**Distribution Agreement**

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature: Date:

_____ _____

Wenyi Fan

# Latent Trajectory Analysis of Youth at High Risk of Psychosis

By

Wenyi Fan

Master of Science in Public Health

Emory University

Rollins School of Public Health

Department of Biostatistics and Bioinformatics

---

John Hanfelt, Ph.D.

Thesis Advisor

---

Robert Lyles, Ph.D.

Thesis Reader

# Latent Trajectory Analysis of Youth at High Risk of Psychosis

By

Wenyi Fan

B.S.

Beijing University of Technology

2015

Thesis Committee Chair: John Hanfelt, Ph.D.

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2017

Abstract

Latent Trajectory Analysis of Youth at High Risk of Psychosis

By Wenyi Fan

Psychotic disorders are a group of serious illnesses that affect the mind and cause abnormal thinking and perceptions. These disorders affect over 5% of the population. Among illnesses that affect people ages 15 to 44, schizophrenia is the 8th leading cause of disability worldwide. The first goal of this analysis was to conduct a latent class analysis(LCA) on the longitudinal trajectories of clinical characteristics of youths at Clinical High-Risk(CHR) of psychosis. A second purpose of this study was to examine whether there was an association between two-year clinical outcome and empirically derived CHR subgroups. In this analysis, we used LCA to analyze variables collected as part of the North American Prodrome Longitudinal Study 2 (NAPLS2). The results showed that the two-class model was preferred according to the model selection criteria, AIC, BIC, and ICL-BIC. Based on these two subgroups, a multinomial logistic regression analysis indicated that there was a significant relation between the participants' classifications into latent subgroups and their clinical outcomes.

# Latent Trajectory Analysis of Youth at High Risk of Psychosis

By

Wenyi Fan

B.S.

Beijing University of Technology

2015

Thesis Committee Chair: John Hanfelt, Ph.D.

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2017

## Acknowledgements

I want to thank the professors, advisors, and staffs of the Biostatistics Department at Rollins School of Public Health for the two years of learning that I have had. I would like to express my extreme thanks for my thesis advisor, Dr. Hanfelt, whose encouragement, guidance, and support made this thesis possible. I am also thankful for Dr. Lyles for taking time to read my thesis. Finally, I would like to thank my parents for their loving support and encouragement through my years of study.

# Table of Contents

**I. Introduction ..... 1**

1. Study topic: The Prodromal Symptoms of Psychosis ..... 1

2. Statistical topic: Latent class analysis ..... 3

**II. Method ..... 5**

1. Coding the study variables ..... 5

2. Latent class membership probability ..... 6

3. Class-specific marker trajectory ..... 7

4. The missing weight estimation ..... 7

5. Maximum Likelihood estimation ..... 9

6. Model selection criteria ..... 9

7. Multinomial Logistic Regression ..... 10

8. Proportional Odds Model ..... 11

**III. Results ..... 12**

**IV. Discussion ..... 18**

1. Discussion of the results ..... 18

2. Future work ..... 18

**V. Reference ..... 20**

**VI. Appendix ..... 23**

1. Sample and Two-year clinical outcome criteria ..... 23

2. Proportional odds model results ..... 23

## I. Introduction:

1. Study topic: The Prodromal Symptoms of Psychosis

Psychotic disorders are a group of serious illnesses that affect the mind and cause abnormal thinking and perceptions. The symptoms include having difficulties in thinking clearly, making good judgements, responding emotionally, communicating effectively, understanding reality and behaving appropriately. When severe symptoms occur, people with psychotic disorders have trouble living in the real world and are often unable to handle everyday life. There are many types of psychotic disorders, including Schizophrenia, Bipolar illness, Depression with psychotic features, Organic psychosis, etc. They are differentiated by their symptoms and length of the illness.

The causes of most cases of psychotic disorders are unclear. Previous studies indicate that a combination of biological factors, such as genetic factors, may make a person at a greater risk. Several brain chemicals may play a role in developing psychosis. Stressful events may trigger psychotic symptoms in a person who is vulnerable.

Scientists have been investigating effective ways to detect and prevent schizophrenia and other psychotic disorders in their early stages, by studying young people who may be at risk of developing a psychotic illness. Identifying predictors and mechanisms of conversion to psychosis among such individuals ascertained to be in a clinical high risk (CHR) or prodromal clinical state are critical steps in the search for preventive strategies for psychosis (Addington et al., 2015).

In this study, we explored the prodromal symptoms of 764 CHR participants recruited in the multi-site North American Prodromal Longitudinal Study (NAPLS 2). This is a consortium of eight programs focusing on psychosis prodrome in North

America. The sites are located at Emory University, Harvard University, University of Calgary in Canada, University of California at Los Angeles, University of California at San Diego, University of North Carolina at Chapel Hill, Yale University, and Zucker Hillside Hospital. Assessment areas of the study include psychopathology, early risk factors, social functioning, social cognition, neuropsychology, treatment monitoring, neuroimaging, electrophysiology, stress and hormones, and genomics (Addington et al., 2012).

