The natural history of smoking: A pattern-mixture random-effects regression model

Article · January 2000

CITATIONS
11

READS
22

2 authors:

Donald Hedeker
University of Chicago
237 PUBLICATIONS **11,071** CITATIONS
SEE PROFILE

Jennifer S Rose
Wesleyan University
86 PUBLICATIONS **2,241** CITATIONS
SEE PROFILE

The Natural History of Smoking: A Pattern-Mixture Random-effects Regression Model

Donald Hedeker, Ph.D.
Division of Epidemiology and Biostatistics
and Health Research and Policy Centers
University of Illinois at Chicago

Jennifer S. Rose, Ph.D.
Department of Psychology
Indiana University

To appear in J.S. Rose, L. Chassin, C.C. Presson, S.J. Sherman (eds.), *Multivariate Applications in Substance Use Research.* Hillsdale, NJ: Lawrence Erlbaum.

Abstract

This article describes and illustrates use of random-effects regression models (RRM) to examine the natural history of smoking from adolescence to adulthood. For longitudinal data analysis, RRM are useful because they allow for the presence of missing data, time-varying or invariant covariates, and subjects measured at different timepoints. Thus, a key advantage of RRM is that it can accomodate "unbalanced" longitudinal data, where a sample of subjects are not all measured at each and every timepoint. Also, variants of RRM have been developed to model dichotomous and ordinal outcomes, which are common in substance use research. A pattern-mixture approach (Little, 1995) can also be accommodated within RRM to further handle and describe the influence of missing data in longitudinal studies. For this approach, subjects are first divided into groups depending on their missing-data pattern, and then variables based on these groups are used as model covariates. Researchers are then able to examine the effect of missing-data patterns on the outcome(s) of interest. In this article we will illustrate these methods using an example from a study examining smoking status from early adolescence to young adulthood.

Introduction

Longitudinal studies play a prominent role in investigations of substance use. In these studies the same individuals are repeatedly measured on a number of variables over a series of timepoints. For example, individuals may be assessed in terms of their substance use across repeated weeks, months, or years. The aim of these studies is often to examine whether individuals change in their substance use across time, and whether other variables can produce a change in an individual's substance use across time (*e.g.*, treatment group). While data from longitudinal substance use studies are potentially very useful, investigators do not always make the best use of these data in their statistical analysis. One reason for this is that missing data invariably occur in longitudinal studies, and investigators may not know how to analyze such incomplete data. For this reason, individuals with incomplete data across time are sometimes removed from the analysis. This complete-data approach can provide biased results to the degree that individuals with complete data are not representative of the total population that was sampled for the study. Clearly, in substance use research individuals with incomplete data may be very different from those with complete data across time. Also, even if complete-data individuals are representative of the larger population, there is a loss in statistical power for hypothesis testing if the total dataset is not used in the analysis.

One approach for analysis of incomplete longitudinal data is random-effects regression models (RRM). For RRM, individuals contribute as many repeated observations to the analysis that they have data on. Furthermore, because time can be treated as a continuous variable in RRM, individuals do not have to be measured at the same timepoints. This is a useful feature for longitudinal studies which have follow-up times that are not uniform across all subjects. Covariates can be included in the model and can be either time-varying or invariant. Thus, changes in the repeated outcomes may be due to both stable characteristics of the subject (*e.g.*, their gender or race) as well as characteristics that change across time (*e.g.*, life-events). Finally, although most traditional approaches to longitudinal data analysis estimate

average change (across time) in a population, RRM can also estimate change for each subject when there is sufficient data.

Variants of RRM have been developed under a variety of names including: random-effects models (Laird & Ware, 1982), variance component models (Dempster, Rubin, & Tsutakawa, 1981); hierarchical linear models (Bryk & Raudenbush, 1987), multilevel models (Goldstein, 1995), two-stage models (Bock, 1989), random coefficient models (DeLeeuw & Kreft, 1986), mixed models (Longford, 1987), empirical Bayes models (Strenio, Weisberg, & Bryk, 1983), unbalanced repeated-measures models (Jennrich & Schlucter, 1986), and random regression models (Bock, 1983a and 1983b). Generalizations of RRM have also been developed for the case of dichotomous response data (Stiratelli, Laird, & Ware, 1984; Gibbons & Bock, 1987; Goldstein, 1991) and ordinal response data (Jansen, 1990; Ezzet & Whitehead, 1991; Hedeker & Gibbons, 1994), thus allowing a general framework for analysis of both continuous and categorical outcome variables. Recent articles that describe RRM usage for longitudinal substance use and/or mental health data include Gibbons *et al.*, (1993) and Hedeker and Mermelstein (1996). Several book-length texts (Bryk & Raudenbush, 1992; Diggle, Liang, & Zeger, 1994; Goldstein, 1995; Jones, 1993; Longford, 1993) further describe use of RRM.

For longitudinal studies with missing data across time, a further development that can be implemented in RRM is the "pattern-mixture" approach for incomplete longitudinal data described by Little (1993, 1994, 1995). In this approach, subjects are divided into groups based on their missing-data pattern across time. These groups can then be used, for example, to examine the effect of the missing-data pattern on the outcome(s) of interest. Also, interactions of these groups with other important model terms (*i.e.*, group and group by time interaction) can be included to examine whether changes across time, for example, depend on a subject's missing-data pattern. Overall estimates can be obtained by averaging over the missing-data patterns. Hedeker and Gibbons (1997) illustrate use of this approach applied to psychiatric clinical trials data.

In this paper we will first describe use of RRM applied to substance use data and then indicate how the pattern-mixture approach can be implemented. In addition to illustrating how the missing-data patterns are used as grouping variables in the analysis, we will also describe how overall estimates are obtained by averaging over patterns ("mixing" the patterns). Our aim is to illustrate an approach for analysis of incomplete longitudinal data that is both practical and informative. While we will apply the pattern-mixture approach within RRM, it can also be used with other longitudinal models that allow for incomplete data across time (*e.g.*, structural equation models or GEE-based models). For use within structural equation models, interested readers should consult Muthén, Kaplan, and Hollis (1987) or McArdle and Hamagami (1992).

Random-effects Regression Models (RRM)

To introduce RRM, consider a simple linear regression model for the measurement y of individual i ($i = 1, 2, \ldots, N$ subjects) on occasion j ($j = 1, 2, \ldots, n_i$ occasions):

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij} \tag{1}$$

This model represents the regression of the outcome variable y on the independent variable time (denoted t). The subscripts indicate whose observation it is (subscript i) and the relative timing of the observation (the subscript j). The actual timing is represented by the independent variable t which may represent time in weeks, months, etc. Both y and t carry the i and j subscripts, and so they are allowed to vary both by individuals and occasions. In a linear regression model, like (1), the errors ε_{ij} are assumed to be normally and *independently* distributed in the population with zero mean and common variance σ^2. This assumption of independence is generally unreasonable for longitudinal data. Since outcomes y are observed repeatedly from the same individuals, it is unlikely that the model errors of a given individual will be independent. It is much more likely to assume that errors within

an individual are correlated. In RRM, individual-specific effects are added to the model to account for this data dependency, as in

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \upsilon_{0i} + \varepsilon_{ij} \ , \tag{2}$$

where the additional term υ_{0i} indicates the influence of individual i on his/her repeated observations. Specifically, β_0 is the overall population intercept, υ_{0i} is the intercept deviation for subject i, and β_1 is the overall population slope. Thus, individuals deviate from the regression of y on t in a parallel manner in this model.

Since individuals in a sample are usually thought to be representative of a larger population of individuals, the individual-specific effects υ_{0i} are treated as random effects. This population distribution is usually assumed to be a normal distribution with mean 0 and variance σ^2_υ. With the random effects υ_{0i} in model (2), the errors ε_{ij} are now assumed to be normally and *conditionally independently* distributed in the population with zero mean and common variance σ^2. That is, the errors are independent conditional on the random individual-specific effects υ_{0i}. Since the errors now have an influence due to individuals removed from them, this conditional independence assumption is much more reasonable than the ordinary independence assumption associated with (1).

In model (2), individuals deviate from the regression of y on t in a parallel manner. A slightly more sophisticated model allows both the intercept and trend due to time to vary by individuals:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \upsilon_{0i} + \upsilon_{1i} t_{ij} + \varepsilon_{ij} \tag{3}$$

where, β_0 is the overall population intercept, β_1 is the overall population slope, υ_{0i} is the intercept deviation for subject i, υ_{1i} is the slope deviation for subject i, and ε_{ij} is an independent error term distributed normally with mean 0 and variance σ^2. The errors are independent conditional on both υ_{0i} and υ_{1i}. With two random individual-specific effects, the population distribution of intercept and slope deviations is assumed to be a bivariate normal $\mathcal{N}(0, \mathbf{\Sigma}_{\boldsymbol{\upsilon}})$, where $\mathbf{\Sigma}_{\boldsymbol{\upsilon}}$ is a 2×2 variance-covariance matrix. This model can be thought of as a personal trend or change

model because it represents the measurements of y as a function of time, both at the individual (υ_{0i} and υ_{1i}) and population (β_0 and β_1) levels. The intercept parameters indicate the starting point, and the slope parameters indicate the degree of change over timepoints. The population intercept and slope parameters represent the trend for the population, whereas the individual parameters express how the individual deviates from the population trend.

