MULTIVARIATE BEHAVIORAL RESEARCH
2023, VOL. 58, NO. 6, 1057–1071



# Reorienting Latent Variable Modeling for Supervised Learning

Booil Jo[a], Trevor J. Hastie[a], Zetan Li[a], Eric A. Youngstrom[b], Robert L. Findling[c], and Sarah McCue Horwitz[d]

[a]Stanford University; [b]University of North Carolina; [c]Virginia Commonwealth University; [d]New York University

**ABSTRACT**

Despite its potentials benefits, using prediction targets generated based on latent variable (LV) modeling is not a common practice in supervised learning, a dominating framework for developing prediction models. In supervised learning, it is typically assumed that the outcome to be predicted is clear and readily available, and therefore validating outcomes before predicting them is a foreign concept and an unnecessary step. The usual goal of LV modeling is inference, and therefore using it in supervised learning and in the prediction context requires a major conceptual shift. This study lays out methodological adjustments and conceptual shifts necessary for integrating LV modeling into supervised learning. It is shown that such integration is possible by combining the traditions of LV modeling, psychometrics, and supervised learning. In this interdisciplinary learning framework, generating practical outcomes using LV modeling and systematically validating them based on clinical validators are the two main strategies. In the example using the data from the Longitudinal Assessment of Manic Symptoms (LAMS) Study, a large pool of candidate outcomes is generated by flexible LV modeling. It is demonstrated that this exploratory situation can be used as an opportunity to tailor desirable prediction targets taking advantage of contemporary science and clinical insights.

**KEYWORDS**

Latent variable modeling; growth mixture modeling; model-based clustering; prediction; psychometrics; supervised learning; clinical validators

## Introduction

With growing interest in personalized medicine and the use of machine learning in medicine, developing risk prediction and prognostic models has been drawing renewed attention. If these models come with good accuracy, they may facilitate faster and better informed prognosis and treatment decisions, and therefore improved patient care. They may also lead to optimized use of resources by minimizing the number of cases that require closer examination and prognosis by clinical experts.

In developing prediction models to be used in clinical practice, the usual goal is to predict a single outcome, which is observed in the sample used for model development, but is unobserved and to be predicted among new patients. In the sample used in model development, abundant information is often available not only on the baseline side, but also on the outcome side. That is, we may have rich multivariate and longitudinal outcome information gathered through various outlets such as research studies and health care services. However, having a simple univariate outcome is critical, especially in using supervised learning methods, because it lets us focus on handling of a large pool of possible predictors of the outcome without worrying about the outcome itself. What is ignored here is that a single observed measure can be unreliable and can be far from a good representation of a particular patient's true outcome status. An ideal solution to this situation would be to create more reliable and valid outcome variables using multivariate outcome information without losing the simplicity of a single observed outcome.

The current practice of building prediction models has much room for improvement in terms of utilization of valuable outcome data. How do we effectively organize and simplify complex outcome data and still preserve its rich information? We consider latent variable (LV) modeling as a promising way to achieve these seemingly conflicting goals. Characterizing individuals using latent subgroups is particularly attractive in the clinical context as they can serve as clinically meaningful and interpretable summary measures of complex multivariate information (e.g., HbA1c patterns in Bayliss et al., 2011; substance use patterns in Beseler et al., 2012; systolic blood pressure patterns in

Joo et al., 2020). LV models' flexible nature is a big advantage in organizing complex multivariate information. However, the same flexibility can also make resulting models esoteric and subjective, which is not desirable in risk prediction in medicine. Further, LV modeling has not been used as a common tool for building prediction models in the supervised learning framework, and therefore does not have established conventions that guide the process.

There have been some developments to improve LV modeling and structural equation modeling by incorporating the concepts and strategies from predictive modeling and machine learning. Cole and Bauer (2016) discussed the importance of examining the individual-level predicted values in the longitudinal context to improve understanding (inference) about the predictive relationship in theory-driven models. In Brandmaier et al. (2013), regression tree methods were introduced to combine exploratory and confirmatory approaches with the goal of improving model building. Our proposed learning framework is unique in that it not only aims to benefit from machine learning strategies, but also aims to provide a framework that will facilitate integration of LV modeling in machine learning. That is, we incorporate machine learning strategies to accelerate and improve exploration using LV modeling. The LV modeling results are then validated and organized to provide better prediction targets that are ready to be used in any supervised learning contexts.

In our previous studies (Jo et al., 2017, 2018), we explored the possibility of building prediction models by utilizing a LV strategy known as growth mixture modeling (GMM, Muthén & Shedden, 1999). These studies showed the benefits and possibilities of using GMM in unsupervised and supervised learning, although much work is ahead to shift the interest from inference to prediction, not only in GMM, but also in LV modeling in general. Not to mention, a systematic framework needs to be established for this transition to be successful. In this study, building on our previous work, we explore systematic ways of integrating LV modeling into the supervised learning framework and discuss necessary adjustments in both frameworks. We will focus on LV modeling of longitudinal outcomes as a concrete example motivated by clinical research and practice. To demonstrate the generality of the proposed learning framework, we will include an unsupervised learning method called model-based clustering (Bouveyron et al., 2019; Scrucca et al., 2016) in our investigation. We will also include K-means clustering, which is not model based, but is the most commonly used clustering method. The overall pipeline of our approach is shown in Figure 1, which will be detailed in following sections (a beta version of the matching R program will be made available upon request).

## Motivating example: the LAMS study

One of the studies that motivated our investigation is the Longitudinal Assessment of Manic Symptoms Study (LAMS). LAMS was designed to investigate phenomenology, psycho-pathologic evolution, related conditions and predictors of functional outcomes in children with elevated manic symptoms (Findling et al., 2010; Horwitz et al., 2010; Youngstrom et al., 2008). Children aged 6–12 years and their parents were recruited from outpatient clinics. Their primary outcome was a parent-completed measure of a child's mood symptoms: Parent General Behavior Inventory-10-item Mania Form (PGBI-10M) (Youngstrom et al., 2008). The study's emphasis on dimensional symptoms of mania and their changes over time fundamentally differentiates LAMS from other studies that have focused on diagnosis of bipolar disorder and its risk. The key outcomes, including manic (PGBI-10M) and anxiety (SCARED-P) symptoms, were measured every six months for 10 years, leading to rich longitudinal outcome data. In developing prognostic models, such longitudinal information is largely underutilized, which is an unfortunate situation as it may contribute to improved prognosis and personalized care of future patients.

In providing care for pediatric patients who present to outpatient clinical care with concerns in elevated manic symptoms, it is of great clinical importance to predict the long-term pattern of their symptoms. Whereas calculating risk of conversion to bipolar disorder has been previously conducted (Birmaher et al., 2018), such attempt has not been made for predicting progression of manic symptoms over time. Ironically, the first hurdle to this development is the richness of longitudinal data. It is not self-evident how to formulate a simple prediction target that best captures individuals' longitudinal symptom patterns. Our study has been motivated by this situation, where constructing a reliable and valid prediction target that captures longitudinal symptom patterns is the critical first step in developing useful prediction models.

We are particularly interested in predicting manic symptom patterns within the first two years, which is regarded as a clinically useful and reasonable range of prediction. Within this prediction range, Table 1 shows sample statistics of repeatedly measured PGBI-

