Published in final edited form as:

J Appl Stat. 2014 January 1; 41(5): 1040–1052. doi:10.1080/02664763.2013.859236.

Analysis of Ordinal Outcomes with Longitudinal Covariates Subject to Missingness

Melody S. Goodman^{a,*}, Yi Li^b, Anne M. Stoddard^c, and Glorian Sorensen^d

^aDivision of Public Health Sciences, Department of Surgery, Washington University in St. Louis School of Medicine, St. Louis, MO, U.S.A

^bDepartment of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, U.S.A

^cCenter for Statistical Analysis & Research, New England Research Institute, Watertown, MA, U.S.A

^dCenter for Community Based Research, Dana Farber Cancer Institute, Department of Society, Human Development, and Health, Harvard School of Public Health, Boston, MA, U.S.A

Abstract

We propose a mixture model for data with an ordinal outcome and a longitudinal covariate that is subject to missingness. Data from a tailored telephone delivered, smoking cessation intervention for construction laborers are used to illustrate the method, which considers as an outcome a categorical measure of smoking cessation, and evaluates the effectiveness of the motivational telephone interviews on this outcome. We propose two model structures for the longitudinal covariate, for the case when the missing data are missing at random, and when the missing data mechanism is non-ignorable. A generalized EM algorithm is used to obtain maximum likelihood estimates.

Keywords

ordinal outcomes; longitudinal covariates; missingness

1. Introduction

Epidemiological evidence suggest that there is a disparity in smoking status by socioeconomic position; the prevalence of smoking is higher in blue-collar workers than in their white-collar counterparts [17]. Although the prevalence of smoking is on the decline, the rate of decline is lower among blue-collar workers increasing disparities in smoking prevalence between white-collar and blue-collar workers. Blue-collar workers are less likely to be employed at work sites that offer health promotion programs, and are less likely to participate in such programs when they are available [17]. Due to increasing disparities in smoking prevalence among blue-collar workers, researchers are interested in developing

smoking cessation programs specifically for this population. Researchers at the Center for Community Based Research at the Dana Farber Cancer Institute collaborated with the Laborers' International Union of North America to evaluate the effectiveness of a tailored, telephone delivered, smoking cessation intervention for construction laborers. The construction industry poses challenges for worksite-based interventions, as workers move frequently from one worksite to another. To circumvent this issue, the intervention was primarily conducted by telephone [17].

Formative research of the sample was conducted through a background survey. For those participants that completed the background survey and were eligible and consented to be in the randomized trial, the study consisted of two additional surveys, pre-intervention (baseline) and post-intervention (final). The intervention group received a tailored intervention that consisted of motivational telephone interviewing and targeted written material sent via mail. Tailored interventions increase the relevance of health information, by personalizing the messages based on individual data, and have proven effective among blue-collar workers [17]. Despite the usefulness of tailored interventions, there has been limited work on the development of methods to analyze the unique data that is obtained from such approaches. We are interested in assessing the effectiveness of the motivational telephone interviewing in helping the construction laborers quit smoking, and limit our analysis to participants that were in the intervention group.

The time to smoking cessation was recorded as an ordinal variable. Ordered categorical responses are frequently encountered in survey data. However, these measures are often a proxy or a coarsely measured version of some unobservable continuous variable. Methods for the analysis of ordinal data are well established [1–3, 5, 6, 12]. The telephone intervention is measured as a longitudinal binary covariate, an indicator (yes/no) for each completed session. To examine the effect of the intervention on the time to smoking cessation, we propose a mixture model with an ordinal outcome and longitudinal covariate.

Models for the joint distribution of longitudinal and event-time variables have a wide variety of applications in clinical trials, prospective studies, and community-based interventions. Previous work on mixture models for longitudinal and time to event data have assumed the time to event is measured continuously (see, for example, [11, 18]). We extend the existing methodology, and develop models where the time to event is an ordered categorical variable. We propose four model structures and discuss the advantages and disadvantages of each.

