixture modeling. In: Little, TD., editor. The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2: Statistical Analysis. Oxford University Press; 2013. p. 550-610.
19. Muthén BO. Statistical and substantive checking in growth mixture modeling: Comment on Bauer and Curran. Psychological Methods. 2003; 8:369–377. [PubMed: 14596497]
20. Nagin DS, Tremblay RE. Developmental trajectory groups: Fact or a useful statistical fiction. Criminology. 2005; 43:873–904.
21. Nylund KL, Asparouhov T, Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural Equation Modeling. 2007; 14:535–569.
22. Petras, H.; Masyn, K. General growth mixture analysis with antecedents and consequences of change. In: Piquero, A.; Weisburd, D., editors. Handbook of Quantitative Criminology. New York: Springer; 2010. p. 69-100.
23. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer; 2009.
24. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. An Introduction to Statistical Learning with Applications in R. Springer; 2013.
25. Findling RL, Youngstrom EA, Fristad MA, Birmaher B, Kowatch RA, Arnold E, Frazier TW, Axelson DA, Ryan N, Demeter CA, Gill MK, Fields B, Depew J, Kennedy SM, Marsh LE, Rowles BM, Horwitz SM. Characteristics of children with elevated symptoms of mania: the Longitudinal Assessment of Manic Symptoms (LAMS) study. Journal of Clinical Psychiatry. 2010; 71:1664–1672. [PubMed: 21034685]
26. Horwitz SM, Demeter CA, Pagano ME, Youngstromd EA, Fristad MA, Arnold E, Birmaher B, Gill MK, Axelson DA, Kowatch RA, Frazier TW, Findling RL. Longitudinal Assessment of Manic Symptoms (LAMS) Study: background, design, and initial screening results. Journal of Clinical Psychiatry. 2010; 71:1511–1517. [PubMed: 21034684]
27. Youngstrom EA, Birmaher B, Findling RL. Pediatric bipolar disorder: Validity, phenomenology, and recommendations for diagnosis. Bipolar Disorders. 2008; 10:194–214. [PubMed: 18199237]
28. Muthén, LK.; Muthén, BO. Mplus User's Guide. 7. Los Angeles, CA: Muthén and Muthén; 1998–2012.
29. Dempster A, Laird N, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. 1977; 39:1–38.
30. Little, RJA.; Rubin, DB. Statistical analysis with missing data. New York: Wiley; 2002.
31. McLachlan, GJ.; Krishnan, T. The EM algorithm and extensions. New York: Wiley; 1997.
32. Tanner, M. Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions. New York: Springer; 1996.
33. Schwartz G. Estimating the dimension of a model. The Annals of Statistics 1978. 1978; 6:461–464.
34. McLachlan, G.; Peel, D. Finite mixture models. New York: Wiley; 2000.
35. Tofighi, D.; Enders, CK. Identifying the correct number of classes in a growth mixture model. In: Hancock, GR., editor. Mixture models in latent variable research. Greenwich, CT: Information Age; 2007. p. 317-341.
36. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5. Arlington, VA: American Psychiatric Publishing; 2013.

37. Bandeen-Roche K, Miglioretti DL, Zeger SL, Rathouz PJ. Latent variable regression for multiple discrete outcomes. Journal of the American Statistical Association. 1997; 92:1375–1386.

38. Wang CP, Brown CH, Bandeen-Roche K. Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. Journal of the American Statistical Association. 2005; 100:1054–1076.

39. Stone M. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, Series B. 1974; 36:111–147.

40. Stone M. An asymptotic equivalence of choice of model by crossvalidation and Akaikes criterion. Journal of the Royal Statistical Society, Series B. 1977; 39:44–47.

