# *Reorienting Latent Variable Modeling for Supervised Learning and Prediction*

Booil Jo

Department of Psychiatry & Behavioral Sciences

Stanford University

C-DIAS & PSMG Virtual Grand Rounds

February 7, 2023

# *BACKGROUND*

# Based on

- Jo B, Hastie JT, Li, Z, Youngstrom EA, Findling RL, Horwitz SM (in press). Reorienting latent variable modeling for supervised learning. *Multivariate Behavioral Research*.
- Jo B, Findling RL, Hastie JT, Youngstrom EA, Wang C-P, Arnold LE, Fristad MA, Frazier TW, Birmaher B, Gill MK, & Horwitz SM. (2018). Construction of longitudinal prediction targets using semi-supervised learning. *Statistical Methods in Medical Research*, *27,* 2674-2693.
- Jo B, Findling RL, Wang C-P, Hastie JT, Youngstrom EA, Arnold LE, Fristad MA, Horwitz SM. (2017). Targeted use of growth mixture modeling: A learning perspective. *Statistics in Medicine, 36*, 671-686.
- Funded by MH123443 from the National Institute of Mental Health and DA031698 from the National Institute on Drug Abuse

# Prediction in supervised learning

- Prediction: In developing prediction algorithms to be used in clinical practice, the usual goal is to accurately predict an outcome, which is observed in the sample used for model development, but is unobserved and needs to be predicted among new patients.
- Supervised learning methods mostly assume a simple, clear univariate outcome. It lets us focus on handling of a large pool of possible predictors of the outcome without worrying about the outcome itself.
- A single observed measure can be unreliable and can be far from a good representation of a particular patient's true outcome status.

# Improving the output side of prediction

- In the sample used in model development, abundant information is often available not only on the predictor side, but also on the outcome side.
- An ideal solution would be to create more reliable and valid outcome variables using multivariate information without losing the simplicity of a single observed outcome.
- How do we effectively organize (unsupervised learning) complex outcome data and still preserve its rich information?
- Latent variable (LV) modeling is a promising way to achieve these seemingly conflicting goals.

# Why didn't this happen already?

- In behavioral/psychological science research, LV modeling is mostly used for inference and building theoretical models
  - Developing prediction models/algorithms that will be deployed in clinical practice (or industrial) has not been a typical goal.
  - Predictive relationships are studied for inference/interpretation.
  - ➢ No need for reframing LV modeling as an unsupervised learning tool
  - ➢ No benefits of using supervised learning (LV modeling is sufficient to study predictive relationships)
- Despite great potentials, little interest in how LV modeling can play a major role in this promising venue, prediction.

# Why didn't this happen already?

- In supervised learning, possible benefits of incorporating LV modeling is basically unknown.
- In unsupervised learning, there are various established techniques such as K-means. Why bother with mysterious LV modeling?
- How LV modeling is conducted in behavioral science is not quite compatible with machine learning (cf. model-based clustering by Raftery's group, a successful exception).
- No established framework that shows how LV modeling can be incorporated into the standard process of building prediction algorithms in the machine learning framework

# Some signs of integration

- There have been some developments to improve LV modeling and SEM 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.
- Brandmaier et al. (2013) introduced regression tree methods to combine exploratory and confirmatory approaches with the goal of improving SEM model building.
- One way integration – the goal is to improve inference and model building.

# Motivating/tricky situation

- Theoretical psychologists and modelers
- Machine learners (unsupervised and supervised)
- Statisticians
- Medicine (clinicians and clinical researchers)
  - Psychiatry
  - Diabetes
  - CVD
  - Developmental and aging related outcomes
  - Many other where multivariate/longitudinal patterns matter
  - Both prognosis and diagnosis

# Motivating example: LAMS

- Longitudinal Assessment of Manic Symptoms Study (Findling et al., 2010; Horwitz et al., 2010; Youngstrom et al., 2008)
- Children aged 6–12 years at baseline
- Focused on elevated symptoms of mania over time, which fundamentally differentiates it from other studies that have focused on diagnosis of bipolar disorder and its risk.
- Primary outcome: PGBI–10M (Parent General Behavior Inventory–10–item Mania Form) (Youngstrom et al., 2008)
- Outcomes were measured every 6 months for 10 years, leading to an impressive collection of rich longitudinal data.

# Clinical need for prediction: LAMS

- Early prediction of outcome progression is crucial in treatment decision making and patient care.
- In particular, separating out patients who would maintain moderate levels of symptoms (low risk) is critical in planning optimal treatments, better allocating resources, and reducing patient burdens.
- Great interest in developing prediction models using the LAMS data that can be deployed in clinical practice to aid prognosis among incoming patients.