This article is structured as follows, we define models where the missing data mechanism on the longitudinal covariate is ignorable in Section 2. In Section 3, we present models when the missing data mechanism on the longitudinal covariate is non-ignorable. The methods used for model estimation are described in Section 4. We illustrate the proposed methodology through an application to the data from the laborers' study in Section 5, and conclude with a general discussion in Section 6.

2. Repeated Measure is Missing At Random

We propose a likelihood-based mixture model approach for modeling the joint distribution of a vector Y of repeated binary measurements and Q an event time, whose joint distribution

can be expressed as a mixture, $f(Q, Y|Z) = f(Q|Y, Z) f(Y|Z)$, where Z is a collection of covariates. We first define the distribution of the repeated measure. Let Y_{im} be the m th measurement on participant i , and M_i the number of measurements made on participant i before event or censoring, the data for participant i can be written as $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{iM_i})$, and Z_i as an $M_i \times P$ design matrix of covariates. We propose a model with first-order Markov dependence for the vector of repeated outcomes (Y) [16]. A first-order Markov process is a random process whose future probabilities are determined by its most recent value, i.e., $P(Y_{im} = b | Y_{i(m-1)}, \dots, Y_{i1}) = P(Y_{im} = b | Y_{i(m-1)})$. Let

$$P(Y_{i1}=1|Z_i)=P_{i1}=\frac{\exp(\eta+\nu'Z_i)}{1+\exp(\eta+\nu'Z_i)}, \quad (1)$$

and

$$P(Y_{im}=1|Y_{i(m-1)}=a, Z_i)=P_{aim}=\frac{\exp(\varphi a+\eta+\nu'Z_i)}{1+\exp(\varphi a+\eta+\nu'Z_i)}, \quad (2)$$

where η is the effect of all things constant, ν is the effect of the covariates, and φ is the effect of the previous measurement, $P(Y_{i1}=0|Z_i) = 1 - P_{i1}$ and $P(Y_{im}=0|Y_{i(m-1)}=a, Z_i) = 1 - P_{aim}$.

The longitudinal covariate Y is subject to missingness, and for now, we assume that the data are missing at random. Define Y_i^{obs} and Y_i^{mis} as vectors of observed and missing observations for the i th participant, respectively, we can re-order Y_i where, $Y_i = (Y_i^{obs}, Y_i^{mis})$, and the first r_i observations of Y_i are observed and the remaining $M_i - r_i$ observations are missing. Under these assumptions, the likelihood for the unknown parameters $\theta = (\eta, \nu, \varphi)$, given Y^{obs} and Z , is obtained by integrating the missing data over the marginal distribution as shown in Equation 3.

$$\begin{aligned} L(\theta|Y^{obs}, Z) &= \int f(Y|Z) dY_{miss} \\ &= \sum_{Y_{i(r_i+1)}=0}^1 \sum_{Y_{i(r_i+2)}=0}^1 \dots \sum_{Y_{iM_i}=0}^1 P_{i1}^{Y_{i1}} (1-P_{i1})^{1-Y_{i1}} \\ &\quad \times \prod_{m=2}^{M_i} \prod_{a=0}^1 \prod_{Y_{im}=0}^1 P_{aim}^{Y_{im}} (1-P_{aim})^{1-Y_{im}} \end{aligned} \quad (3)$$

Equation 3 defines the distribution of the repeated measure Y ; we also must define the distribution for the time to event, Q , and the likelihood for the mixture model. We propose two ways to examine the effect of the repeated measure on the time to event. In our first approach, we model the effect of the last measurement before the event or censoring, and in our second approach we model the effect of the cumulative sum of the binary repeated measures on the time to event or censoring.

