Published in final edited form as:

Stat Methods Med Res. 2018 September ; 27(9): 2674–2693. doi:10.1177/0962280216684163.

Construction of longitudinal prediction targets using semisupervised learning

Booil Jo¹, Robert L Findling², Trevor J Hastie¹, Eric A Youngstrom³, Chen-Pin Wang⁴, L Eugene Arnold⁵, Mary A Fristad⁵, Thomas W Frazier⁶, Boris Birmaher⁷, Mary K Gill⁷, and Sarah McCue Horwitz⁸

¹Stanford University, Stanford, USA

²Johns Hopkins University, Baltimore, USA

³University of North Carolina, Chapel Hill, USA

⁴University of Texas Health Science Center, San Antonio, USA

⁵Ohio State University, Columbus, USA

⁶Cleveland Clinic, Cleveland, USA

⁷University of Pittsburgh Medical Center, Pittsburgh, USA

⁸New York University, New York, USA

Abstract

In establishing prognostic models, often aided by machine learning methods, much effort is concentrated in identifying good predictors. However, the same level of rigor is often absent in improving the outcome side of the models. In this study, we focus on this rather neglected aspect of model development. We are particularly interested in the use of longitudinal information as a way of improving the outcome side of prognostic models. This involves optimally characterizing individuals' outcome status, classifying them, and validating the formulated prediction targets. None of these tasks are straightforward, which may explain why longitudinal prediction targets are not commonly used in practice despite their compelling benefits. As a way of improving this situation, we explore the joint use of empirical model fitting, clinical insights, and cross-validation based on how well formulated targets are predicted by clinically relevant baseline characteristics (antecedent validators). The idea here is that all these methods are imperfect but can be used together to triangulate valid prediction targets. The proposed approach is illustrated using data from the longitudinal assessment of manic symptoms study.

Reprints and permissions: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Booil Jo, Stanford University, 401 Quarry Road, Room 1364, MC 5795, Stanford, CA 94305-5795, USA. booil@stanford.edu.

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.

Keywords

prognostic model; clinical threshold; latent trajectory class; cross-validation; semisupervised learning

1 Introduction

There has been growing interest in building prognostic models in various fields of medicine to improve individual-level prediction and treatment decisions. In establishing prognostic models, often aided by machine learning methods, much effort is often concentrated in developing algorithms that facilitate effective utilization of large amount of potential predictor variables. Interestingly, much less attention has been given to improving the outcome side of prognostic models. In this study, we focus on this rather neglected aspect of model development. We are particularly interested in the joint use of longitudinal information, empirical model fitting, and clinical insights to formulate valid and reliable prediction targets.

In building prognostic models, it is common to use an observed outcome measured at a particular time point as a prediction target. This approach has much room for improvement when repeated outcome measures are available. First of all, observations have measurement error as most outcomes are affected by noise and imperfect reliability. Even a simple summary measure (e.g. average) based on longitudinal observations is more likely to be reliable compared to a single observation. Needless to say, predicting more reliable outcomes is more likely of high quality. Another important aspect of having longitudinal data is that it provides information on how the outcome changes over time, which can be a key component that characterizes individual patients' true outcome status. Given these reasons, using longitudinally observed outcomes as prediction targets seems obviously a better choice. However, how to characterize individuals' outcome status based on repeated measures and to formulate a prediction target is not so obvious.

Another common practice in building prognostic models is to use fixed thresholds based on clinical insights and science. To make prevention, intervention, or treatment decisions, categorical classification is necessary and practical even when the outcome is dimensional and there is no reason to believe the existence of distinct classes of people. The conventional method of using a single outcome and a fixed threshold may work well for some outcomes such as death, cancer, and infection status. However, most clinical outcomes, both biological (e.g. blood pressure, cholesterol, glucose, BMI) and psychiatric (e.g. depression, anxiety, manic disorder, autism spectrum disorder), do not have clear-cut points. We mostly rely on contemporary science, which is likely imperfect. Further, clinical insights may vary across different experts and are likely subjective at least to some extent. This task becomes even more challenging when we attempt to classify individuals based on longitudinal outcome measures.

Utilizing longitudinal information seems to be a promising way of improving the outcome side of clinical prognostic models. However, this involves optimally characterizing individuals' outcome status, classifying them, and validating the formulated prediction

targets. None of these tasks are straightforward, which may explain why longitudinal prediction targets are not commonly used in practice despite their compelling benefits. As a way of improving this situation, this study explores the joint use of empirical model fitting, clinical insights, and cross-validation based on how well formulated prediction targets are predicted by clinically relevant baseline characteristics (antecedent validators). The idea here is that all these methods are likely imperfect but can be used together to triangulate valid prediction targets.

This article is organized as follows. Section 2 describes the motivating example. In Section 3, we present three ways of formulating prediction targets using longitudinal information. They are (1) using observed outcome measures, (2) using estimated trajectory classes, and (3) using estimated trajectories. In Section 4, we present our strategy of validating formulated prediction targets on the basis of clinical insights, model fit, and the quality of prediction by relevant baseline features. In Section 5, the proposed approach is applied to the longitudinal study of manic symptoms. Section 6 presents Monte Carlo simulation study. Section 7 provides conclusions.

2 The longitudinal assessment of manic symptoms (LAMS) study

Our investigation has been motivated by the LAMS study,^{1–3} designed to investigate phenomenology, the development of bipolar disorder and related conditions, and to establish predictors of functional outcomes in children with elevated manic symptoms. Children aged 6–12 years at screening and their parents were recruited from nine outpatient clinics associated with four university affiliated LAMS sites. The primary outcome was elevated symptoms of mania as measured by the Parent General Behavior Inventory 10-item Mania Form (PGBI-10M),⁴ which will be analyzed as the outcome in this study. A total of 707 children were enrolled in the study, and 685 were eligible to be longitudinally assessed during the first 24 months of the study (at baseline, 6, 12, 18, and 24 months). We included 682 cases in our analyses, excluding ineligible cases and four cases with missing PGBI-10M information at all five assessment points. Table 1 shows sample statistics of the PGBI-10M outcome and a small set of baseline variables that will be used in our analyses. Figure 1 shows the distribution of the PGBI-10M outcome at each assessment.

In the LAMS context, there are two important categories of patients we want to identify early on—those we can pay less attention to and those we need to focus on. In particular, separating out patients who would maintain moderate levels of symptoms is of great clinical significance as this would be the first step toward 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 some antimanic treatments. Being able to identify patients who will follow the low-risk trajectory also has critical implications for their treatment. For example, if some of the patients in the low-risk trajectory have attention-deficit hyperactivity disorder (ADHD), they could be treated with a stimulant with greater confidence that they would not develop mania. Similarly, if they have depression or anxiety, then antidepressants could be considered with more confidence that the patient will not suffer treatment-emergent mania.

In the LAMS study, the manic symptom outcome, PGBI-10M, is measured five times during the two-year period (at baseline, 6, 12, 18, and 24 months). This provides a valuable opportunity to develop prognostic models targeting the longitudinal course of manic symptoms among children who present to outpatient clinical care with concerns in manic symptoms. Previous studies have suggested a clinical threshold that sets PGBI-10M ≥ 12 as elevated in manic symptoms.²⁻⁴ Based on this threshold, we can classify patients into elevated and nonelevated (PGBI-10M < 12) manic symptom categories. However, it is not clear how to characterize and classify individual patients' manic symptom status based on their long-term outcome patterns, which is actually what clinicians and parents would want to know early on to make better treatment decisions with more confidence.

3 Formulation of prediction targets

The first step toward building good prognostic models is to formulate reliable and valid prediction targets. For a prognostic model to be useful, the target outcome should be carefully formulated on the basis of the purpose of prediction, clinical insights, and how the prognostic model will be used in practice. However, even with these guiding principles, formulating prediction targets, in particular with repeated outcome measures, is not a straightforward task. As a way of improving this situation, we consider the joint use of clinical insights and empirical model fitting.

In line with our LAMS example, we will focus on a situation where we can utilize repeated measures of a continuous outcome (manic symptoms measured by PGBI-10M) to construct a prediction target. The ultimate goal is of course to develop a prognostic model that accurately predicts each individual's long-term manic symptom course without the full longitudinal information at the time of prediction. In this paper, we take a step back to develop methods to improve the formulation of prediction targets. Once we successfully construct a prediction target that is worthy of predicting, we can move on to the next step with more confidence and focus on finding the best combination of predictors for that prediction target.

Our focus is in formulating prediction targets that classify individuals into different diagnostics groups. For example, in LAMS, we are interested in developing a prognostic model to categorize each future patient into a long-term manic symptom trajectory type that is either elevated or low risk. Clinical thresholds are often used to define a desirable or undesirable outcome status, which is practical in many situations where treatment, intervention, or prevention choices need to be made on the basis of the projected outcome status. Deciding on clinically meaningful thresholds can be arbitrary as having absolutely established science is an unlikely scenario in most fields of medicine. Nonetheless, clinical thresholds, especially those based upon agreement across clinical experts, are extremely valuable in evaluating the clinical meaningfulness of the formulated prediction targets. In our approach, clinical thresholds serve as the basis not only for formulation but also for validation of prediction targets.

3.1 Classification based on observed outcome measures

3.1.1 Using an outcome observed at a single time point—A commonly used way of defining a prediction target is to simply use a single observation. For example, one may be interested in predicting depression, cholesterol, blood pressure, or BMI two years from the baseline. In this approach, we focus on the outcome two years later instead of its whole longitudinal profile. In conjunction with the use of observed outcomes, clinical thresholds are often used to define a desirable or undesirable outcome status, which we will focus on in this study. This approach may not ideally capture individuals' outcome status or profiles but are widely used in building prognostic models because of their simplicity in definition. In the context of LAMS, previous studies^{2–4} have suggested a clinical threshold that sets PGBI-10M = 12 as having elevated symptoms of mania. This definition will guide our investigation using the LAMS data. As this threshold is not absolutely fixed, either scientifically or empirically, we will also consider adjacent scores (e.g. PGBI-10M score of 11 or 13) as possible alternative cut points. These thresholds will be used to classify individuals based on the outcome observed at each assessment.