# What to predict: LAMS

- Constructing a simple, reliable, and valid prediction target is the critical first step in developing useful prediction models.
- Uniqueness of clinical prediction: unlike in industrial prediction models, using clinically meaningful and interpretable outcomes and predictors is important in clinical prediction models.
- Given the richness of longitudinal data, it is not self-evident how to formulate a simple prediction target that best captures individuals' longitudinal symptom patterns.

# What to predict: LAMS

- Difficulties in effectively characterizing longitudinal progression at the individual level.
- Ad hoc approaches: predict outcome at each time point, an average, a peak, dichotomized using a cutpoint, etc. – not the best way to characterize/utilize rich patient data.
- Standard single-class linear mixed effects modeling does not provide good individual level summary measures that can be used as prediction outcomes.
- Using latent classes seems to be a promising way of effectively organizing complex outcome data – also well aligned with categorical decision making in clinical practice.

# Why categorize: LAMS

- In clinical practice, clinicians need to make swift decisions.
- Clinical decisions are often made in a categorical manner (e.g., surgery or not, medicine or not, treatment A or B).
- Categorical decision making is facilitated by categorized outcomes (disease or not, high or low risk) often using clinical cutpoints. E.g.,
  - Elevated symptoms of mania if PGBI-10M >=12
  - PHQ9 total (moderately severe depression if 15-19, severe depression if 20-27)
  - Cholesterol (total < 200 mg/dL, Triglycerides < 150, LDL < 130, HDL > 55 (females) & > 45 (males))
  - A1C (normal if < 5.7%, prediabetes if between 5.7-6.5, diabetes if >= 6.5)

# Why LV modeling: LAMS

- Why not use simple clinical cutpoints that are well accepted?
  - To better characterize individuals utilizing multivariate outcome information.
  - To generate prediction outcomes with improved reliability and validity
  - To develop prediction models/algorithms with improved accuracy and generalizability

# *3-STEP LEARNING FRAMEWORK*

# 3-Step Learning Framework

- Step 1. Generate latent outcome labels
  - Using LV modeling, generate outcomes to be used as output in prediction models. In line with LAMS and common clinical decision making, we focus on generating binary risk labels using latent class and clustering results.
- Step 2. Systematically validate generated risk labels
  - Following the psychometrics tradition, we validate a large pool of candidate labels using well-structured explicit validators (concurrent, antecedent, consequent validators).
- Step 3. Using the validated risk labels, develop prediction models in the supervised learning framework

# A 3-step learning pipeline with LV modeling

**Observed Full Data**

Variables to be used to generate latent labels

Variables to be used to validate latent labels

| Table A | Multivariate Outcome | Baseline | Validators | | |
|---|---|---|---|---|---|
| Observation | $d = (1, 2, \ldots, D)$ | Covariates | Concurrent | Consequent | Antecedent |
| $i=1$ | $Y_{11}, Y_{12,\ldots}, Y_{1D}$ | $\boldsymbol{X}_1$ | $Z_1$ | $Q_1$ | $\boldsymbol{W}_1$ |
| $i=2$ | $Y_{21}, Y_{22,\ldots}, Y_{2D}$ | $\boldsymbol{X}_2$ | $Z_2$ | $Q_2$ | $\boldsymbol{W}_2$ |
| . | . | . | . | . | . |
| $i=N$ | $Y_{N1}, Y_{N2,\ldots}, Y_{ND}$ | $\boldsymbol{X}_N$ | $Z_N$ | $Q_N$ | $\boldsymbol{W}_N$ |

**1. Generate latent labels**

1a. Cluster individuals based on multivariate outcomes using various clustering methods such as growth mixture modeling (GMM), model-based clustering (MBC), or K-means clustering.

1b. Construct simple labels based on a practical purpose (e.g., low vs. high risk groups):
- In each GMM, MBC, or K-means clustering *(m = (1, 2,..., M))*, latent classes (clusters) are regrouped (e.g., split into two) using all possible ways of splits *(s = (1, 2,..., S))*.
- For individual $i$, a latent label $(L_i^{ms})$ is created based on a parent model $m$ and a splitting method $s$ . All labels $(L_i^{ms})$ each individual $i$ can have are described in Table B.

| Table B | Grouping (splitting) of latent classes (clusters) | | | |
|---|---|---|---|---|
| Parent clustering model | $s = 1$ | $s = 2$ | ... | $s = S$ |
| $m = 1$ | $L_i^{11}$ | $L_i^{12}$ | | $L_i^{1S}$ |
| $m = 2$ | $L_i^{21}$ | $L_i^{22}$ | | $L_i^{2S}$ |
| . | . | . | . | . |
| $m = M$ | $L_i^{M1}$ | $L_i^{M2}$ | | $L_i^{MS}$ |

**2. Validate clustering-based risk labels in the prediction framework**

- Validate each label ($L^{ms}$) using clinical validators (concurrent $Z$, consequent $Q$, antecedent $\boldsymbol{W}$).
- Using logistic regression, get predicted $\widehat{L^{ms}}$ using V, where $\mathrm{V} \in \{Z, Q, \boldsymbol{W}\}$. I.e., $logit(\pi_{L^{ms}}) = \beta_0 + \beta_1 V$.
- Calculate Cohen's kappa between $L^{ms}$ and $\widehat{L^{ms}}$ as a measure of association between $L^{ms}$ and V.
- If K-fold cross-validation is used, average kappa estimates across K-folds.