2.1 Model 1: last measurement

We are interested in the effect of the last measurement before event or censoring, as this provides insight into the immediate effect of the longitudinal covariate. Let Q_i be the time to event for participant i , where Q is an ordered categorical response. Y_{iM_i} is the outcome of the last measurement before event or censoring. Let the probability of having time to event in or before category k , given the last measurement before event or censoring Y_{iM_i} , and covariate vector Z_i be defined by

$$P(Q_i \leq k | Z_i, Y_i) = \frac{\exp(\alpha_k - \beta' Z_i - \gamma Y_{iM_i})}{1 + \exp(\alpha_k - \beta' Z_i - \gamma Y_{iM_i})}. \quad (4)$$

The likelihood function, conditional on the covariate vector Z , is defined by

$$L_c = \prod_{i=1}^n f(Q_i, Y_i | Z_i) = \prod_{i=1}^n f(Q_i | Y_i, Z_i) f(Y_i | Z_i),$$

where $f(Q_i | Y_i, Z_i) = \prod_{k=1}^K p_{ik}^{d_{ik}}$, $p_{ik} = P(Q_i \leq k | Z_i, Y_i) - P(Q_i \leq k-1 | Z_i, Y_i)$, $d_{ik} = I(Q_i = k)$, and $f(Y | Z)$ is defined in Equation (3). The resulting likelihood function is

$$\begin{aligned} & \prod_{i=1}^n \prod_{k=1}^K \left(\frac{\exp(\alpha_k - \beta' Z_i - \gamma Y_{iM_i})}{1 + \exp(\alpha_k - \beta' Z_i - \gamma Y_{iM_i})} - \frac{\exp(\alpha_{k-1} - \beta' Z_i - \gamma Y_{iM_i})}{1 + \exp(\alpha_{k-1} - \beta' Z_i - \gamma Y_{iM_i})} \right)^{d_{ik}} \\ & \times \left[\sum_{Y_{i(r_i+1)}=0}^1 \sum_{Y_{i(r_i+2)}=0}^1 \cdots \sum_{Y_{iM_i}=0}^1 P_{i1}^{Y_{i1}} (1 - P_{i1})^{1-Y_{i1}} \prod_{m=2}^{M_i} \prod_{a=0}^1 \prod_{Y_{im}=0}^1 P_{aim}^{Y_{im}} (1 - P_{aim})^{1-Y_{im}} \right]. \end{aligned}$$

2.2 Model 2: Cumulative sum of the binary repeated measures

Here we model the effect of the cumulative sum of the repeated binary measures on the time to event or censoring. Let $\tilde{Y}_i = \sum_{m=1}^{M_i} Y_{im}$. The joint distribution of Q and Y can be expressed as a mixture, where $f(Q, Y | Z) = f(Q | \tilde{Y}, Z) f(Y | Z)$. Let the probability of having a time to event in or before category k , given the cumulative sum of repeated measures $\tilde{Y}_i$, and covariate vector Z_i be defined by

$$P(Q_i \leq k | Z_i, \tilde{Y}_i) = \frac{\exp(\alpha_k - \beta' Z_i - \gamma \tilde{Y}_i)}{1 + \exp(\alpha_k - \beta' Z_i - \gamma \tilde{Y}_i)}, \quad (5)$$

and $f(Q_i | Z_i, \tilde{Y}_i) = \prod_{k=1}^K \{P(Q_i \leq k | Z_i, \tilde{Y}_i) - P(Q_i \leq k-1 | Z_i, \tilde{Y}_i)\}^{d_{ik}}$.

The likelihood function is

$$\times \left[\sum_{Y_{i(r_i+1)}=0}^1 \sum_{Y_{i(r_i+2)}=0}^1 \cdots \sum_{Y_{iM_i}=0}^1 P_{i1}^{Y_{i1}} (1-P_{i1})^{1-Y_{i1}} \prod_{m=2}^{M_i} \prod_{a=0}^1 \prod_{Y_{im}=0}^1 P_{aim}^{Y_{im}} (1-P_{aim})^{1-Y_{im}} \right].$$

β and γ are the parameters of primary interest, as they assess the effect of the covariate Z and the repeated measure Y , respectively. They are interpreted in the same manner as the parameters of a proportional odds model with a logit link, and are the log of the cumulative odds ratio, with the interpretation that the odds of being in a category k are $\exp(\beta)$ and $\exp(\gamma)$ respectively.