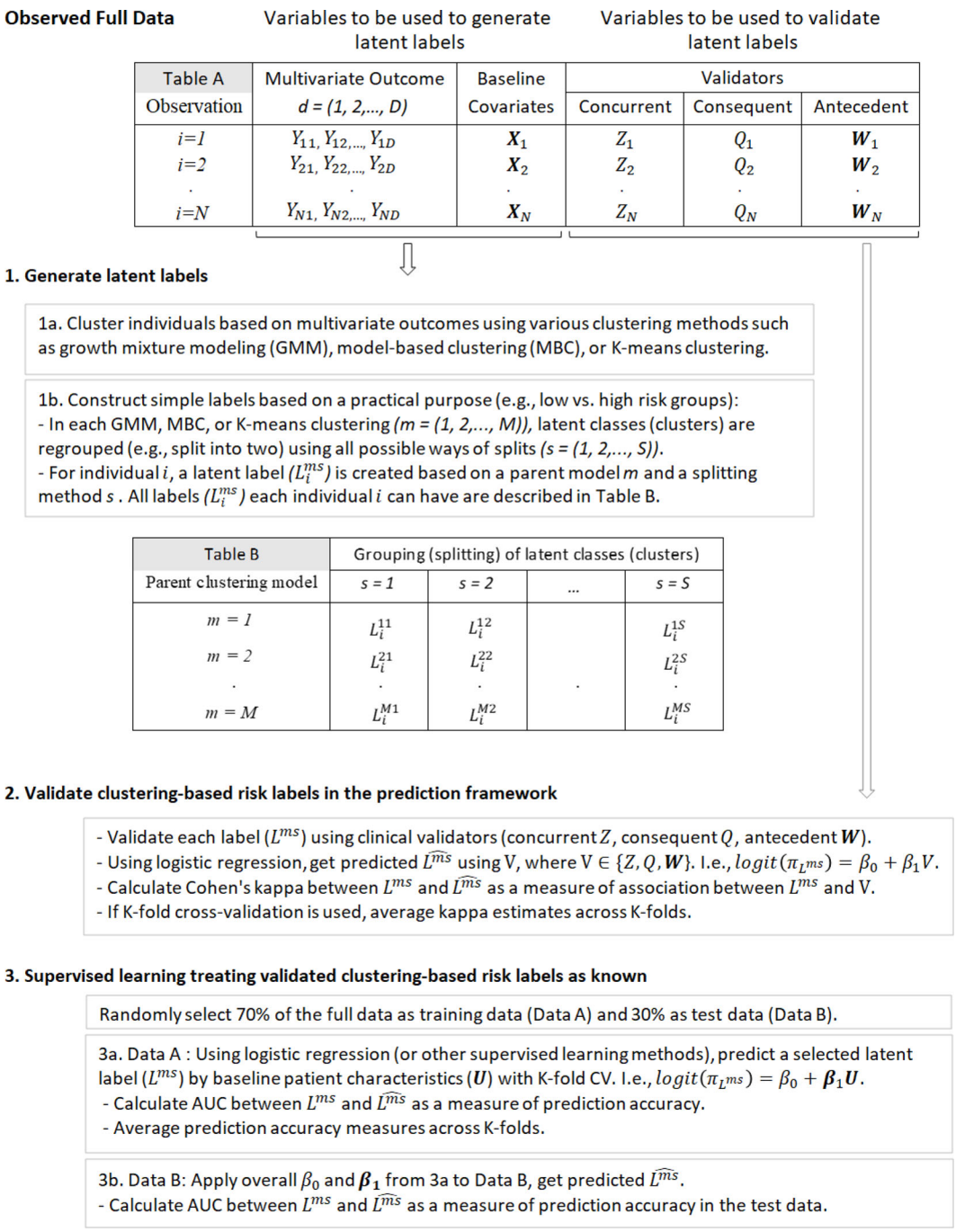


**Figure 1.** A 3-step learning pipeline with latent variable modeling.

10M. Previous studies (Findling et al., 2010; Horwitz et al., 2010; Youngstrom et al., 2008) have suggested a clinical threshold that sets PGBI-10M $\geq$ 12 as having elevated symptoms of mania (ESM). A currently recommended way to construct a summary measure is to apply this established threshold to each repeated measure of PGBI-10M. The resulting summary label, $Z$, is coded as 1 if any ESM is observed during 6 to 24 months (i.e., elevated risk) and 0 otherwise (i.e., low risk). Taking advantage of long-term observations in LAMS, we can create another summary measure, $Q$, by defining future risk (consequence) as having any ESM between 30 and 48 months, which is outside the targeted prediction range. Table 1 also shows clinically relevant baseline measures, $\boldsymbol{W}$ (bipolar diagnosis, depression by CDRS-R, anxiety by SCARED-P) and some demographic measures.

## Step 1: LV modeling of multivariate outcomes

The first step in the proposed learning approach is to generate simple outcomes using LV modeling. The

goal here is to create more reliable and valid outcomes using multivariate outcome information without losing the simplicity of a single observed outcome. We are particularly interested in LV modeling methods that classify individuals into clusters or latent classes, which has been motivated by clinical diagnosis and prognosis in practice. We employed two LV modeling strategies, growth mixture modeling and model-based clustering, which have been developed and used in different fields without much overlap or comparison between the two. We used these two distinct methods with the intention of demonstrating the generality of the proposed learning framework. We are also including K-means clustering, which is not based on LV modeling, but is the most commonly used clustering method. We will not discuss details of K-means clustering as it is a commonly used method that is well covered in machine learning text books.

Both growth mixture modeling and model-based clustering utilize finite mixture modeling, which makes it possible to use the common framework for the two methods. Let us consider data with $d$ multivariate measures for the $i^{th}$ unit (individual $i$ in our application). That is, $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{id})$, which applies to both LV modeling strategies we consider. In a finite mixture model, the probability distribution of $Y_i$ with $J$ mixture components $(j = 1, 2, ..., J)$ can be expressed as

$$p(Y_i|\boldsymbol{\theta}, \pi) = \pi_1 f(Y_i|\boldsymbol{\theta}_1) + \pi_2 f(Y_i|\boldsymbol{\theta}_2) + \cdots + \pi_J f(Y_i|\boldsymbol{\theta}_J), \quad (1)$$

where $\boldsymbol{\theta}_j$ is a vector of model parameters for the $j^{th}$ class or mixture component, and $\pi_j$ is the proportion of the population from the $j^{th}$ component with $\sum_{j=1}^{J} \pi_j = 1$.

### Growth mixture modeling (GMM)

One of the LV modeling methods we will use to summarize multivariate outcome data is growth mixture modeling (GMM, e.g., Muthén & Shedden, 1999), which is a popular method of discovering latent trajectory types. We will use a simple GMM setting given our intention to demonstrate proof of concept of the proposed learning framework. Focusing on five repeated measures of manic symptoms (PGBI-10M) in LAMS, we used a quadratic growth specification. The outcome $Y$ for individual $i$ $(i = 1, 2, ..., N)$ at time point $t$ $(t = 1, 2, ..., d)$ conditioned on trajectory class $C_i = j$ can be expressed as

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

$$\eta_{1ij} = \eta_{1j} + \boldsymbol{\alpha}_1'\ \boldsymbol{X}_i + \zeta_{1ij}, \quad (3)$$

$$\eta_{2ij} = \eta_{2j} + \boldsymbol{\alpha}_2'\ \boldsymbol{X}_i + \zeta_{2ij}, \quad (4)$$

$$\eta_{3ij} = \eta_{3j} + \boldsymbol{\alpha}_3'\ \boldsymbol{X}_i + \zeta_{3ij}, \quad (5)$$

where there are $J$ possible classes $(j = 1, 2, ..., J)$. There are three intercept parameters to capture change: initial status $(\eta_{1j})$, linear growth $(\eta_{2j})$, and quadratic growth $(\eta_{3j})$ for trajectory class $j$. The time measure $T_t$ reflects linear and $T_t^2$ quadratic growth. The measurement errors $\varepsilon_{ij} = (\varepsilon_{ij1}, ..., \varepsilon_{ijd})$ are assumed to be normally distributed with $\varepsilon_{ij} \sim MN(0, \Sigma_\varepsilon)$. For the random effects associated with growth parameters, we used two variations: models with random intercepts only $(Var(\zeta_{2i}) = Var(\zeta_{3i}) = 0)$ and models with no random effects $(Var(\zeta_{1i}) = Var(\zeta_{2i}) = Var(\zeta_{3i}) = 0)$. We assumed that $\zeta_{1i} \sim MN(0, \Sigma_\zeta)$. To maintain identifiability in models with larger numbers of classes, we imposed restrictions that $\Sigma_\varepsilon$ is diagonal and that $\Sigma_\varepsilon$ and $\Sigma_\zeta$ do not vary across classes. The relationship

**Table 1.** Sample statistics of manic symptoms (PGBI-10M) and clinical validators in the LAMS study ($N = 616$ based on everyone who has at least one PGBI-10M measure).

| Variable | N | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| Repeated outcome measures within the prediction range | | | | | |
| PGBI-10M at baseline | 609 | 0 | 30 | 12.60 | 7.14 |
| PGBI-10M at 6 m | 538 | 0 | 30 | 10.52 | 6.73 |
| PGBI-10M at 12 m | 520 | 0 | 30 | 8.60 | 6.55 |
| PGBI-10M at 18 m | 479 | 0 | 27 | 8.43 | 6.10 |
| PGBI-10M at 24 m | 464 | 0 | 30 | 8.15 | 6.60 |
| Elevated risk within the prediction range based on a clinical cutpoint (concurrent validator *Z*) | | | | | |
| Any PGBI-10M $\geq$ 12 during 6 − 24 m | 607 | 0 | 1 | 0.53 | |
| Elevated risk outside the prediction range based on a clinical cut-point (consequence *Q*) | | | | | |
| Any PGBI-10M $\geq$ 12 during 30 − 48 m | 503 | 0 | 1 | 0.38 | |
| Clinically relevant baseline variables (antecedents ***W***) | | | | | |
| Bipolar diagnosis | 616 | 0 | 1 | 0.23 | |
| Depression symptom (CDRS-R) | 616 | 17 | 73 | 34.77 | 10.86 |
| Anxiety symptom (SCARED-P) | 607 | 0 | 69 | 17.82 | 13.50 |
| Other baseline characteristics | | | | | |
| Female | 616 | 0 | 1 | 0.32 | |
| Medicaid | 616 | 0 | 1 | 0.43 | |
| Age | 616 | 6.1 | 13.2 | 9.39 | 1.92 |