# A 3-step learning pipeline with LV modeling

**3. Supervised learning treating validated clustering-based risk labels as known**

Randomly select 70% of the full data as training data (Data A) and 30% as test data (Data B).

3a. Data A : Using logistic regression (or other supervised learning methods), predict a selected latent label ($L^{ms}$) by baseline patient characteristics ($\boldsymbol{U}$) with K-fold CV. I.e., $logit(\pi_{L^{ms}}) = \beta_0 + \boldsymbol{\beta_1 U}$.
- Calculate AUC between $L^{ms}$ and $\widehat{L^{ms}}$ as a measure of prediction accuracy.
- Average prediction accuracy measures across K-folds.

3b. Data B: Apply overall $\beta_0$ and $\boldsymbol{\beta_1}$ from 3a to Data B, get predicted $\widehat{L^{ms}}$.
- Calculate AUC between $L^{ms}$ and $\widehat{L^{ms}}$ as a measure of prediction accuracy in the test data.

# STEP 1. Generate LV-based risk labels

# 1.1. Identify latent classes and clusters

- We used 2 LV approaches so far
  - Growth mixture modeling (GMM)
  - Model-based clustering (MBC)
- We used 1 model-free, non-LV approach
  - K-means clustering (K-means): not model based, but is the most commonly used clustering method, well-covered in any machine learning textbooks. We used `kmeans` function in R.
- All these approaches can be seen as clustering
- All these approaches can be seen as unsupervised learning (no direct measures of success such as prediction accuracy)

# GMM & MBC

- Both utilize finite mixture modeling. We can use the common framework for the two methods.
- Let us consider data with $d$ multivariate measures for unit $i$ (individual $i$ in LAMS application). i.e., $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{id})$
- The probability distribution of $Y_i$ with J mixture components ( j = 1,2,...,J) can be expressed as

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

where $\theta_j$ 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 $\sum_{j=1}^{J} \pi_j = 1$.

- GMM focuses on modeling the trend, whereas MBC focuses on modeling the covariance structure.

# Growth mixture modeling (GMM)

- A popular method of discovering latent trajectory types in various areas of research (Muthén & Shedden, 1999), although not incorporated in the unsupervised learning tool box yet.
- Focus on inference/interpretation about longitudinal patterns – relatively simple and interpretable parametrization (explicitly model where they start, how they change).
- Unlike in industrial prediction models, using clinically meaningful and interpretable outcomes and predictors is important in clinical prediction models.
- GMM seems to be a promising unsupervised learning tool to generate good prediction output (also input).

# GMM for LAMS

- The outcome Y for individual i (i=1, 2,..., N) at time t (t=1, 2,..., d) conditioned on trajectory class $C_i = j$ can be expressed as

$$\begin{aligned} Y_{it}|(C_i = j) &= \eta_{1ij} + \eta_{2ij} T_t + \eta_{3ij} T_t^2 + \varepsilon_{ijt}, \\ \eta_{1ij} &= \eta_{1j} + \zeta_{1ij}, \\ \eta_{2ij} &= \eta_{2j} + \zeta_{2ij}, \\ \eta_{3ij} &= \eta_{3j} + \zeta_{3ij}, \end{aligned}$$

  where there are three growth parameters to capture change: initial status ($\eta_{1j}$), linear ($\eta_{2j}$) and quadratic growth ($\eta_{3j}$) for trajectory class j. The time measure $T_t$ reflects linear and $T_t^2$ quadratic growth. We assumed $\varepsilon_{ij} \sim MN(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$ and $\zeta_{1i} \sim MN(0, \Sigma_\zeta)$ .

- Complex models with fewer or simple models with more classes – for LAMS, we used simple models without covariates. Radom intercept and no random effect models are included. In principle, a large pool of models can be used (Step1 can be automated).