To decide the risk status u for individual i , we apply a clinical threshold τ (e.g. PGBI-10M=12) to the observed outcome y at time point t ($t = 1, 2, \dots, T$). That is

$$u_{it}^{obs} = \begin{cases} 1 \text{ (low risk)} & \text{if } y_{it}^{obs} < \tau \\ 0 \text{ (elevated risk)} & \text{if } y_{it}^{obs} \geq \tau \end{cases} \quad (1)$$

3.1.2 Using an outcome observed at multiple time points—For outcomes that tend to change continuously, it is not clear how these thresholds should be applied. For example, in the LAMS study, parents completed the PGBI-10M at five assessments during the two-year period. With five repeated measures of manic symptoms, we may classify individuals in many different ways. In line with previous studies,^{2–4} we will use the PGBI-10M score of 12 as a primary threshold. In principle, we should classify individuals as developing low-risk profiles if they stay below these thresholds at all assessments. However, it is reasonable to assume some random fluctuation and noise in observed repeated measures. As a way of taking this into account, we employ a wide range of thresholds and multiple ways of classification.

Based on observed repeated outcome measures, we determine the risk status L for individual i as

$$L_i^{obs} = \begin{cases} 1 \text{ (low risk)} & \text{if } \sum_{t=1}^T u_{it}^{obs} \geq T - \delta \\ 0 \text{ (elevated risk)} & \text{if } \sum_{t=1}^T u_{it}^{obs} < T - \delta \end{cases} \quad (2)$$

where u_{it}^{obs} is defined in equation (1). We use δ to adjust the degree of deviation from the threshold τ . For example, if we do not want any deviation from the threshold at any assessments, we use $\delta = 0$. If we want to allow some deviation, we could use $\delta = 1$ or $\delta = 2$ depending on the total number of assessments. In our LAMS application, where there are five assessments ($T=5$), we consider three simple definitions of a low-risk profile. That is, PGBI-10M below the threshold at all five times ($\delta=0$), four or more times ($\delta=1$), or at least three times ($\delta=2$).

3.2 Classification based on empirical model fitting

The classification approach described above is based on observed raw individual outcome data. In an effort to capture individual patients' true outcome status in the face of possible random fluctuation and noise in observed repeated measures, we proposed to use ad hoc methods that allow some deviation from the clinical thresholds (e.g. PGBI-10M below the thresholds at three or four times instead of at all five times in LAMS). An alternative approach we will employ is based on empirical model fitting, which is closely tied with the ad hoc approach as it uses the same clinical thresholds. The difference is that we now apply clinical thresholds to smoothed individual curves instead of raw repeated measures. One advantage of this approach is in improved precision by utilizing all cases with any outcome data instead of only using cases without missing measurements. Another advantage of using empirical model fitting is that it allows us to effectively utilize the information about longitudinal changes, which is possible but more challenging to capture using ad hoc algorithms based on observed data.

One way of estimating individual patients' trajectory curves is to use mixed effects modeling,^{5,6} which is a popular method of modeling longitudinal changes, often using polynomial growth models, allowing for random variation in trajectory shapes. In this study, we will employ growth mixture modeling (GMM),⁷⁻¹² which can be thought of as an extended version of standard single-class mixed effects modeling. In GMM, we relax the assumption that all individual trajectories are from a single population, expanding the capacity to accommodate longitudinal heterogeneity. GMM has been mostly used as a method of identifying latent subpopulations that develop heterogeneous trajectory strata.¹³⁻²¹ The focus in this study is in constructing longitudinal prediction targets that will be used in prognostic models. Therefore, our primary reason for using GMM is not to make inference about true trajectory classes but to better estimate individual trajectories. Here we briefly describe the GMM procedure focusing on our LAMS application.

In line with LAMS with five repeated outcome measures (at baseline, 6, 12, 18, and 24 months), we used a simple quadratic growth specification. In the presence of baseline covariates, the manic symptom outcome Y (PGBI-10M) for individual i ($i = 1, 2, \dots, N$) at time t ($t = 1, 2, \dots, T$) conditioned on the trajectory class (C_{ij}) is expressed as

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

$$\eta_{1ij} = \alpha_{1j} + \lambda'_1 \mathbf{X}_i + \zeta_{1ij} \quad (4)$$

$$\eta_{2ij} = \alpha_{2j} + \lambda'_2 \mathbf{X}_i + \zeta_{2ij} \quad (5)$$

$$\eta_{3ij} = \alpha_{3j} + \lambda'_3 \mathbf{X}_i + \zeta_{3ij} \quad (6)$$

where there are J possible trajectory classes ($j = 1, 2, \dots, J$). This model includes three mean growth parameters: initial status (α_{1j}), linear growth (α_{2j}), and quadratic growth (α_{3j}) for class j . We used two sets of time scores, S_t and S_t^2 to capture potentially nonlinear trends over time. The relationship between the growth factors and the covariates $\mathbf{X}$ is captured by the vectors of regression coefficients λ_1 , λ_2 , and λ_3 . In the LAMS example, we included age, sex, Medicaid, and children's depression rating scale-revised (CDRS-R) as covariates. To maintain identifiability in models with larger numbers of classes, we imposed a restriction that these coefficients do not vary across classes.

Conditioned on the trajectory class ($C_{F=j}$), the residuals $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}, \dots, \varepsilon_{ijT})$ are assumed to be normally distributed with $\boldsymbol{\varepsilon}_{ij} \sim MN(0, \Sigma_{\boldsymbol{\varepsilon}})$ and are allowed to vary over time. The residuals ζ_{1ij} , ζ_{2ij} , and ζ_{3ij} are allowed to be correlated and assumed to be normally distributed as $MN(0, \Sigma_{\zeta})$. In response to skewed distributions, GMM assuming a normal distribution tends to extract more trajectory classes, which can be problematic when the goal is in recovering the true trajectory classes. It is less of a concern in our case, where the goal is in estimating individual-level trajectories and separating out low-risk individuals. Alternatively, we assumed a skew-t distribution¹⁰ for Σ_{ζ} , although this led to serious convergence problems in most models in our LAMS example. We used four variations of Σ_{ζ} : models allowing for all three random effects, allowing for random linear slope/intercept ($Var(\zeta_{3ij}) = 0$), allowing for random intercept only ($Var(\zeta_{2ij}) = Var(\zeta_{3ij}) = 0$), and allowing for no random effects ($Var(\zeta_{1ij}) = Var(\zeta_{2ij}) = Var(\zeta_{3ij}) = 0$). With these variations, we conducted GMM with and without allowing Σ_{ζ} and $\Sigma_{\boldsymbol{\varepsilon}}$ to vary across trajectory classes.

The probability of individual i belonging to a certain trajectory class ($C_{F=j}$) can depend on the influence of covariates. The multinomial logit model of π_{ij} conditioned on covariates subsumed in vector $\mathbf{X}_i$ is described as

$$\text{logit}(\pi_{ij} | \mathbf{X}_i) = \beta_{0j} + \beta'_{1j} \mathbf{X}_i \quad (7)$$

for $j = 1, 2, \dots, J-1$, where β_{1j} is a vector of multinomial logistic regression coefficients. The probability $\pi_{ij} = Pr(C_i = j)$, where $\text{logit}(\pi_{ij}) = \log(\pi_{ij} / \pi_{ij})$.

Based on the model specifications described above, we conducted a series of GMM with varying numbers of classes. We used maximum likelihood estimation using the expectation maximization algorithm^{22–25} implemented in the *Mplus* program.²⁶ We increased the number of classes until the latent variable covariance matrix in any of the classes was not positive definite, any of the classes has less than 10 individuals (using the most likely class membership), or the model could not be identified.

From equations (3) to (7), the observed data log-likelihood can be expressed as

$$\sum_{i=1}^N \log \left\{ \sum_{j=1}^J \pi_{ij}(\boldsymbol{\beta}) f(\mathbf{y}_i | C_i = j, \mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}_{\zeta}, \boldsymbol{\Sigma}_{\epsilon}) \right\} \quad (8)$$

where $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iT})$, $\boldsymbol{\alpha} = (\alpha_{1j}, \alpha_{2j}, \alpha_{3j})$, and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$. For $i = 1, \dots, N$, $\sum_{j=1}^J \pi_{ij}(\boldsymbol{\beta}) = 1$, where $\pi_{ij}(\boldsymbol{\beta}) = \Pr(C_i = j | \mathbf{X}_i, \boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{11}, \dots, \boldsymbol{\beta}_{0J}, \boldsymbol{\beta}_{1J})$ is the likelihood of $\mathbf{Y}_i$ arising from mixture component j given $\mathbf{X}_i$.

The log-likelihood of the complete data ($\mathbf{y}_i, C_i: i = 1, \dots, N$) can be written as

$$\text{Log}L_C = \sum_{i=1}^N \text{Log}L_{Ci} = \sum_{i=1}^N \sum_{j=1}^J \{ \log[f(\mathbf{y}_i | C_i = j, \mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}_{\zeta}, \boldsymbol{\Sigma}_{\epsilon})] + \log[\pi_{ij}(\boldsymbol{\beta})] \} \quad (9)$$

To maximize equation (9), the E step computes the expected values of the log-likelihood in equation (8) given observed data and the current parameter estimates ($\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*, \boldsymbol{\Sigma}_{\zeta}^*, \boldsymbol{\Sigma}_{\epsilon}^*$).

In this step, latent trajectory class C is treated as missing data. The E step computes

$$\sum_{i=1}^N \sum_{j=1}^J \log[f(\mathbf{y}_i | C_i = j, \mathbf{X}_i, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*, \boldsymbol{\Sigma}_{\zeta}^*, \boldsymbol{\Sigma}_{\epsilon}^*)] p_{ij}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*, \boldsymbol{\Sigma}_{\zeta}^*, \boldsymbol{\Sigma}_{\epsilon}^*)$$

where the posterior class probability of individual i belonging to class j is calculated as

$$p_{ij}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*, \boldsymbol{\Sigma}_{\zeta}^*, \boldsymbol{\Sigma}_{\epsilon}^*) = \frac{\pi_{ij}(\boldsymbol{\beta}^*) f(\mathbf{y}_i | C_i = j, \mathbf{X}_i, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*, \boldsymbol{\Sigma}_{\zeta}^*, \boldsymbol{\Sigma}_{\epsilon}^*)}{\sum_{j'=1}^J \pi_{ij'}(\boldsymbol{\beta}^*) f(\mathbf{y}_i | C_i = j', \mathbf{X}_i, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*, \boldsymbol{\Sigma}_{\zeta}^*, \boldsymbol{\Sigma}_{\epsilon}^*)} \quad (10)$$

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.