Symptoms were rated on the Scale of Prodromal Symptoms (SOPS) at baseline and 6-, 12-, 18-, and 24-month follow-ups. The SOPS is a 19-item scale designed to measure the severity of prodromal symptoms. The SOPS contains four subscales for Positive (5 items), Negative (6 items), Disorganization (4 items) and General Symptoms (4 items). Each item has a severity scale rating from 0 (Never, Absent) to 6 (Severe/ Extreme – and Psychotic, for the positive items).

Most measures of functioning in schizophrenia research have been designed to study individuals with an established psychotic illness, and hence are ill suited for studying the generally younger and less-disabled CHR population. Two new measures of global functioning have been developed for the CHR period, namely the Global Functioning: Social (GFS) and the Global Functioning: Role (GFR) scales (Cornblatt et al. 2007a). These measures were designed to represent parallel, well-anchored scales that account for age and phase of illness, distinguish social functioning from role performance, and detect functional changes over time. The GFS scale emphasizes age-appropriate social contacts and interactions outside the family, with a particular focus on social withdrawal and isolation. Ratings on the GFR scale are based on demands of role

and the level of support provided to the individual, and they reflect the individual's overall performance in the role given the level of support (Addington et al., 2012).

2. Statistical topic: Latent class analysis

Latent class analysis (LCA) is a statistical procedure for defining the structure underlying a set of measurements based on their empirical (or observed) associations. According to this method, the population of subjects are considered as heterogeneous, with homogeneous latent subgroups of subjects that share the same marker trajectory and the same risk of distal events (Proust-Lima et al., 2014). Latent variable methods would allow the data to guide our understanding of how prodromal symptoms are related and how they might best be combined to form an overall measure of prodromal symptoms. Latent class models involve the identification of mutually exclusive groups, or classes, of individuals on the basis of the pattern of response to a set of measurements (Reboussin et al., 2002). By defining classes, we can detect whether there is evidence supporting qualitative differences between classes and assess the trajectories in each class. Also, we can investigate the link between latent classes and clinical outcomes.

To study changes in prodromal symptoms by using the data from NAPLS 2, we should consider the possibility that non-response in this sample of CHR participants is related to measured and/or unmeasured health outcomes. This is a concern especially for risk factors, such as depression, that influence dropout. By using a standard method of dealing with the missing data, inverse-probability-of-attrition weights, we adjusted the latent class trajectory model to make it more representative of the overall CHR population.

In this study, we used the LCA approach to analyze variables collected as part of

the North American Prodrome Longitudinal Study (NAPLS 2). We hypothesized that incorporating information concerning Scale of Prodromal Symptoms (SOPS), Role Functioning Scale, Social Functioning Scale, life events, daily stress level and substance abuse status to distinguish subgroups of CHR would allow us to expand the phenotype. A second purpose of this study was to examine whether there was an association between clinical outcome and empirically derived CHR subgroups.

## II. Method

1. Coding the study variables

To characterize the diversity of prodromal subgroups, we selected some relevant longitudinal clinical variables from the NAPLS2 dataset. The feature variables include SOPS (Negative, Disorganization, General), Alcohol and Drug usage (Total, Tobacco, Alcohol, Marijuana), Life events (Dependent, Independent, Desirable, Undesirable), Daily stress (Total score and Total number of events), GFS and GFR. We did not include the SOPS Positive score as a longitudinal feature in the latent class analysis, since SOPS Positive is a crucial component in clinical status assessment. In each of the other SOPS symptom categories (Negative, Disorganization, General), we used the sum of all the individual item scores to summarize the participants' status; larger score indicates more severe impairment. The alcohol and drug usage (AUSDUS) scale was included and coded by total score, tobacco, alcohol and marijuana usage. We included a life events scale into the model, since previous studies have shown a link between life events and the emergence of psychotic symptoms (Bebbington et al., 1993). Life events were categorized into 4 groups: Dependent, Independent, Desirable, Undesirable. Daily stress was included in the model using two measures, the total stress severity score and the total number of stress events. We also included GFS and GFR in the model. These two assessments of social functioning and role functioning possess good psychometric properties; prior studies have shown that, relative to non-psychiatric control subjects, CHR individuals display significantly impaired social and role functioning at the time of initial clinical assessment and that social impairments persist over time and are predictive of later psychosis (Cornblatt et al., 2007).

In addition to above features used to conceptualize the prodromal subgroups, we considered the covariates of age, gender, parental education (coded as the number of parents who completed high school) and first-degree relatives' history of psychotic disorders and mood disorders.

The clinical outcomes were categorized into 4 groups: 1) Remission; 2) Symptomatic; 3) Prodromal Progression; 4) Psychotic. The detailed criteria can be found in the Appendix.