3. Non-ignorable Missingness

When the missingness depends on the values of the missing data, the missing data mechanism is non-ignorable. Correct likelihood analysis must be based on the full likelihood of the joint distribution of the repeated measure $Y = (Y_{i1}, \dots, Y_{iM_i})$ and the missing data mechanism $X = (X_{i1}, \dots, X_{iM_i})$ [13], where

$$X_{im} = \begin{cases} 0 & Y_{im} \text{ missing} \\ 1 & Y_{im} \text{ observed.} \end{cases}$$

We propose a model with first-order Markov dependence for the intended binary repeated measure (Y), and first-order Markov dependence for the missing data mechanism (X) that depends on the current binary measurement (Y_{im}) [4]. Under these assumptions,

$$P(Y_i, X_i) = P(Y_{i1}) \prod_{m=2}^{M_i} P(Y_{im} | Y_{i(m-1)}) \times P(X_{i1} | Y_{i1}) \prod_{m=2}^{M_i} P(X_{im} | X_{i(m-1)}, Y_{im}).$$

Let P_{i1} be defined as in Equation 1, P_{aim} be defined as in Equation 2,

$$P(X_{i1}=1 | Y_{i1}) = R_{i1} = \frac{\exp(\kappa + \rho Y_{i1})}{1 + \exp(\kappa + \rho Y_{i1})},$$

and let

$$P(X_{im}=1 | X_{i(m-1)}=a, Y_{im}) = R_{aim} = \frac{\exp(\kappa + \rho Y_{im} + \psi a)}{1 + \exp(\kappa + \rho Y_{im} + \psi a)},$$

and thus, $P(Y_{i1}=0 | Z) = 1 - P_{i1}$, $P(Y_{im}=0 | Y_{i(m-1)}=a, Z) = 1 - P_{aim}$, $P(X_{i1}=0 | Y_{i1}) = 1 - R_{i1}$, and $P(X_{im}=0 | X_{i(m-1)}=a, Y_{im}) = 1 - R_{aim}$. Here, κ is the constant effect, ρ is the effect

of the current longitudinal measurement, and ψ is the effect of the previous missing indicator on the probability that the current repeated measure is observed.

Accordingly, we integrate out the missing repeated measures from the marginal distribution,

$$\int f(Y, X|Z) dY_{miss} = \sum_{Y_{i(r_i+1)}=0}^1 \sum_{Y_{i(r_i+2)}=0}^1 \cdots \sum_{Y_{iM_i}=0}^1 \left[P_{i1}^{Y_{i1}} (1-P_{i1})^{1-Y_{i1}} R_{i1}^{X_{i1}} (1-R_{i1})^{1-X_{i1}} \right. \\ \left. \times \prod_{m=2}^{M_i} \prod_{a=0}^1 \prod_{Y_{im}=0}^1 P_{aim}^{Y_{im}} (1-P_{aim})^{1-Y_{im}} \prod_{m=2}^{M_i} \prod_{a=0}^1 \prod_{X_{im}=0}^1 R_{aim}^{X_{im}} (1-R_{aim})^{1-X_{im}} \right]$$

and develop the likelihood function for the mixture model,

$$\prod_{i=1}^n \prod_{k=1}^K P(Q_i=k|Z_i, Y_i)^{d_{ik}} \left[\sum_{Y_{i(r_i+1)}=0}^1 \sum_{Y_{i(r_i+2)}=0}^1 \cdots \sum_{Y_{iM_i}=0}^1 P_{i1}^{Y_{i1}} (1-P_{i1})^{1-Y_{i1}} \right. \\ \left. \prod_{m=2}^{M_i} \prod_{a=0}^1 \prod_{Y_{im}=0}^1 P_{aim}^{Y_{im}} (1-P_{aim})^{1-Y_{im}} \times R_{i1}^{X_{i1}} (1-R_{i1})^{1-X_{i1}} \prod_{m=2}^{M_i} \prod_{a=0}^1 \prod_{X_{im}=0}^1 R_{aim}^{X_{im}} (1-R_{aim})^{1-X_{im}} \right],$$

