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:

_____ _____

MENG SHI Date

The Structure of the Prodrome of Psychosis

By

Meng Shi

Master of Science in Public Health

Emory University

Rollins School of Public Health

Department of Biostatistics and Bioinformatics

_____ [Thesis Advisor's signature]

John J. Hanfelt, Ph.D

Thesis Advisor

_____ [Reader's signature]

Eugene Huang, Ph.D

Reader

The Structure of the Prodrome of Psychosis

By

Meng Shi

B.S.
Nankai University
2011

Thesis Committee Chair: John J. 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
2013

Abstract

The Structure of the Prodrome of Psychosis

By Meng Shi

Psychotic disorder is a group of serious illness that affects the mind. The symptoms are severe and it affects over 5% of the population. Among illness that affect people aged in 15 to 44, schizophrenia is the 8th leading cause of the disability worldwide. The first aim of this analysis is to conduct a latent class analysis on the clinical characteristics of adolescents at high risk of psychosis using the software Latent Gold. The second aim is to model the time to onset of psychosis in high-risk youth based on latent classes of comprehensive clinical information and socio-demographic variables. In this analysis, we used the Latent Class Analysis (LCA) approach to analyze variables collected as the North American Prodrome Longitudinal Study (NAPLS). The results showed that the four-class model was preferred according to the model selection criteria, such as AIC, BIC, and ICL-BIC. Based on these four subgroups, a proportional hazards model was used to characterize the relationship. In comparing the proportional hazards regression models with and without covariate measurement error, we found that the standard errors of the coefficients in the model with measurement error are smaller than the ones without measurement error.

The Structure of the Prodrome of Psychosis

By

Meng Shi

B.S.
Nankai University
2010

Thesis Committee Chair: John J. 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
2013

Acknowledgements

I want to thank the faculty, advisors, and staff of the Biostatistics Department at Rollins School of Public Health for the dynamic two years of learning that I have had. This thesis is only a sample of the vast knowledge that was attained and applied through my two years here at Rollins.

I am extremely grateful for my thesis advisor, Dr. John J. Hanfelt, whose encouragement, guidance, and support made this thesis possible. I am also thankful for Dr. Eugene Huang for taking time to read my thesis.

Finally, I would like to thank my parents for their loving support and encouragement that motivated me to pursue a degree at the graduate level.

Contents

1. Introduction

Psychotic disorders, such as schizophrenia and affective psychotic disorders, are a group of serious illnesses that affect the mind. They are associated with impairment in emotional, cognitive and social functioning, potentially leading to long-term disability. Two instruments were developed by the PRIME prodromal research team at Yale University to rate and track the phenomena cross-sectional and over time. These instruments are the Structured Interview for Prodromal Syndromes (SIPS) and the Scale of Prodromal Symptoms (SOPS). The SOPS consist of five positive symptom items, six negative symptom items, four disorganization symptom items, and four General symptom items. Each item has a severity scale rating from 0 (Never, Absent) to 6 (Severe/ Extreme – and Psychotic, for the positive items). The severity of the prodromal stage is judged according to the sum of the ratings from each of the SOPS items and ranges from 0 to 114. Thus, there are severity ratings from the overall scale, each domain of pathology, as well as individual items. The SIPS includes 29 major probes organized according to each positive symptom item in the SOPS. (Miller, McGlashan et al. 1999). In the SIPS, patients are also rated according to their Global Assessment of Functioning (GAF), a DSM IV Schizotypal Personality Disorder criterion checklist, and family history of mental illness. The SIPS is used to diagnose the prodromal syndromes and may be thought of as analogous to the Structured Clinical Interview for DSM-IV (SCID) or other structured diagnostic interviews (Miller, McGlashan et al. 2003). The SIPS includes the SOPS, the Schizotypal Personality Disorder Checklist, a family history questionnaire(Andreasen, Endicott et al. 1977), and a well-anchored version of the Global Assessment of Functioning scale (GAF) (Hall and Parks 1995).

Accurate assessment of the individual risk for psychotic disorders has great value. The risk for psychosis is always defined primarily on the basis of attenuated positive symptoms; studies showed that early impaired social and role functioning appeared to be risk factors for psychosis (Cornblatt, Carrion et al. 2012). Meanwhile, family history, gender and age are suggested to be the examples of baseline risk factors for psychotic disorders (Heckers 2009).

The symptoms of a psychotic disorder vary from person to person and may change over time. The major symptoms are hallucinations and delusions. When symptoms are severe, people with psychotic disorders have difficulty staying in touch with reality and often are unable to meet ordinary demands of daily life. Approximate 5% of the population is affected by psychotic disorders. They are some of the most distressing and costly diseases, and schizophrenia is the 8th leading cause of disability worldwide in 15 to 44 years of age (Schultze-Lutter, Frauke, Ruhrmann et al. 2008). In most cultures, these disorders are highly stigmatized which makes treatment and integration into the community difficult.

Psychosis nearly always emerges in late adolescence or early adulthood, with a peak between ages 18 and 25, when the prefrontal cortex is still developing. The risk factors for schizophrenia in adolescents include a schizotypal personality, sub-threshold psychotic symptoms, functional decline and a family history of schizophrenia (Yung, Phillips et al. 2004, Owens and Johnstone 2006). The identification and prevention of individuals prodromal for schizophrenia and other psychotic disorders are significant public health challenges.