3.2.1 Classifying estimated trajectory classes—One way to classify individual patients using GMM solutions is to classify trajectory classes. That is, in each GMM

solution, we apply the suggested clinical threshold (e.g. PGBI-10M<12 in LAMS) to the estimated mean outcome trajectory of each class. In this method, each individual's risk status is decided based on the trajectory class she or he is most likely to belong to. What is nice about this approach is that the validity of classification is supported not only by a priori clinical thresholds but also by empirical clustering of individuals.

To decide the risk status u for the trajectory class j at time point t , we apply a clinical threshold τ to the estimated mean outcome at each time point. That is

$$u_{jt}^{gmm} = \begin{cases} 1 \text{ (low risk)} & \text{if } \{ \sum_i \hat{Y}_{it} \mid (\hat{C}_i = j) \} / \sum_i \mathbf{1}_{(\hat{C}_i = j)} < \tau \\ 0 \text{ (elevated risk)} & \text{if } \{ \sum_i \hat{Y}_{it} \mid (\hat{C}_i = j) \} / \sum_i \mathbf{1}_{(\hat{C}_i = j)} \geq \tau \end{cases} \quad (11)$$

where Y_{it} is estimated according to equations (3) to (6) using each person's covariate values and the parameter estimates conditioned on the class membership.

Then, we can determine the risk type of individual i in class j as

$$L_i^{gmm1} \mid (\hat{C}_i = j) = \begin{cases} 1 \text{ (low risk)} & \text{if } \sum_{t=1}^T u_{jt}^{gmm} \geq T - \delta \\ 0 \text{ (elevated risk)} & \text{if } \sum_{t=1}^T u_{jt}^{gmm} < T - \delta \end{cases} \quad (12)$$

where T is the total number of assessments ($t = 1, 2, \dots, T$). We may apply δ to adjust the degree of deviation from the threshold τ as we did with observed measures in equation (2). In this method, the risk status is the same for everyone in the same trajectory class j . Specifically, from equation (10), a set of posterior class probabilities ($\hat{p}_{i1}, \hat{p}_{i2}, \dots, \hat{p}_{iJ}$) are generated for each person. The most likely class ($\hat{C}_i$) is the class with the highest posterior probability. For example, in a three-class model, if person i has the posterior probabilities of $\hat{p}_{i1} = 0.2$, $\hat{p}_{i2} = 0.3$, and $\hat{p}_{i3} = 0.5$, her most likely class is the third class ($\hat{C}_i = 3$).

3.2.2 Classifying estimated trajectories—Another way of classifying individual patients using GMM solutions is to apply the clinical threshold directly to each individual's estimated outcome curves. When trajectory classes are not clearly divided between the low- and elevated-risk categories, the results of this method can be quite different from those of indirectly classifying through the most likely trajectory class described above. The idea behind this approach is that we rely less on empirical clustering (i.e. unsupervised learning) and more on a priori clinical threshold. However, when trajectory classes are well divided between the low- and elevated-risk categories, the results are likely similar as will be shown in the LAMS example.

To decide the individual risk status u at each assessment, we now directly apply a clinical threshold τ to the estimated outcome for individual i at time point t . That is

$$u_{it}^{gmm2} | (\hat{C}_i = j) = \begin{cases} 1 \text{ (low risk)} & \text{if } \hat{Y}_{it} | (\hat{C}_i = j) < \tau \\ 0 \text{ (elevated risk)} & \text{if } \hat{Y}_{it} | (\hat{C}_i = j) \geq \tau \end{cases} \quad (13)$$

where $\hat{Y}_{it}$ is estimated according to equations (3) to (6) using each person's covariate values and the parameter estimates conditioned on the class membership.

Then, we determine the overall risk status for individual i as

$$L_i^{gmm2} | (\hat{C}_i = j) = \begin{cases} 1 \text{ (low risk)} & \text{if } \sum_{t=1}^T u_{it}^{gmm2} | (\hat{C}_i = j) \geq T - \delta \\ 0 \text{ (elevated risk)} & \text{if } \sum_{t=1}^T u_{it}^{gmm2} | (\hat{C}_i = j) < T - \delta \end{cases} \quad (14)$$

where δ can be applied to adjust the degree of deviation from the threshold τ as is done with observed measures in equation (2).

4 Validation of prediction targets

Formulating reliable and valid prediction targets is a critical step in developing good prognostic models. As a way of improving reliability, we proposed utilizing repeated outcome measures, which seems to be an obvious choice in the presence of such measures. However, establishing the validity of prediction targets, showing that they are worthy of predicting, is a more challenging task. As the true individual trajectories are unknown (latent), how much variation is due to individual heterogeneity and how much is due to measurement error (noise) is also unknown. Therefore, deciding on the degree and methods of smoothing longitudinal data is difficult regardless of whether we use observed outcome measures or model-based estimated trajectories. We can come up with various ad hoc ways of summarizing/characterizing observed longitudinal data. Similarly, we can come up with different estimated individual trajectories depending on the model choice in empirical fitting. In other words, regardless of whether we use ad hoc or model fitting methods, we will always have some risk of misspecification and misclassification of individuals (in particular, those in the gray zone). As a way of minimizing this risk, we are proposing the joint use of multiple strategies, by using clinical insights (science), empirical fitting/clustering (fit measures), validation through antecedent validators (i.e. by examining prediction quality with cross-validation), and by examining the relationship across prediction targets formulated by multiple methods and models.

4.1 Clinical thresholds

Clinical insights play crucial roles throughout our validation procedure. The first role is to provide clinical thresholds, which can be flexibly applied in various ways in the presence of repeated outcome measures. As described in Section 3, we use clinical thresholds to formulate prediction targets with both observed and estimated individual trajectories. Clinical thresholds are usually formulated based upon agreement across multiple clinical experts. Once established, clinical thresholds work as a convenient tool that can be easily understood and consistently used across different research and clinical applications. For example, in the context of LAMS, previous studies^{2–4} suggested a clinical threshold that sets PGBI-10M ≥ 12 as having elevated symptoms of mania. Of course, clinical insights based on contemporary science may not be perfect, and therefore corresponding clinical thresholds may need adjustment, for example, in line with advancement in science or insights gained from empirical data. As will be shown in the LAMS application, insights gained from investigating empirical data are likely complementary than contradictory to existing clinical insights. This is particularly true when dealing with repeated measures as it is not self-evident how to apply clinical thresholds to best characterize individual trajectory types in the face of unknown measurement errors.

4.2 Model fit measures

When empirical model fitting is used, using fit measures is a widely practiced method of model selection. This approach can be considered exploratory as there is no guarantee that we will recover true individual trajectories. However, note that, regardless of whether we use model-estimated or observed repeated measures, we will always have some risk of misestimation and misclassification of individuals as true individual trajectories are unknown. We propose the joint use of clinical thresholds and fit measures as a way of reducing this risk. In this approach, the resulting model selection is supported not only by clinical insights but also by empirical fit.

Specifically, we used two popular methods of model selection in GMM. One is to compare Bayesian information criteria (BIC)²⁷ and the other is to apply bootstrapped likelihood ratio test (BLRT),²⁸ both of which have been examined as preferred methods of model selection in GMM.^{29,30} Within each type of variance/covariance restrictions, best fitting models were selected based on BIC and BLRT. If no model can be selected by BIC or BLRT (i.e. BIC keeps decreasing, or BLRT never reaches $p > .05$), the model with the largest number of classes was selected. In the LAMS application, using the proposed model specifications, 40 candidate models reached normal convergence. Among these, 10 models were selected based on BIC and/or BLRT.

4.3 Cross-validation using antecedent validators

As another way of establishing the validity of prediction targets, we examine the quality of prediction by relevant predictors. The idea is that a good prediction target should be well predicted by relevant (demographically and clinically) baseline characteristics, known as antecedent validators.^{31,32} This method is completely separated from formulation of prediction targets, whereas clinical thresholds and fit measures are used in the process of formulating prediction targets. To assess the prediction quality in this setting, we employ K-

fold cross-validation,^{33–35} which is a widely used validation method in statistical learning. Our use of cross-validation is unique in that it is applied to selection of prediction targets instead of selection of predictors. As we utilize the benefits of both unsupervised learning (using GMM to formulate candidate targets) and supervised learning (selecting prediction targets based on prediction quality), the proposed approach can be viewed as a type of semisupervised learning.

We will focus on a prediction setting with a simple binary target in line with LAMS, where separating out children who would maintain low-risk profiles of mania symptoms is considered the critical first step toward efficient clinical practice. We now treat the risk status formulated based on observed outcome measures and GMM approaches as known outcomes and assess the quality of prediction by preselected baseline covariates (antecedent validators). The prediction relationship is modeled using a simple logistic regression as

$$\begin{aligned} P(L_i = 1 \mid \mathbf{W}_i) &= \pi_{L_i}(\mathbf{W}_i) \\ P(L_i = 0 \mid \mathbf{W}_i) &= 1 - \pi_{L_i}(\mathbf{W}_i) \\ \text{logit}(\pi_{L_i}(\mathbf{W}_i)) &= \gamma_0 + \gamma_1' \mathbf{W}_i \end{aligned} \quad (15)$$

where the binary outcome L_i (low versus elevated risk) represents one of L_i^{obs} , L_i^{gmm1} , and L_i^{gmm2} , defined in equations (2), (12), and (14). The probability of belonging to the low-risk category varies as a function of baseline covariates $\mathbf{W}_i$. In the LAMS application, these covariates include relevant demographic (sex, age, smoking status, health insurance as a proxy for family SES) and clinical (baseline PGBI-10M, depression, bipolar diagnosis) variables. Note that several baseline variables were used in GMM and again in the validation process. This decision was made based on the fact that these variables will help characterize trajectory patterns in GMM and that the prediction targets are formulated by reclassifying the estimated trajectories or trajectory classes based on clinical thresholds (therefore binary prediction targets are not the same as the estimated trajectory classes). To improve the credibility of our validation process given the redundant use of baseline predictors, we omitted the baseline bipolar diagnosis in formation of trajectory classes using GMM and used it only in the validation step. The baseline bipolar diagnosis is clinically considered the most relevant antecedent validator.