where $P(Q_i=k|Z_i, Y_i) = P(Q_i=k|Z_i, Y_i) - P(Q_i=k-1|Z_i, Y_i)$, and $P(Q_i=k|Z_i, Y_i)$ is defined as in (4) or (5), depending on the approach used to examine the effect of the repeated measures on the time to event.

4. Model Estimation

The repeated measure Y is subject to missingness and thus incomplete. Full use of the data is made by obtaining maximum likelihood estimates through an application of the generalized EM algorithm [8], an iterative procedure for finding maximum likelihood estimates from incomplete data. At each iteration, the algorithm updates the parameter estimates by maximizing the expected value of the complete data log-likelihood, given the observed data and the current parameter estimates.

When the repeated measures are missing at random, the complete data log-likelihood can be expressed as

$$\ell(\theta) = \sum_{i=1}^n \left\{ \sum_{k=1}^K d_{ik} \log(P(Q_i=k|Z_i, Y_i)) + Y_{i1} \log(P_{i1}) + (1-Y_{i1}) \log(1-P_{i1}) + \sum_{m=2}^{M_i} Y_{im} \log(P_{aim}) + (1-Y_{im}) \log(1-P_{aim}) \right\}, \quad (6)$$

where θ is a vector of the model parameters.

When the missingness mechanism is non-ignorable, the complete data log-likelihood is

$$\begin{aligned} \ell(\theta) = & \sum_{i=1}^n \left\{ \sum_{k=1}^K d_{ik} \log(P(Q_i=k|Z_i, Y_i)) + Y_{i1} \log(P_{i1}) + (1-Y_{i1}) \log(1-P_{i1}) \right. \\ & + \sum_{m=2}^{M_i} Y_{im} \log(P_{aim}) + (1-Y_{im}) \log(1-P_{aim}) + X_{i1} \log(R_{i1}) + (1-X_{i1}) \log(1-R_{i1}) \\ & \left. + \sum_{m=2}^{M_i} X_{im} \log(R_{aim}) + (1+X_{im}) \log(1-R_{aim}) \right\}, \end{aligned} \quad (7)$$

where again θ represents the model parameters.

The objective function to be maximized at each iteration of the EM algorithm is the expected value of either (6) or (7), given the observed data and the current update of the parameter estimates $\theta^{(g)}$. To calculate the expected value of the complete data log-likelihood, we need the expected value of Y_{im} given the observed data and the current update of the parameter estimates, $E[Y_{im}|Y_i^{obs}, Z_i, \hat{\theta}^{(g)}]$. For those participants with Y_{im} observed, the conditional expectation $E[Y_{im}|Y_i^{obs}, Z_i, \hat{\theta}^{(g)}]$ is simply Y_{im} (equal to either zero or one). For those with incomplete Y_i , $E[Y_{i1}|Y_i^{obs}, Z_i, \hat{\theta}^{(g)}] = P(Y_{i1}=1|Y_i^{obs}, Z_i, \hat{\theta}^{(g)}) = P_{i1}$ and

$$E[Y_{im}|Y_i^{obs}, Z_i, \hat{\theta}^{(g)}] = P(Y_{im}=1|Y_i^{obs}, Z_i, \hat{\theta}^{(g)}) = \sum_{b_m=0}^1 P_{i1}^{b_1} (1-P_{i1})^{1-b_1} \times \prod_{m=2}^{m-1} P_{abm}^{b_m} (1-P_{abm})^{1-b_m},$$

where b_m is an indicator equal to one if $Y_{im} = 1$ and zero otherwise. Variance estimates of the maximum likelihood parameter estimates were obtained from the expected value of the negative Hessian matrix evaluated at the final parameter estimates. For ignorable cases EM algorithm can be helpful. However, in models with non-ignorable missingness the EM algorithm may take longer to converge to a maximum due to large amount of missing information. In addition, it is necessary to check for the multiple maxima of the likelihood function after convergence to maximum [14].