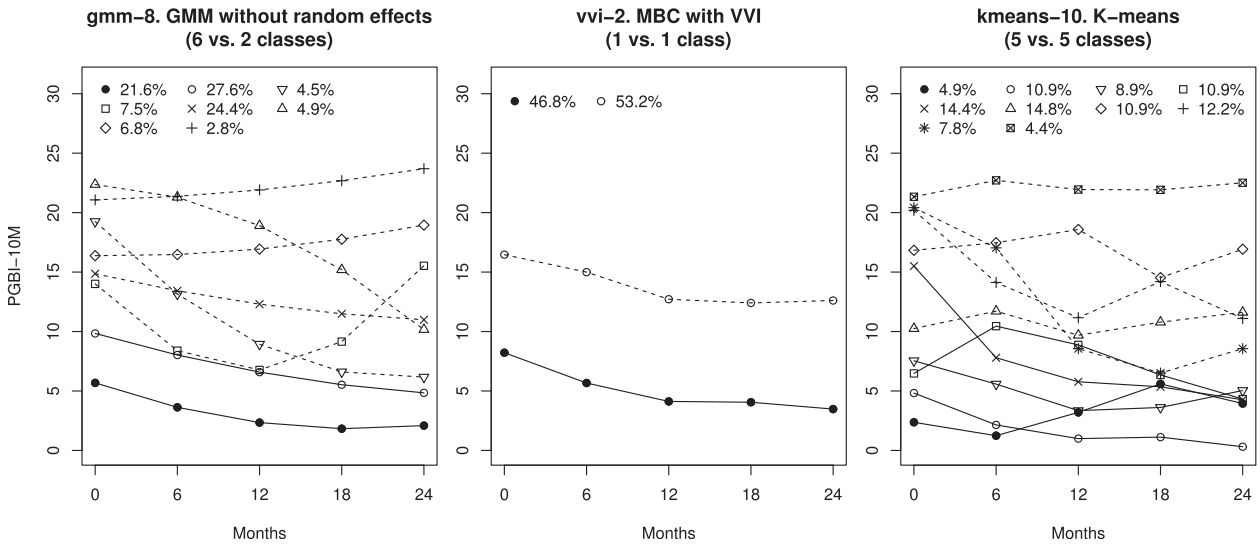


**Figure 2.** Examples of clustering-based risk labels and their parent models.

between the growth factors and covariates $\boldsymbol{X}$ is captured by the vectors of regression coefficients $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, and $\boldsymbol{\alpha}_3$, which may also vary across trajectory classes.

The probability of subject $i$ belonging to a certain trajectory class ($\pi_{ij} = Pr(C_i = j)$) may depend on the influence of covariates. The multinomial logit model of $\pi_{ij}$ conditioned on baseline covariates $\boldsymbol{X}_i$ is described as

$$logit(\pi_{ij} \mid \boldsymbol{X}_i) = \gamma_{0j} + \boldsymbol{\gamma}'_{1j}\, \boldsymbol{X}_i \tag{6}$$

for $j = 1, 2, ..., (J-1)$, where $\boldsymbol{\gamma}_{1j}$ is a vector of multinomial logit regression coefficients. In the current example, we focus on models without covariates, although we keep $\boldsymbol{X}_i$ in all equations for the generality in our presentation. The array of candidate models will simply expand in our framework as we include covariates and/or consider various model specifications with regard to random effects, residual variances, trajectory shapes, and other auxiliary variables.

For estimation of GMM models, we used maximum likelihood (ML) estimation via the expectation maximization (EM) algorithm (Dempster et al., 1977; McLachlan & Krishnan, 1997). For ML-EM estimation, we used Mplus program (Muthén & Muthén, 1997–2017), a popular latent variable modeling program. Estimating GMM models with ML-EM is a well-established practice, especially with the common models used in our application. For interested readers, directly relevant details can be found in our previous papers (Jo et al., 2017, 2018).

In the proposed learning framework, the only information we will actually use from GMM estimation is the posterior class probabilities generated in the E step, where latent trajectory class $C_i$ is handled as missing data. The posterior class probability of subject $i$ belonging to class $j$ conditioned on observed data $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ and the current estimates of model parameters $(\boldsymbol{\gamma}^*, \boldsymbol{\alpha}^*, \boldsymbol{\eta}_j^*, \Sigma_\zeta^*, \Sigma_\varepsilon^*)$ in the iterative procedure is expressed as

$$p_{ij}(\boldsymbol{\gamma}^*, \boldsymbol{\alpha}^*, \boldsymbol{\eta}_j^*, \Sigma_\zeta^*, \Sigma_\varepsilon^*) = \frac{\pi_{ij}(\boldsymbol{\gamma}^*) f(\boldsymbol{Y}_i \mid C_i = j,\ \boldsymbol{X}_i, \boldsymbol{\alpha}^*, \boldsymbol{\eta}_j^*, \Sigma_\zeta^*, \Sigma_\varepsilon^*)}{\sum_{j'=1}^{J} \pi_{ij'}(\boldsymbol{\gamma}^*) f(\boldsymbol{Y}_i \mid C_i = j',\ \boldsymbol{X}_i, \boldsymbol{\alpha}^*, \boldsymbol{\eta}_{j'}^*, \Sigma_\zeta^*, \Sigma_\varepsilon^*)}, \tag{7}$$

where $\boldsymbol{\eta}_j = (\eta_{1j}, \eta_{2j}, \eta_{3j})$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3)$, $\sum_{j=1}^{J} \pi_{ij}(\boldsymbol{\gamma}) = 1$ for $i = 1, ..., N$, and $\pi_{ij}(\boldsymbol{\gamma}) = Pr(C_i = j | \boldsymbol{X}_i, \gamma_{01}, \gamma_{11}, ..., \gamma_{0J}, \gamma_{1J})$. Once the ML-EM procedure reaches the optimal status, the above posterior class probability of subject $i$ belonging to class $j$ can be seen as valuable summary information that characterizes each individual.

Using the full LAMS data, we estimated a series of GMM models with varying numbers of classes. We increased the number of classes until any of the $\Sigma_\varepsilon$ and $\Sigma_\zeta$ estimates in any of the classes was not positive definite or any of the classes had less than 10 individuals based on their most likely latent class membership. We used ample starting values to avoid potential convergence at local maxima. Excluding one-class models, a total of 13 models with (6 models with 2–7 classes) and without (7 models with 2–8 classes) random intercepts met these conditions. An example of GMM results (mean trajectories) with 8 classes without random effects is shown in Figure 2.

### *Model-based clustering (MBC)*

Another method of LV modeling we will use to summarize multivariate outcomes is model-based clustering (Bouveyron et al., 2019; Fraley & Raftery, 2002;

Scrucca et al., 2016). Model-based clustering (MBC) is a line of method that focuses on identification of latent classes (clusters) based on finite mixture modeling of multivariate normal distributions. In principle, GMM is also a type of model-based clustering. However, GMM has not been incorporated into the tool box of unsupervised learning strategies, and therefore, it is not well known or commonly used in machine learning. MBC is a more widely known tool for unsupervised learning, although it is not commonly used in psychological and behavioral sciences. Whereas the signature feature of GMM is modeling of longitudinal trends, the signature feature of the currently known MBC is the use of geometric constraints on the covariance matrix of multivariate data. Other than this difference, GMM and MBC basically share the same analytical (finite mixture modeling) and estimation (ML-EM or Bayesian) strategies.

In MBC, without any parameters to model the longitudinal trend, the multivariate data $Y_{it}$ $(t = 1, 2, ..., d)$ conditioned on class $C_i = j$ can be simply expressed as

$$Y_{it}|(C_i = j) = \eta_{jt} + \varepsilon_{ijt}, \tag{8}$$

where various geometric constraints on the variance/-covariance matrix of $\varepsilon_{ijt}$ are the key to the identification of latent classes (clusters). If the distribution is univariate, $\eta_j$ is interpreted as the mean for the $j$th mixture component, and $\sigma_j^2$ as the associated variance. For the covariance matrix $(\Sigma_j)$ of multivariate data, there are several types of geometric constraints that can be imposed on volume (equal, variable), shape (equal, variable), and orientation (equal, variable, axis-aligned) of the ellipsoidal distribution. Such parameterization is based on an eigen-decomposition of the form $\Sigma_j = \lambda_j \boldsymbol{D}_j \boldsymbol{A}_j \boldsymbol{D}'$, where $\lambda_j$ is a scalar controlling the volume of the ellipsoid, $\boldsymbol{A}_j$ is a diagonal matrix specifying the density contours, and $\boldsymbol{D}_j$ is an orthogonal matrix which determines the orientation of the ellipsoid (from Scrucca et al., 2016). More details on these types can be found elsewhere (Bouveyron et al., 2019; Lebret et al., 2015; Scrucca et al., 2016). Recall that we used relatively simple constraints for the covariance matrix for the GMM model described in (2)–(5). Instead, we used multiple intercept (or mean) parameters (i.e., $\eta_{1j}, \eta_{2j}, \eta_{3j}$) to explicitly model the longitudinal trend.