2. Latent class membership probability

Assume a population of N subjects that can be divided into a finite number G of latent homogeneous subgroups. The latent class membership for each subject i (i=1, …, N) is defined using a categorical latent variable $c_i$, which equals to g if subject i belongs to latent class g (g=1, …, G). An individual has a probability $\pi_{ig}$ of belonging to latent class g, which is modelled using multinomial logistic regression and can incorporate covariates $X_i$

$$\pi_{ig} = P(c_i = g|X_i) = \frac{e^{\zeta_{0g}+X_i^T\zeta_{1g}}}{\sum_{l=1}^{G} e^{\zeta_{0l}+X_i^T\zeta_{1l}}} \qquad (1)$$

where $\zeta_{0g}$ is the intercept for class g, and $\zeta_{1g}$ is the vector of class-specific parameters associated with the vector of time-independent covariates $X_i$.

Each latent class is characterized by a class-specific marker trajectory and a class-specific risk of the event, and the marker and the time-to-event are assumed to be

conditionally independent given these latent classes. This conditional independence is a central assumption of joint latent class model.

3. Class-specific marker trajectory

Given the latent class g, the vector of repeated measures of the longitudinal marker $Y_i = (Y_i(t_{i1}), \dots, Y_i(t_{ij}), \dots, Y_i(t_{in_i}))$ is described at the different times of measurements $t_{ij}$ (j=1,… $n_i$) by a standard linear mixed model:

$$Y_i(t_{ij})|_{c_i=g} = Z_i(t_{ij})^T u_{ig} + X_i(t_{ij})^T \beta_g + \epsilon_i(t_{ij}) \qquad (2)$$

where the p-vector of class-specific random-effects $u_{ig} = u_i|_{c_i=g} \sim N(\mu_g, B_g)$. The $n_i$-vector of measurement errors $\epsilon_i = \left(\epsilon_i(t_{i1}), \dots, \epsilon_i(t_{in_i})\right)^T \sim N(0, \Sigma_i)$. The variance-covariance matrix $B_g$ can be common over classes or class-specific. However, when considered as class-specific, it is usually assumed that $B_g = \omega_g^2 B$ with B unstructured and $\omega_G = 1$ to limit the number of parameters and identifiability concerns. The variance and covariance matrix $\Sigma_i$ is usually restricted to the diagonal matrix $\sigma^2 I_{n_i}$ for homoscedastic independent errors, but $\epsilon_i$ can also include a correlation process such as an auto-regressive process (Proust-Lima et al., 2014).

Bayes Rule can be used to predict each individual's latent class membership probabilities based on all the observed data:

$$\tau_{ig} = P(c_i = g|Y_i, X_i) = \prod_{ig} f(Y_i|c_i = g) \,/\, \prod_{h=1}^{G}\{\prod_{ih} f(Y_i|c_i = h)\} \qquad (3)$$

4. Missingness weight estimation

The retention of subjects over a long study period is difficult, particularly in a study of psychosis like NAPLS2. In NAPLS2, participants were followed in a 2-year

assessment, with complicated patterns of missing data. Based on a marginal structural model approach(Weuve et al., 2012), we implemented an artificially monotone pattern of missing data that excludes from analyses interviews after the first incomplete interview.

To account for potentially informative missingness in our analysis, we estimated weights to apply to each observation in the latent class analysis. For each wave of visits contributing to the analysis, the weights were based on the inverse of wave-specific probability of being observed at the wave, and thus of being uncensored at the wave. The intuition behind these weights is that respondents with characteristics similar to the observations missing due to attrition are up-weighted in the analysis of their prodromal symptoms and psychosis outcome, so as to represent their original contribution as well as their missing contributions.

Let $C_{ik}$ indicate whether person i is no longer in the study by wave k for loss to follow-up. Each weight represents the reciprocal of individual i's probability of remaining in the study at wave k. For each wave of follow-up, we modeled and estimated the probability of being observed in that wave, using pooled logistic regression, conditional on remaining observed in the previous wave. We defined as predictors in the missingness model a set of variables L, some of which varied over time, which we thought were likely to influence drop-out and affect psychosis outcomes: age, sex, education in years, SOPS positive/negative/disorganization/general subscales, AUSDUS (an alcohol and drug scale) and CDSS(Calgary Depression Scale for Schizophrenia) at previous visit. We fitted models that included following as predictors both the baseline time-covariates in L and the most recent prior values of time-varying covariates $L_{i(k-1)}$. These models were combined to calculate the cumulative probability of remaining on

study up to a given follow-up wave and of participating in the assessment at that wave. These weights can be obtained by the formula (Weuve et al., 2012):

$$wt_{ij} = \prod_{k=0}^{j} \frac{1}{\widehat{Pr}[C_{ik}=0|C_{i(k-1)}=0,\bar{L}_{i(k-1)}]} \quad (4)$$