# GMM for LAMS

- ML-EM estimation using Mplus (Muthén & Muthén, 1997-2017)
- The posterior class probability of subject i belonging to class j conditioned on observe data ($Y_i$) and the current estimates of model parameters ( $\boldsymbol{\eta}_j^*, \boldsymbol{\Sigma}_\zeta^*, \boldsymbol{\Sigma}_\varepsilon^*$ ) in the iterative procedure is expressed as

$$p_{ij}(\boldsymbol{\eta}_j^*, \boldsymbol{\Sigma}_\zeta^*, \boldsymbol{\Sigma}_\varepsilon^*) = \frac{\pi_{ij} f(Y_i \mid C_i = j, \boldsymbol{\eta}_j^*, \boldsymbol{\Sigma}_\zeta^*, \boldsymbol{\Sigma}_\varepsilon^*)}{\sum_{j'=1}^{J} \pi_{ij'} f(Y_i \mid C_i = j', \boldsymbol{\eta}_{j'}^*, \boldsymbol{\Sigma}_\zeta^*, \boldsymbol{\Sigma}_\varepsilon^*)},$$

where $\boldsymbol{\eta}_j = (\eta_{1j}, \eta_{2j}, \eta_{3j})$, $\sum_{j=1}^{J} \pi_{ij} = 1$ for $i = 1, \cdots, N$, and $\pi_{ij} = Pr(C_i = j)$, the probability of subject $i$ belonging to a certain trajectory class.

- $p_{ij}$ is valuable summary information that characterizes each person.
- Excluding models with classes with <10 individuals, a total of 13 models reached normal convergence (6 random intercept models with 2-7 classes, 7 no random effect models with 2-8 classes).

# Model-based clustering (MBC)

- MBC is a line of method that focuses on identification of latent classes (clusters) based on finite mixture modeling of multivariate normal distributions (Bouveyron et al., 2019; Fraley & Raftery, 2002; Scrucca et al., 2016).
- GMM and MBC basically share the same analytical (finite mixture modeling) and estimation (ML–EM or Bayesian) strategies.
- MBC is a more widely known tool for unsupervised learning (e.g., imaging data, microarray data, retail barcode data).
- 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.

# MBC for LAMS

- 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

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

  where various geometric constraints on the variance/covariance matrix of $\varepsilon_{ijt}$ are the key to the identification of latent clusters.

- Geometric constraints are imposed on volume, shape, and orientation of the ellipsoidal distribution (Bouveyron et al., 2019; Lebret et al., 2015; Scrucca et al., 2016).

- ML-EM estimation using 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).

# MBC for LAMS

- The posterior class probability of subject i belonging to cluster j conditioned on observed data $(Y_i)$ and the current parameter estimates $(\boldsymbol{\eta}_j^*, \boldsymbol{\Sigma}_j^*)$ in the iterative procedure can be expressed as

$$p_{ij}(\boldsymbol{\eta}_j^*, \boldsymbol{\Sigma}_j^*) = \frac{\pi_{ij} f(Y_i \mid C_i = j, \boldsymbol{\eta}_j^*, \boldsymbol{\Sigma}_j^*)}{\sum_{j'=1}^{J} \pi_{ij'} f(Y_i \mid C_i = j', \boldsymbol{\eta}_{j'}^*, \boldsymbol{\Sigma}_{j'}^*)}$$

- $p_{ij}$ is valuable summary information that characterizes each individual.

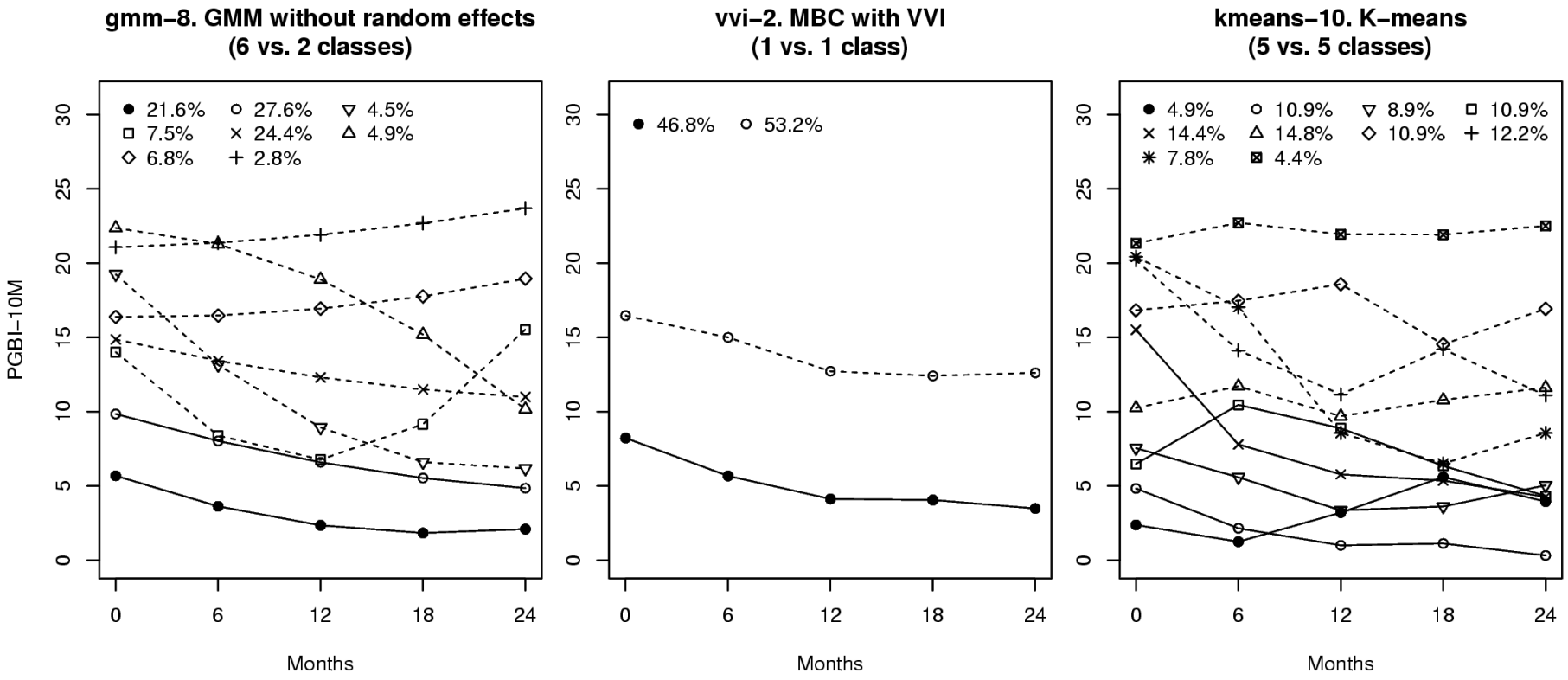
- Using the LAMS data, we estimated a series of MBC models using all 14 types of geometric constraints allowing up to 19 classes. We obtained a total of 135 MBC models with 2–19 classes.