For ML-EM estimation of MBC models, we used R package `mclust` (Scrucca et al., 2016), which has 14 types of constraints on the covariance matrix (EEE, EEI, EEV, EII, EVE, EVI, EVV, VEE, VEI, VEV, VII, VVE, VVI, VVV). The posterior class probability of subject $i$ belonging to class $j$ conditioned on observed data $(\boldsymbol{Y}_i)$ and the current parameter estimates $(\boldsymbol{\eta}_j^*, \Sigma_j^*)$ in the iterative procedure can be expressed as

$$p_{ij}(\boldsymbol{\eta}_j^*, \Sigma_j^*) = \frac{\pi_{ij} f(\boldsymbol{Y}_i \mid C_i = j,\ \boldsymbol{\eta}_j^*, \Sigma_j^*)}{\sum_{j'=1}^{J} \pi_{ij'} f(\boldsymbol{Y}_i \mid C_i = j',\ \boldsymbol{\eta}_{j'}^*, \Sigma_{j'}^*)}. \tag{9}$$

Using the LAMS data, we estimated a series of MBC models using all 14 types of geometric constraints allowing up to 19 classes. We obtained a total of 135 MBC models with 2–19 classes. Figure 2 shows an example of MBC results with 2 clusters with the VVI restriction, where VVI stands for a diagonal distribution with varying (V) volume across classes, varying (V) shape, and coordinate axis-aligned (I) orientation.

## Reorienting LV modeling for prediction

Behavioral and social sciences have a long history of using LV modeling as a flexible tool for building theoretical models focusing on inference. In this tradition, using LV models in developing prediction algorithms and deploying them in actual clinical practice has not been a common goal. Focusing on prediction instead of inference implies a significant shift in LV modeling, requiring a system that effectively connects LV modeling with supervised learning, a dominating approach for developing prediction algorithms.

In this article, we intend to outline a foundational framework that will facilitate integration of LV modeling in supervised learning. In particular, we are interested in the utility of LV modeling as a way of generating improved prediction targets (outcomes, responses, outputs). Focusing on this utility, we lay out necessary adjustments to reorient LV modeling for supervised learning. Specifically, we aim for (1) generation of simple outcomes and (2) systematic validation and selection of generated outcomes.

### *Simple latent outcomes*

A promising role of LV modeling is to generate simple outcomes that effectively summarize complex multivariate information. In the examples shown in Figure 2, various latent class solutions differently summarize repeated outcome measures. However, even this level of organization may not be enough in developing prediction models. For example, in LAMS, separating out patients who would maintain moderate levels of symptoms (low risk) early on is critical in planning optimal treatments, better allocating resources, and reducing patient burdens. In this context,

categorizing patients into fine-grained symptom trajectory patterns and predicting them is neither practical nor necessary. Therefore, this step may require further targeting and simplification so that the generated outcomes are better aligned with practical purposes. Not to mention, such simplification will make the generated outcomes easier to handle in supervised learning.

In line with LAMS, we will focus on dividing individuals into two coarsened groups of trajectory classes. For example, in the 8-class GMM solution in Figure 2, the top six classes can be coarsened into one group and the bottom two classes into the other group. The total number of possible two-group splits ($s = 1, 2, 3, ..., S$) in each GMM or MBC model ($m = 1, 2, 3, ..., M$) can be simply calculated as $S = 2^{J-1} - 1$, where $J$ is the number of classes in the model.

At the individual level, splitting or coarsening of clusters is straightforward when each person belongs to only one cluster. When using K-means, each person can be categorized into one of the coarsened groups his or her cluster belongs to. When using a cutpoint, each person can be categorized by simply applying a cutpoint (e.g., PGBI-10M $\geq 12$ as elevated risk in LAMS) to one of the observed outcome measures (e.g., at 24 months), or to the maximum or average of all targeted outcome measures (e.g., at 6, 12, 18, and 24 months). When using LV modeling or model-based clustering, splitting of clusters can be done using the posterior class probabilities from (7) or from (9). For example, in a 8-class model, person $i$ has a set of posterior class probabilities $(p_{i1}, p_{i2}, p_{i3}, p_{i4}, p_{i5}, p_{i6}, p_{i7}, p_{i8})$ that can be grouped into two in 127 ways ($S = 2^{8-1} - 1 = 127$). Based on model $m$ and splitting method $s$ ($s = 1, 2, 3, ...S$), let $p_i^{ms}$ stand for the coarsened posterior probability of person $i$ belonging to the first group (e.g., $p_{i3} + p_{i4} + p_{i5} + p_{i6} + p_{i7} + p_{i8}$ in gmm-8 in Figure 2) and $1 - p_i^{ms}$ for the second group (e.g., $p_{i1} + p_{i2}$ in gmm-8 in Figure 2).

One simple way to utilize coarsened posterior class probabilities is to create a binary label $L_i^{ms}$ by dichotomizing $p_i^{ms}$. That is,

$$L_i^{ms} = \begin{cases} 1 & \text{if} \quad p_i^{ms} \geq 0.5 \\ 0 & \text{if} \quad p_i^{ms} < 0.5 \end{cases}. \tag{10}$$

which results in $S$ binary variables that can be used as output variables in subsequent supervised learning and in building prediction models. We will use this strategy to simplify comparisons across soft clustering (GMM, MBC), hard clustering (standard K-means), and cutpoint-based categorization methods. However, when using soft clustering, it is in principle possible to account for uncertainty in cluster assignment (Jo et al., 2017).

## Step 2: Systematic validation with explicit validators

The targeted utility of LV modeling in our context is to generate simple outcomes to be used in prediction. The results of LV models can be further organized in line with our clinical intention (e.g., low vs. elevated risk), leading to simple and practical outcomes. Having simple output variables is a big step toward supervised learning. However, being simple does not guarantee the validity of the outcomes, which in fact applies to both observed and LV-based outcomes. As shown in the LAMS example, LV modeling can generate a large pool of candidate outcomes in the absence of known truth, which can be viewed as a big drawback. However, the same situation can be viewed as an opportunity to tailor the most desirable prediction targets based on multiple criteria. Observed and cutpoint-based measures do not possess such flexibility, which is exactly why we propose the use of LV modeling as a way of generating improved output variables.

### *Combining supervised learning and psychometrics traditions*

In supervised learning, a large number of candidate models are systematically evaluated in terms of direct measures of success such as prediction or classification accuracy (Hastie et al., 2009). This is possible due to the simple structure of considered models (predictors and the predicted). It is typically assumed that the outcome to be predicted is clear, simple, and readily available, which lets us focus on the predictor side and assessing how accurately and stably various combinations or subsets of predictors predict the outcome. What is different about our scenario is that we are trying to validate outcomes generated based on LV modeling. From the perspective of supervised learning, validating outcomes before predicting them is a foreign concept and an unnecessary step.

In psychometrics, it is a long tradition to question the validity of measured outcomes. Many different concepts of validation have been developed in psychometrics to enhance validation of tests or measures that are intended to capture true status of outcomes that are hard to quantify such as intelligence, aptitude,

and various psychiatric outcomes. An LV-based outcome can be simply thought of as a new test or a measure that needs validation before it gets presented as a competitive alternative. Validation is particularly important here as it gives LV-based outcome measures concrete meanings by connecting them with scientifically or clinically meaningful validators. Further, validation can serve as a selection tool, which is critical in sorting out the best among the large pool of candidate outcomes. However, using validation as a selection tool is not a common practice in developing new tests because we do not normally develop many tests at the same time. In our context, a large pool of LV-based outcomes are generated, creating a new situation for the traditional validation practice in psychometrics.

In the new validation framework, we combine the traditions of supervised learning and psychometrics. The relationship between clinical validators and LV-based outcomes fits well in the prediction framework, meaning that we can approach outcome validation as a supervised learning task. What is nice about this approach is that it naturally uses validation as a selection tool, which is critical in our scenario with a large pool of candidate outcomes. Selecting the best out of many possible output variables is not a typical use of supervised learning, which focuses more on the predictor side. However, with some conceptual shifts, the existing machinery of supervised learning is basically ready for validation of LV-based outcomes. Making the validation process automation-ready is also important in that it will encourage outcome validation and the use of LV-based outcomes in building prediction models and in supervised learning in general.