5. Maximum Likelihood estimation

For a fixed number of latent classes G, the log-likelihood $L(\theta_G)$ of the observed data can be decomposed using the conditional independence assumption so that:

$$L(\theta_G) = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} \log \left( \sum_{g=1}^{G} \pi_{ig} f(Y_i|c_i = g; \theta_G) \right) \quad (5)$$

where $\theta_G$ is the entire vector of parameters for a joint latent class model with G classes; the class-membership probability $\pi_{ig}$ is defined in (1); and the density $f(Y_i|c_i = g; \theta_G)$ of the longitudinal marker in class g is defined in (2) but also includes the missingness weight $wt_{ij}$ defined in (4). The log-likelihood (4) is maximized using the EM algorithm (McLachlan & Peel, 2000).

6. Model selection criteria

We considered three model selection criteria: 1) the Akaike Information Criterion (AIC); 2) the Bayesian Information Criterion (BIC); and 3) the Integrated Classification Likelihood-BIC (ICL-BIC). The latent class model that minimized the value of AIC, BIC, or ICL-BIC was selected as the best-fitting model (McLachlan & Peel, 2000).

The Akaike Information Criterion (AIC) is a measure of the goodness of fit of a model that considers the number of model parameters q, where AIC = $2q - 2 \log L(\theta_G)$. Schwarz's Information Criterion (also called the Bayesian Information Criterion) is a

measure of the goodness of fit of a model that considers the number of parameters q and the number of observations N, where BIC = q log(N) − 2 log $L(\theta_G)$.

The integrated classification likelihood criteria is another model fit index. Its calculation is often estimated using a BIC-type approximation (ICL-BIC). The ICL-BIC is calculated as ICL-BIC = q log(N) − 2 log $L(\theta_G)$+ $2O_k$ , where $O_k$ is the entropy of the fuzzy classification matrix (McLachlan & Peel, 2000). The only difference between BIC and ICL-BIC is that the ICL-BIC takes entropy into concern as a component of the model selection criteria. The entropy of a model is defined to be a measure of classification uncertainty. The relative entropy is defined on [0, 1], with values near one indicating high certainty in classification and values near zero indicating low certainty.

We used ICL-BIC as our primary criterion for model selection, as it has been shown to more accurately estimate the number of latent classes, based on previous simulation studies (McLachlan & Peel, 2000).

7. Multinomial Logistic Regression

After the participants in the study were assigned to latent classes, we built the multinomial logistic regression model using their demographic information and latent class as predictors of the two-year clinical outcome (either remission, symptomatic, prodromal progression or psychotic). Multinomial logistic regression is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable.

Specifically, let Y be a categorical response with J categories, and let X be a vector of explanatory variables. It follows that Y is multinomial with probabilities

$\{\pi_1(X), \pi_2(X) \ \ldots, \pi_j(X)\}$, where $\sum_j \pi_j(X) = 1$. Multinomial logistic models compare each response category with a reference category:

$$\log\left(\frac{\pi_j(X)}{\pi_J(X)}\right) = \alpha_j + \beta_j^T X, for\ j = 1,2, \ldots . J-1 \qquad (6)$$

This model describes the effects of X on these J-1 logits. These J-1 logits also can be used to determine the logits comparing any with other pairs of response categories:

$$\log\left(\frac{\pi_a(X)}{\pi_b(X)}\right) = \log\left(\frac{\pi_a(X)}{\pi_J(X)}\right) - \log\left(\frac{\pi_b(X)}{\pi_J(X)}\right) = (\alpha_a + \beta_a^T X) - (\alpha_b + \beta_b^T X) \qquad (7)$$

8. Proportional Odds Model

As an alternative to multinomial logistic analysis considered in Section II.7, we could regard clinical outcome as an ordinal response in our study, utilizing its ordinal nature results in more parsimonious models and potentially more powerful inference. One way to use the ordinal nature of the response is to form logits of cumulative probabilities:

$$logit[\Pr(Y \le j|X)] = log\frac{\Pr(Y \le j|X)}{1-\Pr(Y \le j|X)} = log\frac{\Pr(Y \le j|X)}{\Pr(Y > j|X)} = log\frac{\sum_{i=1}^{j}\pi_i(X)}{\sum_{i=j+1}^{J}\pi_i(X)} \qquad (8)$$

As seen in (8), a model for $logit[\Pr(Y \le j)]$ for a single j is an ordinary logit model for a binary response, Y ≤ j vs Y> j. The binary response is formed by collapsing levels 1 through j of the response into a single category and collapsing levels j+1 through J into a second single category.

The proportional odds model simultaneously uses all cumulative logits:

$$logit[\Pr(Y \le \text{j}|\text{X})] = \alpha_j + \beta^T X, j = 1, \ldots, J-1. \qquad (9)$$

Each cumulative logit has its own intercept. The intercepts are increasing in j since $logit[\Pr(Y \le \text{j}|\text{X})]$ is increasing in j for a fixed value of X, and the logit is an increasing

function of this probability. This model assumes the same covariate effect β for each cumulative logit. Hence the shape of each response curve is the same, only shifted horizontally.