The North American Prodrome Longitudinal Study (NAPLS) is a consortium of clinical research programs dedicated to the early detection and prevention of psychotic disorders and other forms of serious mental illness. It is a consortium of eight independent NIMH-funded prodromal studies located at Emory University, Harvard University, University of California (UCLA), University of California San Diego (UCSD), University of North Carolina Chapel Hill, University of Toronto, Yale University, and Zucker Hillside Hospital. The study combined previously collected prospective, longitudinal data into a common federated database. The NAPLS dataset constitutes the largest currently available longitudinal set of data on potentially prodromal patients and is currently being utilized to address a series of scientific questions about the nature of the currently defined prodromal syndromes (Cannon, Cadenhead et al. 2008). All the subjects in NAPLS database were evaluated using the SPIS; and they were assessed clinically every six months over a two years follow-up period, and tested yearly on laboratory procedures at baseline, 12 and 24 months. In addition, data were collected on demographic, academic/work, and diagnostic characteristics of all subjects (Addington, Cadenhead et al. 2007).

One goal of the current study was to incorporate information concerning scores of Structured Interview for Prodromal Syndromes (SIPS), Role Functioning Scale, and Social Functioning Scale to arrive at empirically defined subgroups of the prodrome of psychosis. Another goal was to assess whether sociodemographic factors were covariates affect the relative frequency of the latent classes. This was accomplished by using latent class analysis (LCA), a statistical method to investigate empirically the structure of heterogeneous syndromes. LCA is a likelihood-based approach to elucidate the underlying structure or subgroups of the study population based on the set of observed feature variables (Hanfelt, Wuu et al. 2011). Typically, these latent classes

cannot be observed directly, but they are also meaningful. It is a powerful statistical tool in detecting subgroups or subcategories.

In the current study, we used the LCA approach to analyze variables collected as the North American Prodrome Longitudinal Study (NAPLS). A total of 888 subjects enrolled in North American prodromal schizophrenia research projects between 1998 and 2005 were included in the baseline database. The specific variables we used in the LCA are the total SIPS positive score, the total SIPS negative score, the total SIPS disorganization score, the total SIPS general score, global social functioning scale score and global role functioning scale score. In our study, the total Positive Symptom score, the total Negative Score, total Disorganization Score, and the total General Score were obtained by adding up all the sub-symptoms scores in each category respectively.

A second purpose of this study was to construct survival analysis by using proportional hazards model to characterize the relationship between time to psychosis and the prodromal subgroups as well as demographic factors. In a standard survival analysis model, true values of the covariates are required to implement the partial likelihood inference procedure. However, in our study, some of the covariates might be measured with error. In the presence of covariate measurement error, special methods are acquired to fit the proportional hazards model. In this study, we used the corrected score approach which can be further classified as parametric correction proposed by Nakamura (Nakamura 1992). Nakamura's approach applies the corrected score function method to the proportional hazards model when measurement errors are additive.

The covariates in this study to be included in a proportional hazards model are a combination of covariates with and without measurement error. The prodromal subgroup covariates, which were obtained by latent class analysis, are measured with error, whereas the demographic covariates, such as gender and race, are measured without error. For the prodromal subgroup covariates, since the measurement made with a measuring device is approximate, so if same object was measured two different times, the two measurements may not be exactly the same. The difference between two measurements is called an "error" in the measurements. It represents the uncertainty in measurement. The error of measurement is a mathematical way to show the uncertainty in the measurement; it is the different between the result of the measurement and the true value.

2. Method

2.1 Coding Psychosis

To characterize the diversity of prodromal subgroups, we selected some relevant variables from the NAPLS baseline dataset. The Structured Interview for Prodromal Syndromes (SIPS), which was developed by the research team at Yale University, is a structured diagnostic interview used to diagnose the prodromal syndromes that was developed by Miller et al. and McGlashan et al (2001). The validity of the SIPS in the diagnosis of prodromal syndrome for psychosis has been confirmed by several reports (Miller and Cicchetti 2004). We assessed the SIPS positive, negative, disorganized and general symptoms for all the participants. In each symptom category, we used the sum of all the sub-symptoms scores to summarize the participants' status. Larger score indicates more severe impairment.

We also included the global social functioning scale and the global role functioning scale. The influence of social functioning and role functioning in individuals at clinical high risk for psychosis was examined in some studies (Carrion, Goldberg et al. 2011). Based on the original data, the larger social functioning score and role functioning score, the less severe condition. In order to be consistent, we change the sign of the scores to present the participants' social and role functioning status.

In addition to these variables, we considered the covariates of age, gender, race and parental education in our polytomous logistic regression model. In building the model, we treated the covariate race as a categorical variable with two categories, black and others. Meanwhile, we divided the years of parental education into 2 groups by the mean value.

Psychiatric diagnoses are categorized by the Diagnostic and Statistical Manual of Mental Disorder, 4th Edition. It is known as the DSM-IV, it includes all currently recognized mental health disorders. The code of DSM-IV are used to describe the features of a given mental disorders and indicate how the disorder can be distinguished from others. The DSM-IV code for psychosis for this study see Appendix 1.