In line with the psychometrics tradition, we will use well-structured validation with multiple criteria. Specifically, we will use clinically meaningful validators (e.g., antecedent, concurrent, and consequent validators) selected by experts in clinical and psychometrics fields. There are several advantages of validating and selecting outcomes based on these clinically meaningful validators. Given the exploratory use of LV modeling in our context, using explicit validators is probably the simplest and fastest way to evaluate and narrow a large pool of constructed outcomes. The selected LV-based outcomes will be closely aligned with contemporary science and clinical practice, leading to easy interpretation and clear communication across all involved parties (outcome developers, prediction model developers, clinical researchers, practitioners, and patients). Focusing on the LAMS context, we chose three types of validators targeting to identify latent outcomes that well capture long-term progression of manic symptoms.

i. Concurrent validator ($Z$): This is a primary validator that ensures that developed measures are closely related to what is currently used and well-accepted. In developing new tests or measures, it is a typical practice to first examine the concurrent validity by correlating a new test (e.g., a geriatric depression test) with a widely used test (e.g., a general depression test). In the LAMS example, we created a concurrent validator ($Z$ in Table 1), by applying an established clinical cutpoint (any or maximum PGBI-10M $\geq$ 12 as elevated risk) to all repeated measures within the prediction range (6 to 24 months).
ii. Consequence ($Q$): Consequences are future outcomes that are supposed to be correlated with the developed tests or measures. In our LAMS example, this is distal future risk beyond the timeframe of prediction interest. To create this variable, we applied the same clinical cutpoint (any or maximum PGBI-10M $\geq$ 12 as elevated risk) to all repeated measures between 30 to 48 months ($Q$ in Table 1).
iii. Antecedents ($\boldsymbol{W}$): These are clinically relevant variables that precede and are supposed to be correlated with the measures or tests that are being validated. In the LAMS example, our clinical experts identified three variables (bipolar diagnosis, anxiety by SCARED-P, depression by CDRS-R, shown in Table 1) as directly relevant clinical antecedents, which will further validate candidate risk labels.

### Toward automation

In developing prediction models, using LV-based outcomes is not a well-accepted practice despite its great potentials. The flexible nature of LV modeling is a big advantage in organizing complex multivariate information, although can also make resulting LV-based outcomes look subjective and esoteric. Having a structured validation plan using explicit clinical validators dramatically changes this situation. Integrating this validation concept from psychometrics into supervised learning further solidifies the possibility of systematic validation of LV-based outcomes. This means that automation of the validation process is possible guided by experts' knowledge and clinical practice.

The association between an LV-based output variable from (10) and each set of validators can be put

in the prediction framework using logistic regression as

$$logit(\pi_{L_i^{ms}(Z_i)}) = \alpha_{Z0}^{ms} + \alpha_{Z1}^{ms} Z_i, \quad (11)$$

$$logit(\pi_{L_i^{ms}(Q_i)}) = \alpha_{Q0}^{ms} + \alpha_{Q1}^{ms} Q_i, \quad (12)$$

$$logit(\pi_{L_i^{ms}(\boldsymbol{W}_i)}) = \alpha_{W0}^{ms} + \boldsymbol{\alpha}_{W1}^{ms} \boldsymbol{W}_i, \quad (13)$$

where $\pi_{L_i^{ms}(Z_i)} = Pr(L_i^{ms} = 1|Z_i)$, $\pi_{L_i^{ms}(Q_i)} = Pr(L_i^{ms} = 1|Q_i)$, and $\pi_{L_i^{ms}(\boldsymbol{W}_i)} = Pr(L_i^{ms} = 1|\boldsymbol{W}_i)$ denote the probability of person $i$ belonging to the first category of binary label $L_i^{ms}$ as a function of $Z_i$, $Q_i$, or $\boldsymbol{W}_i$. With a single clinical cut-point (PGBI-10M $\geq 12$ as elevated risk), both $Z_i$ and $Q_i$ are binary (1 = elevated risk, 0 = low risk). For simplicity, we put $Q_i$ on the right side of equation like the other validators. In principle, $Q_i$ should be predicted by $L_i^{ms}$, although which one becomes the predictor does not matter here as we are looking at one to one association.

The estimated $\pi_{L_i^{ms}(Z_i)}$ from (11) can be categorized to form a predicted binary label for individual $i$ as

$$\hat{L}_i^{ms}(Z_i) = \begin{cases} 1 \quad \text{(elevated risk)} & \text{if} \quad \hat{\pi}_{L_i^{ms}(Z_i)} \geq 0.5 \\ 0 \quad \text{(low risk)} & \text{if} \quad \hat{\pi}_{L_i^{ms}(Z_i)} < 0.5, \end{cases} \quad (14)$$

where we can now label the two categories as low and elevated risk. This is possible because of the use of a concurrent validator ($Z$) based on observed outcomes. In the same manner, we can also formulate $\hat{L}_i^{ms}(Q_i)$ from (12) and $\hat{L}_i^{ms}(\boldsymbol{W}_i)$ from (13).

Then, we can calculate the degree of agreement between the candidate label from (10) and the estimated label from (14). We used Cohen's $\kappa$ (Cohen, 1960) as a conservative measure of agreement between two binary variables taking into account the agreement occurring by chance, which is particularly important when evaluating candidate labels with considerably different proportions. To enhance our validation, we used K-fold cross-validation (CV) to take into account generalization error (variation across samples), which is a common practice in supervised learning, although has not been used until recently in the psychometric validation context (Jo et al., 2017, 2018). Combining these traditions, cross-validated Cohen's $\kappa$ for a candidate label $L^{ms}$ using $Z$ as a validator can be calculated averaging across $K$ folds ($f = 1, 2, 3, ..., K$) as

$$CV_\kappa^{Zms} = \sum_{f=1}^{K} \kappa_f^{Zms}/K \ . \quad (15)$$

where $\kappa_f^{Zms}$ is Cohen's $\kappa$ for the $f^{th}$ fold when we use $Z$ as a validator. In the same manner, we can calculate $CV_\kappa^{Qms}$ and $CV_\kappa^{\boldsymbol{W}ms}$ when using $Q$ or $\boldsymbol{W}$ as a validator.

Specifically, in the LAMS example, we used 10-fold CV. The full sample is randomly divided into 10 equal size subsamples ($f = 1, 2, 3, ..., 10$). We set aside one subsample ($f^{th}$ fold) to be used as a validation sample. The rest of the subsamples (training data) are used to estimate the association between each set of clinical validators and each candidate label, shown in (11)–(13). The parameter estimates (logit coefficients) were then applied to the validation sample ($f^{th}$ fold) to obtain each person's predicted label when the model estimates using the training data are applied to a data set that is not used to get those estimates. The degree of agreement between a candidate label from (10) and the estimated label from (14) was measured using $\kappa$. This process is repeated and averaged over K (=10) folds as shown in (15).

The associated standard error for (15) can be calculated considering the variance across $K$ folds as

$$SE_\kappa^{Zms} = \sqrt{Var(\kappa_1^{Zms}, \kappa_2^{Zms}, ..., \kappa_K^{Zms})}/\sqrt{K}, \quad (16)$$

where $\kappa_K^{Zms}$ is Cohen's $\kappa$ for the $K^{th}$ fold when using model $m$, splitting method $s$, and validator $Z$. In the same manner, $SE_\kappa^{Qms}$ and $SE_\kappa^{\boldsymbol{W}ms}$ can be calculated when using $Q$ or $\boldsymbol{W}$ as a validator. It is also possible to account for the uncertainty in cluster assignment when soft clustering methods, such as GMM and MBC, are used (Jo et al., 2017).

### Validation results in the LAMS example

We applied the proposed validation method to the LAMS example, where GMM generated 367 binary outcome labels with the intention of capturing low and elevated risk trajectory types. Based on MBC with 14 types of covariance constraints, 954,755 binary outcome labels were generated. Based on K-means clustering with up to 13 clusters, 6,142 labels were generated. Additionally, we validated four cutpoint-based labels including the concurrent validator $Z$. We used three sets of validators ($Z, Q, \boldsymbol{W}$) to triangulate good outcome labels that are aligned with expert knowledge and intended clinical utility. In LAMS, we are particularly interested in separating out low-risk patients early on.

Given the defining role of the concurrent validator ($Z$), we first selected 10 best candidate labels from each method based on their association with $Z$ (i.e., $CV_\kappa^{Zms}$). Then, we eliminated those that are worse than the best based on all accounts (i.e., association with $Z$, $Q$, and $\boldsymbol{W}$). Using this simple rule, we selected two best outcome labels from each method. An alternative would be to average association measures

across three validators with equal weights. Another alternative would be to focus more on $Q$ (future risk), which will lead to selection of labels that are good predictors of distal outcomes. The choice among these rules depends on the intended utility of generated labels. In the LAMS example, we focus more on $Z$ given our interest in generating outcome labels to be used as output variables in developing prognostic models and algorithms.