For a fixed j and at two different levels of the predictors, X1 and X2, we have

$$logit[\Pr(Y\leq \mathrm{j}|X_2)] - logit[\Pr(Y\leq \mathrm{j}|X_1)] = log\frac{\Pr(Y \leq j|X_2)\Big/\Pr(Y > j|X_2)}{\Pr(Y \leq j|X_1)\Big/\Pr(Y > j|X_1)}$$

$$= \beta^T(X_2 - X_1) \qquad (10)$$

This is an odds ratio of cumulative probabilities, also called a cumulative odds ratio. The name proportional odds model comes from the fact that the log cumulative odds ratio is proportional to the distance between X1 and X2.

**III. Results**

The 764 Clinical High-Risk participants' demographic characteristics in NAPLS2 are shown as Table1. Family history information was missing for 61 participants; we regarded these individuals as not having family history of psychosis or mood disorders.

| Variable | Mean (SD); Count (Frequency) |
|---|---|
| Age (years) | 18.45 (4.23) |
| Education (years) | 11.28 (2.82) |
| Sex | |
| Male | 436 (57.1%) |
| Female | 328 (42.9%) |
| Race* | |
| First nations | 13 (1.7%) |
| Asian | 54 (7.1%) |
| Black | 118 (15.5%) |
| Latin America/ Middle East/ White | 478 (62.6%) |
| Native Hawaiian or Pacific Islander | 3 (0.4%) |
| Interracial | 97 (12.7%) |
| Hispanic or Latino* | |
| Yes | 142 (18.6%) |
| No | 621 (81.3%) |
| Father's highest level of formal education** | |
| No or primary school | 35 (4.9%) |
| Some high school | 72 (10.0%) |
| High school and/or some college | 280 (38.9%) |
| College graduate | 333 (46.2%) |
| Mother's highest level of formal education*** | |
| No or primary school | 32 (4.3%) |
| Some high school | 59 (7.9%) |
| High school and/or some college | 283 (38.1%) |
| College graduate | 370 (49.7%) |
| History of psychotic illness in the first-degree relatives | |
| Yes | 129 (16.9%) |
| No | 635 (83.1%) |
| History of mood disorder in the first-degree relatives | |
| Yes | 297 (38.9%) |
| No | 467 (61.1%) |

Note: *: 1 missing value; **: 44 missing values; ***: 20 missing values.

TABLE 1. Demographic Characteristics of Participants in NAPLS2

In the first part of the analysis, we investigated subgroups of the CHR participants by using the latent trajectory class analysis. We fitted a series models with two or three latent classes based on 5 baseline covariates and 15 longitudinal features. We included weights to adjust for informative missing data. The EM algorithm did not converge after

200 iterations, so we chose the best two-class and three-class solutions after by using 50 random starting values and and considering up to 200 iterations. We found that the two-class model was preferred according to the following model selection criteria: AIC, BIC and ICL-BIC. All these three criteria were smaller in the two-class model compared with the three-class model. The results are summarized in Table 2.

| | Number of Parameters | -2LogL | AIC | BIC | ICL-BIC |
|---|---|---|---|---|---|
| Two-Class | 126 | 38490.061973 | 38742.062 | 39318.522 | 39355.687 |
| Three-Class | 192 | 42251.446126 | 42635.446 | 43513.861 | 43553.501 |

TABLE 2. Model Selection

The results indicated that the alcohol/drug usage, score of life events and especially total score of daily stress level all contributed to the classification of the subgroups. The first subgroup of CHR had a relative frequency of 45% and was characterized by relatively low levels of alcohol/drug abuse and life events, happened and much lower daily stress level. The second subgroup of CHR had a relative frequency of 55% and was characterized by relatively high levels of substance abuse and life events, and much higher daily stress level. The severity of daily stress level was more pronounced than the diversity of different daily stress events in this subgroups.

| | Low Daily Stress (45%) | High Daily Stress with higher substance usage and more life event (55%) |
|---|---|---|
| SOPS | | |
| SOPS Negative | 12.24, -0.96 | 11.54, -0.75 |
| SOPS Disorganization | 5.13, -0.42 | 5.16, -0.34 |
| SOPS General | 8.63, -0.67 | 9.47, -0.67 |
| AUSDUS | | |
| Total | 0.15, 0.03 | 3.41, 0.24 |
| Tobacco | 0.36, 0.00 | 0.77, 0.05 |

| | | |
|---|---|---|
| Alcohol | 0.15, 0.03 | 1.41, 0.07 |
| Marijuana | 0.00, 0.00 | 1.14, 0.05 |
| Life Events | | |
| Dependent | 13.50, -1.68 | 19.86, -2.51 |
| Independent | 2.95, -0.51 | 4.43, -0.73 |
| Desirable | 5.90, -0.59 | 8.07, -0.86 |
| Undesirable | 8.19, -1.21 | 13.21, -1.80 |
| Daily Stress | | |
| Total Score | 1.55, 2.69 | 85.63, -5.67 |
| Total number of events | 21.99, -1.31 | 35.36, -1.36 |
| GFS | 5.82, 0.16 | 6.43, 0.13 |
| GFR | 6.23, 0.03 | 6.04, 0.03 |