In this model, occasions range from $j = 1$ to n_i, with each person measured at n_i timepoints. Since n carries the subject subscript i, this indicates an important feature of the model, namely, that each subject may vary in terms of the number of measured occasions. The underlying assumption of the model is that the data that are available for a given individual are representative of how that individual deviates from the population trend across the timeframe of the study. Later, more detail will be given about the assumption of the model with regard to missingness.

If an individual deviates from the population trend in only a random manner, the individual parameters (*i.e.*, the intercept υ_{0i} and slope υ_{1i}) will be approximately equal to zero for that person. In this case, that persons' trend across time is well represented by the population trend parameters: the intercept β_0 and slope β_1. More likely, though, is that each individual will deviate in a personal (*i.e.*, systematic) manner from the population trend, and so the individual parameters are necessary in the model to characterize these personal trends. The variability in the individual parameters is assessed by the intercept variance $\sigma^2_{\upsilon_0}$, slope variance $\sigma^2_{\upsilon_1}$, and the covariance of the intercept and the slope $\sigma_{\upsilon_0\upsilon_1}$. If each individual's deviation from the population trend is only due to random error, then the individual intercepts υ_{0i} and slopes υ_{1i} equal 0 for all subjects, and these variance terms will also equal zero. Alternatively, as each individual's deviation from the population trend is non-random (*i.e.*, υ_{0i} and/or υ_{1i} are non-zero) these variance terms increase from zero.

Often a researcher is interested in assessing the influence of covariates, such as treatment group, on the responses across time. For this, covariates that either do not change over time (time invariant) or that vary across measured occasions (time

varying) can be added to the model:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_i + \beta_3 x_{ij} + \upsilon_{0i} + \upsilon_{1i} t_{ij} + \varepsilon_{ij} \ . \tag{4}$$

Here, β_2 is the coefficient for the time invariant covariate x_i, and β_3 is the coefficient for the time varying covariate x_{ij}. Interactions between the covariates can be included in the same way as interactions are included into an ordinary multiple regression model. For example, x_i might represent the treatment group that a subject is assigned to (for the course of the study), and x_{ij} might be the treatment by time interaction that is obtained as the product of x_i by t_{ij}.

RRM for Longitudinal Categorical Data

Generalizations of RRM for categorical outcomes have typically used either a probit or logistic regression model for the categorical outcomes, and various methods for incorporating and estimating the influence of the random effects. A recent review article (Pendergast *et al.*, 1996) presents and describes many of these approaches. In general, parameter estimation is computationally more intensive for these models than for models of continuous outcome data, and so, often only a single random subject effect is assumed. In this article we will describe RRM for categorical data with a single random effect; for examples with multiple random effects see Hedeker and Gibbons (1994) or Gibbons and Hedeker (1994).

For dichotomous outcomes, denoted Y_{ij} (with values, say, of 0 or 1), a logistic regression model can be used to relate the log odds of the outcome in terms of explanatory variables:

$$\log\left[\frac{P(\mathrm{Y}_{ij} = 1 \mid \boldsymbol{x}_{ij})}{1 - P(\mathrm{Y}_{ij} = 1 \mid \boldsymbol{x}_{ij})}\right] = \beta_0 + \beta_1 t_{ij} + \beta_2 x_i + \beta_3 x_{ij} \ , \tag{5}$$

where, throughout the article, $\boldsymbol{x}_{ij}$ will denote the vector of covariates (*e.g.*, t_{ij}, x_i, x_{ij}). The numerator is the probability of a 1 response, and the denominator $1 - P(\mathrm{Y}_{ij} = 1)$ equals the probability of a 0 response. The ratio of these probabilities is the odds of a 1 response, and the log of this ratio is the log odds of a 1 response (sometimes called the logit of P). Notice that the log odds is equal to 0 when the probability

of a 1 response equals .5 (*i.e.*, equal odds of a response in category 0 and category 1), is negative when the probability is less than .5 (*i.e.*, odds favoring a response in category 0), and is positive when the probability is greater than .5 (*i.e.*, odds favoring a response in category 1).

Again, because longitudinal data are not independent, use of this ordinary logistic regression model would be inappropriate. As in the linear regression case, we can augment the logistic regression model with random subject-effects to account for the data dependency, namely,

$$\log\left[\frac{P(\mathrm{Y}_{ij}=1 \mid \boldsymbol{x}_{ij}, \upsilon_i)}{1-P(\mathrm{Y}_{ij}=1 \mid \boldsymbol{x}_{ij}, \upsilon_i)}\right] = \beta_0 + \beta_1 t_{ij} + \beta_2 x_i + \beta_3 x_{ij} + \upsilon_i \quad . \tag{6}$$

The individual-varying effect υ_i represents an individual's influence on the log odds of a response for that individual. As in the model for continuous outcomes, these subject-varying effects are assumed to be normally distributed in the population of subjects with mean 0 and variance σ_υ^2. The subject effects υ_i reflect the influence that subject i has on his/her repeated observations. To the degree that subjects exert little influence on their responses, over and above the influence of the other model terms, values of υ_i, and correspondingly σ_υ^2, will not deviate from zero. However, as subjects exert influence on their repeated observations, values of υ_i will deviate from zero, resulting in a population variance σ_υ^2 that is greater than zero.

Pattern-mixture models

When missing data are observed across time, Laird (1988) has shown that RRM provide valid statistical tests if the missing data are ignorable. By ignorable, it is meant that the missingness can depend on observed covariates *and* previous values of the dependent variable from the subjects with missing data. Thus, if missingness is related to previous performance, in addition to other observable subject characteristics (*i.e.*, covariates), then RRM provides valid statistical tests for the model parameters. However, by itself, RRM does not provide valid tests if the missing data are nonignorable. Nonignorable missing data result when the missingness depends

on the value of the dependent variable that would have been observed if the observation was not missing. For example, in smoking research, missing smoking status observations are often recoded to smoking under the belief that the observation is missing because the individual is smoking (*i.e.*, the missingness is nonignorable). Besides recoding missing to smoking, and therefore assuming that missingness is always indicative of smoking, more statistical methods for dealing with nonignorable missingness are becoming available. Little and Schenker (1995) describe a variety of techniques for dealing with nonignorable missingness; here, we will describe one of these techniques: the pattern-mixture approach.

In a series of articles, Little (1993, 1994, 1995) has described "pattern-mixture models" for analysis of missing data, and has provided a general statistical treatment of these models. Using a longitudinal psychiatric dataset, Hedeker and Gibbons (1997) describe and illustrate use of pattern-mixture RRM for continuous outcomes. Here, we will describe use of a pattern-mixture RRM for dichotomous substance use outcomes. As will be illustrated, the pattern-mixture approach is appealing because it provides useful results and is relatively easy to implement.

To apply the pattern-mixture approach, subjects are classified into groups based on their missing-data pattern across time. For example, if subjects are measured at three timepoints, then eight (2^3) missing-data patterns are possible. One of these patterns (missing at all three timepoints), however, provides no data on the outcome variable, and although this pattern can be included in the general pattern-mixture approach by making various missing-data assumptions (see Little, 1993), in what follows we will exclude it. To include the missing-data pattern information in a statistical model like RRM, the remaining seven patterns can be represented by six dummy-coded variables, for example the codings D1 to D6 given in Table 1.

Insert Table 1 about here

These six dummy-codes represent deviations from the complete-data pattern (OOO); D1 compares pattern MOO to completers, D2 compares pattern OMO to completers,

etc. Depending on substantive considerations, alternative coding schemes that provide other comparisons among the seven pattern groups can be used (see Darlington, 1990, pp. 232-241; or Cohen and Cohen, 1983, chapter 5). These variables are then entered into the model as main effects and as interactions with other model variables. Thus, the pattern-mixture approach allows one to examine: (a) the degree to which the outcome variable differs across missing-data patterns (*i.e.*, a main effect of the missing-data pattern dummy-coded variables), and (b) the degree to which the missing-data pattern moderates the influence of other model terms (*i.e.*, interactions with missing-data pattern). From the full model, submodels can be obtained for each of the missing-data pattern groups, and overall averaged estimates (*i.e.*, averaging over the missing-data patterns) derived for the model parameters.

In some cases, it may not be feasible, theoretically interesting, or interpretable to model differences between all potential missing-data pattern groups. Also, some of the patterns, either by design or by chance, may not be realized in the sample. For example, in some studies once a subject is missing at a given wave they are missing at all later waves. In this case, with three waves, three monotone missing-data patterns (OMM, OOM, and OOO) would result. Then, two dummy-coded variables M1 and M2 given in Table 1 could be used to represent differences between each of the two dropout groups and the group of subjects observed at all timepoints. An investigator might want to combine some missing-data patterns to increase interpretability. For example, groups might be formed based on the last available measurement wave. Table 1 lists dummy-codes L1 and L2 for this purpose: L1 contrasts individuals who were not measured after the first timepoint with those who were measured at the last timepoint, and L2 contrasts those subjects not measured after the second timepoint with those who were measured at the last timepoint. Other recodings that may be used include a simple grouping of complete data vs. incomplete data, as given by contrast I1 in Table 1, or missing at the final timepoint vs. available at the final timepoint, as given by contrast F1 in Table 1. From an analytic perspective, missing-data pattern is simply an additional between-subjects factor. In determining which

set of contrasts to use, similar considerations arise as when dealing with the coding of other between-subjects factors (*e.g.*, race or marital status). Thus, substantive considerations and data sparseness are important factors in determining how many missing-data patterns to consider and which set of contrasts to use.