The validation results are shown in Table 2. First of all, the results clearly show the benefits of using clustering, both LV-based (GMM, MBC) and K-means-based. The cutpoint-based labels generally show weaker association with all clinical validators. Even the primary validator ($Z$) shows weaker association with the rest of validators ($Q, \boldsymbol{W}$). It is also shown that cutpoint-based labels categorize much fewer patients as elevated risk. The difference in the proportion is 20% or more from the primary validator $Z$, indicating that too many patients are categorized as low risk. Given the goal of safely separating out low-risk patients, cutpoint-based labels are considerably misaligned with the clinical intention, implying missed opportunities for proper early treatments. As discussed earlier, LV modeling and clustering methods can generate a large pool of candidate labels, which makes it possible to select tailored labels that are well-aligned with clinical validators. Cutpoint-based labels lack such flexibility.

Table 2 also shows that the validation results are remarkably comparable across different clustering methods despite their distinct approaches. Based on the top binary labels (gmm-8, vvi-2, kmeans-10), the agreement between GMM and MBC is 94.6%, between K-means and GMM is 94.6%, between MBC and K-means is 96.1%. Across the three methods, 92.7% of individuals are consistently labeled (either as elevated or as low risk). Such agreement is not surprising given that the labels shown in Table 2 have been already selected out of a very large pool of candidate labels based on the same clinical validators ($Z$, $Q$, and $\boldsymbol{W}$). One may still attempt to choose one best label for the intended purpose, perhaps by examining how mean trajectories are divided into low and elevated risk (as shown in Figure 2), or by examining individual patients that showed any disagreement in labeling across methods (7.3% of the LAMS sample).

Table 3 shows some examples of disagreement across different methods. We also included experts' opinion, which is based on the majority vote from three clinical experts. Patients A and B are labeled as elevated risk by all methods except in the cutpoint-based method using the average PGBI-10M. Their averages are less than 12 even though some scores are well over 12. Their scores are also trending up, which is a concerning pattern from the experts' point of view. Patients C and D are labeled as low risk by all

**Table 2.** Validation of risk labels based on their association with clinical validators.

| Clustering- or cutpoint-based risk labels** | | %Elevated risk | Association with clinical validators ($\kappa^*$) | | |
|---|---|---|---|---|---|
| | | | $Z$: concurrent | $Q$: consequent | $\boldsymbol{W}$ : antecedent |
| GMM-based | | | | | |
| gmm-8 | (6 vs 2 classes) | 50.8 | 0.77 (0.75, 0.80) | 0.47 (0.44, 0.50) | 0.33 (0.31, 0.36) |
| gmm-7 | (5 vs 2 classes) | 46.4 | 0.75 (0.72, 0.78) | 0.51 (0.48, 0.54) | 0.31 (0.28, 0.34) |
| MBC-based | | | | | |
| vii-2 | (1 vs 1 class) | 53.2 | 0.80 (0.78, 0.82) | 0.46 (0.43, 0.50) | 0.33 (0.30, 0.36) |
| vei-9 | (5 vs 4 classes) | 48.7 | 0.77 (0.75, 0.78) | 0.50 (0.47, 0.53) | 0.30 (0.27, 0.34) |
| K-means-based | | | | | |
| kmeans-10 | (5 vs 5 classes) | 50.0 | 0.76 (0.73, 0.78) | 0.48 (0.45, 0.51) | 0.33 (0.30, 0.37) |
| kmeans-2 | (1 vs 1 class) | 44.3 | 0.70 (0.68, 0.73) | 0.50 (0.47, 0.53) | 0.33 (0.29, 0.37) |
| Cutpoint-based | | | | | |
| PGBI-10M at 12 m $\geq$ 12 | | 31.7 | 0.57 (0.51, 0.62) | 0.41 (0.36, 0.45) | 0.18 (0.13, 0.23) |
| PGBI-10M at 24 m $\geq$ 12 | | 26.5 | 0.44 (0.42, 0.46) | 0.40 (0.35, 0.45) | 0.11 (0.06, 0.15) |
| $Z^{\dagger}$ (any PGBI-10M, 6–24 m $\geq$ 12) | | 53.0 | . | 0.41 (0.37, 0.45) | 0.26 (0.21, 0.32) |
| average $^{\ddagger}$ PGBI-10M, 6–24 m $\geq$ 12 | | 30.3 | 0.56 (0.53, 0.59) | 0.48 (0.44, 0.52) | 0.18 (0.15, 0.22) |

$^*$ Cohen's $\kappa$ was calculated as a measure of association according to (15) by averaging across 10 folds ($CV_\kappa^{Zms}$, $CV_\kappa^{Qms}$, $CV_\kappa^{\boldsymbol{W}ms}$). In parentheses are shown $CV_\kappa^{Zms} \pm SE_\kappa^{Zms}$, $CV_\kappa^{Qms} \pm SE_\kappa^{Qms}$, and $CV_\kappa^{\boldsymbol{W}ms} \pm SE_\kappa^{\boldsymbol{W}ms}$.

**Parent GMM (growth mixture modeling) models used to generate risk labels were estimated using the Mplus program (Muthén & Muthén, 1997–2017). For estimation of parent MBC (model based clustering) models, we used R package `mclust` (Scrucca et al., 2016). For K-means clustering, we used the `kmeans` function in R. In parentheses are shown how the latent classes or clusters are split into two categories to generate binary risk labels. For example, gmm-8 (see Figure 2) means a label generated based on a GMM model with 8 classes, and 6 vs. 2 classes means that the 8 classes are split into two groups with 6 classes in one group (elevated risk) and 2 classes in the other. In the cutpoint-based methods, individuals are simply divided into two groups by applying a single cutpoint to their observed scores or to an average of observed scores. The best two labels from each method based on their association with a priori clinical validators are presented.

$^{\dagger}Z$ is the concurrent validator. A clinical cutpoint is applied to each PGBI-10M score within the prediction range (6, 12, 18, and 24 months). The label takes the value of 1 if any PGBI-10M $\geq$ 12 and 0 otherwise.

$^{\ddagger}$PGBI-10M scores are first averaged across 6, 12, 18, and 24 months, and then a cutpoint is applied to the average score. The label takes the value of 1 if the average $\geq$ 12 and 0 if $<$ 12.

**Table 3.** Examples of disagreement across risk labels.

| | PGBI-10M | | | | | Risk labels | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Patient | 0m | 6m | 12m | 18m | 24m | gmm-8 | vvi-2 | kmeans-10 | $Z^{\dagger}$ | average $^{\ddagger}$ $\geq 12$ | Experts $^{*}$ |
| A | 21 | 9 | 7 | 14 | 17 | 1 | 1 | 1 | 1 | 0 | 1 |
| B | 18 | 6 | 3 | 10 | 15 | 1 | 1 | 1 | 1 | 0 | 1 |
| C | 7 | 12 | 7 | 5 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| D | 6 | 7 | 12 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| E | 9 | 6 | 8 | 9 | 10 | 0 | 0 | 1 | 0 | 0 | 0 |
| F | 7 | 7 | 12 | 13 | 6 | 1 | 1 | 0 | 1 | 0 | 1 |
| G | . | . | 1 | 9 | 8 | 0 | 1 | 0 | 0 | 0 | 0 |
| H | 14 | 10 | . | . | . | 0 | 1 | 0 | 0 | 0 | 1 |
| I | 19 | 14 | . | . | . | 1 | 0 | 0 | 1 | 1 | 1 |
| J | 19 | 6 | 4 | 12 | . | 1 | 0 | 0 | 1 | 0 | 1 |

$^{\dagger}Z$ is the concurrent validator. A clinical cutpoint is applied to each PGBI-10M score within the prediction range (6, 12, 18, and 24 months). The label takes the value of 1 if any PGBI-10M $\geq 12$ and 0 otherwise.

$^{\ddagger}$PGBI-10M scores are first averaged across 6, 12, 18, and 24 months, and then a cutpoint is applied to the average score. The label takes the value of 1 if the average $\geq 12$ and 0 if $< 12$.

$^{*}$Based on the majority vote from three clinical experts who independently rated the patients.

methods except by $Z$. These patients have one PGBI measure at the cutpoint, although the rest are safely below the cutpoint. Patients E and F show examples of disagreement between kmeans-10 and the other risk labels. Patient E is labeled as low risk by all methods except by kmeans-10. Patient F has two scores that are 12 or greater, although labeled as low risk by kmeans-10. Patients G and H show examples of disagreement between vvi-2 and the others. Patient G is labeled as elevated risk only by vvi-2, which seems overly conservative even with some missing measurements. Patient H has one score above the cutpoint at baseline, and has several missing measurements. Only vvi-2 and experts conservatively labeled this patient as elevated risk. Patients I and J show that vvi-2 is not necessarily the most conservative of the three clustering methods.