TABLE 3. Maximum Likelihood Estimates of Intercepts and Slopes from a Model with Two Latent Subclasses of CHR Participants

With regard to covariates, the results indicated an association between a family history of mood disorders and the empirically derived subgroups of CHR (Table 4). Participants in "High daily stress with higher substance usage and more life events" were 1.12 times as likely (95% CI: 1.05- 1.19) compared with the "Low daily stress" subgroup to have mood disorder in their first-degree relatives' history. The other covariates considered were not associated with the empirically defined subgroups of CHR patients.

| Risk Factors | Low Daily Stress (45%) | High Daily Stress with higher substance usage and more number of life event (55%) |
|---|---|---|
| Family History | | |
| Psychotic illness | 1.00 | 0.98 (0.63, 1.54) |
| Mood disorder | 1.00 | 1.12 (1.05, 1.19) |
| Sex | 1.00 | 0.75 (0.42, 1.32) |
| Age | 1.00 | 1.19 (0.78, 1.83) |
| Parents Education | 1.00 | 0.95 (0.72, 1.24) |

TABLE 4. Estimated Odds Ratios (95% Confidence Intervals) for the Associations Between Prodromal Risk Factors and CHR Subgroup classifications

The relative frequencies of two-year cohorts were: remission, 22.7%; symptomatic, 24.1%; prodromal progression, 21.1%; and psychotic, 13.7% (missing, 18.4%). We fitted multinomial logistic and proportional odds models to explore whether there was an association between two-year clinical outcome and empirically derived CHR subgroups. We also tried entering demographic information as covariates, but this information did not significantly improve the model fit.

Maximum likelihood estimates for the multinomial logistic regression indicated that, compared with the low daily stress group, the higher daily stress group participants were less likely to be in remission, but only the Prodromal Progression vs. remission outcome attained statistical significance (Table 5). “Prodromal Progression” participants were 1.68 times as likely (95% CI: 1.075-2.626) to be in the High daily stress group compared with “In Remission” participants.

Results for a proportional odds model were not significant, and can be found in the Appendix.

| Parameter | Two Year Clinical Outcome (In Remission as Reference) | Estimate | Standard Errors | p-value |
|---|---|---|---|---|
| Latent class (Low Daily Stress Group as Reference) | Symptomatic | 0.3351 | 0.2191 | 0.1261 |
| | Prodromal Progression | 0.5188 | 0.228 | 0.0229 |
| | Psychotic | 0.3072 | 0.2564 | 0.2308 |

TABLE 5. Analysis of Maximum Likelihood Estimates in Multinomial Logistic Regression Model

| Effect | Two Year Clinical Outcome (In Remission as Reference) | Point Estimate | 95% Confidence Interval |
|---|---|---|---|
| Latent class (Low Daily Stress Group as Reference) | Symptomatic | 1.398 | (0.910, 2.148) |
| | Prodromal Progression | 1.680 | (1.075, 2.626) |
| | Psychotic | 1.360 | (0.823, 2.247) |

TABLE 6. Estimated Odds Ratios and 95% Confidence Intervals for the Associations Between Empirically Based Subclassifications and Clinical Outcomes

## IV. Discussion

1. Discussion of the results

This study aimed to detect the structure of the prodrome of psychosis and construct a multinomial logistic regression model to characterize the relationship between prodromal subgroups and two-year clinical outcomes. In this study, the clinical high-risk(CHR) youths could be classified into two subgroups based on the longitudinal information, especially the severity of daily stress. Interestingly, participants with a family history of mood disorders had greater odds to be in the high daily stress group. There was a significant relation between the participants' latent classifications and their 2-year clinical outcomes compared with the low daily stress group, the high stress group had greater odds of being in prodromal progression rather than remission.

2. Future work

For studying how repeated marker data and the risk of events are linked, there are two kinds of methods that are usually used. The first one is the shared random-effect model (SREM), also called a selection model in missing data problems. A shared random-effect model (Henderson, Diggle et al. 2000) models the repeated quantitative outcome with a mixed model and includes the individual random coefficients as covariates in the model for the event. In contrast, a latent class model considers the population of subjects as heterogeneous, and assumes that it consists of homogeneous latent subgroups of subjects that share the same marker trajectory and the same risk of event (Proust-Lima et al., 2014). We prefer latent class analysis because the assumptions underlying SREM are not always met. SREM assumes the random-effects come from a common Gaussian distribution. Latent class analysis does not rely on such strict

assumptions (Proust-Lima, Letenneur et al. 2007).