Based on equation (15), the predicted risk status of individual i is defined as

$$\hat{L}_i(\mathbf{W}_i) = \begin{cases} 1 \text{ (low risk)} & \text{if } \hat{\pi}_{L_i}(\mathbf{W}_i) > 0.5 \\ 0 \text{ (elevated risk)} & \text{otherwise} \end{cases} \quad (16)$$

As we focus on prediction, a wide array of candidate targets can be evaluated based on one specific standard, prediction quality. The prediction quality is evaluated on the basis of

common measures of classification performance, sensitivity (S) and specificity (P), as defined as

$$S = TP/(TP + FN) \quad (17)$$

$$S = TN/(TN + FP) \quad (18)$$

where true positives (TP) are individuals who are in the low-risk class ($L=1$) and correctly classified as belonging to the low-risk class by our prediction model ($\hat{L}_A(\mathbf{W}_i) = 1$), true negatives (TN) are individuals who are not low risk ($L=0$) and correctly predicted as low risk ($\hat{L}_A(\mathbf{W}_i) = 0$), false positives (FP) are individuals who are not low risk ($L=0$) and incorrectly classified as low risk ($\hat{L}_A(\mathbf{W}_i) = 1$), and false negatives (FN) are individuals who are low risk ($L=1$) and incorrectly classified as not belonging to the low-risk class by our prediction model ($\hat{L}_A(\mathbf{W}_i) = 0$).

In addition, we will use Cohen's kappa,^{36,37} which is a measure of agreement between two categorical variables taking into account the agreement occurring by chance. We use Cohen's kappa as an overall measure of prediction performance as we focus on situations where sensitivity and specificity are equally important (as in LAMS). That is

$$\kappa = [A - \{D \times Q + (1 - D) \times (1 - Q)\}] / [1 - \{D \times Q + (1 - D) \times (1 - Q)\}] \quad (19)$$

where $A = (TP + TN)/(TP + TN + FP + FN)$,
 $D = (TP + FN)/(TP + TN + FP + FN)$,
and $Q = (TP + FP)/(TP + TN + FP + FN)$

To assess the prediction quality taking into account variation across samples, we use K-fold cross-validation, where we randomly divide the total sample into K equal size subsamples ($k = 1, 2, \dots, K$). We set aside one subsample (k th fold) to be used as a validation sample. With the rest of the subsamples (training data), we obtain parameter estimates of the prediction model shown in equation (15). The k th fold is then used to estimate the expected prediction quality when the model estimates are applied to a data set that is not used to obtain the model estimates. This process is repeated K times and then the results are averaged over K results. That is

$$CV_S = \left(\sum_{k=1}^K S_k \right) / K \quad (20)$$

$$CV_P = \left(\sum_{k=1}^K P_k \right) / K \quad (21)$$

$$CV_{\kappa} = \left(\sum_{k=1}^K \kappa_k \right) / K \quad (22)$$

where S_k , P_k , and κ_k are sensitivity, specificity, and Cohen's kappa in the k th fold, and the standard errors of CV_S , CV_P , and CV_{κ} are calculated as

$$SE_S = \sqrt{\text{var}(S_1, \dots, S_K)} / \sqrt{K} \quad (23)$$

$$SE_P = \sqrt{\text{var}(P_1, \dots, P_K)} / \sqrt{K} \quad (24)$$

$$SE_{\kappa} = \sqrt{\text{var}(\kappa_1, \dots, \kappa_K)} / \sqrt{K} \quad (25)$$

4.4 Combining insights from multiple strategies

Even with the guidance by clinical thresholds and model fit measures, prediction targets can be formulated in many different ways. Therefore, we can still end up with several candidate targets that seem similar and clinically meaningful. This is particularly true when dealing with repeated measures of outcomes that tend to change continuously. In contrast, cross-validation of the prediction quality based on antecedent validators provides a straightforward answer to the question of target selection. However, relying solely on the prediction quality can be dangerous as it represents only one of the many aspects of validity. As will be shown in the LAMS application, it is possible to reach a high level of prediction accuracy with a clinically meaningless prediction target. As none of these methods are likely perfect, it is important to jointly use them and combine insights gained within and across different methods.

5 Application to the LAMS study

In the LAMS context, it is of great clinical importance to develop prognostic models that will help to predict the long-term pattern of manic symptoms among pediatric patients who present to outpatient clinical care with concerns in elevated manic symptoms. In particular, separating out patients who would maintain moderate levels of symptoms is considered critical in clinical practice. This intention seems obvious, although it is not self-evident how to formulate prediction targets that best capture the longitudinal outcome patterns. Here we

focus on constructing reliable and valid prediction targets, which is the first step toward developing good prognostic models.

We first formulated prediction targets using a single observed outcome measure as defined in equation (1). We used three different thresholds ($\tau = 11, 12, 13$) to define the low-risk status at each assessment point. In this approach, the use of clinical thresholds and the prediction quality based on relevant baseline measures (antecedent validators) support the validity of formulated targets. Table 2 shows the performance of formulated prediction targets. We aim to identify prediction targets that will lead to good prediction quality, preferably with sensitivity and specificity of at least 0.7, and Cohen's kappa above 0.6. The results show that specificity and overall prediction quality (Cohen's κ) are quite low and further decrease when predicting targets farther away from the baseline. It is also shown that specificity and overall prediction quality decrease with higher thresholds.

We then formulated prediction targets using observed repeated outcome measures as defined in equation (2). Table 3 shows that the prediction quality is considerably better than when we used prediction targets formulated based on a single assessment (Table 2). To take into account potential fluctuation due to measurement error, we allowed some deviation from the threshold (δ of 0–2). With $\tau = 11$ or $\tau = 12$, we reached the best prediction quality ($\kappa = 0.65$ or 0.60) with a small deviation ($\delta = 1$) from the threshold. We also examined the prediction quality with τ of 13–17, which is higher than the clinically recommended threshold.^{2–4} We did not allow any deviation (i.e. $\delta = 0$) from these high thresholds. The prediction quality turned out to be better than when we used lower thresholds. However, one needs to be careful with such classification as it may not safely separate individuals with potentially problematic patterns of manic symptoms. For example, having PGBI-10M score of 12 throughout the two-year period meets the low-risk standard defined with $\tau = 13$, although consistently having this level of manic symptoms is concerning from the clinical standpoint. The highest prediction quality is achieved with $\tau = 16$, although categorizing individuals approaching such high threshold at one or more times as low risk would be a risky choice. Closely monitoring individual cases based on clinically meaningful thresholds seems to be essential when using this method.

Another way of characterizing longitudinal patterns of manic symptoms is to utilize empirical model fitting. For this purpose we used GMM with relatively simple model specifications as described in Section 3. We used baseline variables such as age, sex, Medicaid, and CDRS-R (depression) as predictors of both growth parameters and the trajectory class membership. We did not include the baseline bipolar diagnosis due to its potentially low frequencies in the trajectory classes with mild to moderate manic symptoms. More importantly, we can avoid using the same set of predictors in formation of trajectory classes (GMM) and in the validation process by excluding bipolar diagnosis, which is considered a key antecedent validator. With this setting, 40 models with 1–8 trajectory classes reached normal convergence. Among these, 10 models were selected based on BIC and/or BLRT. Table 4 lists these 10 models that will be further considered as candidate models to be used to generate prediction targets.

One way to classify individual patients using GMM solutions is to classify trajectory classes based on a clinical threshold. In this approach, all individuals in the same trajectory class are assigned the same risk status (low or elevated risk) as described in equation (12). Based on the previously suggested clinical threshold (e.g. PGBI-10M<12 as low risk), we used $\tau=11$ and $\tau=12$. We did not allow any deviation from the threshold (i.e. $\delta=0$) given that model-based smoothing has been already applied and that there are only five assessments. Figure 2 shows examples of how trajectory classes are divided into the low- and elevated-risk categories based on their estimated mean trajectories. As the low-risk trajectories are clearly separated from the rest, the choice of threshold (11 or 12) did not affect the categorization results (the horizontal line in Figure 2 represents $\tau=11$).

Table 5 shows how the categorization of trajectory classes described above is related to the baseline bipolar diagnosis, which is considered the most relevant clinical antecedent. It is shown that, in all 10 models, bipolar diagnosis at baseline is clearly less frequent in the trajectory classes that are categorized as low risk, supporting the validity of prediction targets generated based on GMM and the clinical threshold.

We now formally examine the prediction quality based on the prediction model in equation (15) and the full set of baseline variables that were preselected as demographically and clinically relevant antecedent validators (sex, age, smoking status, health insurance status, baseline PGBI-10M, depression, bipolar diagnosis). Table 6 shows the prediction quality when prediction targets were formulated by classifying estimated trajectory classes (i.e. L_i^{gmm1} in equation (12)). The choice of threshold ($\tau=11$ or 12) did not affect the categorization results. It is shown that the six-class random intercept model (IN6i^{gmm1}) performed the best with both sensitivity and specificity above 0.8 and Cohen's kappa above 0.7. The next tier of models (VN3i^{gmm1}, VN5i^{gmm1}, VN4i^{gmm1}, IN4i^{gmm1}) also performed quite well with sensitivity and specificity above 0.7 and Cohen's kappa above 0.6.

Focusing on the top three GMM models that led to the best performing prediction targets (Table 6), we examined how different classification methods affect the results. We now compare the prediction quality when prediction targets are formulated by classifying estimated trajectory classes (i.e. L_i^{gmm1} defined in equation (12)) and by directly applying the clinical threshold to estimated individual trajectories (i.e. L_i^{gmm2} defined in equation (14)).

We used the same clinical threshold of 11 or 12 for both methods (shown in the subscript as $\tau=11$ or $\tau=12$ in Table 7). In Table 7, the best prediction quality (sensitivity=0.83, specificity=0.89, $\kappa=0.72$) is now shown when using the prediction target formulated by directly applying the clinical threshold of $\tau=12$ to estimated individual trajectories based on the six-class model (IN6i^{gmm2} _{$\tau=12$}). However, across the three GMM models, the two classification methods produced very comparable results in terms of the prediction quality and the proportion of individuals categorized as low risk.