Utilizing multiple clustering methods provides an opportunity to identify a small portion of patients that are difficult to classify. This strategy effectively narrowed the LAMS sample (7.3% based on 3 clustering methods), which makes careful examination by clinical experts feasible. The results of such examination can be incorporated to improve the validation process, for example, by formulating a more elaborate $Z$, or by modifying the labeling process based on experts' ratings. In the LAMS example, a GMM-based label (gmm-8) turned out to be somewhat better aligned with experts' labeling. One possible explanation would be that clinical experts will consider not only the clinical cutpoint (PGBI-10M $\geq 12$), but also how the scores change over time, which is a key modeling component in GMM. However, note that the three methods were largely consistent in labeling the patients (92.7% agreement across three clustering methods). This is an area that needs further investigation in various application contexts.

## Step 3: Prediction of generated risk labels

Once the best label is selected based on clinical validators and practical utilities (Step 2), we can focus on developing prediction models in Step 3. There are many well-established supervised learning strategies for predicting a known outcome with a large pool of possible predictors (e.g., Hastie et al., 2009). In principle, a selected label from Step 2 can be used as a known input or output variable with any supervised learning methods. However, note that the validation step is closely aligned with the intended clinical utility. In the LAMS example, we focused on a concurrent validator ($Z$) given our interest in generating a risk label to be used as an output variable in developing prognostic models and algorithms. In other words, it is not ideal to use the best label from Step 2 as a predictor (input) variable in Step 3. If generating risk labels as input variables is the goal, the validation process should put more emphasis on $Q$ (a consequent validator) than on $Z$.

Let $\boldsymbol{U}$ represent the set of baseline variables to be used as predictors. In the clinical context, $\boldsymbol{U}$ is expected to provide not only good prediction, but also good interpretation (i.e., using them as predictors of risk should make sense clinically). In that sense, $\boldsymbol{U}$ can be thought of as an expanded version of $\boldsymbol{W}$ (antecedents). In our previous studies (Jo et al., 2017, 2018), we in fact conducted validation using $\boldsymbol{U}$, and then used the concurrent validator ($Z$) more qualitatively after narrowing candidate labels. From the automation point of view, we find the currently proposed framework more straightforward, where the validation step (Step 2) only uses a minimal set of core validators ($\boldsymbol{W}$) carefully selected by clinical experts. Once the validation step is completed and the best risk labels are selected, one can explore with a wider array of possible predictors in Step 3.

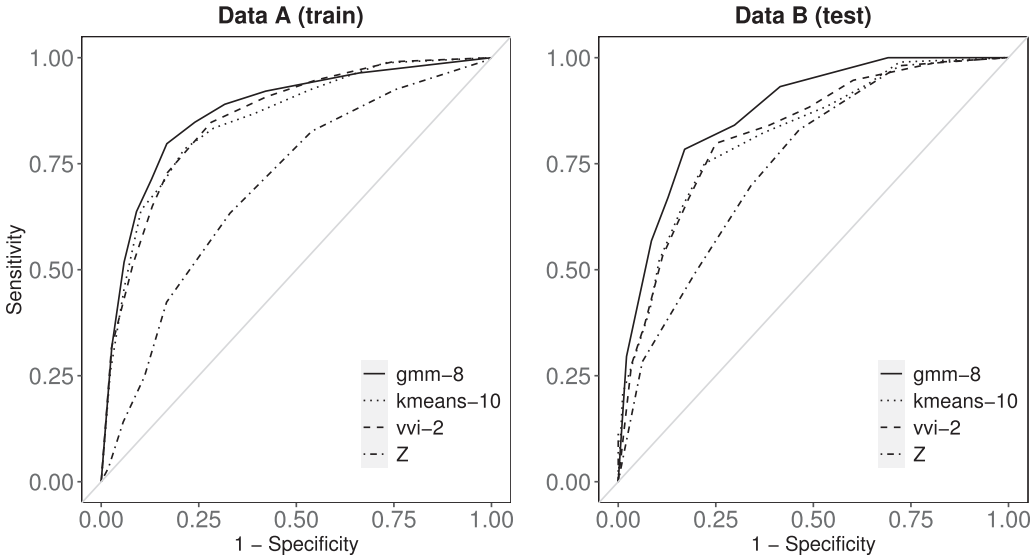


**Figure 3.** Prediction of clustering-based risk labels by baseline covariates.

**Table 4.** Clustering-based risk labels used as prediction output.

| Risk label | Data * | Sensitivity | Specificity | Accuracy | AUC † |
|---|---|---|---|---|---|
| Each clustering-based risk label predicted by 7 baseline covariates‡ | | | | | |
| gmm-8 | A | 0.83 (0.81, 0.85) | 0.80 (0.78, 0.81) | 0.82 (0.81, 0.83) | 0.81 (0.80, 0.82) |
| | B | 0.83 | 0.78 | 0.81 | 0.81 |
| vvi-2 | A | 0.80 (0.78, 0.82) | 0.77 (0.75, 0.78) | 0.78 (0.77, 0.80) | 0.78 (0.77, 0.79) |
| | B | 0.80 | 0.75 | 0.77 | 0.77 |
| kmeans-10 | A | 0.79 (0.76, 0.82) | 0.78 (0.76, 0.80) | 0.79 (0.77, 0.80) | 0.79 (0.77, 0.80) |
| | B | 0.78 | 0.75 | 0.76 | 0.76 |
| *Z* | A | 0.74 (0.71, 0.76) | 0.70 (0.67, 0.73) | 0.72 (0.71, 0.73) | 0.72 (0.70, 0.73) |
| | B | 0.70 | 0.84 | 0.76 | 0.77 |

*Data A (70% of the full data) was used to train prediction models with K-fold cross-validation. Data B (30% of the full data) was used as a test data to examine whether the prediction results are generalizable.

†Prediction performance measures including AUC (area under the curve), sensitivity, specificity, and accuracy were calculated using the same method in (15) by averaging across 10 folds. In parentheses are shown their values ±1 standard errors calculated using the same method in (16). In the test step using Data B, standard errors are not reported because K-fold cross-validation was not used.

‡Prediction input variables include seven baseline patient measures (manic symptoms by PGBI-10M, anxiety by SCARED-P, depression by CDRS-R, bipolar diagnosis, age, sex, and health insurance).

### *Prediction results in the LAMS example*

In the LAMS context, the developed prediction model is expected to aid 2-year outcome prognosis (elevated or low risk patterns within 2 years) for pediatric patients who present to outpatient clinical care with concerns in elevated manic symptoms. In addition to the antecedents used in the validation step (anxiety, depression, bipolar diagnosis), four more variables were included in $\boldsymbol{U}$. They are baseline PGBI-10M and key demographic variables that are typically correlated with psychiatric outcomes including age, sex, and health insurance as a proxy for socio economic status. We used simple logistic regression, the same method used in the validation step. Prediction performance measures including AUC (area under the curve) and their standard errors were calculated in the same way described in (15) and (16). We used 70% of the full data (Data A) to train prediction models using 10-fold cross-validation. The rest 30% of the data was used as a test data (Data B) to examine whether the prediction algorithm built based on Data A would be generalizable outside Data A.

Table 4 and Figure 3 show some preliminary results on how well we can predict clustering-based risk labels used as outcome (output) variables. We also included our primary validator ($Z$) as a reference outcome label that is not based on clustering methods. Note that the goal of Step 3 is not to compare different labels, but to develop prediction models using already validated and selected labels from Step 2. Also note that the results shown here should be considered preliminary. Fuller investigation with a larger pool of input variables using various supervised learning strategies is in order to formally develop prediction models that are ready to be deployed in clinical practice. Table 4 show that the results are highly comparable across different clustering-based labels, both LV-based

(GMM, MBC) and K-means-based, which was expected given their good agreement due to our validation method. All three clustering-based labels also show stable results between the train and test data, which is an important property in prediction. The GMM-based label (gmm-8) is slightly better predicted in the test data (see Figure 3), although the differences are small and the results may change as we introduce more covariates and use various supervised learning methods. Overall, prediction based on clustering-based labels showed promising results with AUC around 0.8, which is practically meaningful.