Example

We first illustrate RRM applied to longitudinal substance use data, and then augment the model to accommodate the pattern-mixture approach. The data for this example are taken from an ongoing cohort-sequential study of the natural history of cigarette smoking from adolescence to adulthood (Chassin *et al.*, 1984; Chassin *et al.*, 1990, Chassin *et al.*, 1996). The current example expands upon the results reported by Chassin *et al.*, (1996), which investigated the natural history of cigarette smoking from adolescence to adulthood in a community sample.

The following description is taken from Chassin *et al.*, (1996). Between 1980 and 1983, all consenting 6-12th graders in a midwestern county school system completed up to four annual, classroom administered surveys (a total sample of 8556 students were assessed at least once. Follow-up surveys were conducted in 1987 (young adulthood) and 1993 (adulthood) with retention rates of 73% of the original adolescent sample at each follow-up. For this study, we used the five oldest cohorts to ensure that sample participants had graduated high school by 1987, and thus could be considered young adults (n=3897). At all three time points, participants were defined as current regular smokers if they reported smoking weekly or more often. For the measure of adolescent smoking status, we used reports of 11th or 12th grade smoking status so that this measurement could best represent participants' "final" adolescent smoking status. A very small number of infrequent smokers (less often than weekly) were eliminated from the analyses. Participants were primarily Caucasian (96%) and gender was split evenly. In 1993, 72% reported having had some post high school education. The mean age at the adolescent, young adult, and adult time points were 17, 23, and 29 years, respectively; Table 2 lists the frequency

distribution of age for each of the three timepoints.[1]

Insert Table 2 about here

For data analysis, Chassin *et al.*, (1996) used RRM to identify age-related trajectories in smoking across time (*i.e.*, adolescence, young adulthood, and adulthood). The outcome of interest was a dichotomous variable measuring regular (at least weekly) smoking at the three timepoints. The observed prevalences of smoking were .20 (adolescence), .27 (young adulthood), and .26 (adulthood). RRM results indicated that smoking increased significantly between adolescence and young adulthood, but remained stable from young adulthood to adulthood (*i.e.*, a nonsignificant decrease between these 2 timepoints). The Chassin *et al.*, (1996) study included participants with missing data in the RRM analyses (*i.e.*, individuals with smoking data on at least one of the three time points were included), but did not specifically explore the influence of different patterns of missingness. The current example extends the previous analyses by using a pattern-mixture approach to explore the effects of patterns of missing data on smoking trajectories from adolescent to adulthood.

In this example, time is a categorical variable that takes on three values (adolescence, young adult, and adult). Subjects may have different numbers of observations, but the measurement timepoints are the same for all subjects. With three timepoints, there are two degrees of freedom available to represent changes across time in our model. Thus in performing the analysis one must consider how to represent change attributable to time in the model. One choice would be to perform a trend analysis and use polynomials to examine whether there are linear and/or quadratic trends in smoking across time. Alternatively, one of the timepoints could be designated as the reference time period (*e.g.*, adolescence) and tests for differences between this time period and each of the other two time periods (*i.e.*, young adult and adult) could be performed. Since the levels of time are ordinal, Helmert

[1]©(1996) by the American Psychological Association. Adapted with permission.

contrasts provide a useful choice (Bock, 1975). Here, with three levels, the first Helmert contrast compares adolescence (timepoint 1) to the average of young adult and adult time periods (average of timepoints 2 and 3). The second Helmert contrast compares the young adult (timepoint 2) to the adult (timepoint 3) time period. Thus, in simple terms, the first contrast is between adolescence and adulthood, and the second is a comparison within adulthood. Another, somewhat similar, choice would be to use sequential contrasts and compare (a) adolescence to young adult and (b) young adult to adult. In general, the choice of contrasts should be specified prior to the analysis and should be determined by the research question(s) of interest. As noted in Chassin *et al.*, (1996), specific questions of interest in this study were (1) whether there was an increase in smoking following adolescence, and (2) whether smoking began to decline between young adulthood and later adulthood. These two questions are directly examined by the two successive Helmert contrasts.

In order to examine the possibility of a cohort effect in these data, we created a variable indicating age in 1993. Individuals were separated into two age cohort groups: those 28 or younger in 1993 versus those older than 28 in 1993. Of the 3613 individuals, 1250 (34.6%) were 28 or younger, and 2363 (65.4%) were older than 28. Table 3 lists the missing-data patterns for these two cohorts.

Insert Table 3 about here

As can be seen from Table 3, there is an approximate linear relationship between the amount of missing data and proportion smoking. Clearly, subjects with complete data at all timepoints have the lowest proportion of smoking. Also, there appears to be more missing data for the younger cohort. Classifying missing-data patterns as either completion (*i.e.*, pattern OOO) or non-completion (*i.e.*, all other patterns) gives completion rates of 48% (596 of 1250) and 55% (1303 of 2363) for the younger and older cohorts, respectively. This difference in data completion differs significantly between the two cohorts ($\chi^2 = 18.3$, d.f. $= 1$, $p < .001$), indicating

that missingness was more pronounced for the younger cohort.[2]

Though the association of missingness and cohort is interesting, a more vital question is whether missingness is related to the outcome variable, smoking status. To the degree that missingness and smoking are related, it will be important to control for missingness when assessing the cohort effect on smoking. It will also be important to examine potential interactive effects of missingness (with time and cohort-related influences) on smoking status. For example, the interaction of missingness and cohort will indicate whether the cohort effect on smoking differs between missing data patterns.

Separate Analysis of Completers and All Available Cases

Let us first consider a model for changes in smoking status across time as a function of cohort and the Helmert time-related contrasts:

$$\log\left[\frac{P(\mathrm{Y}_{ij}=1 \mid \boldsymbol{x}_{ij}, \upsilon_i)}{1-P(\mathrm{Y}_{ij}=1 \mid \boldsymbol{x}_{ij}, \upsilon_i)}\right] = \beta_0 + \beta_1 T1_j + \beta_2 T2_j + \beta_3 C_i + \beta_4(C_i \times T1_j) + \beta_5(C_i \times T2_j) + \upsilon_i \quad . \qquad (7)$$

Since $Y = 1$ designates smoking and $Y = 0$ designates abstinence, this is a model for the log-odds of smoking. The explanatory variables and their codings are as follows: T1 represents the Helmert contrast comparing adolescent versus the average of the young adult and adult timepoints (coded as -1, .5, .5, for the three timepoints, respectively), T2 represents the second Helmert contrast comparing young adult versus adult timepoints (coded as 0, -1, 1, for the three timepoints, respectively),

[2]Participants in the younger cohort were less likely to have been in 11th or 12th grade during adolescent data collection. Specifically, 61.8% and 84.5% were in either 11th or 12th grade during adolescent data collection for the young and old cohorts, respectively, the remaining 38.2% and 15.4% were measured somewhere between grades 7 to 10 for this assessment. Consequently, adolescent smoking status was ambiguous for those participants measured in grades 7 to 10, because we could not be sure that they did or did not take up smoking before leaving high school. These participants were coded as missing on the measure of adolescent smoking status, resulting in greater missingness among younger cohort participants.

and C represents Cohort (0 for ≤ 28 and 1 for > 28). Due to the coding of cohort, β_1 and β_2 represent the time effects (*i.e.*, T1 and T2) for the younger individuals, and $\beta_1 + \beta_4$ and $\beta_2 + \beta_5$ represent the time effects for the older individuals. Thus, β_4 and β_5 represent differences in the two time effects between age cohort groups. Specifically, if there is no difference in smoking between cohorts when comparing the adolescent versus the two adult timepoints, β_4 will equal 0. Similarly, if the difference in smoking between the two adult timepoints is the same for the two cohorts, then β_5 will equal 0.

This logistic random-effects regression model was first fit including only completers (*i.e.*, only the 1899 subjects with data at all timepoints) and then re-estimated using all 3613 subjects. Parameter estimates are presented in the first two sets of columns in Table 4.

Insert Table 4 about here

Conclusions based on these two analyses agree in terms of indicating that there is a significant increase in smoking between adolescence and the two later timepoints ($p < .001$ for both analyses). Calculating odds ratios (OR) based on the all-subjects analysis, subjects are 2.08 ($= \exp[1.5 \times .489]$) and 1.59 ($= \exp[1.5 \times (.489 - .181)]$) times as likely to be smoking at the adult timepoints, relative to adolescence, for the young and old cohorts, respectively.[3] The multiplication factor of 1.5 is necessary to account for the Helmert contrast coding that differed by 1.5 units between adolescent (-1) and adult timepoints (both .5).