We formulated a series of prediction targets by jointly utilizing clinical insights, empirical model fitting, and validation through antecedent validators. We selected two best prediction targets formulated based on observed repeated measures (OBS ^{δ_1} _{$\tau=11$} and OBS ^{δ_1} _{$\tau=12$} in Table 3)

and two based on GMM ($IN6i^{gmm1}$ and $IN6i_{\tau12}^{gmm2}$ in Table 7). Given that both approaches used the same threshold of 11 or 12 and that GMM-based prediction targets performed better in terms of prediction by antecedent validators, it seems reasonable to prefer GMM-based prediction targets. However, the difference in prediction quality was not serious, leaving some uncertainty in our choice. As a way of ensuring that we select a clinically more sensible prediction target, we visually examined the individual cases that show discrepancy in classification.

The prediction targets using the results of the six-class GMM model are consistent across different classification methods ($IN6i^{gmm1}$ or $IN6i_{\tau12}^{gmm2}$) and different thresholds ($\tau=11$ or $\tau=12$). The classification results are in fact identical between $IN6i^{gmm1}$ and $IN6i_{\tau11}^{gmm2}$. There are only five out of 671 patients who were differently classified between $IN6i^{gmm1}$ ($= IN6i_{\tau11}^{gmm2}$) and $IN6i_{\tau12}^{gmm2}$, which is not surprising given the clear separation of low-risk trajectory classes around the threshold of 11 or 12. The prediction targets using the observed repeated measures ($OBS_{\tau11}^{\delta1}$ versus $OBS_{\tau12}^{\delta1}$) showed more discrepancies between the thresholds used ($\tau=11$ versus $\tau=12$). There were 23 out of 352 patients with complete data who were differently categorized. Between the prediction targets based on empirical model fitting and observed measures (e.g. $IN6i_{\tau11}^{gmm2}$ versus $OBS_{\tau11}^{\delta1}$), there were 33 out of 352 patients with complete data who were differently categorized.

Figure 3 shows a few examples of patients categorized differently depending on the method of formulating prediction targets. Patient A started with a high level of manic symptoms. Based on his or her declining trend, it may be still fine to classify the patient as having a low-risk pattern. However, the initial level is very high (observed baseline PGBI-10M>20), which makes classifying as low risk a somewhat risky choice. Patient B also qualifies as low risk using the prediction targets formulated based on observed measures. From the clinical point of view, this can be another risky choice as the person's PGBI-10M score is near 20, especially at a later assessment. Patient C starts with a high PGBI-10M and stays around the threshold. Given the bipolar diagnosis and the high depression level (CDRS=56) at baseline, this patient is a good candidate for close monitoring, which makes classifying as elevated risk more reasonable. Patient D is a more ambiguous case and is classified as low risk using the targets based on GMM, but as elevated risk using the targets based on observed measures. As shown in Figure 3, the GMM-based prediction targets tend to classify uncertain cases as elevated risk. Overall, the discrepancy in classification is quite low among the prediction targets that were selected as good targets based on the joint use of clinical insights, empirical model fitting, and validation through baseline antecedents. An additional step of carefully examining cases that are differently classified across different methods will lead to a better informed choice.

6 Monte Carlo simulation study

Our strategy of selecting prediction targets does not solely rely on evaluation based on antecedent validators. However, good performance of this method has an important implication that a large pool of candidate targets based on multiple models can be swiftly

and reliably processed, significantly narrowing the number of candidate targets to be closely examined. The key assumption in this method is that the relationship between antecedent validators and the prediction targets constructed based on GMM estimation (L_i^{gmm1} or L_i^{gmm2}) reflects the relationship between antecedent validators and the true target (true L_i).

Having consistency between the two relationships is important as it assures that the proposed strategy will not misguide our target selection in real data applications, where the true target (i.e. true longitudinal risk status) is unknown or latent. Given that both the true and GMM-based surrogate targets are predicted by the same antecedent validators, this assumption seems reasonable, although it is not clear how strongly and consistently the two relationships are connected. We conducted Monte Carlo simulations with 1000 replications, each with a sample size of 500, to examine the consistency between the two relationships under different scenarios.

6.1 Simulation design

We considered four different scenarios illustrated in Figure 4. The true scenarios have either two trajectory classes ($j=1, 2$ in Scenarios A to C) or a single trajectory class ($j=1$ in Scenario D). Under these four scenarios, the outcome data were generated based on equations (3) to (7). The outcome residual ε_{ijt} is diagonal and normally distributed with class-invariant variance of 0.5 at all time points. We used time scores of $S_t = (0.0, 0.25, 0.5, 0.75, 1)$ and $S_t^2 = (0.0, 0.0625, 0.25, 0.5625, 1.0)$ for the linear and quadratic growth. We used simple random intercept models with class-invariant residual random effect ζ_{1j} distributed as $N(0, 0.5)$. In all scenarios, we used three uncorrelated covariates (X_1, X_2, X_3) that are normally distributed as $N(0, 1)$.

In two class scenarios (A–C), all three covariates (antecedent validators) predict the trajectory type (low versus elevated risk) with an odds ratio of 0.3 or 3.33. That is, the logit coefficients $\beta_1 = (1.204, -1.204, 1.204)$ for $X = (X_1, X_2, X_3)$ and the logit intercept $\beta_0 = 0$ in equation (7). According to this setting, about 38% of individuals belong to the low-risk class ($C=1$). In Scenario A, the two classes are 1–3 standard deviations apart (for class 1, $a_{11} = 11, a_{21} = -1, a_{31} = 0$; for class 2, $a_{12} = 14, a_{22} = -8, a_{32} = 6$). In Scenario B, the two classes are 0–2 standard deviations apart (for class 1, $a_{11} = 11, a_{21} = -1, a_{31} = 0$; for class 2, $a_{12} = 13, a_{22} = -8, a_{32} = 6$). The two classes substantially overlap in Scenario C (for class 1, $a_{11} = 11.5, a_{21} = -1, a_{31} = 0$; for class 2, $a_{12} = 13, a_{22} = -8, a_{32} = 6$). In the one class scenario (D), the three covariates directly affect the growth trajectories with a large effect size (Cohen's $d=0.7$). That is, the regression coefficients $\lambda_1 = (-0.7, 0.0, 0.0)$, $\lambda_2 = (0.0, -0.7, 0.0)$, and $\lambda_3 = (0.0, 0.0, 0.7)$ for $X = (X_1, X_2, X_3)$ in equations (4) to (6).

6.2 Trajectory estimation using GMM

For model estimation, we used GMM with one, two, and three classes. In analyzing the data generated under the two-class scenarios (A to C), we used random intercept models for GMM with one and two classes, but used a three-class model with no random effects to avoid convergence and/or identification problems. Similarly, in analyzing the data generated under the one-class scenario (D), we allowed a random intercept for the one-class model, but not in GMM with two and three classes. We used two out of three covariates as predictors of

class membership and growth parameters. That is, $\mathbf{X} = (X_1, X_3)$ in model estimation. We excluded one covariate (X_2) in model estimation so that the covariates used in the validation step are not completely redundant with those used as predictors of growth trajectories. This is similar to our LAMS application, where the baseline bipolar diagnosis was used only as an antecedent validator. With these deviations from the data generation models, the analysis models follow the model specifications in equations (3) to (7).

6.3 Evaluation of prediction targets

Using the GMM estimation results, we constructed prediction targets (L_i^{gmm1} or L_i^{gmm2}) based on the definitions in equations (12) and (14). For the one-class analysis model, only the definition in equation (14) was used. We used 12 as the threshold for both definitions, which was intended in the true scenarios. These prediction targets were then evaluated based on the quality of prediction by antecedent validators ($\mathbf{W} = (X_1, X_2, X_3)$) using a simple prediction model in equation (15). We used 10-fold cross-validation for this evaluation. Note that the performance of a prediction target should be interpreted in relation to the performance of other candidate targets. In our simulation study, the best Cohen's kappa estimates are below 0.55 compared to above 0.7 in the LAMS application. The weaker association between antecedent validators and prediction targets was intended so that the simulation results can be applied to a wide range of applications with moderate to high association.

We first examined the predictive relationship between antecedent validators and the prediction targets constructed based on GMM estimation. In this method, GMM-based targets, which can be thought of as surrogates of the true risk status, are used as known outcomes for both training and validation in the cross-validation. As demonstrated in the LAMS application, this is the method that can be used in practice in the absence of the information on the true risk status (true L_j). The cross-validation results are summarized in Table 8. The prediction targets constructed based on two-class GMM consistently showed the best performance when the true scenarios involve two trajectory classes (A–C). The prediction targets based on the definition in equation (12) performed better than those based on the definition in equation (14), which is expected given that two classes were assumed in the data generation process. It is also shown that the performance of the targets based on the definition in equation (14) deteriorates as the distance between the two trajectory classes narrows, whereas the performance of the targets based on the definition in equation (12) remains at the same level. This is also expected because the two trajectory classes overlap more as the distance narrows, weakening the connection between the definition in equation (14) and the trajectory class membership. If the relationship between the class membership and the baseline predictors is weak, the same would happen with the definition in equation (12). In such situations, however, the idea itself of using antecedent validators would be meaningless to begin with. This finding also emphasizes the importance of using good predictors of class membership in the GMM step, in particular, when the separation of classes is not strong.

Table 8 also shows the results when the true scenario (D) has one trajectory class. The prediction targets based on the definition in equation (14) performed the best using GMM

with one or three classes. The performance of the target based on the two-class model was somewhat worse. Note that, in analyzing the data from Scenario D, we did not allow random effects in GMM with two or three classes. To compensate for the restriction and to capture the longitudinal heterogeneity, allowing for more classes is one solution, which explains the better performance by the target based on the three-class model. These results will be considered a failure when the goal is to recover the true model (i.e. one class), although a success when the goal is to predict the outcome at the individual level.