2.2 Latent Class Analysis

In this analysis, we used the Latent Gold 4.0 software package (Statistical Innovations, Inc., Belmont, MA) to conduct the Latent Class Analysis. Latent Gold 4.0 Basic implements the most important types of latent class and finite mixture models in three modules called Cluster, DFactor, and Regression. The differences between these three modules arise from the fact that the application types of latent class analysis differ with respect to the required data organization and nature of the latent variables (Vermunt and Magidson 2005). This analysis constructed a Cluster Module. The Cluster Module can be used to estimate standard Latent Class models for categorical indicators, as well as mixture-based clustering models for continuous and mixed indicators. Here is a brief introduction of the basic concepts of Latent Class Cluster Module.

2.2.1 Basic components of a Latent Class Cluster model

We let $y_{1t}, y_{2t}, \dots, y_{nt}$ denote a random sample of size n, T response variables or indicators, $1 \leq t \leq T$. In the Latent Class Module, the exogenous variables that vary between cases and that may be used to predict class membership are called covariates, and are denoted as z_{ir}^{cov}, $1 \leq r \leq R$, where R is the number of covariates. There are direct relationships between

indicators and /or direct effects of covariates on indicators. It is assumed that there is a single nominal latent variable x with G categories, $1 \leq x \leq G$. The categories of this nominal latent variable are called Clusters or Classes.

We used the bold face for vectors, that is, the symbols $\boldsymbol{Y}_i \; and \; \boldsymbol{Z}_i^{cov}$ refer to the entire set of responses and covariate values of case i. Also, symbol $\boldsymbol{Y}_{ih}$ denoted one of the H subsets of y_{it} variables, and T_h^* denoted the number of variables in subset h.

2.2.2 Probability Structure

The Latent Gold Cluster Module is based on the mixture model probability structure that defines the relationships between the covariates, latent, and response variables:

$$f(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i^{cov}) = \sum_{x=1}^{G} P(x|\boldsymbol{Z}_i^{cov})\, f(\boldsymbol{Y}_i|x, \boldsymbol{Z}_i^{cov}) = \sum_{x=1}^{G} P(x|\boldsymbol{Z}_i^{cov}) \prod_{t=1}^{T} f(y_{it}|x) \quad (1)$$

In this equation, the covariates affect the latent variable but have no direct effects on the indicators, and indicators are assumed to be mutually independent given cluster membership.

The most general probability structure used is the one that allows the inclusion of direct effects of covariates on indicators and association / correlation between indicators within clusters. In this probability structure, the T indicators have to be grouped into H sets, where the indicators belonging to the same set may be correlated within clusters. The most general probability structure is

$$f(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i^{cov}) = \sum_{x=1}^{G} P(x|\boldsymbol{Z}_i^{cov})\, f(\boldsymbol{Y}_i|x, \boldsymbol{Z}_i^{cov}) = \sum_{x=1}^{G} P(x|\boldsymbol{Z}_i^{cov}) \prod_{h=1}^{H} f(\boldsymbol{Y}_{ih}|x, \boldsymbol{Z}_i^{cov}). \quad (2)$$

As can be seen, $f\,(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i^{cov})$ is the probability density corresponding to a particular set of $\boldsymbol{Y}_i$ values given a particular set of $\boldsymbol{Z}_i^{cov}$ values. The middle part of equation (2) shows the $P(x|\boldsymbol{Z}_i^{cov})$ is the mixing weights, which can be denoted by $\pi_{x|\boldsymbol{Z}_i^{cov}}, 1 \leq x \leq G$, that is the probability of belonging to certain latent class given an individual's realized covariates values. And the $f\,(\boldsymbol{Y}_i|x, \boldsymbol{Z}_i^{cov})$ is the probability density of $\boldsymbol{Y}_i$ given x and $\boldsymbol{Z}_i^{cov}$ which denotes the mixture densities. Thus, the latent class variable x can be influenced by the covariates of z variables, and the latent class variable x and the covariate variables z may influence the response variable y.

Noticed in the last part of equation (2) above, it implies that the unobserved variable x intervenes between the $\boldsymbol{Z}_i^{cov}$ and the $\boldsymbol{Y}_i$ variables:

$$f(\boldsymbol{Y}_i|x, \boldsymbol{Z}_i^{cov}) = \prod_{h=1}^{H} f(\boldsymbol{Y}_{ih}|x, \boldsymbol{Z}_i^{cov}).$$

It is assumed that the y variables are mutually independent given the latent and covariates variables. Moreover, the y's belonging to the same set h may be correlated within clusters.

2.2.3 Conditional distributions

A particular distributional form is assumed for $\boldsymbol{Y}_{ih}$ based on the scale types of the variables in a set. A set may consist of one or more categorical variables, one or more continuous variables, or a single count variable. When the variables are categorical, a multinomial distribution is assumed for $\boldsymbol{Y}_{ih}$. For continuous variables, normal distribution is often used. Counts can be modeled via Poisson or binomial. Here we only talk about the linear predictors and corresponding regression models for categorical and continuous response variables.

For continuous and count indicators, we get the linear predictor

$$\eta^t_{x,\boldsymbol{Z}_i^{cov}} = \beta^t_0 + \beta^t_{x0} + \sum_{r=1}^{R} \beta^t_r \cdot z_i^{cov},$$

where β^t_0 is the intercept, β^t_{x0} the effect of the Clusters on y_{it}, and β^t_r the direct effect of covariate r on the indicator concerned. β^t_{x0} have to be imposed by effect or dummy coding constraints.