In terms of the comparison between young adult and adult timepoints (*i.e.*, T2), the completer analysis, but not the all-subjects analysis, indicates a significant decline in smoking between these two timepoints for the young cohort ($p < .036$); the odds of smoking is 1.64 times as likely at the young adult, relative to adult timepoint (OR for adult relative to young adult smoking $= \exp[2 \times (-.247)] = .61$; OR for

[3] All odds ratios reported in this article are subject-specific odds ratios, that is, they are adjusted for the level of the random subject effect (see appendix).

young adult relative to adult smoking = 1/.61 = 1.64). A few marginally significant results are also obtained and differ somewhat between the two analyses. The completer analysis suggests that the increase in smoking between adolescence and adult timepoints (*i.e.*, the T1 contrast) is not as great for the old cohort ($p < .074$; OR = $\exp[1.5 \times (-.245)] = .69$), whereas the analysis using all data instead suggests an overall decrease in smoking for the old cohort ($p < .078$; OR = $\exp[-.304] = .74$).

The p-values in the table correspond to the so-called "Wald test" (Wald, 1943), based on the ratio of the parameter estimate to its standard error. These test statistics (*i.e.*, z = ratio of the parameter estimate to its standard error) are compared to a standard normal frequency table to test the null hypothesis that a given parameter equals 0. These z-statistics can also be squared, in which case the resulting test statistic is distributed as chi-square on 1 degree of freedom. In either case, the p-values are identical. There are some concerns in using the standard errors in constructing a hypothesis test for the random-effect variance term (*i.e.*, the subject standard deviation σ_v), particularly when the variance is near-zero and the number of subjects is small (Bryk & Raudenbush, 1992). As a result, we do not provide p-values for the subject standard deviations listed in Table 4.

Figures 1 and 2, plot the observed and estimated group proportions across the three timepoints corresponding to these two analyses. Later we will describe how these estimated group proportions were obtained. For now, we note that, as the figures illustrate, the models fit the observed data reasonably well.

Insert Figures 1 and 2 about here

A Comparison of Completers versus Incompleters

As stated earlier, we are also interested in examining whether missingness is related to smoking, and whether missingness moderates the effect of cohort on smoking. For this, we will define a variable Miss with two values: 0 if the person was measured at all timepoints (completers), and 1 if the person was not measured at all timepoints (incompleters). Dividing subjects into these two groups is a simple

characterization of the missing-data patterns, however it provides a direct way of assessing whether subjects who were assessed at all timepoints differ from those who were not. Notice that based on the data presented in Table 3, there does seem to be a lower level of smoking for completers relative to incompleters. For these data, this dichotomization of the missing-data patterns appears reasonable, although a more sophisticated analysis could be used to further distinguish the incompleter missing-data patterns.

Figure 3 depicts the observed proportions across time for incompleters.

Insert Figure 3 about here

Notice that the sample size used to determine these observed proportions varies across timepoints. Contrasting Figure 3 to Figure 1 (the figure for completers) clearly illustrates a higher level of smoking for incompleters relative to completers. Regarding cohort, Figures 1 and 3 differ in their interpretation. For completers, Figure 1 suggests a higher smoking level at the two adult timepoints for the younger cohort (relative to the older cohort), whereas, for incompleters, Figure 3 suggests the opposite.

To more formally examine the effect of missing-data pattern on smoking, we augment the model given by equation (7) in the following way:

$$\begin{aligned}\log\left[\frac{P(\mathrm{Y}_{ij}=1 \mid \boldsymbol{x}_{ij},\upsilon_i)}{1-P(\mathrm{Y}_{ij}=1 \mid \boldsymbol{x}_{ij},\upsilon_i)}\right] &= \beta_0+\beta_1 T1_j+\beta_2 T2_j+\beta_3 C_i \\ &\quad +\beta_4(C_i{\times}T1_j)+\beta_5(C_i{\times}T2_j) \\ &\quad +\beta_6 Miss_i+\beta_7(Miss_i{\times}T1_j)+\beta_8(Miss_i{\times}T2_j) \\ &\quad +\beta_9(Miss_i{\times}C_i)+\beta_{10}(Miss_i{\times}C_i{\times}T1_j) \\ &\quad +\beta_{11}(Miss_i{\times}C_i{\times}T2_j)+\upsilon_i \qquad (8)\end{aligned}$$

Because of the interactions with time in the model and the use of the (centered) Helmert contrasts for time, β_3 and β_6 represent main effects (*i.e.*, averaged across

time) of Cohort and Miss, respectively, and β_9 is the averaged two-way interaction of these two variables. With the coding of Cohort and Miss as described, the regression coefficients β_1 and β_2 represent the time effects for younger completers (both Cohort and Miss equal 0). Parameters β_4 and β_5 compare the time-related effects between old and young completers, whereas β_7 and β_8 compare the time-related effects between young completers and young incompleters. Finally, β_{10} and β_{11} represent the interaction of Cohort by Miss on the time-related effects. The results of this analysis are listed in the last set of columns in Table 4 (*i.e.*, labeled "pattern-mixture").

The likelihood ratio test for the joint significance of the Miss-related model terms (β_6 through β_{11}) yields $\chi^2 = 7771.9 - 7648.2 = 123.7$, which on 6 degrees of freedom is highly significant ($p < .001$). This test confirms that missingness and interactions with missingness are significantly related to smoking, over and above the influences of Cohort, Time, and Cohort by Time. In terms of the significance of the individual regression coefficients, from Table 4 we see that subjects with incomplete data have a significantly higher smoking level than completers ($p < .001$). Expressed as an odds ratio, an incomplete-data subject is more than five times as likely as a completer to be smoking (OR = exp[1.633] = 5.12). Also, a significant 3-way interaction term is observed for Miss by Cohort by T1 ($p < .013$). Comparing Figures 1 and 3 help to explain this interaction. Basically, the adult versus adolescence increase in smoking (averaging the young adult and adult timepoints) operates in an opposite manner for completers and incompleters: for completers it is more pronounced for younger subjects, whereas for incompleters it is more pronounced for older subjects. There is also a marginally significant Miss by T2 interaction ($p < .080$) suggesting that the significant drop in smoking from young adult to adult timepoints that is observed in young completers ($p < .038$) does not occur for young incompleters.

Notice that the results for β_0 through β_5 are almost exactly the same for the pattern-mixture analysis and the analysis done only on completers. Because of our coding of Miss (0 = no and 1= yes), these parameters have exactly the same meaning in these two models. Namely, the intercept, Cohort, Time, and Cohort by

Time parameters for completers. The estimates are not exactly the same because the model variance terms differ. In Model 1, σ_v represents the subject standard deviation for completers only, whereas in Model 3 this same parameter represents subject variation for completers and incompleters. This coding of Miss also allows us to address an important substantive point: does the inclusion of subjects with incomplete data matter? If their inclusion does not make any difference, then from a statistical perspective, $\beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$. As mentioned above, the likelihood-ratio test of this null hypothesis was easily rejected. Since statistical significance is influenced by sample size, it is also important to consider the practical significance of separating complete and incomplete data subjects. In this regard, comparison of Figures 1 and 3 strongly conveys the importance of distinguishing between those with complete and incomplete data.

Thus, missingness does matter for these data and our analysis has uncovered two interesting and statistically significant results. First and foremost, there is a large significant increase in smoking that is observed for subjects with incomplete data, relative to complete data subjects. Second, the cohort difference across time varies depending on missingness. Our first result supports the often-reported finding of increased baseline use of tobacco, alcohol, and other drugs for study dropouts compared to completers in adolescent substance use studies (Tebes, Snow, & Arthur, 1992). Our second result is more difficult to interpret, however, it does indicate that the cohort by time interaction is not robust across missing data patterns.

Estimating trend lines

Results from the analyses, which are based on modeling the log-odds of smoking, can be used to obtain estimated probabilities of smoking for various subgroups. In particular, we can use the model estimates to derive the estimated trend lines for the four subgroups that are depicted in Figures 1 and 3. To do this, some additional calculations using the parameter estimates presented in Table 4 are necessary. First, note that equation (8) can be algebraically re-expressed to provide the probability of smoking in terms of the model parameters (*e.g.*, see Kleinbaum, 1994, pages 17-18):

$$P(Y_{ij} = 1 \mid \boldsymbol{x}_{ij}, \upsilon_i) = \frac{1}{1 + \exp(-z_{ij})} \tag{9}$$

where

$$\begin{aligned} z_{ij} &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \upsilon_i \qquad (10) \\ &= \beta_0 + \beta_1 T1_j + \beta_2 T2_j + \beta_3 C_i + \beta_4 (C_i \times T1_j) + \beta_5 (C_i \times T2_j) \\ &\quad + \beta_6 Miss_i + \beta_7 (Miss_i \times T1_j) + \beta_8 (Miss_i \times T2_j) \\ &\quad + \beta_9 (Miss_i \times C_i) + \beta_{10} (Miss_i \times C_i \times T1_j) \\ &\quad + \beta_{11} (Miss_i \times C_i \times T2_j) + \upsilon_i \ . \end{aligned}$$