In the process of fitting latent class analysis, the two-class and three-class models did not converge. This may be the result of our model assumptions. We treated all the longitudinal feature variables as multivariate Gaussian variables. But some of the variables, like alcohol usage and drug usage, had a small range. Treating them as binary or categorical variables may be a better way of dealing with this problem. Although the latent class approach was statistical, it also involved clinical judgment in choosing variables and covariates. Another potential problem that may cause the convergence problem was our use of the weights for missingness: we did not incorporate the missingness weights into the EM algorithm. More research is needed on the proper way to incorporate missingness weights in latent class trajectory analysis.

Another important problem raised during the analysis is missing data. Since the youth under study were at clinical high-risk of psychosis, they were inconsistent in showing up for medical and mental examinations. When constructing weights, we often needed to resort to the classic last observation carried forward(LOCF) method, which could be improved by using more advanced methods. When we were constructing the multinomial logistic and proportional odds model, the missingness of two-year clinical outcome may be more serious.

**V. Reference**

Addington, J., Liu, L., Buchy, L., Cadenhead, K. S., Cannon, T. D., Cornblatt, B. A., ... & Woods, S. W. (2015). North American prodrome longitudinal study (NAPLS 2): the prodromal symptoms. The Journal of nervous and mental disease, 203(5), 328.

Addington, J., & Heinssen, R. (2012). Prediction and prevention of psychosis in youth at clinical high risk. Annual review of clinical psychology, 8, 269-289.

Cornblatt, B. A., Auther, A. M., Niendam, T., Smith, C. W., Zinberg, J., Bearden, C. E., & Cannon, T. D. (2007). Preliminary findings for two new measures of social and role functioning in the prodromal phase of schizophrenia. Schizophrenia bulletin, 33(3), 688-702.

Reboussin, B. A., Miller, M. E., Lohman, K. K., & Ten Have, T. R. (2002). Latent class models for longitudinal studies of the elderly with data missing at random. Journal of the Royal Statistical Society: Series C (Applied Statistics), 51(1), 69-90.

Bebbington, P., Wilkins, S., Jones, P., Foerster, A., Murray, R., Toone, B., & Lewis, S. (1993). Life events and psychosis. Initial results from the Camberwell Collaborative Psychosis Study. The British Journal of Psychiatry, 162(1), 72-79.

Van Winkel, R., Stefanis, N. C., & Myin-Germeys, I. (2008). Psychosocial stress and psychosis. A review of the neurobiological mechanisms and the evidence for gene-stress interaction. Schizophrenia bulletin, 34(6), 1095-1105.

Cornblatt, B. A., Lencz, T., Smith, C. W., Olsen, R., Auther, A. M., Nakayama, E., ... & Kane, J. M. (2007). Can antidepressants be used to treat the schizophrenia prodrome? results of a prospective, naturalistic treatment study of adolescents. The Journal of clinical psychiatry, 68(4), 546-557.

Proust-Lima, C., Séne, M., Taylor, J. M., & Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. Statistical methods in medical research, 23(1), 74-90.

Weuve, J., Tchetgen, E. J. T., Glymour, M. M., Beck, T. L., Aggarwal, N. T., Wilson, R. S., ... & de Leon, C. F. M. (2012). Accounting for bias due to selective attrition: the example of smoking and cognitive decline. Epidemiology (Cambridge, Mass.), 23(1), 119.

Hanfelt, J. J., Wuu, J., Sollinger, A. B., Greenaway, M. C., Lah, J. J., Levey, A. I., & Goldstein, F. C. (2011). An exploration of subgroups of mild cognitive impairment based on cognitive, neuropsychiatric and functional features: Analysis of data from the National Alzheimer's Coordinating Center. The American Journal of Geriatric Psychiatry, 19(11), 940-950.

McLachlan GJ, Peel D: Finite Mixture Models. New York, Wiley, 2000

Agresti A. Categorical Data Analysis, Second Edition, Wiley, 2002.

Everitt, B. S. (1985). Mixture Distributions—I. John Wiley & Sons, Inc.

Henderson, R., et al. (2000). "Joint modeling of longitudinal measurements and event time data." Biostatistics(1): 465-480.

Proust-Lima, C., et al. (2007). "A nonlinear latent class model for joint analysis of multivariate longitudinal data and a binary outcome." Stat Med 26(10): 2229-2245.

McGlashan T, Walsh BC, Woods SW (2010) The Psychosis Risk Syndrome: Handbook for Diagnosis and Follow-up. New York: Oxford University Press, New York.