# 1.2. Generate simple outcome labels

- Clinical decisions are often made in a categorical manner (e.g., surgery or not, medicine or not, treatment A or B).
- Categorical decision making is facilitated by categorized outcomes (disease or not, high or low risk).
  - E.g., 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.
- It is practical to further simplify generated latent classes and clusters in line with the intended clinical utility.
- This will also make the generated outcome labels easier to handle in supervised learning – note that we are shifting our interest from inference to prediction.

# 1.2. Generate simple outcome labels

- We focus on regrouping of individuals into two groups
  - In all possible ways, and therefore results in a large pool of candidate binary outcome labels. E.g., from LAMS,



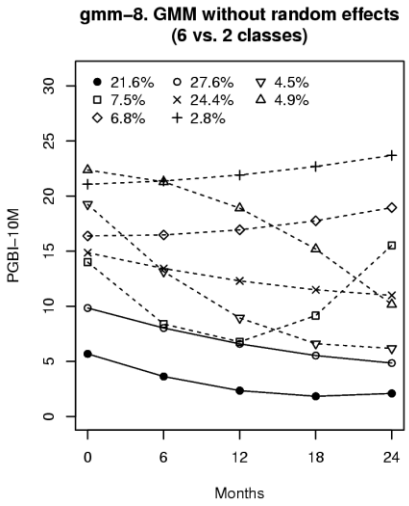
# 1.2. Generate simple outcome labels

- 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 applying a cutpoint (e.g., PGBI-10M >=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).

# 1.2. Generate simple outcome labels

- When using LV modeling, regrouping at the individual level based on the posterior class probability ($p_{ij}$).
- 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 and $1 - p_i^{ms}$ for the second group.

e.g., from gmm–8, person *i* has a set of posterior class probabilities, $p_i = (p_{i1}, p_{i2}, p_{i3}, p_{i4}, p_{i5}, p_{i6}, p_{i7}, p_{i8})$. One possible split would be into $p_i^{ms} = (p_{i3} + p_{i4} + p_{i5} + p_{i6} + p_{i7} + p_{i8})$ and $1 - p_i^{ms} = (p_{i1} + p_{i2})$.



# 1.2. Generate simple outcome labels

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

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

- We will use this strategy to simplify comparisons across soft clustering (GMM, MBC), hard clustering (standard K-means), and cutpoint-based categorization methods.
- With the LAMS data, we obtained 367 binary labels based on GMM models, 954,755 labels based on MBC models, and 6,142 labels based on K-means. All automated in our program.
- When using LV modeling (soft clustering), it is in principle possible to account for uncertainty in cluster assignment (e.g., Jo et al., 2017).

# 1.2. Generate simple outcome labels

- Having simple output variables is a big step toward supervised learning. However, being simple does not guarantee the validity of the outcomes.
- LV modeling can generate a large pool of candidate outcomes in the absence of known truth, which can be viewed as a big drawback.
- The same exploratory situation can be viewed as an opportunity to tailor the most desirable prediction targets based on good validation strategies.
- Observed and cutpoint–based measures do not possess such flexibility.