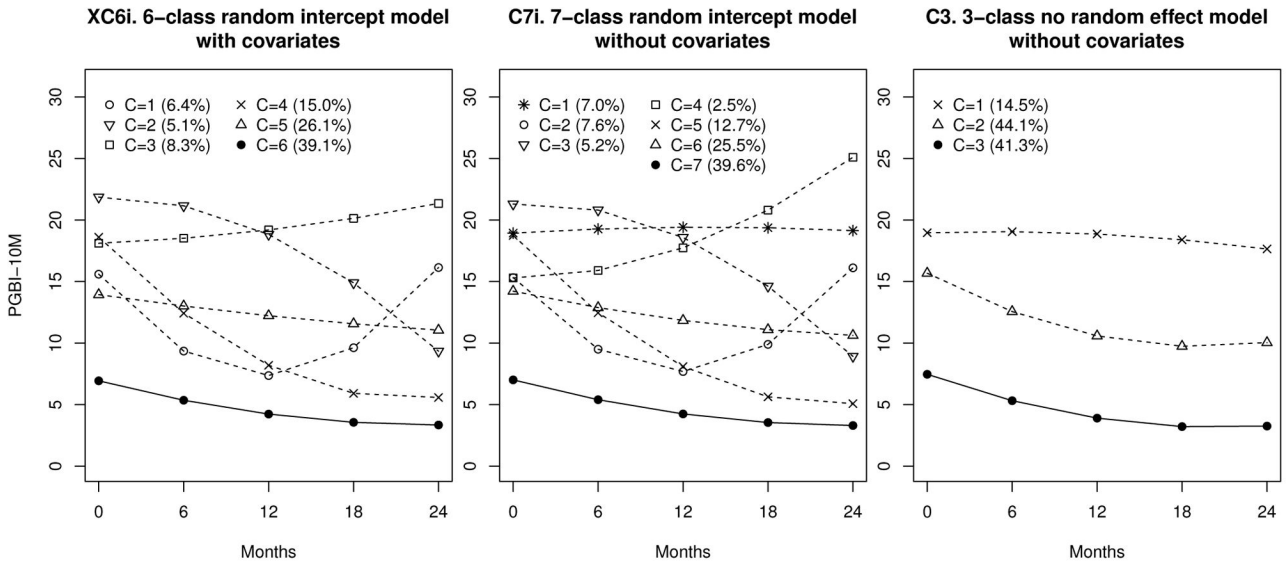

Figure 1.
Examples of candidate GMM solutions

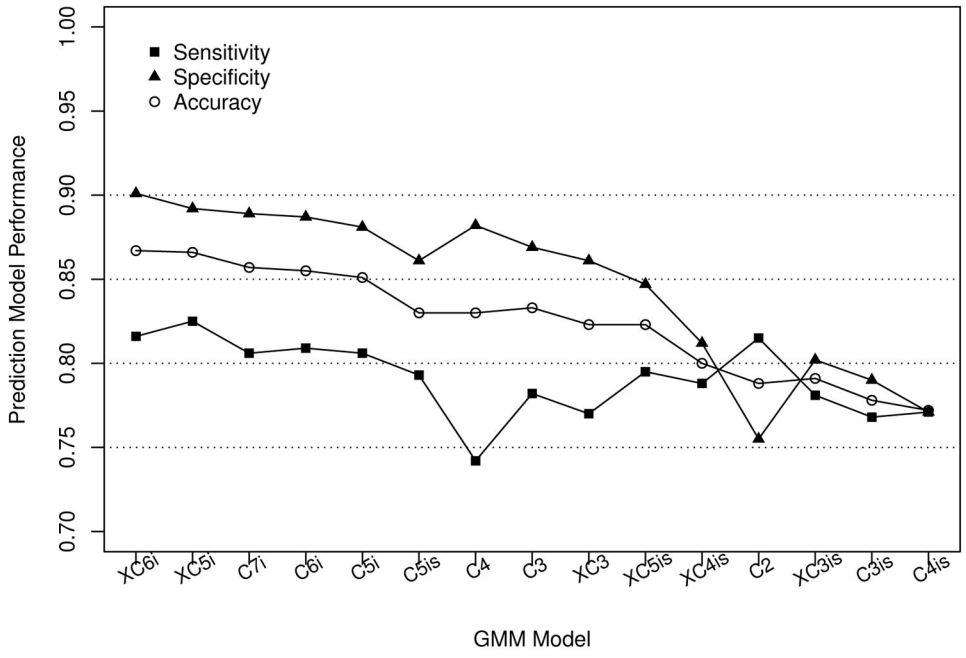

Figure 2.
Quality of prediction when uncertainties in classification are not taken into account.