Using the parameter estimates and equation (9) we can derive estimated probabilities for the four subgroups (based on the codings of C, Miss, T1, and T2). For this, estimated values of z_{ij} for the four subgroups are obtained as:

$$\begin{aligned} \text{Young Completers} \qquad & (C = 0 \text{ and } Miss = 0) \\ \hat{z}_{ij} &= \hat{\beta}_0 + \hat{\beta}_1 T1_j + \hat{\beta}_2 T2_j + \upsilon_i \\ &= -3.343 + .513\, T1_j - .243\, T2_j + \upsilon_i \\ \text{Old Completers} \qquad & (C = 1 \text{ and } Miss = 0) \\ \hat{z}_{ij} &= (\hat{\beta}_0 + \hat{\beta}_3) + (\hat{\beta}_1 + \hat{\beta}_4) T1_j + (\hat{\beta}_2 + \hat{\beta}_5) T2_j + \upsilon_i \\ &= (-3.343 - .222) + (.513 - .242)\, T1_j + (-.243 + .169)\, T2_j + \upsilon_i \\ &= -3.565 + .271\, T1_j - .074\, T2_j + \upsilon_i \\ \text{Young Incompleters} \qquad & (C = 0 \text{ and } Miss = 1) \\ \hat{z}_{ij} &= (\hat{\beta}_0 + \hat{\beta}_6) + (\hat{\beta}_1 + \hat{\beta}_7) T1_j + (\hat{\beta}_2 + \hat{\beta}_8) T2_j + \upsilon_i \\ &= (-3.343 + 1.633) + (.513 - .355)\, T1_j + (-.243 + .286)\, T2_j + \upsilon_i \\ &= -1.710 + .158\, T1_j + .043\, T2_j + \upsilon_i \\ \text{Old Incompleters} \qquad & (C = 1 \text{ and } Miss = 1) \end{aligned}$$

$$
\begin{aligned}
\hat{z}_{ij} &= (\hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_6 + \hat{\beta}_9) + (\hat{\beta}_1 + \hat{\beta}_4 + \hat{\beta}_7 + \hat{\beta}_{10})T1_j \\
&\quad + (\hat{\beta}_2 + \hat{\beta}_5 + \hat{\beta}_8 + \hat{\beta}_{11})T2_j + \upsilon_i \\
&= (-3.343 - .222 + 1.633 + .258) + (.513 - .242 - .355 + .685)\, T1_j \\
&\quad + (-.243 + .169 + .286 - .369)\, T2_j + \upsilon_i \\
&= -1.674 + .601\, T1_j - .157\, T2_j + \upsilon_i
\end{aligned}
$$

To apply these equations, we still need to indicate values for the Helmert time contrasts $T1$ and $T2$, and also a value for the random subject effect υ_i. For the Helmert contrasts, $T1 = -1, .5, .5$ and $T2 = 0, -1, 1$ for the three timepoints, respectively. For the random subject effects, based on our analysis, these came from a normal distribution with mean 0 and standard deviation estimated as 3.697. Thus, we could use specific values or sampled values from this normal distribution to yield probabilities that are specific to the particular subject effect υ_i used. For this reason, estimates from the random-effects model are sometimes termed subject-specific estimates (Neuhaus, Kalbfleisch, & Hauck, 1991; Bryk, Raudenbush, & Congdon, 1996). To reiterate, using equation (9) yields the subject-specific, or conditional, probability of the outcome for a specific value of the random effect υ_i. The observed proportions, alternatively, are marginal estimates that indicate the estimated probability of the outcome for a group of subjects, not a specific subject. For non-linear models like the logistic model used here, marginal and subject-specific estimates are not on the same scale, and so cannot be directly compared. However, marginal estimates can be obtained from the subject-specific estimates, but additional calculations are required (see Appendix). Essentially, the subject-specific estimates need to be averaged across the population of subjects to obtain marginal estimates. The trend lines depicted in the figures can then be obtained from these marginal estimates.

Averaging across patterns

Thus far, the analysis has indicated that the Cohort and Time effects vary by

missing-data pattern (*i.e.*, whether a subject has complete data or not). Trying to understand the reasons that effects vary by missing-data pattern is often of primary importance. For example, in the pattern-mixture analysis in Hedeker and Gibbons (1996) of schizophrenic patients' severity scores across time, the significant drug by time by dropout interaction indicated that placebo dropouts experienced little change in severity levels across time, whereas active drug dropouts experienced dramatic improvement across time. In this case, the reasons for dropout might have been that placebo subjects who did not improve were removed from the study and given active drug treatment, whereas active drug patients who dropped out left the study because of their dramatic improvement. In the present example, the meaning of the significant Miss by Cohort by T1 and marginally significant Miss by T2 interactions is less apparent. Clearly, when important interactions with the missing-data patterns exist, interpretation of the interactions can be enlightening.

It is also useful, in many cases, to obtain overall population estimates averaging over the missing-data patterns. Little (1993 and 1995) and Hogan and Laird (1997) describe this additional step of the pattern-mixture approach. In the present case, based on the final model, we can obtain estimates for the six fixed effects (Intercept, T1, T2, Cohort, Cohort by T1, and Cohort by T2) separately for completers

$$\hat{\boldsymbol{\beta}}^{(c)} = \begin{bmatrix} -3.343 & .513 & -.243 & -.222 & -.242 & .169 \end{bmatrix}' \tag{11}$$

and incompleters

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(ic)} &= \hat{\boldsymbol{\beta}}^{(c)} + \begin{bmatrix} 1.633 & -.355 & .286 & .258 & .685 & -.369 \end{bmatrix}' \\ &= \begin{bmatrix} -1.710 & .158 & .043 & .036 & .443 & -.200 \end{bmatrix}' . \end{aligned} \tag{12}$$

Averaged estimates for these four parameters (denoted $\hat{\bar{\boldsymbol{\beta}}}$) are then equal to:

$$\hat{\bar{\boldsymbol{\beta}}} = \pi^{(c)}\hat{\boldsymbol{\beta}}^{(c)} + \pi^{(ic)}\hat{\boldsymbol{\beta}}^{(ic)} \tag{13}$$

where $\pi^{(c)}$ and $\pi^{(ic)}$ represent the population weights for completers and incompleters, respectively. Although these weights are not usually known, they can be estimated by the sample proportions (1899/3613 and 1714/3613 for completers and incompleters, respectively). This yields:

$$\hat{\bar{\boldsymbol{\beta}}} = \begin{bmatrix} -2.568 & .344 & -.108 & -.099 & .083 & -.006 \end{bmatrix}' \tag{14}$$

as the averaged overall estimates. To obtain corresponding estimates of the standard errors, the delta method as described in Hogan and Laird (1997) can be used:

$$\hat{V}(\hat{\bar{\beta}}_h) = (\hat{\pi}^{(c)})^2\, \hat{V}(\hat{\beta}_h^{(c)}) \;+\; (\hat{\pi}^{(ic)})^2\, \hat{V}(\hat{\beta}_h^{(ic)}) \;+\; \frac{\hat{\pi}^{(c)}\hat{\pi}^{(ic)}}{N}(\hat{\beta}_h^{(c)} - \hat{\beta}_h^{(ic)})^2 \tag{15}$$

where $h = 1, \ldots, 6$ denotes the six fixed effects, $N = 3613$ is the total number of subjects, and $\hat{V}(\hat{\beta}_h)$ denotes the estimate of the variance of $\hat{\beta}_h$ (*i.e.*, the square of its estimated standard error). The last term in the sum is the contribution to the variance that is added because the proportion of completers (and incompleters) is estimated in the sample. The estimated standard errors for these six overall terms are then .163, .112, .082, .178, .135 and .115. Calculating Wald statistics (*i.e.*, estimate divided by its standard error) for these six parameters yield the following p-values: .001, .002, .189, .575, .539 and .961.

It is interesting to compare these pattern-mixture averaged estimates and standard errors to those obtained from the RRM analysis ignoring missing-data patterns (*i.e.*, the RRM analysis of all 3613 subjects presented in the second set of columns in Table 4). Comparing these two indicates that the estimates for the young subjects (β_0, β_1, and β_2) are very similar. However, the estimates for the Cohort effect (β_3) and Cohort by Time interaction effects (β_4 and β_5) are not so similar between these approaches. Most notably, although a marginally significant Cohort effect ($\hat{\beta}_3 = -.304, p < .078$) is indicated by ordinary RRM analysis, the pattern-mixture averaged result is highly non-significant for this parameter ($\hat{\bar{\beta}}_3 = -.108, p < .575$). Essentially, in computing the averaged estimates across patterns the pattern-mixture

approach gives more relative weight to the subgroup of subjects with missing data (whose Cohort estimate $\hat{\beta}_3 = .036$) than does the ordinary RRM which weights each subject's data by its degree of precision. All other things equal, the degree of precision is largely determined by the amount of data. Thus, in ordinary RRM, because incompleters provide less data (than completers) their data receives less influence than in the pattern-mixture approach. The bottom line is that misleading results can be obtained by ignoring the missing-data patterns when they have influence on the outcome variable or when they moderate the effects of other model variables.

Discussion

For incomplete longitudinal data, RRM provide a useful approach because subjects are not assumed to be measured at the same number of timepoints. As a result, subjects who are missing at a given interview wave are not excluded from the analysis. The assumption of the model is that the data that are available for a given subject are representative of that subject's deviation from the (average) trend lines based on the model covariates. Also, RRM can include both time-varying and time-invariant covariates and are available for continuous and categorical outcomes.