# STEP 2. Systematic validation of generated outcome labels

# Validating candidate outcomes

- 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.
- This is possible due to the simple structure of considered models (predictors and the predicted).
- It is typically assumed that the outcome is clear, simple, and readily available in the data, which lets us focus on assessing how accurately and stably various combinations or subsets of predictors predict the outcome.
- From the perspective of supervised learning, validating outcomes before predicting them is a foreign concept and an unnecessary step.

# Validating candidate outcomes

- In psychometrics, it is a long tradition to question the validity of measured outcomes.
- Various concepts have been developed 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 as it gives LV-based outcomes concrete meanings by connecting them with scientifically or clinically meaningful validators.

# Clinically meaningful validators

- In line with the psychometrics tradition, we will use well-structured validation with multiple criteria. Specifically, we will use clinically meaningful validators selected by experts in clinical and psychometrics fields.
- Validation: 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 algorithm developers, clinical researchers, practitioners, and patients).
- Selection: 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.

# Clinically meaningful validators

- Focusing on the LAMS context, we chose three types of validators targeting to identify outcome labels that well capture long-term progression of manic symptoms.
  - Concurrent validator (Z): This is a primary validator that ensures that developed measures are closely related to what is currently used and well-accepted (the incumbent). An established clinical cutpoint (any PGBI-10M >=12 as elevated risk) is applied to all repeated measures within the prediction range (6 to 24 months).
  - Consequence (Q): Consequences are future outcomes that are supposed to be correlated with the developed measures. In LAMS, distal future risk beyond the timeframe of prediction interest. The same cutpoint is applied to all PGBI-10M measures between 30 to 48 months.
  - Antecedents (**W**): Clinically relevant variables that precede and are supposed to be correlated with the new measure. In LAMS, our clinical experts identified 3 variables (bipolar diagnosis, anxiety by SCARED-P, depression by CDRS-R) as directly relevant clinical antecedents.

# Towards Automation

- The flexible nature of LV modeling can make resulting LV–based outcomes look subjective and esoteric. Not attractive as prediction output, especially with a large number of candidate outcomes and apparently cumbersome extra steps.
- Integrating the validation concept from psychometrics and having a structured validation plan using explicit validators dramatically changes this situation.
- This means that automation of the validation process is possible guided by experts' knowledge and clinical practice.
- This makes LV modeling a more adoptable option in formulating prediction outcomes despite some extra steps.

# Validation in the prediction framework

- The association between a candidate outcome label from Step 1 and each set of validators can be put in the prediction framework, e.g., using logistic regression, as

$$
\begin{aligned}
logit(\pi_{L_i^{ms}(Z_i)}) &= \alpha_{Z0}^{ms} + \alpha_{Z1}^{ms} Z_i, \\
logit(\pi_{L_i^{ms}(Q_i)}) &= \alpha_{Q0}^{ms} + \alpha_{Q1}^{ms} Q_i, \\
logit(\pi_{L_i^{ms}(\boldsymbol{W}_i)}) &= \alpha_{W0}^{ms} + \boldsymbol{\alpha}_{W1}^{ms} \boldsymbol{W}_i,
\end{aligned}
$$

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 $\boldsymbol{W}_i$.

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

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

# Validation in the prediction framework

- To measure the degree of agreement between the candidate label ($L_i^{ms}$) and the estimated label ( $\widehat{L}_i^{ms}(Z_i)$, $\widehat{L}_i^{ms}(Q_i)$, $\widehat{L}_i^{ms}(\boldsymbol{W}_i)$), we used Cohen's $\kappa$ (Cohen, 1960).
- 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 not in the psychometric validation context (Jo et al., 2017, 2018).
- Combining these traditions, cross-validated kappa for a candidate label $L_i^{ms}$ using validator $Z$ can be calculated averaging across K folds (f=1,2,3,...,K) as

$$CV_{\kappa}^{Zms} = \sum_{f=1}^{K} \kappa_f^{Zms} / K \, .$$

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.

# Validation in the prediction framework

- The associated standard error for can be calculated considering the variance across K folds as

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

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.

# Validation results in the LAMS example

- We first selected 10 best candidate labels from each method based on their association with primary validator Z.
- Then, we eliminated those that are worse than the best based on all accounts (i.e., association with Z, Q, and **W**).
- Using this simple rule, we selected two best outcome labels from each method. The choice among selection rules depends on the intended utility of generated labels.
- One alternative would be to average association measures with equal weights (all-around). 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.
- In LAMS example, we focus more on Z given our interest in generating risk labels to be used as outputs in developing prognostic algorithms.

# Validation results in the LAMS example

| 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 12m $\geq 12$ | | 31.7 | 0.57 (0.51, 0.62) | 0.41 (0.36, 0.45) | 0.18 (0.13, 0.23) |
| PGBI-10M at 24m $\geq 12$ | | 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, 6-24m $\geq 12$) | | 53.0 | . | 0.41 (0.37, 0.45) | 0.26 (0.21, 0.32) |
| average$^{\ddagger}$ PGBI-10M, 6-24m $\geq 12$ | | 30.3 | 0.56 (0.53, 0.59) | 0.48 (0.44, 0.52) | 0.18 (0.15, 0.22) |

- The results clearly show the benefits of using clustering, both LV-based (GMM, MBC) and K-means-based.