## VI. Appendix

1. Sample and Two-year clinical outcome criteria

The CHR sample met the Criteria of Prodromal Syndromes (COPS), which is based on the Structured Interview for Prodromal Syn- dromes (SIPS) (McGlashan et al., 2010). The COPS has three possible prodromal syndromes - attenuated positive symptom syndrome (APSS), genetic risk and deterioration (GRD) and/or brief intermittent psychotic syndrome (BIPS).

Two-year clinical outcome criteria: 1) Remission: remission from syndromes which means score of 2 or less on all five positive symptoms on the SOPS scale, or for those who have only GRD, in "remission" will require global assessment of functioning(GAF) to have returned to 90% of previous best GAF; 2) Symptomatic which means not currently meeting the criteria for a prodromal risk syndromes but having ratings of 3 to 5 on any one of five positive symptoms on SOPS, or with no change in the GAF; 3) Prodromal progression which is defined as meeting criteria for one of the at risk syndromes APSS, GRD, BIPS; and 4) Psychotic: meet criteria for a psychotic disorder or evidencing scores of 6 on one or more positive symptoms of SOPS.

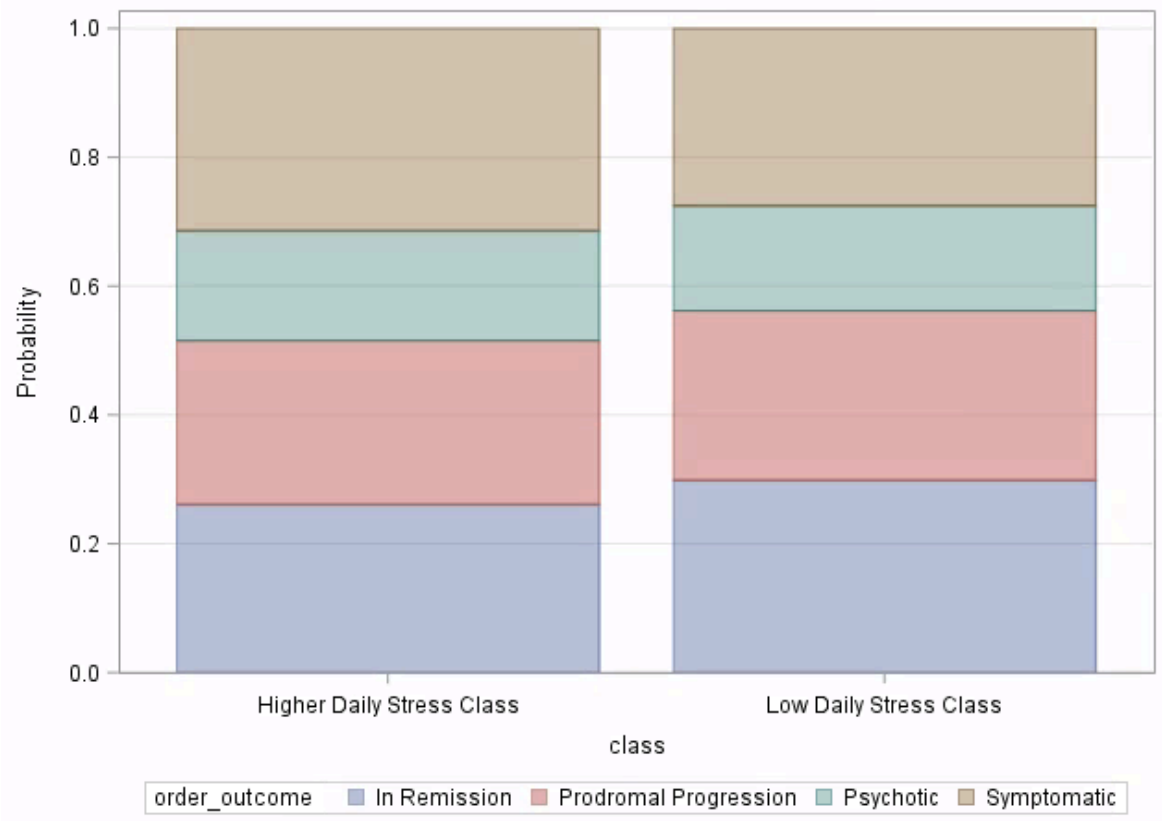
2. Proportional odds model results

| Parameter | Estimate | Standard Errors | p-value |
|---|---|---|---|
| Latent class (Low Daily Stress Group as Reference) | − 0.0932 | 0.0744 | 0.2105 |

APPENDIX TABLE 1. Analysis of Maximum Likelihood Estimates in Proportional Odds Model

| Effect | Point Estimate | 95% Confidence Interval |
|---|---|---|
| Latent class (Low Daily Stress Group as Reference) | 0.830 | (0.620, 1.111) |

APPENDIX TABLE 2. Estimated Odds Ratios and 95% Confidence Intervals for the Associations Between Empirically Based Subclassifications and Clinical Outcomes




APPENDIX FIGURE Predictive Cumulative Probabilities for Four Ordered Outcomes