Table 9 shows the cross-validation results when the prediction models were trained with prediction targets based on GMM estimation as in Table 8, but validated against the true trajectory type (true L_j), which is known in Monte Carlo simulations. Under the two-class scenarios, the true targets were best predicted using the predictive models trained with prediction targets based on two-class GMM estimation. As in Table 8, the proximity between the two trajectory classes negatively affects the prediction quality when the definition in equation (14) is used, although its impact does not appear until the two trajectory classes substantially overlap (Scenario C). In real data applications without the information in Table 9, we will miss L_i^{gmm2} under Scenario B (see Table 8). This is fine for the practical purpose of prediction as the prediction models trained using the preferred (L_i^{gmm1}) and missed (L_i^{gmm2}) targets have similar relationships with the true target. Under Scenario D, we treated the prediction target based on estimated trajectories using the one-class model as the true target (same as one-class L_i^{gmm2} in Table 8). In this scenario, the prediction models trained with the prediction targets based on the definition in equation (14) showed the best prediction quality regardless of the number of classes used in GMM estimation. However, in real data applications, we will miss two-class L_i^{gmm2} (see Table 8) as we will not have the information provided in Table 9. This is harmless for the purpose of prediction because the prediction models trained using the selected (one- or three-class L_i^{gmm2}) and missed (two-class L_i^{gmm2}) targets have almost the same relationship with the true target. The simulation results in Tables 8 and 9 showed that the relationship between antecedent validators and the prediction targets constructed based on the definition in equation (12) and GMM estimation (i.e. L_i^{gmm1}) perfectly mirrors the relationship between antecedent validators and the true targets (true L_j). The connection between the two relationships is less strong when using the definition in equation (14) (i.e. L_i^{gmm2}), although it is unlikely to misguide our target selection as long as we select prediction targets with the best cross-validation results.

7 Conclusions

With growing interest in building prognostic models in various fields of medicine, much effort has been concentrated in developing algorithms that facilitate effective utilization of large amount of potential predictor variables. In this study, we took a step back and focused on constructing reliable and valid prediction targets, which is the critical first step toward establishing good prognostic models. On the basis of the joint use of clinical insights, empirical model fitting, and cross-validation of prediction by antecedent validators, this

study demonstrated the possibilities of constructing prediction targets that are not only clinically sensible but also empirically supported.

In line with LAMS, we focused on a situation where the goal is to develop a prognostic model that will accurately predict each future patient's long-term outcome course. To align with this intention, we constructed prediction targets based on repeated outcome measures by using clinical thresholds and empirical model fitting (unsupervised learning via GMM). These prediction targets are connected through the same clinical thresholds, and therefore the agreement between them can be interpreted as supporting the validity of each other. To further refine our validation procedure and to help narrow our options, we evaluated using cross-validation how well the formulated targets are predicted (supervised learning) by demographically and clinically relevant baseline characteristics (antecedent validators). Our use of K-fold cross-validation is unique in that it is applied to selection of prediction targets instead of selection of predictors.

Validation of prediction targets based on their association with relevant variables can be expanded by utilizing other available information in the data, which we did not pursue in this study. For example, a good prediction target should well predict future outcomes they are supposed to predict (predictive validity) and should be highly correlated with concurrent outcomes they are supposed to be correlated with (concurrent validity). If possible, we should also aim to examine the generalizability of our findings by testing with independent data. Given the importance of constructing valid prediction targets, there is no reason why we should not maintain the same level of rigor in construction of prediction targets as in selection of predictors.

Once we successfully construct a prediction target that is worthy of predicting, we can move on to the next step with more confidence and focus on developing the intended prognostic model. Variable selection methods used in statistical learning can be employed in search of various features (beyond the preselected antecedent validators) that maximize the prediction of the selected target. The LAMS study collected rich data in multiple domains using a relatively large sample from nine diverse clinics in four geographical areas, which provides an opportunity for developing generalizable prognostic models with practical level of accuracy. We leave this topic for future investigation.

Acknowledgments

We appreciate insightful discussion and feedback from the Prevention Science Methodology Group (PSMG). We also thank Helena Kraemer and Bengt Muthén for their valuable input. Dr. Findling receives or has received research support, acted as a consultant and/or served on a speaker's bureau for Alcobra, American Academy of Child & Adolescent Psychiatry, American Physician Institute, American Psychiatric Press, AstraZeneca, Bracket, Bristol-Myers Squibb, CogCubed, Cognition Group, Coronado Biosciences, Dana Foundation, Elsevier, Forest, GlaxoSmithKline, Guilford Press, Johns Hopkins University Press, Johnson and Johnson, Jubilant Clinsys, KenPharm, Lilly, Lundbeck, Merck, NIH, Neurim, Novartis, Noven, Otsuka, Oxford University Press, Pfizer, Physicians Postgraduate Press, Purdue, Rhodes Pharmaceuticals, Roche, Sage, Shire, Sunovion, Supernus Pharmaceuticals, Transcept Pharmaceuticals, Validus, and WebMD. Dr. Arnold has received research funding from Curemark, Forest, Lilly, Neuropharm, Novartis, Noven, Shire, and YoungLiving (as well as NIH and Autism Speaks) and has consulted with or been on advisory boards for Gowlings, Neuropharm, Novartis, Noven, Organon, Otsuka, Pfizer, Roche, Seaside Therapeutics, Sigma Tau, Shire, and Tris Pharma and received travel support from Noven. Dr. Fristad receives royalties from Guilford Press, American Psychiatric Press, and Child & Family Psychological Services. Dr. Youngstrom has consulted with Lundbeck and Otsuka about assessment.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the National Institute on Drug Abuse (DA031698) and National Institute of Mental Health (Johns Hopkins University: R01 MH073967-06A1, Cincinnati Children's Hospital Medical Center: MH073816-06A1, The Ohio State University: MH073801-06A1, University of Pittsburgh: MH073953-06A1).

References

- Findling RL, Jo B, Frazier TW, et al. The 24-month course of manic symptoms in children. *Bipolar Disord.* 2013; 15:669–679. [PubMed: 23799945]
- Findling RL, Youngstrom EA, Fristad MA, et al. Characteristics of children with elevated symptoms of mania: the Longitudinal Assessment of Manic Symptoms (LAMS) study. *J Clin Psychiatry.* 2010; 71:1664–1672. [PubMed: 21034685]
- Horwitz SM, Demeter CA, Pagano ME, et al. Longitudinal Assessment of Manic Symptoms (LAMS) study: background, design, and initial screening results. *J Clin Psychiatry.* 2010; 71:1511–1517. [PubMed: 21034684]
- Youngstrom EA, Birmaher B, Findling RL. Pediatric bipolar disorder: validity, phenomenology, and recommendations for diagnosis. *Bipolar Disord.* 2008; 10:194–214. [PubMed: 18199237]
- Singer JD, Willett JB. *Applied longitudinal data analysis: modeling change and event occurrence.* New York, NY: Oxford University Press; 2003.
- Raudenbush SW, Bryk AS. *Hierarchical linear models: applications and data analysis methods.* Thousand Oaks, CA: Sage; 2002.
- 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. 291–322.
- 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. 345–368.
- Muthén BO, Asparouhov T. Growth mixture modeling: analysis with non-Gaussian random effects. In: Fitzmaurice G, Davidian M, Verbeke G, et al., editors *Longitudinal data analysis.* Boca Raton, FL: Chapman and Hall/CRC Press; 2009. 143–165.
- Muthén BO, Asparouhov T. Growth mixture modeling with non-normal distributions. *Stat Med.* 2015; 34:1041–1058. [PubMed: 25504555]
- Muthén BO, Brown CH, Masyn K, et al. General growth mixture modeling for randomized preventive interventions. *Biostatistics.* 2002; 3:459–475. [PubMed: 12933592]
- Muthén BO, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics.* 1999; 55:463–469. [PubMed: 11318201]
- Dunn LB, Cooper BA, Neuhaus J, et al. Identification of distinct depressive symptom trajectories in women following surgery for breast cancer. *Health Psychol.* 2011; 30:683–692. [PubMed: 21728421]
- Gueorguieva R, Mallinckrodt C, Krystal JH. Trajectories of depression severity in clinical trials of duloxetine insights into antidepressant and placebo responses. *Arch Gen Psychiatry.* 2011; 68:1227–1237. [PubMed: 22147842]
- 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. 193–214.
- Jo B, Wang C-P, Jalongo NS. Using latent outcome trajectory classes in causal inference. *Stat Interf.* 2009; 2:403–412.
- Kellam SG, Brown CH, Poduska JM, et al. Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. *Drug Alcohol Depend.* 2008; 95:S5–S28. [PubMed: 18343607]

18. Rodriguez D, Audrain-McGovern J. Team sport participation and smoking: analysis with general growth mixture modeling. *J Pediatr Psychol*. 2004; 29:299–308. [PubMed: 15148352]
19. Stulz N, Thase ME, Klein DN, et al. Differential effects of treatments for chronic depression: a latent growth model reanalysis. *J Consult Clin Psychol*. 2010; 78:409–419. [PubMed: 20515216]
20. van Lier PAC, Muthén BO, van der Sar RM, et al. Preventing disruptive behavior in elementary schoolchildren: impact of a universal classroom-based intervention. *J Consult Clin Psychol*. 2004; 72:467–478. [PubMed: 15279530]
21. Wang C-P, Jo B, Brown CH. Causal inference in longitudinal comparative effectiveness studies with repeated measures of a continuous intermediate variable. *Stat Med*. 2014; 33:3509–3527. [PubMed: 24577715]
22. Dempster A, Laird N, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B*. 1977; 39:1–38.
23. Little RJA, Rubin DB. Statistical analysis with missing data. New York, NY: Wiley; 2002.
24. McLachlan GJ, Krishnan T. The EM algorithm and extensions. New York, NY: Wiley; 1997.
25. Tanner M. Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions. New York, NY: Springer; 1996.
26. Muthén LK, Muthén BO. Mplus user's guide. 7. Los Angeles, CA: Muthén and Muthén; 1998. 2012
27. Schwartz G. Estimating the dimension of a model. *Ann Stat*. 1978; 6:461–464.
28. McLachlan G, Peel D. Finite mixture models. New York, NY: Wiley; 2000.
29. 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. *Struct Eq Model*. 2007; 14:535–569.
30. 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. 317–341.
31. Muthén BO. Statistical and substantive checking in growth mixture modeling: comment on Bauer and Curran. *Psychol Methods*. 2003; 8:369–377. [PubMed: 14596497]
32. 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, NY: Springer; 2010. 69–100.
33. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York, NY: Springer; 2009.
34. James G, Witten D, Hastie T, et al. An introduction to statistical learning with applications in R. New York, NY: Springer; 2013.
35. Stone M. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc Ser B*. 1974; 36:111–147.
36. Kraemer HC, Periyakoil VS, Noda A. Tutorial in biostatistics: kappa coefficients in medical research. *Stat Med*. 2002; 21:2109–2129. [PubMed: 12111890]
37. Landis JR, Koch GG. The measure of observer agreement for categorical data. *Biometrics*. 1977; 33:159–174. [PubMed: 843571]

Figure 1.
Distribution of the PGBI-10M score.
PGBI-10M: Parent General Behavior Inventory 10-item Mania Form.