# Validation results in the LAMS example: clustering vs. cutpoint

- The results clearly show the benefits of using clustering, both LV-based (GMM, MBC) and K-means-based.
- Even the primary validator (Z) shows weaker association with the rest of validators (Q, **W**).
- Cutpoint-based labels categorized much fewer patients as elevated risk, misaligned with the clinical intention of safely separating out low risk patients, implying missed opportunities for proper early treatments for elevated risk patients.
- 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.

# Validation results in the LAMS example

| 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 12m $\geq 12$ | | 31.7 | 0.57 (0.51, 0.62) | 0.41 (0.36, 0.45) | 0.18 (0.13, 0.23) |
| PGBI-10M at 24m $\geq 12$ | | 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, 6-24m $\geq 12$) | | 53.0 | . | 0.41 (0.37, 0.45) | 0.26 (0.21, 0.32) |
| average$^{\ddagger}$ PGBI-10M, 6-24m $\geq 12$ | | 30.3 | 0.56 (0.53, 0.59) | 0.48 (0.44, 0.52) | 0.18 (0.15, 0.22) |

- The results show that the validation results are remarkably comparable across different clustering methods (GMM, MBC, K-means).

# Validation results in the LAMS example:
## across different clustering methods

- 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 these labels have been already selected out of a very large pool of candidate labels based on the same clinical validators.
- One may still attempt to choose one best label for the intended purpose, e.g., by examining individual patients that showed any disagreement in labeling across methods (7.3% of the LAMS sample).

# Examples of disagreement across outcome 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 |

- 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, a concerning pattern from the experts' point of view.

# Examples of disagreement across outcome 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 |

- Patients C and D are labeled as low risk by all methods except by Z. These patients have one PGBI measure at the cutpoint, although the rest are safely below the cutpoint.

# Examples of disagreement across outcome 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 |

- 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.

# Examples of disagreement across outcome 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 |

- 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 earlier missing data. Patient H has one score above the cutpoint at baseline, and has several missing measurements. Only vvi−2 and experts labeled this patient as elevated risk.

# Examples of disagreement across outcome 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 |

- Patients I and J show that vvi-2 is not necessarily the most conservative of the three clustering methods.

# Utilizing multiple clustering methods

- Utilizing multiple clustering methods seems to be an effective way of narrowing patients that are difficult to classify. This makes careful examination by clinical experts feasible.
- Such examination results can be incorporated into and improve the validation process, e.g., 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 cutpoint, but also how the scores change over time.
- Note that the three methods were largely consistent in labeling the patients (92.7% agreement in LAMS). Further investigation is necessary in various application contexts.

# STEP 3. Prediction of validated and selected outcome labels

# Supervised learning with selected labels

- Once the best label is validated and selected based on clinical validators and practical utilities, we can focus on developing prediction models in Step 3.
- 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 prognostic models. 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.

# Supervised learning with selected labels

- Once the validation step is completed and the best risk labels are selected, one can focus on supervised learning with a wider array of possible predictors.
- Let $U$ represent the set of baseline variables to be used as predictors. In the clinical context, $U$ is expected to provide not only good prediction, but also good interpretation. In that sense, $U$ can be thought of as an expanded version of $W$, a minimal set of core antecedent validators carefully selected by clinical experts.
- In the LAMS example, in addition to the antecedents used in the validation step (anxiety, depression, bipolar diagnosis), four more variables were included in $U$. They are baseline PGBI-10M and key demographic variables, typically correlated with psychiatric outcomes including age, sex, and health insurance as a proxy for socio economic status.