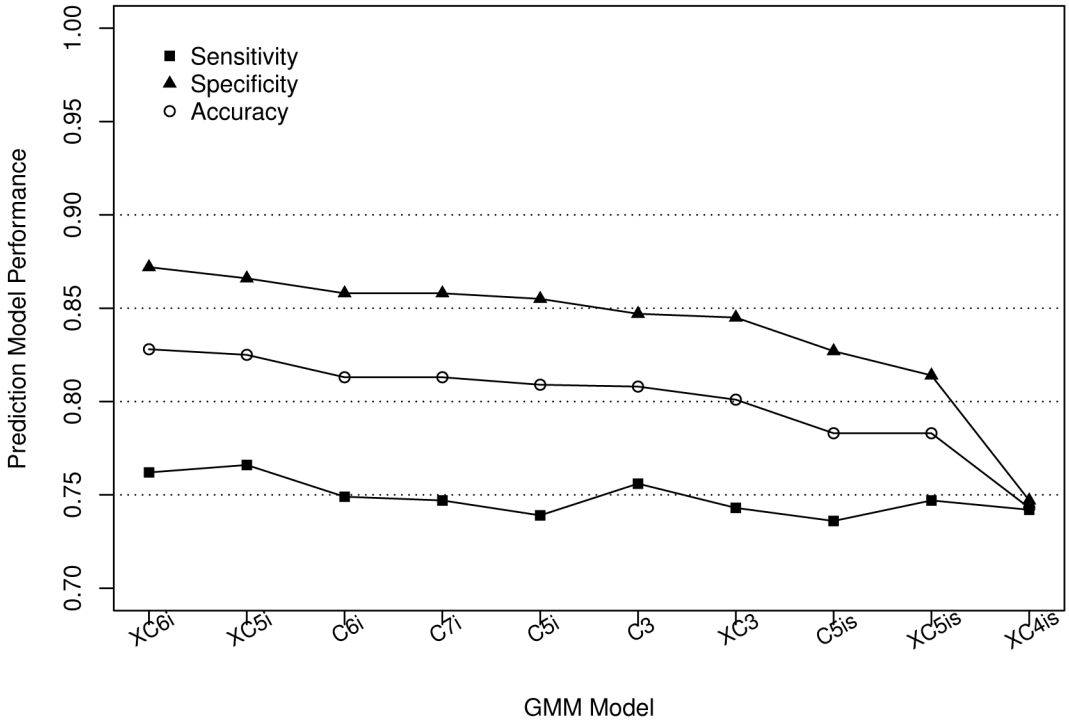

Figure 3.
Quality of prediction when uncertainties in classification are taken into account using pseudoclass draws.

Table 1

LAMS: Sample Statistics (N=682)

Variable	N	Min	Max	Mean	SD
PGBI-10M at baseline	671	0	30	12.78	7.15
PGBI-10M at 6 months	536	0	30	10.51	6.71
PGBI-10M at 12 months	520	0	30	8.60	6.55
PGBI-10M at 18 months	476	0	27	8.42	6.10
PGBI-10M at 24 months	445	0	30	8.07	6.51
Age	682	6.06	13.19	9.39	1.93
Depression (CDRS-R)	682	17	73	34.73	10.73
Female	682	0	1	0.33	
Medicaid	682	0	1	0.42	
Bipolar	682	0	1	0.23	
Smoking	682	0	1	0.05	

Table 2

Average sensitivity (CV_S) and specificity (CV_P) from 10-fold cross-validation in prediction of low risk status defined based on a single observed outcome and a fixed threshold (Listwise deletion has been applied. $CV_S \pm SE_S$ and $CV_P \pm SE_P$ are shown in parentheses).

Observed Outcome	%Low Risk	Sensitivity	Specificity
PGBI-10M 6m < 12	58.2	0.791 (0.764, 0.818)	0.602 (0.571, 0.633)
PGBI-10M 12m < 12	68.3	0.894 (0.879, 0.908)	0.401 (0.341, 0.461)
PGBI-10M 18m < 12	71.0	0.882 (0.860, 0.905)	0.360 (0.324, 0.395)
PGBI-10M 24m < 12	73.9	0.914 (0.895, 0.933)	0.213 (0.173, 0.252)