2.2.4 Latent variable

In equation (2), the mixing weights $P(x|\boldsymbol{Z}_i^{cov})$ which can be denoted as $\pi_{x|\boldsymbol{Z}_i^{cov}}$ are nonnegative quantities that sum to one, that is,

$$0 \leq \pi_{x|\boldsymbol{Z}_i^{cov}} \leq 1 \ \ (x = 1, \dots, G)$$

and

$$\sum_{x=1}^{G} \pi_{x|\boldsymbol{Z}_i^{cov}} = 1.$$

The values of the latent variables given the covariates values are assumed to come from a joint multinomial distribution. The multinomial probability $P(x|\boldsymbol{Z}_i^{cov})$ is parameterized as follows:

$$P(x|\boldsymbol{Z}_i^{cov}) = \pi_{x|\boldsymbol{Z}_i^{cov}} = \frac{\exp\left(\eta_{x|\boldsymbol{Z}_i^{cov}}\right)}{\sum_{x'=1}^{G} \exp\left(\eta_{x'|\boldsymbol{Z}_i^{cov}}\right)} \quad .$$

2.2.5 Local independence

The local independence assumption is the basic assumption of the Latent Class model. Lack of fit of a Latent Class model is caused by violation of this assumption. Increasing the number of classes can help to get an acceptable fit model. An alternative way is to relax the local independence assumption by allowing for associations between indicators as well as direct effects of covariates on the indicators (Hagenaars 1988).

Latent Gold calculates bivariate z-y and y-y residuals, which can be used to detect which pairs of observed variables are more strongly related than can be explained by the formulated model. Latent Gold starts by setting up a probability structure corresponding to a local independence model. When users include local dependencies using information on bivariate residuals, the program automatically sets up the correct and most parsimonious probability structure for the situation concerned.

The most general Latent Class Cluster model is the model for mixed mode data. This model is used when one has y variables of different scale types. The structure that serves as the starting

point is again the local independence structure that we also used for categorical and continuous variables (see equation (1)). For each indicator, the user has to specify whether it is nominal, ordinal, continuous, or a count. It is possible to include covariates in LC Cluster models for mixed mode data. These covariates can also have direct effects on the various types of indicators.

2.3 Proportional hazards model

The Cox proportional hazards model, which was introduced by Cox in 1972, is a broadly applicable and the most widely used method of the survival analysis.

For the data with sample size n, for $j = 1, \dots, n$ let T_j represents the time on study for the jth patient, Δ_i is the event indicator for the jth patient, which $\Delta_j = 1$ means the event has occurred. $\boldsymbol{Z}_j(t) = \boldsymbol{Z}_j = (Z_{j1}, \dots, Z_{jp})^t$ is the vector of covariates for the jth individual at time t.

Let $h(t|\boldsymbol{Z})$ be the hazard rate at time t for an individual with risk vector **Z.** The basic model due to Cox (1972) is:

$$h(t|\boldsymbol{Z}) = h_0(t)\exp(\boldsymbol{\beta}^t\boldsymbol{Z}) = h_0(t)\exp\left(\sum_{k=1}^{p}\beta_k z_k\right)$$

where $h_0(t)$ is an unspecified baseline hazard at t, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^t$ is a parameter vector. This is a semiparametric model because a parametric form is assumed only for the covariate effect, and the baseline hazard rate is nonparametric.

The Cox model is often called a proportional hazards model because, if we look at two individuals with covariate values $\mathbf{Z}$ and $\boldsymbol{Z}^*$, the ratio of the hazard rate is:

$$\frac{h(t|\boldsymbol{Z})}{h(t|\boldsymbol{Z}^*)} = \frac{h_0(t)\exp\left(\sum_{k=1}^{p}\beta_k Z_k\right)}{h_0(t)\exp\left(\sum_{k=1}^{p}\beta_k Z_k^*\right)} = \exp\left[\sum_{k=1}^{p}\beta_k(Z_k - Z_k^*)\right]$$

which is a constant. Thus, the hazard rates are proportional.

We assume that censoring is noninformative such that, given $\boldsymbol{Z}_j$, the event and censoring time for the jth patient are independent. Suppose there are no ties between the event times. Let $t_1 < t_2 < \cdots < t_D$ denote the ordered event times and $Z_{(i)k}$ be the kth covariate associated with the individual whose failure time is t_i. Define the risk set at time t_i, and the set of all individuals who are still under study at a time just prior to t_i is R(t_i). The probability that an individual dies at time t_i with covariates $\boldsymbol{Z}_{(i)}$, given one of the individuals in R(t_i) dies at this time, is given by

$$P[individual\ dies\ at\ time\ t_i | one\ death\ at\ t_i\ ]$$

$$= \frac{P[individual\ dies\ at\ t_i | survival\ to\ t_i]}{P[one\ death\ at\ t_i | survival\ to\ t_i]} = \frac{h[t_i|\boldsymbol{Z}_{(i)}]}{\sum_{j\in R(t_i)} h[t_i|\boldsymbol{Z}_j]}$$

$$= \frac{h_0(t_i)\exp\left(\boldsymbol{\beta}^t\boldsymbol{Z}_{(i)}\right)}{\sum_{j\in R(t_i)} h_0(t_i)\exp\left(\boldsymbol{\beta}^t\boldsymbol{Z}_j\right)} = \frac{\exp\left(\boldsymbol{\beta}^t\boldsymbol{Z}_{(i)}\right)}{\sum_{j\in R(t_i)}\exp\left(\boldsymbol{\beta}^t\boldsymbol{Z}_j\right)}.$$

Multiplying these conditional probabilities over all deaths, then the partial likelihood function is:

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp\left(\boldsymbol{\beta}^t \boldsymbol{Z}_{(i)}\right)}{\sum_{j \in R(t_i)} \exp\left(\boldsymbol{\beta}^t \boldsymbol{Z}_j\right)} = \prod_{i=1}^{D} \frac{\exp\left(\sum_{k=1}^{p} \beta_k Z_{(i)k}\right)}{\sum_{j \in R(t_i)} \exp\left(\sum_{k=1}^{p} \beta_k Z_{jk}\right)}$$

Note that the numerator of the likelihood depends only on information from the individual who experiences the event, meanwhile the denominator is about the information of all individual who have not yet experienced the event.