In this article, we have illustrated how to augment these random-effects models by including variables defined by a subject's pattern of missing data. The pattern-mixture approach provides assessment of degree to which (the influence of) model terms vary by the missing-data patterns, and provides a way of obtaining estimates averaged over the missing-data patterns. In our example, we considered a very simple characterization of the pattern-mixture approach, namely, was the subject measured at all study timepoints or not. Though simple, it showed that subjects with incomplete data had higher levels of smoking than those measured at all timepoints. Also, there was evidence that the cohort by time interaction varied between completers and incompleters. Thus, the addition of missing data pattern led to substantive findings that were overlooked in our ordinary RRM analysis. In particular, our finding of increased smoking for incompleters from the pattern-mixture analysis

supports previous research from adolescent substance use studies (Tebes, Snow, & Arthur, 1992).

In dealing with incomplete longitudinal data, the complexity of the statistical model used may depend on the amount of missing data. For example, if a study has only a few subjects with missing data across time, it may not matter whether the subjects with missing data are ignored or included as a group in the analysis. The statistical power for detecting effects (*i.e.*, main effects and interactions) due to missing-data patterns is lower, all other things equal, as the numbers of subjects in those patterns are reduced. While it is difficult to give general guidelines for the treatment of missing-data patterns in a given analysis (*i.e.*, number of missing-data patterns, or grouping together of missing-data patterns), the considerations are similar to those encountered when dealing with other between-subjects grouping variables.

Another consideration in appying the pattern-mixture approach is the number of interaction terms to form with the missing data patterns. In the present example, we had a fairly small number of model covariates and only one variable to represent missing data patterns, so that including interactions with missingness for all covariates was not problematic. However, if many covariates are included in the model (*e.g.*, gender, education level beyond high school, ethnicity) then the number of possible interactions can become quite large. In this case, a stratified analysis by missing data pattern can be helpful in identifying covariates that have substantially different effects across missing data patterns. Including missing-data pattern interactions for this subset of covariates would then provide a more parsimonious pattern-mixture model.

Methods for appropriately handling missing data have been increasingly developed in the last decade or so, allowing researchers the ability to more effectively deal with this potential threat to validity. Besides the pattern-mixture approach presented in this article, "selection models" have also been proposed to handle missing data in longitudinal studies (Heckman, 1976; Heyting, Tolboom, & Essers,

1992; Leigh, Ward, & Fries, 1993; Diggle and Kenward, 1994). In selection models, whether or not a subject drops out is modeled in terms of variables obtained prior to the dropout, often the variables measured at baseline. Since dropout is dichotomous, a logistic or probit regression model is often used for this, yielding a predicted dropout probability or propensity for each subject. These dropout propensity scores can then be used in the longitudinal data model (*e.g.*, a RRM) as a covariate to adjust for the potential influence of dropout. So instead of using the missing data pattern as a model covariate in the longitudinal data analysis (*i.e.*, the pattern mixture approach), the selection model approach uses the predicted probability of dropout as a covariate. By modeling dropout, selection models provide information about the predictors of study dropout, however, an advantage of pattern-mixture models is that they can be used even when no such predictors are available. Currently, there is much debate in the statistical literature about these two approaches; further discussion on some of the differences between pattern-mixture and selection models can be found in Glynn, Laird, and Rubin (1986), Little (1993, 1994, 1995), and in the discussion of the Diggle and Kenward (1994) article.

In sum, it is important that researchers consider the reasons for missing data in their own datasets, and to choose among statistical methods with these reasons in mind. For example, if it can be assumed that missing observations on the dependent variable y are predicted well by the available y measurements for those individuals with missing data, then an ordinary RRM is reasonable (because the missingness is ignorable). Alternatively, if the missingness cannot be assumed to be ignorable, but there are good predictors of missingness, then a selection model approach may be used to obtain results that are adjusted for the predicted probability or propensity to be missing. As we have illustrated in this article, even when there are no good predictors of missingness, the pattern-mixture approach can be used to examine and adjust for the influence of missing data pattern. Of course, any method of "dealing" with missing data makes some assumptions about the missing data. For instance, in our example, ordinary RRM assumes that Figure 2 is representative

of the population trends in smoking across time for the two age cohorts, whereas pattern-mixture RRM assumes that Figures 1 and 3 are representative of these trends for complete and incomplete-data subjects, respectively. The bottom line is that researchers must choose a statistical model or models to represent both the observed and unobserved data.

Conclusion

As demonstrated, RRM provide a useful way of analyzing longitudinal substance use data. Specifically, RRM allow for the presence of missing data, irregularly-spaced measurements across time, time-varying and invariant covariates, accomodation of individual-specific deviations from the average time trend, and estimation of the population variance associated with these individual effects. Methods and software exist for analysis of continuous, dichotomous, and ordinal outcomes. Additionally, when missing data are present for the outcome variable, the pattern-mixture approach provides a way of assessing the influence of missing data-pattern on the outcome variable and on the effects in the model. Hopefully, this article has demonstrated that the combined approach of pattern-mixture RRM represents an important and useful tool for analyzing incomplete longitudinal data.

End Note: Computer Programs

Software for RRM is increasingly available, especially for normal-theory models (MLn, Rasbash, Yang, M., Woodhouse, G., & Goldstein, 1995; HLM, Bryk, Raudenbush, & Congdon, 1996; VARCL, Longford, 1986; the BMDP 5V procedure, Schluchter, 1988; the SAS procedure MIXED; MIXREG, Hedeker and Gibbons, 1996b). A detailed comparison of some of these programs is included in Kreft, de Leeuw, and van der Leeden (1994). For categorical outcomes, software is available for dichotomous (EGRET, Statistics and Epidemiology Research Corporation, 1991) and dichotomous/ordinal (MIXOR, Hedeker and Gibbons, 1996a) outcomes. Also, MLn, HLM, and VARCL provide facilities for analysis of dichotomous outcomes. MIXOR was used for the analyses presented in this article. To obtain the pattern-mixture averaged results based on the MIXOR estimates, a SAS IML program was written. Both MIXREG and MIXOR can be obtained free of charge through the internet (http://www.uic.edu/~hedeker/mix.html). Readers interested in the specifications that were required to run MIXOR for the examples in this article, or for the SAS IML program to produce the pattern-mixture averaged results, can send a note to the first author at hedeker@uic.edu.

References

Anderson, D. A. & Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society, 47*, 203-210.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research.* McGraw-Hill: New York.

Bock, R. D. (1983a). Within-subject experimentation in psychiatric research. In R. D. Gibbons & M. W. Dysken (Eds.), *Statistical and methodological advances in psychiatric research* (pp. 59-90). New York: Spectrum.

Bock, R. D. (1983b). The discrete Bayesian. In: H. Wainer & S. Messick (Eds.), *Modern advances in psychometric research* (pp. 103-115). Hillsdale, NJ: Earlbaum.

Bock, R. D. (1989). Measurement of human variation: A two stage model. In: R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 319-342). New York: Academic Press.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*, 147-158.

Bryk, A. S., & Raudenbush, S. W., (1992). Hierarchical linear models: Applications and data analysis methods. Newbury Park, CA: Sage Publications, Inc.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs.* Chicago: Scientific Software, Inc.

Chassin, L., Presson, C.C., Rose, J.S., & Sherman, S.J. (1996). The natural history of cigarette smoking from adolescence to adulthood: Demographic predictors of continuity and change. *Health Psychology, 15*, 478-484.

Chassin, L., Presson, C.C., Sherman, S.J., Corty, E., and Olshavsky, R. (1984). Predicting the onset of cigarette smoking in adolescents: a longitudinal study. *Journal of Applied Social Psychology, 14*, 224-243.

Chassin, L., Presson, C.C., Sherman, S.J., & Edwards, D.A. (1990). The natural history of cigarette smoking: predicting young adult smoking outcomes from adolescent smoking patterns. *Health Psychology, 9*, 701-716.

DeLeeuw, J, & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11*, 57-85.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance component models. *Journal of the American Statistical Society, 76*, 341-353.

Diggle, P., & Kenward, M. G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics, 43,* 49-94.

Diggle, P., Liang, K.-Y., & Zeger, S. L. (1994). *Analysis of Longitudinal Data*. New York: Oxford University Press.

EGRET (1991). *Reference manual.* Seattle, WA: Statistics and Epidemiology Research Corporation.

Ezzet, F., & Whitehead, J. (1991). A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine, 10,* 901-907.

Gibbons, R. D., & Bock, R. D. (1987). Trend in correlated proportions. *Psychometrika, 52*, 113-124.

Gibbons, R. D., & Hedeker, D. (1994). Application of random-effects probit regression models. *Journal of Consulting and Clinical Psychology, 62*, 285-296.

Gibbons, R. D., Hedeker, D., Elkin, I., Waternaux, C., Kraemer, H. C., Greenhouse, J. B., Shea, M. T., Imber, S. D., Sotsky, S. M., & Watkins, J. T. (1993). Some

conceptual and statistical issues in analysis of longitudinal psychiatric data. *Archives of General Psychiatry, 50*, 739-750.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In: H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 115-142). New York: Springer-Verlag.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika, 78*, 45-51.

Goldstein, H. (1995). *Multilevel statistical models, 2nd edition.* New York: Halsted Press.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement, 5*, 475-492.

Hedeker, D., & Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics, 50*, 933-944.