Table 3

GMM Models Used for Classification Based on Longitudinal PGBI-10M

Label	Model Description	# Par
C2is	2-class, random intercept & slope	15
C3is	3-class, random intercept & slope	19
C4is †	4-class, random intercept & slope	23
C5is *	5-class, random intercept & slope	27
C2i	2-class, random intercept	13
C3i	3-class, random intercept	17
C4i	4-class, random intercept	21
C5i	5-class, random intercept	25
C6i †	6-class, random intercept	29
C7i *	7-class, random intercept	33
C2	2-class, no random effect	12
C3	3-class, no random effect	16
C4	4-class, no random effect	20
C5	5-class, no random effect	24
C6 †	6-class, no random effect	28
C7	7-class, no random effect	32
C8	8-class, no random effect	36
C9 *	9-class, no random effect	40
C10	10-class, no random effect	44
C11	11-class, no random effect	48
XC2is †*	2-class with covariates, random int & slope	31
XC3is	3-class with covariates, random int & slope	39
XC4is	4-class with covariates, random int & slope	47
XC5is	5-class with covariates, random int & slope	55
XC2i	2-class with covariates, random intercept	29
XC3i †	3-class with covariates, random intercept	37
XC4i	4-class with covariates, random intercept	45
XC5i	5-class with covariates, random intercept	53
XC6i *	6-class with covariates, random intercept	61
XC2	2-class with covariates, no random effect	28
XC3	3-class with covariates, no random effect	36
XC4 †	4-class with covariates, no random effect	44
XC5 *	5-class with covariates, no random effect	52
XC6	6-class with covariates, no random effect	60
XC7	7-class with covariates, no random effect	68
XC8	8-class with covariates, no random effect	76

Par is the number of parameters;

† indicates a model selected by BIC and

* by BLRT within each type of random effect structure Σ_ζ with and without covariates

Table 4

Potential disagreement in classification between two GMM models

	Observed PGBI-10M score					CDRS-R (depression)	$\hat{p}_{iJ}$	
	baseline	6 mo	12 mo	18 mo	24 mo		XC6i	XC5i
Case 1	6	12	9	.	7	60	0.432	0.666
Case 2	13	.	10	1	0	59	0.409	0.576
Case 3	9	11	13	5	7	45	0.443	0.646
Case 4	20	3	1	0	0	30	0.489	0.595
Case 5	17	6	7	6	4	42	0.456	0.578
Case 6	16	8	5	1	0	42	0.372	0.508
Case 7	15	11	1	.	1	40	0.416	0.518
Case 8	8	15	6	.	6	38	0.478	0.531
Case 9	.	9	8	3	5	42	0.491	0.530
Case 10	8	.	.	.	.	36	0.492	0.509
Case 11	9	.	.	.	.	30	0.495	0.503
Case 12	9	6	8	9	10	21	0.531	0.426
Case 13	13	6	.	.	.	22	0.559	0.499
Case 14	11	12	.	9	3	26	0.507	0.472
Case 15	12	14	4	9	0	31	0.507	0.485

Table 5

Average sensitivity (CV_S^{PCD}) and specificity (CV_P^{PCD}) from 10-fold cross-validation in prediction of trajectory type based on various GMM models taking into account uncertainties in classification

Model	# Par	BIC	% Low Risk	Sensitivity	Specificity
XC6i*	61	16510	39.1	0.762 (0.734, 0.791)	**0.872** (0.855, 0.888)
XC5i	53	16495	39.8	**0.766** (0.737, 0.795)	0.866 (0.850, 0.882)
C6i†	29	16398	40.1	0.749 (0.721, 0.777)	0.858 (0.839, 0.877)
C7i*	33	16411	39.6	0.747 (0.716, 0.778)	0.858 (0.839, 0.877)
C5i	25	16407	40.5	0.739 (0.711, 0.768)	0.855 (0.838, 0.872)
C3	16	16501	41.3	0.756 (0.726, 0.786)	0.847 (0.828, 0.866)
XC3	36	16535	41.4	0.743 (0.710, 0.776)	0.845 (0.828, 0.861)
C5is*	27	16410	43.6	0.736 (0.701, 0.771)	0.827 (0.802, 0.852)
XC5is	55	16499	45.4	0.747 (0.712, 0.781)	0.814 (0.796, 0.833)
XC4is	47	16472	49.5	0.742 (0.713, 0.771)	0.747 (0.720, 0.775)

† indicates a model selected by BIC and

* by BLRT within each type of random effect structure Σ_ζ with and without covariates; $CV_S^{PCD} \pm SE_S^{PCD}$ and $CV_P^{PCD} \pm SE_P^{PCD}$ are shown in parentheses