Let $l(\boldsymbol{\beta}) = \ln\left(L(\boldsymbol{\beta})\right)$, then

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{D} \sum_{k=1}^{p} \beta_k Z_{(i)k} - \sum_{i=1}^{D} \ln\left[\sum\nolimits_{j \in R(t_i)} \exp\left(\sum_{k=1}^{p} \beta_k Z_{jk}\right)\right].$$

By taking parial derivatives of above equation with respect to the β's, let $U_m(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_m}, m = 1, \ldots, p$. Then

$$U_m(\boldsymbol{\beta}) = \sum_{i=1}^{D} Z_{(i)m} - \sum_{i=1}^{D} \frac{\sum_{j \in R(t_i)} Z_{jm} \exp\left(\sum_{k=1}^{p} \beta_k Z_{jk}\right)}{\sum_{j \in R(t_i)} \exp\left(\sum_{k=1}^{p} \beta_k Z_{jk}\right)}.$$

The partial maximum likelihood estimates are found by solving the set of p nonlinear equations $U_m(\boldsymbol{\beta}) = 0, m = 1, \ldots, p$.

2.4 Measurement with error

If an observed value, say w, is used as the true value z in a regression model, it possibly underestimates the effect of z. Particularly, in this analysis, the observed value w_i is the estimated probabilities for the ith participant in the subgroups. The true value x_i stand for the subgroup that the ith participant belongs to. For example, if there are 4 subgroups, the value of w_i could be $(0.2, 0.1, 0.1, 0.6)$, an the value of x_i is $(0, 0, 0, 1)$.

Suppose x_i denotes p time-independent covariates with observed value $w_i = x_i + \varepsilon_i$, where ε_i is the error with the known variance matrix Σ. Meanwhile, we assume that $(T_i, C_i, X_i, \varepsilon_i)$ are IID across i, and ε_i is independent of (T_i, C_i, X_i). If the covariates are not error contaminated, the corresponding error variance is 0.

Let **W** denotes the set of w's. Then

$$U(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{Y}) = \sum \{\boldsymbol{w}_{(i)} - \hat{E}_{\boldsymbol{i}}(\boldsymbol{W}|\boldsymbol{\beta})\}$$

is called a native score function if w's are used in equation (3) instead of z's. If $\boldsymbol{\beta}_w$ satisfies $U(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{Y}) = 0$, then $\boldsymbol{\beta}_w$ is even asymptotically biased (Prentice 1982). A correction of this bias is proposed by using a function $U^*(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{Y})$ whose expectation $E^*\{U^*(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{Y})\}$ with respect to the ε's given **Y** and **Z** coincides with $U(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{Y})$. This U^* is called a corrected score function, and β^* such that

$$\sum_{i=1}^{n} \int_0^L \{W_i + \Sigma\beta - \frac{\sum_{j=1}^{n} Y_{j(u)} W_j \exp(\beta^T W_j)}{\sum_{j=1}^{n} Y_{j(u)} \exp(\beta^T W_j)}\} \times dN_i(u) = 0$$

where $N_i(u) = I(V_i \leq u, \Delta_i = 1)$ is the counting process and $Y_i(u) = I(V_i \geq u)$ is the at risk process(Nakamura 1992).

3. Results

The NAPLS participants' demographic and clinical characteristics are given in Table 1.

In the first part of the analysis, we fitted a series of model with one to four latent classes

TABLE 1. Demographic and Clinical Characteristics of 888 High-Risk Teenagers From NAPLS

	Mean ± SD or Frequency (%)
Demographic	
Age, years	17.8 ± 4.4
Gender: male	516 (58)
Race:	
Black	118 (13)
Others	757 (85)
Parental Education:	
Low	348 (39)
High	424 (48)
Clinical Characteristics	
Sum of SIPS Positive Score	8.4 ± 5.9
Sum of SIPS Negative Score	9.8 ± 7.4
Sum of SIPS Disorganization Score	4.8 ± 4.0
Sum of SIPS General Score	6.1 ± 4.8
Global Social Functioning Scale	-6.5 ± 1.6
Global Role Functioning Scale	-6.6 ± 1.8

Notes: Number of subjects for whom data were unavailable: Race, N=13; Parental Education, N=116; SIPS positive, N=117; SIPS negative, N=138; SIPS disorganization, N=131; SIPS general, N=133; Social functioning, N=62; Role functioning, N=63.

and 10 response variables. We found that the four-class model was preferred according to the following model selection criteria: AIC, BIC and ICL-BIC. The four-class model has the minimum values of all these three criteria (Table 2).