The results in Table 4 and Figure 3 also show how well we can predict our primary validator ($Z$). The differences between $Z$ and clustering-based labels are quite noticeable. This does not necessarily mean that $Z$ is a worse label, although knowing that $Z$ will be harder to predict is certainly useful. The goal of Step 3 is not in validating and comparing different labels based on how well they are predicted. However, one thing we could compare here is how well the prediction results can be generalized. Table 4 shows that the prediction results for $Z$ are more variable between the train and test data, especially in terms of specificity, implying possible difficulties in applying prediction models developed based on $Z$. Recall that the validation results in Step 2 (see Table 2) supported the use of clustering-based risk labels instead of $Z$. Although preliminary, the results from Step 3 seem to further support this decision.

## Conclusions

Using LV-based outcomes in developing prediction models is not a well-accepted concept either in LV modeling or in supervised learning. This is an unfortunate situation because LV strategies will facilitate utilization of rich outcome data collected from research and health services, which may lead to improved prognostic or diagnostic models for future patients. As a way of improving this situation, this study proposed a learning framework that combines the traditions of LV modeling, psychometrics, and supervised learning. At the core of this framework is the structured use of clinical validators, which makes systematic validation of LV-based outcomes possible guided by experts' knowledge and clinical practice. The proposed framework, if successfully adopted, will help position LV modeling as a key contributor in developing prediction models and in supervised learning in general.

To demonstrate possible strategies of systematic outcome validation and selection, we applied the proposed method to the LAMS data. We used two clustering methods based on LV modeling (GMM and MBC). To show the generality of our approach, we also included K-means clustering, which is not based on LV modeling, but is a better known clustering method. Using these three clustering methods, a large number of binary risk labels were generated. In the proposed framework, this exploratory situation is viewed as an opportunity to tailor desirable prediction outputs using multiple clinical validators. Cutpoint-based labels lack such flexibility. The example showed the possibility that, with structured sets of validators, a large pool of candidate risk labels can be swiftly validated and selected. This means that it is possible to make the validation process automation-ready, which is important in that it will encourage the use of LV-based outcomes in building prediction models and in supervised learning.

In the LAMS example, the validation results supported the use of clustering-based risk labels instead of cutpoint-based labels including the currently used best label (i.e., concurrent validator $Z$). Among the different clustering methods, the validation results for the selected labels were remarkably comparable despite their distinct approaches (92.7% agreement across 3 methods). Such agreement is not surprising given that a large number of candidate risk labels went through the validation process based on the same selection rules with the same clinical validators. Utilizing multiple clustering methods also provides an opportunity to identify a small portion of cases that are difficult to classify (7.3% disagreement across 3 methods), dramatically narrowing the pool of patients that need to be carefully examined by clinical experts. These cases with disagreement across clustering methods (see Table 3) also show the value of including LV-based methods (GMM, MBC) even though the common K-means clustering does a comparable job.

The validated and selected risk labels are ready to be used in developing prediction models using any types of supervised learning methods. The preliminary results in the LAMS example, based on a minimal set of baseline predictors and logistic regression analysis, showed promising results with AUC around 0.8. Note that our interest in the LAMS example was to generate risk labels to be used as output variables in developing prognostic models and algorithms. Given that, the risk labels were validated and selected focusing more on the concurrent validator ($Z$). However, if the goal is to generate risk labels to be used as prediction

inputs, the focus should be shifted. For example, if we focus more on *Q*, a distal outcome, different risk labels will be selected (other than those listed in Tables 2–4). The choice among the validation rules depends on the intended utility of generated labels. We see this flexibility as an advantage of our framework.

The proposed approach can be fine-tuned and expanded in many different ways. Some immediate extensions include the use of three-category labeling (e.g., low, medium, high risk), joint prediction of multiple outcomes (e.g., manic symptoms and anxiety), and incorporation of broader unsupervised and supervised learning methods. To focus on the conceptual framework of the proposed approach, we ignored the uncertainty surrounding the cluster membership. Extending the previous work (Jo et al., 2017), we are actively exploring practical strategies to smoothly connect LV-based soft clusters with various supervised learning methods. There is much to explore in terms of various application possibilities. We focused on prediction, although using simplified and validated latent variables can be an attractive and practical strategy to deal with complexities in building theoretical models. Applying the proposed framework in developing algorithms to help clinical diagnosis (instead of prognosis) also seems to be a promising direction.

## Article information

**Conflict of interest disclosures:** Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

**Ethical principles:** The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

**Funding:** This work was funded by grant MH123443 from the National Institute of Mental Health and grant DA031698 from the National Institute on Drug Abuse.

**Role of the funders/sponsors:** None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

## References

Bayliss, E. A., Blatchford, P. J., Newcomer, S. R., Steiner, J. F., & Fairclough, D. L. (2011). The effect of incident cancer, depression and pulmonary disease exacerbations on type 2 diabetes control. *Journal of General Internal Medicine*, *26*(6), 575–581. https://doi.org/10.1007/s11606-010-1600-x

Beseler, C. L., Taylor, L. A., Kraemer, D. T., & Leeman, R. F. (2012). A latent class analysis of DSM-IV alcohol use disorder criteria and binge drinking in undergraduates. *Alcoholism, Clinical and Experimental Research*, *36*(1), 153–161. https://doi.org/10.1111/j.1530-0277.2011.01595.x

Birmaher, B., Merranko, J. A., Goldstein, T. R., Gill, M. K., Goldstein, B. I., Hower, H., Yen, S., Hafeman, D., Strober, M., Diler, R. S., Axelson, D., Ryan, N. D., & Keller, M. B. (2018). A risk calculator to predict the individual risk of conversion from subthreshold bipolar symptoms to bipolar disorder I or II in youth. *Journal of the American Academy of Child and Adolescent Psychiatry*, *57*(10), 755–763.e4. https://doi.org/10.1016/j.jaac.2018.05.023

Bouveyron, C., Celeux, G., Murphy, T., & Raftery, A. (2019). *Model-based clustering and classification for data science: With applications in R* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press.

Brandmaier, A. M., Oertzen, T. v., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*(1), 71–86. https://doi.org/10.1037/a0030001

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46. https://doi.org/10.1177/001316446002000104

Cole, V. T., & Bauer, D. J. (2016). A note on the use of mixture models for individual prediction. *Structural Equation Modeling*, *23*(4), 615–631. https://doi.org/10.1080/10705511.2016.1168266

Dempster, A., Laird, N., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Findling, R. L., Youngstrom, E. A., Fristad, M. A., Birmaher, B., Kowatch, R. A., Arnold, L. E., Frazier, T. W., Axelson, D., Ryan, N., Demeter, C. A., Gill, M. K., Fields, B., Depew, J., Kennedy, S. M., Marsh, L., Rowles, B. M., & Horwitz, S. M. (2010). Characteristics of children with elevated symptoms of mania: The longitudinal assessment of manic symptoms (LAMS) study. *The Journal of Clinical Psychiatry*, *71*(12), 1664–1672. https://doi.org/10.4088/JCP.09m05859yel

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*, 611–631. https://doi.org/10.1198/016214502760047131

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Horwitz, S. M., Demeter, C. A., Pagano, M. E., Youngstrom, E. A., Fristad, M. A., Arnold, L. E., Birmaher, B., Gill, M. K., Axelson, D., Kowatch, R. A., Frazier, T. W., & Findling, R. L. (2010). Longitudinal assessment of manic symptoms (LAMS) study: Background, design, and initial screening results.*The Journal of Clinical Psychiatry*, *71*(11), 1511–1517. https://doi.org/10.4088/JCP.09m05835yel

Jo, B., Findling, R. L., Hastie, J. T., Youngstrom, E. A., Wang, C.-P., Arnold, L. E., et al. (2018). Construction of longitudinal prediction targets using semi-supervised learning. *Statistical Methods in Medical Research*, *27*, 2674–2693.

Jo, B., Findling, R. L., Wang, C.-P., Hastie, J. T., Youngstrom, E. A., Arnold, L. E., et al. (2017). Targeted use of growth mixture modeling: A learning perspective. *Statistics in Medicine*, *36*, 671–686.

Joo, Y. S., Lee, C., Kim, H. W., Jhee, J., Yun, H.-R., Park, J. T., et al. (2020). Association of longitudinal trajectories of systolic BP with risk of incident CKD: Results from the Korean Genome and Epidemiology Study. *Journal of the American Society of Nephrology*, *31*, 2133–2144.

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., & Govaert, G. (2015). Rmixmod: The R package of the model-based unsupervised, supervised and semi-supervised classification Mixmod library. *Journal of Statistical Software*, *67*, 241–270.

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley.

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén and Muthén.

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, *55*, 463–469.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 289–317.

Youngstrom, E. A., Birmaher, B., & Findling, R. L. (2008). Pediatric bipolar disorder: Validity, phenomenology, and recommendations for diagnosis. *Bipolar Disorders*, *10*(1 Pt 2), 194–214. https://doi.org/10.1111/j.1399-5618.2007.00563.x