5. Application: Laborers' Study

We use data from the Laborers' Study to illustrate our proposed methods. As this intervention was conducted primarily by telephone, researchers are interested in the effect of the health educator calls on smoking cessation. We limit our analysis to those participants in the intervention group that completed both baseline and final surveys and were current smokers at baseline. 100 participants met this inclusion criteria, 81 of whom had not quit smoking by the end of the study, approximately six months after the intervention began, and are censored at the time of their final survey. The remaining 19 had quit smoking without relapse for at least seven days. Of these, five quit smoking at baseline, eleven quit less than three months into the intervention, and the remaining three between 4 and 6 months after the beginning of the intervention.

The time to quit category, Q , was determined using the date of the baseline survey, the date of the final survey, and two final survey questions. Participants were first asked, "During the past seven days, have you smoked any cigarettes, even a puff?" If they answered no,

they were then asked, “*How long has it been since you had your last cigarette?*” They were given six options; less than one month, 1–2 months, 3–4 months, 5–6 months, 7 months or more but less than one year, or one year or more. We collapsed these groups to create the following categories:

$$Q_i = \begin{cases} 0 & \text{quit at baseline} \\ 1 & \text{quit 0–3 months into intervention} \\ 2 & \text{quit 4–6 months after the intervention began} \\ 3 & \text{never quit (censored),} \end{cases}$$

where Q_i is the time to quit category for the i^{th} participant.

Approximately two weeks after the baseline survey, participants received a tailored feedback report that introduced the program and provided personalized health messages incorporating their responses to the survey. Shortly thereafter, telephone counseling calls were initiated by health educators, which were also tailored to the participants, and relied on motivational interviewing techniques [17]. The outcome of the health educator calls is the binary repeated measure Y , which is subject to missingness (e.g., no answer, answering machine) where,

$$Y_{im} = \begin{cases} 0 & \text{refused/postponed/rescheduled session} \\ 1 & \text{completed session.} \end{cases}$$

The parameters for the non-quitters, those participants that were censored at the final measurement, were set to zero, as this is the reference group for the categorical time to event measurement, Q . Table 1 displays the demographic breakdown of the data, for the variables of race, income, and education which were examined in our models. The population was predominately male (92%) so we were unable to assess the effect of gender. For the purposes of this analysis, non-binary variables were dichotomized in order to ensure sufficient numbers in each group. Indicator variables were created for income and education, with income categories above and below \$50,000 annually, and education categories of beyond a high school diploma or not. As there are very few participants in each of the quit categories, we consider separate models for each of these three covariates.

In the first two models, we assume that the missing data mechanism for the longitudinal covariate is ignorable, the data are missing at random, and modeled as a first order Markov process. In Model 1, we examine the effect of the last call before event or censoring, where the call is a binary covariate subject to ignorable missingness. In Model 2, we model the effect of the cumulative number of calls on the time to smoking cessation. The covariate for the calls is modeled as the sum of binary random variables subject to ignorable missingness. Models 3 and 4 use the assumption that the missing data mechanism is non-ignorable, and model the intended outcome of the call with first-order Markov dependence. In these models, we assume that the missing data mechanism has first-order Markov dependence, and is also dependent on the call. In Model 3 we assess the effect of the last call before event

or censoring, and in Model 4 the effect of the cumulative sum of calls on time to smoking cessation.