The results indicated that the SIPS scales, the role functioning and social functioning all contributed to the classification of the subgroups, the p-value of the Wald tests yielded $p<0.001$

TABLE 2 Model Summary

		Npar	LL	AIC	BIC	ICL-BIC
Model 1	1-Cluster	12	-10772.4145	21568.8290	21624.4761	21624.4671
Model 2	2-Cluster	28	-9339.2557	18734.5115	18864.3547	18951.41373
Model 3	3-Cluster	43	-8808.9975	17703.9950	17903.3971	18029.50742
Model 4	4-Cluster	61	-8421.2972	16964.5944	17247.4671	17533.34665

*The 4-Cluster model has the smallest AIC, BIC and ICL-BIC.

for each variable. The high-risk teenagers can be classified into following 4 subgroups (Table 3): 1) the first subgroup has the relative frequency of 46% that is the largest relative frequency among these four subgroups. This subgroup is the most severely impaired, since it contains the largest mean values of the sum of SIPS positive score, the sum of SIPS disorganization score, the sum of SIPS general score and the role functioning score. 2) The second subgroup has the relative frequency of 23%, the mean values of the response variables in this category are all much smaller than the overall average values of the response variables. 3) The third subgroup has the relative frequency of 18%. The mean values of the SIPS negative score, the SIPS

TABLE 3 Maximum Likelihood Estimates of Means From a Model with Four Latent Subgroups

	Highly Impaired	Medium Impaired	Negative Symptoms and Impaired Social Function	Mildly Impaired
Frequency	46%	23%	18%	12%
BLSumPos Mean	10.6	5.2	9.2	0.4
BLSumNeg Mean	13.1	3.6	13.9	0
BLSumDis Mean	6.4	1.5	6.4	0.1
BLSumGen Mean	8.5	2.3	7.6	0.1
BLSocNow Mean	-6.2	-7.4	-5	-8.9
BLRolNow Mean	-5.8	-7.5	-6.1	-9

* Larger Social Functioning Score and Role Functioning Score mean more severe condition.

* The bootstrap p-value for order restricted 4-cluster model and non-order restricted 4-cluster model is 0.002.

disorganization score and the social functioning score in this subgroup are all the largest ones among 4 subgroups. 4) The fourth subgroup has the relative frequency of 12%. In this subgroup, the mean values of all the variables are the smallest among the 4 subgroups; meanwhile, they are much smaller than the overall average values. The fit of this four classes model was significantly better than the fit under an order-restricted four classes model that posited a unidimensional construct of severity (bootstrap p-value<0.05).

The results showed the association between the covariates and the empirically derived subgroups (Table 4). Compared with participants in the fourth subgroup, participants in the other three subgroups were more likely to be younger and male.

In the original 888 patients, based on the data records, there are only 198 teenagers completed the follow-up interview and had no missing values. We fitted the proportional hazards model for all the cluster of psychosis, with adjustment for covariate measurement error. The calculations of the covariance matrix of the latent subgroups are given in the Appendix 2. Table 5 shows the comparison of the estimate of the coefficients between the naive model and model with measurement error. The only covariate that is real significant is Parental Education. The SE of the model with measurement error is smaller than that of the native model. Since the measurement error is very small among different subgroups, the differences in results between these two models are tiny.

Additionally, this model may be unreliable if the dataset contain few events, which may be the case if either the disease or the event of interest is rare. Compared to the sample size (N=198),

the number of event (n=8) is small in this case. That is main reason for the meaningless estimation of the naïve model, especially for the estimation of Cluster 3. Ambler et al. did some simulation studies suggesting that the Cox models fitted using maximum likelihood can perform poorly when there are few events, and that significant improvement are possible by taking a penalized modeling approach. Meanwhile, he suggested that the ridge method generally performs the best, although lasso is recommended if variables selection is required under this situation (Ambler, Seaman et al. 2012). These methods can be implemented in further analysis.

TABLE 4 Estimated Odds Ratios and 95% Confidence Intervals for the Associations Between Demographic Characteristics and Empirically Based Subclassifications

		Highly Impaired	Medium Impaired	Negative Symptoms and Impaired Social Function	Mildly Impaired
Baseline Age		0.89 (0.84, 0.93)	0.82 (0.76,0.89)	0.86 (0.80, 0.92)	1
Gender	Female	0.46 (0.28, 0.77)	0.77 (0.43, 1.39)	0.33 (0.18, 0.61)	1
	Male	1	1	1	1
Race	Not Black	0.79 (0.31, 2.01)	0.38 (0.15, 1.01)	0.67 (0.24, 1.84)	1
	Black	1	1	1	1
ParentEd	Low	1.45 (0.86, 2.43)	0.86 (0.47, 1.56)	1.07 (0.59, 1.93)	1
	High	1	1	1	1

TABLE 5 Estimated Log Hazard Ratios for the Time of Onset of Psychosis with and without Measurement Error

	β			
	Naive Analysis		Analysis with Adjustment for Measurement Error	
	Estimate	SE	Estimate	SE
Race	0.463	1.080	0.506	0.796
ParentEd	-1.327	0.821	-1.261	0.734
Gender	0.461	0.837	0.490	0.768
Cluster1	-0.549	0.876	-0.115	0.843
Cluster2	-1.297	1.481	-0.812	1.882
Cluster3	-97.506	7184	-0.347	0.985
Cluster4	-0.251	1.300	1.048	1.481

4. Discussion

4.1 Discussion of Results

This study aimed to detect the structure of the prodrome of psychosis and construct a proportional hazards model to characterize the relationship between time to psychosis and prodromal subgroups as well as demographic factors. In this study, the high-risk teenagers could be classified into 4 subgroups based on the prodrome of psychosis. Compared to the subgroup that contained mildly impaired teenagers, the younger and male teenagers tended to have severe psychosis prodrome. No statistically significant results could be concluded based on the proportional hazards model for the subgroups of psychosis with and without adjustment for covariate measurement error. The only covariate that approached statistical significance was Parental Education, but conclusions were limited owing to the small number of adolescents who were diagnosed with psychosis during the course of the study.