Figure 2.
Examples of candidate GMM solutions.
GMM: growth mixture modeling.

Figure 3.

Disagreement in classification across differently formulated prediction targets.

Figure 4.
Four scenarios used in Monte Carlo simulations.

Table 1

LAMS sample statistics.

	N	Min	Max	Mean	SD	Skewness(SE)	Kurtosis(SE)
<i>Outcome</i>							
PGBI-10M 0m	671	0	30	12.8	7.2	0.23 (0.13)	-0.78 (0.26)
PGBI-10M 6m	536	0	30	10.5	6.7	0.49 (0.13)	-0.27 (0.26)
PGBI-10M 12m	520	0	30	8.6	6.6	0.71 (0.13)	0.01 (0.26)
PGBI-10M 18m	476	0	27	8.4	6.1	0.60 (0.13)	-0.21 (0.26)
PGBI-10M 24m	445	0	30	8.1	6.5	0.75 (0.13)	-0.02 (0.26)
<i>Baseline characteristics</i>							
Age	682	6.1	13.2	9.4	1.9		
Depression (CDRS-R)	682	17	73	34.7	10.7		
Female	682	0	1	0.33			
Medicaid	682	0	1	0.42			
Bipolar	682	0	1	0.23			
Smoking	682	0	1	0.05			

CDRS-R: children's depression rating scale-revised; LAMS: longitudinal assessment of manic symptoms; PGBI-10M: Parent General Behavior Inventory 10-item Mania Form; SD: standard deviation; SE: standard error.

Table 2

Average sensitivity, specificity, and Cohen's kappa from 10-fold cross-validation in prediction of low risk defined with a single observed outcome at each assessment point and a fixed threshold (u_{it}^{obs} in equation (1)). Listwise deletion has been applied.

Target	%Low risk	Sensitivity	Specificity	Cohen's κ
<i>Low-risk criteria: observed PGBI-10M < 11</i>				
PGBI-10M 12m	65.0	0.864 (0.849, 0.879)	0.444 (0.397, 0.490)	0.322 (0.272, 0.372)
PGBI-10M 18m	64.9	0.835 (0.807, 0.864)	0.427 (0.389, 0.466)	0.274 (0.246, 0.301)
PGBI-10M 24m	69.9	0.884 (0.867, 0.901)	0.309 (0.263, 0.355)	0.209 (0.159, 0.258)
<i>Low-risk criteria: observed PGBI-10M < 12</i>				
PGBI-10M 12m	68.3	0.894 (0.879, 0.908)	0.401 (0.341, 0.461)	0.306 (0.253, 0.360)
PGBI-10M 18m	71.0	0.882 (0.860, 0.905)	0.360 (0.324, 0.395)	0.266 (0.227, 0.304)
PGBI-10M 24m	73.9	0.914 (0.895, 0.933)	0.213 (0.173, 0.252)	0.153 (0.093, 0.213)
<i>Low-risk criteria: observed PGBI-10M < 13</i>				
PGBI-10M 12m	74.6	0.922 (0.910, 0.934)	0.302 (0.256, 0.347)	0.248 (0.198, 0.299)
PGBI-10M 18m	74.4	0.913 (0.895, 0.932)	0.267 (0.224, 0.310)	0.206 (0.151, 0.260)
PGBI-10M 24m	77.8	0.947 (0.933, 0.962)	0.168 (0.125, 0.211)	0.145 (0.083, 0.207)

PGBI-10M: Parent General Behavior Inventory 10-item Mania Form.

Average CV results $\pm$ SE are shown in parentheses.

Table 3

Average sensitivity, specificity, and Cohen's kappa from 10-fold cross-validation in prediction of low-risk trajectory type based on observed repeated outcome measures and a fixed threshold (L_i^{obs} in equation (2)). Study participants with all five assessments are included (N = 352).

Target	%Low risk	Sensitivity	Specificity	Cohen's κ
<i>Low-risk criteria: observed PGBI-10M < 11</i>				
OBS $_{\tau 11}^{\delta 2}$	62.2	0.841 (0.811, 0.871)	0.645 (0.583, 0.706)	0.485 (0.417, 0.552)
OBS $_{\tau 11}^{\delta 1}$	47.2	0.818 (0.787, 0.849)	0.836 (0.803, 0.868)	0.650 (0.603, 0.698)
OBS $_{\tau 11}^{\delta 0}$	26.1	0.634 (0.570, 0.697)	0.889 (0.873, 0.904)	0.518 (0.464, 0.573)
<i>Low-risk criteria: observed PGBI-10M < 12</i>				
OBS $_{\tau 12}^{\delta 2}$	68.2	0.865 (0.840, 0.890)	0.550 (0.482, 0.617)	0.423 (0.362, 0.484)
OBS $_{\tau 12}^{\delta 1}$	53.7	0.825 (0.797, 0.853)	0.778 (0.733, 0.824)	0.599 (0.548, 0.650)
OBS $_{\tau 12}^{\delta 0}$	30.7	0.712 (0.646, 0.777)	0.875 (0.855, 0.895)	0.576 (0.520, 0.632)
<i>Low-risk criteria: observed PGBI-10M < 13</i>				
OBS $_{\tau 13}^{\delta 0}$	36.9	0.798 (0.761, 0.834)	0.890 (0.867, 0.913)	0.675 (0.631, 0.719)
<i>Low-risk criteria: observed PGBI-10M < 14</i>				
OBS $_{\tau 14}^{\delta 0}$	42.3	0.820 (0.792, 0.848)	0.858 (0.834, 0.881)	0.665 (0.618, 0.712)
<i>Low-risk criteria: observed PGBI-10M < 15</i>				
OBS $_{\tau 15}^{\delta 0}$	47.2	0.850 (0.821, 0.879)	0.840 (0.815, 0.864)	0.680 (0.630, 0.730)
<i>Low-risk criteria: observed PGBI-10M < 16</i>				
OBS $_{\tau 16}^{\delta 0}$	52.0	0.854 (0.832, 0.877)	0.868 (0.837, 0.900)	0.713 (0.665, 0.762)
<i>Low-risk criteria: observed PGBI-10M < 17</i>				
OBS $_{\tau 17}^{\delta 0}$	57.1	0.840 (0.818, 0.863)	0.822 (0.786, 0.858)	0.650 (0.601, 0.700)

OBS: target based on observed repeated measures; PGBI-10M: Parent General Behavior Inventory 10-item Mania Form.

Average CV results $\pm$ SE are shown in parentheses. Subscript $\delta 2$ means that PGBI-10M below the threshold 3–5 times ($\delta = 2$), $\delta 1$ means 4–5 times ($\delta = 1$), and $\delta 0$ means all 5 times ($\delta = 0$).

Table 4

GMM models selected by BIC and BLRT within each type of random effect and residual variance structure.

Model Label	Model description	# Parameter	BIC	%Low risk
<i>Class-invariant, normal distribution</i>				
IN2 _{is} ^{a,b}	Two-class, random int/slope	31	16,451	86.8
IN3 _i ^a	Three-class, random intercept	37	16,463	74.9
IN6 _i ^b	Six-class, random intercept	61	16,510	41.2
IN4 ^a	Four-class, no random effect	44	16,516	34.3
IN5 ^a	Five-class, no random effect	52	16,519	31.5
<i>Class-invariant, skew-t distribution</i>				
IT2 ^a	Two-class, no random effect	36	16,352	57.9
<i>Class-varying, normal distribution</i>				
VN2 _{is} ^{a,b}	Two-class, random int/slope	38	16,269	22.7
VN3 _i ^a	Three-class, random intercept	49	16,235	40.6
VN4 ^a	Four-class, no random effect	59	16,277	38.3
VN5 ^b	Five-class, no random effect	72	16,279	43.7

BIC: Bayesian information criteria; BLRT: bootstrapped likelihood ratio test; GMM: growth mixture modeling.

^aBIC.

^bBLRT.

Table 5

Proportions of bipolar diagnosis at baseline across GMM trajectory classes.

GMM Model	Trajectory class					
	1	2	3	4	5	6
IN6i	0.085	0.304	0.299	0.438	0.481	0.286
IN5	0.060	0.207	0.307	0.433	0.435	
IN4	0.073	0.255	0.405	0.432		
IN3i	0.170	0.388	0.426			
IN2is	0.198	0.433				
VN5	0.024	0.103	0.090	0.317	0.436	
VN4	0.038	0.077	0.298	0.435		
VN3i	0.032	0.102	0.326			
VN2is	0.097	0.268				
IT2	0.139	0.352				

GMM: growth mixture modeling.

Proportions of bipolar in trajectory classes categorized as low risk are in boldface.

Table 6

Average sensitivity, specificity, and Cohen's kappa from 10-fold cross-validation in prediction of low-risk trajectory type formulated by classifying estimated trajectory classes based on GMM (L_i^{gmm1} in equation (12)).

Target	%Low risk	Sensitivity	Specificity	Cohen's κ
IN6i ^{gmm1}	41.2	0.817 (0.795, 0.839)	0.890 (0.875, 0.906)	0.708 (0.681, 0.735)
VN3i ^{gmm1}	40.6	0.772 (0.748, 0.796)	0.867 (0.853, 0.881)	0.640 (0.614, 0.666)
VN5 ^{gmm1}	43.7	0.777 (0.752, 0.803)	0.849 (0.828, 0.869)	0.624 (0.588, 0.661)
VN4 ^{gmm1}	38.3	0.740 (0.718, 0.762)	0.875 (0.851, 0.898)	0.618 (0.585, 0.652)
IN4 ^{gmm1}	34.3	0.724 (0.687, 0.761)	0.886 (0.865, 0.906)	0.616 (0.570, 0.662)
IN5 ^{gmm1}	31.5	0.678 (0.640, 0.716)	0.876 (0.856, 0.897)	0.557 (0.512, 0.601)
IT2 ^{gmm1}	57.9	0.810 (0.787, 0.832)	0.643 (0.624, 0.661)	0.458 (0.431, 0.486)
IN3i ^{gmm1}	74.9	0.922 (0.912, 0.933)	0.499 (0.473, 0.525)	0.453 (0.422, 0.484)
VN2is ^{gmm1}	22.7	0.458 (0.420, 0.495)	0.908 (0.897, 0.919)	0.394 (0.349, 0.438)
IN2is ^{gmm1}	86.8	0.991 (0.988, 0.995)	0.101 (0.069, 0.133)	0.137 (0.091, 0.184)

GMM: growth mixture modeling.