5.1 Model 1: Last call assuming ignorable missingness

In Model 1, summarized in Table 2, we examine the effect of the outcome of the last call before event or censoring assuming the missing data mechanism is ignorable, adjusting separately for race (Model 1a), income (Model 1b), and education (Model 1c). Eleven (15%) white participants quit smoking, versus 9 (30%) non-white participants. We estimate the cumulative log odds ratio (OR) to be 0.948 (OR = 2.582, SE=0.663), non-whites are more than twice as likely as whites to quit in any time category. Participants with successful last calls are more likely to quit smoking in earlier time categories, with the cumulative odds ratio of 1.446 (SE=0.519). Participants earning more than \$50,000 annually are more likely to quit smoking in later time categories or not at all (OR=0.524, SE=0.868), and additionally, those that have more than H.S. education are less likely to quit in each time category (OR=0.571, SE=0.662). In Models 1a, 1b, and 1c, participants in each of the time intervals with a successful last call from a health educator are more likely to quit smoking, OR=1.44, 1.48, and 1.36, respectively. However, the 95% confidence intervals include the null value of one in all models.

5.2 Model 2: Cumulative sum of intended calls assuming ignorable missingness

In Model 2, we examine the effect of the cumulative sum of the outcomes of health educator calls on smoking cessation, assuming that missingness is ignorable. Table 3 displays the results of our analysis for Model 2. Non-whites are more likely to quit in each time category (Model 2a), participants making more than \$50,000 annually are less likely to quit in each time category (Model 2b), and participants having more than a high school diploma are less likely to quit smoking in each time interval (Model 2c). These trends are consistent with those found in Model 1. The more calls a participant received decreases the odds of quitting (Models 2a–2c). However, the 95% confidence intervals for all cases include the null value. It is of interest to note that if people with a successful last call are more likely to quit in earlier time periods (Model 1), then one would assume that the more successful calls a participant has with a health educator, and a larger cumulative sum, the more likely they would be to quit in each time interval. We observe the opposite trend in Model 2, and believe one explanation for this is that the longer a participant takes to quit, the more calls they are likely to receive. Therefore by the nature of the study, participants with larger cumulative sums are those that quit in later time intervals or not at all.

5.3 Model 3: Last call assuming non-ignorable missingness

In Model 3, we examine the effect of the last call before event or censoring, assuming the missing data mechanism is non-ignorable. Table 4 displays the results of our analysis for Model 3. For any quit category, non-whites are more likely to quit smoking in that category or those prior than whites, OR=2.057 (SE=0.834), participants earning more than \$50,000 annually are less likely to quit smoking in each time category, OR=0.59 (SE=1.391), and participants with more than a high school diploma are less likely to quit smoking in each time interval, OR=0.386 (SE=0.808). Adjusting for race, income, or education, participants

with a successful last call before event or censoring are more likely to quit smoking in each time category, $OR=1.433, 1.261, \text{ and } 1.262$, respectively.

5.4 Model 4: Cumulative sum of intended calls assuming non-ignorable missingness

In Model 4, we examine the effects of the cumulative sum of calls from a health educator on smoking cessation assuming the missingness mechanism is non-ignorable. Table 5 displays the results of our analysis for Model 4. Non-whites are more likely to quit smoking in each time category (Model 4a), participants earning more than \$50,000 annually were less likely to quit smoking in each time category (Model 4b), and those having more than a high school diploma were less likely to quit smoking in each time category (Model 4c). These trends are consistent with those seen in the previous three models. Participants receiving a greater number of calls from a health educator are less likely to quit smoking in each time interval. However, as with Model 2, we think this is trend is a function of the data, and not an indicator of the effectiveness of the intervention.

5.5 Model Selection: A practical approach

In our analysis, we were interested in the effectiveness of the health educator's calls on smoking cessation, and thus only considered calls before event or censoring and not calls made after the time of an event. Examining the effect of the intervention on the sustainability of cessation or relapse are interesting questions, but not ones we considered in our analysis. To examine the effectiveness of the intervention, one needs to consider the effect of the calls on time to quit. We observe the counter intuitive result that the larger the cumulative sum of calls, the less likely a participant is to quit, but this is a structural problem in the data, as participants who take longer to quit, and those that do not quit during the course of the study, will receive more calls. Therefore, we do not believe Models 2 or 4 properly address our question of interest.