4.2 Future work

For studying the association between evolution of quantitative outcomes and a clinical event, two kinds of joint models are usually proposed: shared random-effect models and latent class analysis. 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 (Lin, Turnbull et al. 2002) assumes that the population is made of various subpopulations with different longitudinal evolutions modeled by a latent class variable. We chose the way of latent class model in this analysis since it had some advantages over shared random-effect models. Firstly, the assumptions of shared random-effect models that the random-effects come from a common

Gaussian distribution is quite unrealistic when the population consists of several subgroups. Moreover, latent class models are simpler to interpret compared with the shared random-effect models that estimate correlations between the event and the random-effects. In latent class analysis, the impacts of covariates on the probability of each profile are evaluated and the probability of the event in each latent class is estimated (Proust-Lima, Letenneur et al. 2007).

Latent class models for joint analysis of a longitudinal outcome and an event have already been developed. Proust-Lima et al. (2006) proposed a nonlinear model with a latent process to analyze multivariate and non-Gaussian longitudinal data using flexible parameterized nonlinear transformations to model the relationship between the longitudinal outcomes and the latent process. However, in Proust-Lima et al.'s analysis, there were two basic assumptions in their analysis. Firstly, they assumed that the clinical outcome is binary. And, secondly, the conditionally independent between the clinical outcome and the manifest data used to elucidate the latent classes was assumed. Moreover, their approach required special software to fit the unknown parameters in the joint likelihood model (Proust-Lima, Letenneur et al. 2007). In contrast, the method in this analysis extended the first assumption by accommodating the time to the clinical outcome, and did not require the second assumption. Additionally, our method was implemented using existing, separate software programs for latent class analysis and proportional-hazards regression analysis with errors-in-covariates.

In this analysis, we used the standard parametric correction estimating equation to calculate the estimators. But this estimating equation is biased. In the presence of covariate measurement error with the proportional hazards model, several other functional modeling methods have been

proposed, such as the conditional score estimating equation (Tsiatis and Davidian 2001), the second-order parametric correction equation (Nakamura 1992), and the nonparametric correction estimator (Huang and Wang 2000). It is showed that the conditional score estimating equation is unbiased, which suggesting that the conditional score estimator might perform better in the case of finite samples compared to the standard parametric correction estimating equation (Song and Huang 2005). Nakamura (1992) also suggested that the second-order parametric correction equation performs better than the first-order correction under some specific situation.

In addition, the measurement error model in this analysis is limited. Since we assumed that $w_i = x_i + \varepsilon_i$, where ε_i is independent of (T_i, C_i, X_i). Then the range of w_i should be strictly larger than the range of x_i. But based on the assumptions in our analysis, the range of w_i is the same with the range of x_i. A better measurement error model needs to be explored in the future.

The model as presented here includes only clinical information and social demographic variables. In the future study, it can be extended by adding neurophysiological and neurocognitive biomarkers information. Furthermore, we can consider conducting a joint model of the latent class subgroups and the covariates instead of using the combine model of latent class analysis and survival analysis.

5. Reference

Addington, J., et al. (2007). "North American Prodrome Longitudinal Study: a collaborative multisite approach to prodromal schizophrenia research." Schizophr Bull **33**(3): 665-672.

Ambler, G., et al. (2012). "An evaluation of penalised survival methods for developing prognostic models with rare events." Stat Med **31**(11-12): 1150-1161.

Andreasen, N. C., et al. (1977). "The family history method using diagnostic criteria. Reliability and validity." Arch Gen Psychiatry **34**(10): 1229-1235.

Cannon, T. D., et al. (2008). "Prediction of psychosis in youth at high clinical risk: a multisite longitudinal study in North America." Arch Gen Psychiatry **65**(1): 28-37.

Carrion, R. E., et al. (2011). "Impact of neurocognition on social and role functioning in individuals at clinical high risk for psychosis." Am J Psychiatry **168**(8): 806-813.

Cornblatt, B. A., et al. (2012). "Risk factors for psychosis: impaired social and role functioning." Schizophr Bull **38**(6): 1247-1257.

Cox, D. R. (1972). "Regression models and life tables (with discussion)." Journal of the royal Statistical society **Series B 34**: 187-220.

Hagenaars, J. A. (1988). "Latent structure models with direct effects between indicators: local dependence models." Sociological Methods and Research **16**: 379-405.

Hall, R. C. and J. Parks (1995). "The modified global assessment of functioning scale: addendum." Psychosomatics **36**(4): 416-417.

Hanfelt, J. J., et al. (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." Am J Geriatr Psychiatry **19**(11): 940-950.

Heckers, S. (2009). "Who Is at Risk for a Psychotic Disorder?" Schizophr Bull **35**(5): 847–850.

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

Huang, Y. and C. Y. Wang (2000). "Cox regression with accurate covariates unascertainable: A nonparametric correction approach." Journal of the American Statistical Associatoin **95**: 1209-1219.