Hedeker, D., & Gibbons, R.D. (1996a). MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine, 49*, 157-176.

Hedeker, D., & Gibbons, R.D. (1996b). MIXREG: A computer program for mixed-effects regression analysis with autocorrelated errors. *Computer Methods and Programs in Biomedicine, 49*, 229-252.

Hedeker, D., & Gibbons, R.D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods, 2,* 64-78.

Hedeker, D., & Mermelstein, R.J. (1996). Application of random-effects regression models in relapse research. *Addiction, 91 (Supplement),* S211-S229.

Hogan, J.W., & Laird, N.M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine, 16,* 239-258.

Heyting, A., Tolboom, J.T.B.M., & Essers, J.G.A. (1992). Statistical handling of drop-outs in longitudinal clinical trials. *Statistics in Medicine, 11,* 2043-2061.

Jansen, J. (1990). On the statistical analysis of ordinal data when extravariation is present. *Applied Statistics, 39,* 75-84.

Jennrich, R. I. & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics, 42,* 805-820.

Jones, R.H. (1993). *Longitudinal data analysis with serial correlation: a state-space approach.* New York: Chapman and Hall.

Kleinbaum, D. G. (1994). *Logistic regression: A self-learning text.* New York: Springer-Verlag.

Kreft, I. G., de Leeuw, J. & van der Leeden, R. (1994). Comparing five different statistical packages for hierarchical linear regression: BMDP-5V, GENMOD, HLM, ML3, and VARCL. *American Statistician, 48,* 324-335.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*, 963-974.

Leigh, J.P., Ward, M.M., & Fries, J.F. (1993). Reducing attention bias with an instrumental variable in a regression model: Results from a panel of rheumatoid arthritis patients. *Statistics in Medicine, 12,* 1005-1018.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association, 88*, 125-133.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika, 81*, 471-483.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90*, 1112-1121.

Little, R. J. A., & Schenker, N. (1995). Missing Data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39-75). New York: Plenun Press.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage Publications.

Longford, N. T. (1986). VARCL - Interactive software for variance component analysis. *The Professsional Statistician, 74*, 817-827.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika, 74*, 817-827.

Longford, N. T. (1993). *Random coefficient models.* New York: Oxford University Press.

McArdle, J. J. & Hamagami, F. (1991). Modeling incomplete longitudinal data using latent growth structural equation models. In L. Collins & J.L. Horn (Eds.), *Best methods for analysis of change* (pp. 276-304). Washington, DC: APA Press.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*, 431-462.

Neuhaus, J. M., Kalbfleisch, J. D., & Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* **59**, 25-35.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., & Fisher, M. R. (1996). A survey of methods for analyzing clustered binary

response data. *International Statistical Review* **64**, 89-118.

Rasbash, J., Yang, M., Woodhouse, G., & Goldstein, H. (1995). *MLn: command reference guide.* London: Institute of Education, University of London.

Schluchter, M. D. (1988). 5V: Unbalanced repeated measures models with structured covariance matrices. In: W. J. Dixon (Chief Ed.), *BMDP statistical software manual* (vol. 2) (pp. 1081-1114). Berkeley, CA: University of California Press.

Stiratelli, R., Laird, N. M., & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics, 40*, 961-971.

Strenio, J. F., Weisberg, H. I., & Bryk, A. S. (1983). Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates. *Biometrics, 39,* 71-86.

Stroud, A. H. & Sechrest, D. *Gaussian Quadrature Formulas.* Englewood Cliffs, NJ: Prentice Hall, 1966.

Tebes, J. K., Snow, D. L., & Arthur, M. W. (1992). Panel attrition and external validity in the short-term follow-up study of adolescent substance use. *Evaluation Review, 16*, 151-170.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54*, 426-482.

Wong, G. Y., & Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association, 80*, 513-524.

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics, 44*, 1049-1060.

Author Notes

The authors thank Robin Mermelstein for many valuable discussions, comments, and suggestions prior to and during the preparation of this paper. The authors also thank Clark Presson and Michael Seltzer for helpful and constructive comments on an earlier version of this article. For use of the longitudinal data, the authors thank Steven J. Sherman, Laurie Chassin, and Clark Presson. Support for the longitudinal project was provided by National Institute of Child Health and Human Development Grant HD13449. Preparation of this article was supported by National Institutes of Mental Health Grant MH56146.

Correspondence concerning this article should be addressed to Donald Hedeker, Division of Epidemiology and Biostatistics (M/C 922), School of Public Health, University of Illinois at Chicago, 2121 West Taylor Street, Room 510, Chicago, IL, 60612-7260; or mailed electronically to hedeker@uic.edu.

Table 1

Examples of dummy-variable codings for missing-data patterns: a three timepoint study

	dummy-codes by patterns of missing data											
	general						monotone		last wave		incomplete	not at final
pattern	D1	D2	D3	D4	D5	D6	M1	M2	L1	L2	I1	F1
OOO	0	0	0	0	0	0	0	0	0	0	0	0
MOO	1	0	0	0	0	0			0	0	1	0
OMO	0	1	0	0	0	0			0	0	1	0
MMO	0	0	1	0	0	0			0	0	1	0
OOM	0	0	0	1	0	0	0	1	0	1	1	1
MOM	0	0	0	0	1	0			0	1	1	1
OMM	0	0	0	0	0	1	1	0	1	0	1	1

Table 2

Frequency distributions of age across the three study timepoints

		Young			
Adolescence		Adulthood		Adulthood	
Age	%	Age	%	Age	%
10	0.05	17	0.02	23	0.05
11	0.02	18	0.08	24	0.02
12	0.08	19	0.02	25	0.05
13	0.80	20	3.80	26	0.90
14	3.30	21	17.40	27	14.00
15	10.90	22	22.90	28	23.20
16	27.70	23	24.00	29	24.30
17	37.20	24	18.80	30	20.70
18	18.70	25	10.30	31	13.10
19	1.10	26	1.70	32	3.20
20	0.20	27	0.08	33	0.20
21	0.05	28	0.05	34	0.02

Table 3

Missing-data pattern frequencies and proportion smoking[1] by Cohort ($N = 3613$)

	Cohort: age in 1993					
	≤ 28		> 28		total	
pattern	n	$\hat{P}$	n	$\hat{P}$	n	$\hat{P}$
OOO	596	.210	1303	.194	1899	.199
MOO	329	.304	203	.397	532	.339
OMO	83	.361	223	.305	306	.320
MMO	83	.530	69	.449	152	.493
OOM	91	.253	179	.263	270	.259
MOM	66	.439	44	.477	110	.455
OMM	2	.500	342	.295	344	.297
total	1250	.256	2363	.234	3613	.242

[1] These proportions, denoted $\hat{P}$, are estimated across timepoints

Appendix

Marginal and subject-specific probabilities

As discussed in the article, subject-specific probabilities can be obtained using equation (9), namely,

$$P(Y_{ij} = 1 \mid \boldsymbol{x}_{ij}, \upsilon_i) = \frac{1}{1 + \exp\left[-(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \upsilon_i)\right]} \quad . \tag{16}$$

This equation yields the probability of a response for particular groupings of subjects determined by covariate values (*i.e.*, $\boldsymbol{x}_{ij}$) and specific values of the random subject effect υ_i. This is why the probabilities are termed subject-specific, they depend on values of the random subject effect. Alternatively, we might want to obtain probabilities not for a particular subject (*i.e.*, for a specific value of the random effect υ_i), but averaged across all subjects (*i.e.*, averaging across all values of the random effect υ_i) To obtain these marginal or population-averaged probabilities, the following integration over the random-effects distribution υ must be performed:

$$P(Y_{ij} = 1 \mid \boldsymbol{x}_{ij}) = \int_{\upsilon} \frac{1}{1 + \exp[-(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \upsilon_i)]} g(\upsilon) d\upsilon \quad . \tag{17}$$

Performing this integration may seem difficult, however, it can be approximated numerically to any practical degree of accuracy. Specifically, because the assumed distribution of υ is normal, Gauss-Hermite quadrature (Stroud & Sechrest, 1966) can be used to numerically perform the integration. Essentially, the numerical integration replaces the integral with a weighted summation over a finite number of points. Alternatively, another approximation described in Diggle, Liang, and Zeger (1994, page 142), does not require use of numerical integration. Here, the regression coefficients $\boldsymbol{\beta}$ from the random-effects model are divided by $\sqrt{c^2\sigma_\upsilon^2 + 1}$ where $c = 16\sqrt{3}/(15\pi)$. The inverse of the constant c reflects the fact that the standard logistic distribution has variance $\pi/\sqrt{3}$ and that multiplying this variance by 15/16 provides improved empirical fit of the data (see Long, 1997, pages 40-48, or Zeger, Liang, and Albert, 1988). These "marginalized" coefficients, denoted $\boldsymbol{\beta}^*$,

can then be directly used to produce marginal proportions by applying

$$P(Y_{ij} = 1 \mid \boldsymbol{x}_{ij}) = \frac{1}{1 + \exp[-(\boldsymbol{x}'_{ij}\boldsymbol{\beta}^*)]} \tag{18}$$

As an example, for the pattern-mixture analysis in Table 4, the marginalization factor is calculated as $\sqrt{(.346)(3.697)^2 + 1} = 2.394$ (note that $c^2 \approx .346$). Thus, the estimated proportion at adolescence for young completers is equal to

$$\frac{1}{1 + \exp\{-[(-3.343 + .513(-1) - .243(0))/2.394]\}} = \frac{1}{1 + \exp[-(-1.611)]} = .166 \ .$$

Alternatively, using quadrature yields .175 for this proportion. The actual observed proportion for this subgroup at adolescence is .173, so the quadrature estimate is a little closer to the observed data, though this is not always the case. The marginalization approach was used to produce the trend lines depicted in Figures 1-3. A similar set of figures (not shown) was produced using quadrature. These two sets of figures differed very little. In general, because the marginalization approach is simpler and provides very similar results to the more involved quadrature approach, its use is reasonable for examining approximate model fit.