CV results $\pm$ SE are shown in parentheses. Study participants with one or more assessments are included (N = 682).

Table 7

Average sensitivity, specificity, and Cohen's kappa from 10-fold cross-validation in prediction of trajectory type (low risk or not) formulated by classifying estimated trajectory classes (L_i^{gmm1} in equation (12)) and by classifying estimated individual trajectories (L_i^{gmm2} in equation (14)) based on GMM.

Target	%Low risk	Sensitivity	Specificity	Cohen's κ
IN6i ^{gmm1}	41.2	0.817 (0.795, 0.839)	0.890 (0.875, 0.906)	0.708 (0.681, 0.735)
VN3i ^{gmm1}	40.6	0.772 (0.748, 0.796)	0.867 (0.853, 0.881)	0.640 (0.614, 0.666)
VN5 ^{gmm1}	43.7	0.777 (0.752, 0.803)	0.849 (0.828, 0.869)	0.624 (0.588, 0.661)
IN6i ^{gmm2} _{τ_{11}}	41.2	0.817 (0.795, 0.839)	0.890 (0.875, 0.906)	0.708 (0.681, 0.735)
VN3i ^{gmm2} _{τ_{11}}	39.3	0.745 (0.718, 0.772)	0.864 (0.846, 0.882)	0.611 (0.580, 0.642)
VN5 ^{gmm2} _{τ_{11}}	43.6	0.777 (0.752, 0.802)	0.847 (0.827, 0.868)	0.622 (0.588, 0.656)
IN6i ^{gmm2} _{τ_{12}}	41.9	0.832 (0.816, 0.848)	0.886 (0.871, 0.902)	0.717 (0.694, 0.740)
VN3i ^{gmm2} _{τ_{12}}	42.4	0.767 (0.747, 0.787)	0.850 (0.829, 0.871)	0.617 (0.582, 0.653)
VN5 ^{gmm2} _{τ_{12}}	43.7	0.777 (0.752, 0.803)	0.849 (0.828, 0.869)	0.624 (0.588, 0.661)

GMM: growth mixture modeling.

CV results $\pm$ SE are shown in parentheses. Study participants with one or more assessments are included (N = 682).

Table 8

Monte Carlo simulations: Average sensitivity, specificity, and Cohen's kappa from 10-fold cross-validation in prediction of low-risk trajectory type when validated against targets based on GMM estimation.

Training target	%Low	Sensitivity	Specificity	Cohen's κ
<i>Scenario A (two-class)</i>				
One-class L_i^{gmm2}	26.0	0.467 (0.421, 0.512)	0.923 (0.909, 0.937)	0.422 (0.374, 0.469)
Two-class L_i^{gmm1}	37.3	0.662 (0.627, 0.697)	0.861 (0.841, 0.881)	0.528 (0.489, 0.567)
Two-class L_i^{gmm2}	35.5	0.619 (0.582, 0.656)	0.868 (0.849, 0.888)	0.497 (0.457, 0.537)
Three-class L_i^{gmm1}	26.9	0.295 (0.259, 0.331)	0.927 (0.913, 0.940)	0.248 (0.206, 0.289)
Three-class L_i^{gmm2}	28.0	0.366 (0.326, 0.406)	0.915 (0.900, 0.930)	0.310 (0.265, 0.355)
<i>Scenario B (two-class)</i>				
One-class L_i^{gmm2}	38.0	0.545 (0.509, 0.581)	0.837 (0.816, 0.859)	0.391 (0.350, 0.432)
Two-class L_i^{gmm1}	36.9	0.652 (0.616, 0.688)	0.862 (0.842, 0.882)	0.520 (0.481, 0.559)
Two-class L_i^{gmm2}	38.6	0.602 (0.567, 0.637)	0.844 (0.822, 0.865)	0.453 (0.412, 0.493)
Three-class L_i^{gmm1}	26.7	0.137 (0.121, 0.152)	0.910 (0.903, 0.917)	0.058 (0.036, 0.080)
Three-class L_i^{gmm2}	37.0	0.502 (0.466, 0.537)	0.864 (0.845, 0.884)	0.381 (0.340, 0.422)
<i>Scenario C (two-class)</i>				
One-class L_i^{gmm2}	27.4	0.265 (0.230, 0.301)	0.931 (0.917, 0.945)	0.227 (0.184, 0.270)
Two-class L_i^{gmm1}	37.0	0.652 (0.617, 0.687)	0.862 (0.842, 0.882)	0.521 (0.482, 0.559)
Two-class L_i^{gmm2}	32.7	0.424 (0.386, 0.462)	0.886 (0.868, 0.904)	0.333 (0.290, 0.375)
Three-class L_i^{gmm1}	22.1	0.047 (0.035, 0.058)	0.983 (0.978, 0.988)	0.037 (0.021, 0.053)
Three-class L_i^{gmm2}	27.4	0.221 (0.189, 0.253)	0.951 (0.939, 0.962)	0.201 (0.163, 0.240)
<i>Scenario D (one-class)</i>				
One-class L_i^{gmm2}	47.7	0.757 (0.729, 0.785)	0.792 (0.766, 0.817)	0.544 (0.508, 0.580)
Two-class L_i^{gmm1}	49.9	0.640 (0.609, 0.670)	0.643 (0.612, 0.673)	0.282 (0.240, 0.323)

Training target	%Low	Sensitivity	Specificity	Cohen's κ
Two-class L_i^{gmm2}	47.8	0.709 (0.679, 0.739)	0.749 (0.722, 0.777)	0.454 (0.415, 0.492)
Three-class L_i^{gmm1}	49.6	0.585 (0.560, 0.609)	0.595 (0.571, 0.619)	0.209 (0.166, 0.253)
Three-class L_i^{gmm2}	47.7	0.753 (0.725, 0.781)	0.788 (0.762, 0.814)	0.537 (0.500, 0.574)

GMM: growth mixture modeling.

Table 9

Monte Carlo simulations: Average sensitivity, specificity, and Cohen's kappa from 10-fold cross-validation in prediction of low-risk trajectory type when validated against true targets (true L_i).

Training target	%Low	Sensitivity	Specificity	Cohen's κ
<i>Scenario A (two-class)</i>				
One-class L_i^{gmm2}	26.0	0.397 (0.360, 0.433)	0.960 (0.949, 0.971)	0.394 (0.354, 0.433)
Two-class L_i^{gmm1}	37.3	0.668 (0.633, 0.703)	0.867 (0.847, 0.887)	0.541 (0.503, 0.580)
Two-class L_i^{gmm2}	35.5	0.627 (0.592, 0.663)	0.889 (0.871, 0.907)	0.531 (0.492, 0.570)
Three-class L_i^{gmm1}	26.9	0.312 (0.280, 0.344)	0.968 (0.959, 0.977)	0.310 (0.274, 0.346)
Three-class L_i^{gmm2}	28.0	0.378 (0.343, 0.414)	0.963 (0.952, 0.973)	0.376 (0.337, 0.415)
<i>Scenario B (two-class)</i>				
One-class L_i^{gmm2}	38.0	0.596 (0.560, 0.631)	0.865 (0.846, 0.885)	0.473 (0.433, 0.512)
Two-class L_i^{gmm1}	36.9	0.653 (0.618, 0.687)	0.868 (0.849, 0.888)	0.529 (0.490, 0.567)
Two-class L_i^{gmm2}	38.6	0.656 (0.621, 0.691)	0.870 (0.851, 0.889)	0.534 (0.496, 0.572)
Three-class L_i^{gmm1}	26.7	0.131 (0.119, 0.144)	0.912 (0.907, 0.918)	0.049 (0.033, 0.066)
Three-class L_i^{gmm2}	37.0	0.542 (0.507, 0.577)	0.889 (0.872, 0.905)	0.448 (0.408, 0.487)
<i>Scenario C (two-class)</i>				
One-class L_i^{gmm2}	27.4	0.268 (0.237, 0.300)	0.964 (0.954, 0.974)	0.262 (0.225, 0.299)
Two-class L_i^{gmm1}	37.0	0.649 (0.615, 0.684)	0.866 (0.847, 0.886)	0.523 (0.485, 0.562)
Two-class L_i^{gmm2}	32.7	0.473 (0.436, 0.510)	0.941 (0.928, 0.954)	0.445 (0.405, 0.485)
Three-class L_i^{gmm1}	22.1	0.030 (0.023, 0.037)	0.978 (0.974, 0.983)	0.010 (0.000, 0.020)
Three-class L_i^{gmm2}	27.4	0.212 (0.185, 0.239)	0.972 (0.964, 0.980)	0.208 (0.176, 0.240)
<i>Scenario D (one-class)</i>				
One-class L_i^{gmm2}	47.7	0.757 (0.729, 0.785)	0.792 (0.766, 0.817)	0.544 (0.508, 0.580)

Training target	%Low	Sensitivity	Specificity	Cohen's κ
Two-class L_i^{gmm1}	49.9	0.579 (0.548, 0.610)	0.574 (0.544, 0.604)	0.151 (0.109, 0.193)
Two-class L_i^{gmm2}	47.8	0.756 (0.728, 0.784)	0.791 (0.765, 0.816)	0.542 (0.506, 0.578)
Three-class L_i^{gmm1}	49.6	0.532 (0.513, 0.551)	0.539 (0.520, 0.557)	0.071 (0.043, 0.098)
Three-class L_i^{gmm2}	47.7	0.757 (0.729, 0.785)	0.791 (0.765, 0.816)	0.543 (0.507, 0.580)