Insel, T. R. (2010). "Rethinking schizophrenia." Nature **468**(7321): 187-193.

Lin, H., et al. (2002). "Latent class models for joint analysis of longitudinal biomarder and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer." Journal of the American Statistical Associatoin **97**: 53-65.

Miller, T. J., et al. (2003). "Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: predictive validity, interrater reliability, and training to reliability." Schizophr Bull **29**(4): 703-715.

Miller, T. J., et al. (1999). "Symptom assessment in schizophrenic prodromal states." Psychiatr Q **70**(4): 273-287.

Nakamura, T. (1992). "Proportional hazards model with covariates subject to measurement error." Biometrics **48**(3): 829-838.

Owens, D. G. and E. C. Johnstone (2006). "Precursors and prodromata of schizophrenia: findings from the Edinburgh High Risk Study and their literature context." Psychol Med **36**(11): 1501-1514.

Prentice, R. L. (1982). "Covariate measurement errors and parameter estimation in a failure time regression model." Biometrika **69**: 331-342.

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.

Schultze-Lutter , Frauke, et al. (2008). "EARLY DETECTION AND EARLY INTERVENTION IN PSYCHOSIS IN WESTERN EUROPE." Clinical Neuropsychiatry **5**(6): 303-315.

Tsiatis, A. A. and M. Davidian (2001). "A semiparametric estimator for the proportionalhazards model with longitudinal covariates measured with error." Biometrika **88**: 447-458.

Vermunt, J. K. and J. Magidson (2005). Technical Guide for Latent GOLD 4.0: Basic and Advanced, Belmont Massachusetts: Statistical Innovations Inc.

Yung, A. R., et al. (2004). "Risk factors for psychosis in an ultra high-risk group: psychopathology and clinical features." Schizophr Res **67**(2-3): 131-142.

Song, X. and Y. Huang (2005). "On corrected score approach for proportional hazards model with covariate measurement error." Biometrics **61**(3): 702-714.

Miller TJ, Cicchetti D, Markovich PJ et al. The SIPS screen: a brief self-report screen to detect the schizophrenia prodrome. Schizo-phr Res 2004;70(Suppl.):78.

6. Appendices

Appendix 1 DSM-IV Code For Psychosis

Schizophrenia and other psychotic disorders

- Schizophrenia

  295.20 Catatonic type

  295.10 Disorganized type

  295.30 Paranoid type

  295.60 Residual type

  295.90 Undifferentiated type

- 295.40 Schizophreniform disorder
- 295.70 Schizoaffective disorder
- 297.1 Delusional disorder
- 298.8 Brief psychotic disorder
- 297.3 Shared psychotic disorder
- Psychotic disorder due to

  293.81 With delusions

  293.82 With hallucinations

- 298.9 Psychotic disorder NOS

Appendix 2

Consider a model with G latent classes. Let Y be the response variables, and let Z be the covariates affecting the relative frequencies of the class. In addition, let X be an unobserved vector indicating latent class membership. It follows that the posterior membership probabilities are given by

$$\tau(\mathrm{y},\mathrm{z}) = \mathrm{E}(\mathrm{X}|\mathrm{y},\mathrm{z}) = \mathrm{X} + \varepsilon, \quad \mathrm{E}(\varepsilon) = 0,$$

where ε can be regarded as the measurement error.

Claim: If the observed data $(Y_i, Z_i), i = 1, \dots, n$ are i.i.d. then a consistent estimator of $Var(\varepsilon) = \Omega$ (a singular matrix of rank G-1) is given by $\widehat{\Omega}$ with elements

$$\widehat{\Omega}_{jk} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \tau_{ij}(1-\tau_{ij}) & (j = k \in \{1, \dots, G\}) \\ -\frac{1}{n}\sum_{i=1}^{n} \tau_{ij}\tau_{ik} & (j \neq k \in \{1, \dots, G\}) \end{cases}$$

Proof: For notational convenience, we implicitly assume that X is fixed in all probability calculations. We have:

$$\begin{aligned}\Omega &= Var\{\tau(Y) - X\} = Var\{E(X|Y) - X\} \\ &= E[Var\{E(X|Y) - X|Y\}] + Var[E\{E(X|Y) - X|Y\}] \\ &= E\{Var(X|Y)\} + Var\{E(X|Y) - E(X|Y)\} \\ &= E\{Var(X|Y)\} \end{aligned} \quad (1)$$

Using the fact that $X|Y \sim Multinom\,(1;\, \tau_1(Y), \ldots, \tau_G(Y))$, it follows that the conditional variance-covariance matrix of X has elements

$$Var(X|Y)_{jk} = \begin{cases} \tau_j(Y)\{1 - \tau_j(Y)\} & (j = k \in \{1, \ldots, G\}) \\ -\tau_j(Y)\tau_k(Y) & (j \neq k \in \{1, \ldots, G\}) \end{cases}$$

By (1) and the Weak Law of Large Numbers, a consistent estimator of Ω takes the form

$$\widehat{\Omega}_{jk} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \tau_{ij}(1 - \tau_{ij}) & (j = k \in \{1, \ldots, G\}) \\ -\frac{1}{n}\sum_{i=1}^{n} \tau_{ij}\tau_{ik} & (j \neq k \in \{1, \ldots, G\}) \end{cases}$$

which completes the proof.