Table 4

Smoking Status across Time ($N = 3613$)

Logistic RRM parameter estimates (est.), standard error (se), and p-values

	Completers (N=1899)			All Subjects (N=3613)			Pattern-Mixture (N=3613)		
	est.	se	$p <$	est.	se	$p <$	est.	se	$p <$
Intercept β_0	-3.406	.244	.001	-2.528	.150	.001	-3.343	.232	.001
T1 β_1	.519	.109	.001	.489	.094	.001	.513	.109	.001
T2 β_2	-.247	.118	.036	-.111	.081	.171	-.243	.118	.038
Cohort β_3	-.220	.258	.394	-.304	.172	.078	-.222	.252	.378
Cohort by T1 β_4	-.245	.137	.074	-.181	.117	.122	-.242	.137	.077
Cohort by T2 β_5	.172	.147	.243	.030	.110	.785	.169	.147	.250
Miss β_6							1.633	.292	.001
Miss by T1 β_7							-.355	.229	.121
Miss by T2 β_8							.286	.163	.080
Miss by Cohort β_9							.258	.353	.465
Miss by Cohort by T1 β_{10}							.685	.276	.013
Miss by Cohort by T2 β_{11}							-.369	.232	.112
Subject sd σ_υ	3.776	.184		3.667	.146		3.697	.159	
-2 log L	4451.4			7771.9			7648.2		

T1 and T2 represent two Helmert contrasts for time (coded as -1, .5, .5; 0, -1, 1).

Cohort coded as 0 for age ≤ 28 or 1 for age > 28.

Miss coded as 0 for complete-data subjects or 1 for incomplete-data subjects.

p-values not given for variance estimates (see Bryk and Raudenbush, 1992, page 55)

Figure 1. Proportion smoking across time by cohort: complete data

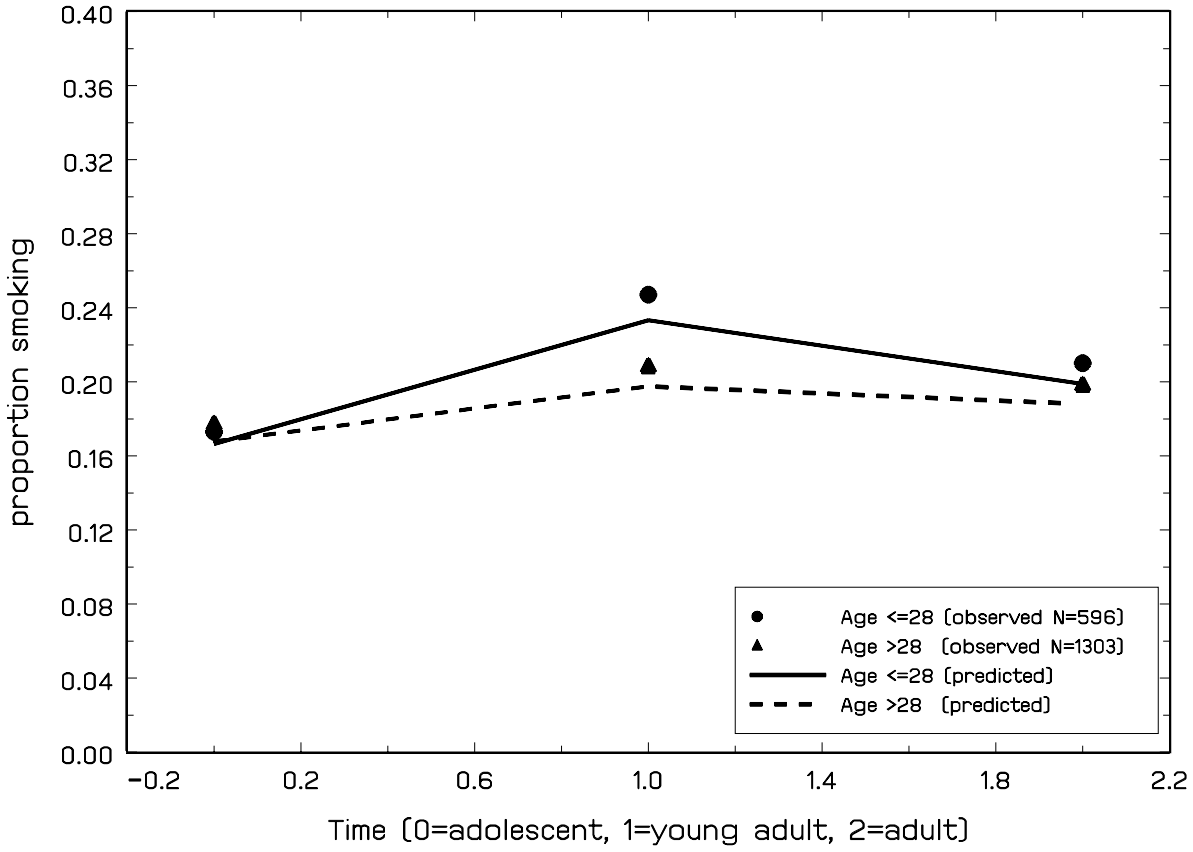

Figure 2. Proportion smoking across time by cohort: available data

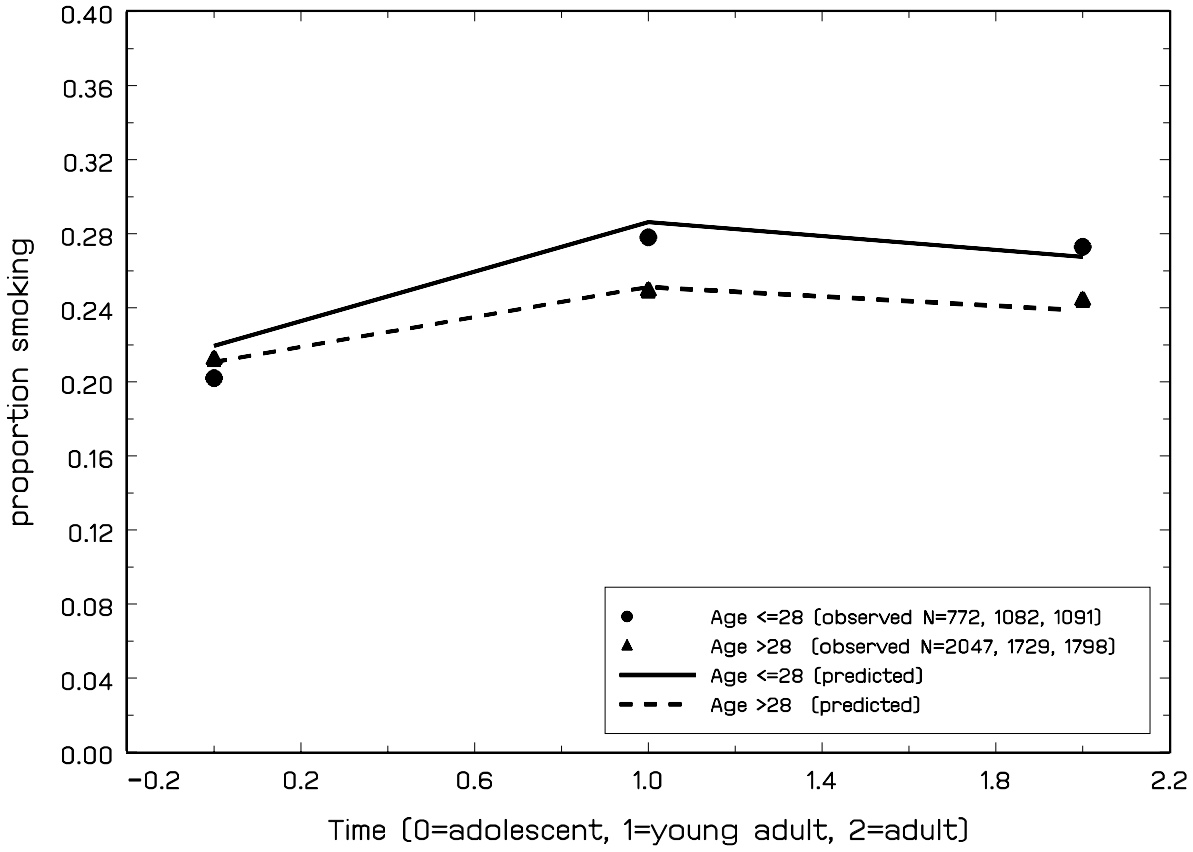

Figure 3. Proportion smoking across time by cohort: incomplete data