# Supervised learning with selected labels

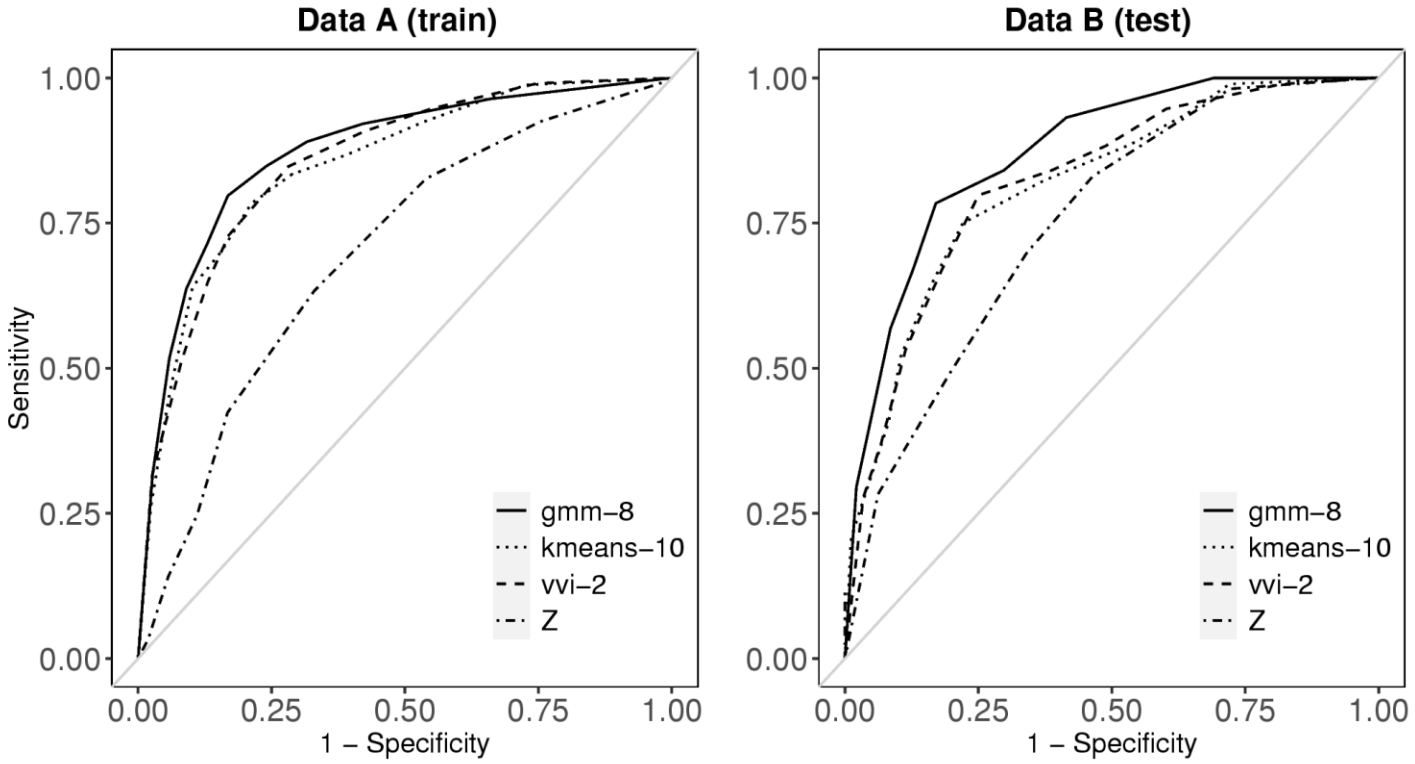
- For the demonstration purpose, we used simple logistic regression, the same method used in the validation step.
- Prediction performance measures including AUC (area under the curve) and their standard errors were calculated in the same way used in the validation step.
- We used 70% of the full data (Data A) to train prediction models using 10-fold CV. 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.
- 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.

# Prediction Results in the LAMS Example

| 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 |

- The results are highly comparable across different clustering-based labels, both LV-based (GMM, MBC) and K-means-based.

# Prediction Results in the LAMS Example




- The results are highly comparable across different clustering-based labels, both LV-based (GMM, MBC) and K-means-based.
- gmm-8 is slightly better predicted in the test data, although the differences are small.

# Prediction Results in the LAMS Example

- The results are 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, although the differences are small and the results may change as we introduce more covariates and use various supervised learning methods.
- Overall, prediction using clustering-based outcome labels showed promising results with AUC around 0.8, which is practically meaningful.

# Prediction Results in the LAMS Example

- 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 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.
- The results based on LAMS 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.

# 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.

# Conclusions

- 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 automate the validation process, which is important in that it will encourage the use of LV-based outcomes in building prediction models and in supervised learning.
- 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.

# Conclusions

- In the LAMS example, the validation results supported the use of clustering-based labels instead of cutpoint-based labels including the currently used best label (i.e., concurrent validator Z).
- It is important to note that the choice among the validation rules depends on the intended utility of generated labels. We see this flexibility as an advantage of our framework.
- Utilizing multiple clustering methods provides an opportunity to identify a small portion of cases that are difficult to classify, dramatically narrowing the pool of patients that need to be carefully examined by clinical experts.
- These cases with disagreement across clustering methods also show the value of including LV-based methods even though the common K-means clustering does a comparable job.

# Ongoing & future work

- 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/supervised methods.
- Dealing with the uncertainty surrounding the cluster/class membership. We are actively exploring practical strategies to smoothly connect LV-based soft clusters with various supervised learning methods.
- Explore various application possibilities. e.g.,
  - Using simplified and validated LVs 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 promising.

# Software

- The initial 3-step learning framework has been automated and will be available soon (as free R package).
- Almost ready!