We also examined the effect of the number of successful calls on time to quit category by fitting a proportional odds model. We found that non-whites were more likely to quit, participants earning more than \$50,000 annually were less likely to quit, and those with more than a high school diploma were less likely to quit. These are the same trends we observed using our proposed methodology. The number of successful calls seems to have very little effect, if any, as we observe odds ratios close to one, $OR=0.91, 0.92, \text{ and } 0.93$, respectively. These models use only the complete data however, and do not take into account the missing data.

The choice between Models 1 and 3 is based on whether the missingness mechanism on the longitudinal covariate is ignorable or non-ignorable. The mechanism is ignorable if the missingness is due to the random call process; participants do not know when to expect a call from their health educator, and may not be home to receive the call. However, there is a possibility that participants may not be interested in the intervention, and purposely do not accept the calls from the health educator, making missingness non-ignorable with respect to the outcome of the intervention. We used Akaike's Information Criteria (AIC) which selects Model 1 in favor of Model 3 in all cases.

Although Model 1 is the simplest of the models that we propose, we feel this model is the most appropriate in determining the effectiveness of this tailored, telephone delivered intervention, as it models the immediate effect of the health educator's call on the participant quitting. We also believe that in this intervention, the missing data mechanism is ignorable as health educators would make several attempts to reach the participants. During the course of the intervention, a participant could have up to six sessions with a health educator, and approximately 60% of our sample completed at least four sessions. In all cases, the health educator calls appear effective, as the odds of quitting are higher for those participants with successful calls, suggesting that this method of intervention is successful in this population.

6. Discussion

It is important to note the limitations of our study. By defining quitters as those persons that have not smoked in seven or more days, the cessation measure does not capture smokers who quit and relapsed prior to the final survey. The outcome relied on self-report, and is therefore subject to reporting bias. Although the results of our analysis could have large implications if generalized to the entire union membership, we recognize that these findings are based on a small number of quitters and may not be representative of this population. We could have used other modeling approaches to analyze these data such as an interval censored Cox proportional hazards model with a time varying covariate [9, 10]. The approach we used here is an extension of models with a continuous outcome [11, 18]; comparison of the approach used in this analysis to approaches where the outcome is modeled continuously is an important area for future research.

The small number of quitters also has effects on the estimation of parameters. When the data are sparse, the estimates of the parameters may be unstable or models may not converge. However, even with data where the number of parameters to be estimated is large in comparison to the number of observations, the usual asymptotic results are quite reliable for the parameters of interest [15]. The small sample size contributes to the large standard errors, large confidence intervals for the estimates, and non-significant results. Although most of the confidence intervals contain the null value, based on our analysis, the health educator calls appear to be effective in helping participants quit smoking.

Of the union members randomly assigned to receive the tailored, telephone delivered intervention, those participants that completed sessions with the health educator were more likely to quit smoking. We recognize the challenge in smoking cessation and are aware that people who quit are often subject to relapse. A larger sample of quitters would allow us sufficient power to examine the effectiveness of the intervention.

When working with survey data from social science research, there are many variables that can not be measured precisely, such as time of smoking cessation, and others that may be subject to missingness, such as calls from a health educator. When information on non-respondents is available ignorable missing data models are frequently used for survey data analysis and often out perform non-ignorable missing data models. We do not posit that missingness mechanism operating in survey research are truly ignorable but that the formulation of non-ignorable models that out perform ignorable models is a difficult task

that is context-specific [14]. Estimates from EM algorithm provide advantages over the traditional approaches. However, estimates of EM algorithm are of unknown precision when the missingness mechanism is non-ignorable [7]. We believe this model structure has multiple applications beyond evaluating a tailored smoking cessation intervention for construction laborers.

Acknowledgments

The research of Dr. Goodman was supported by National Institute of Child Health and Human Development grant 5 F31 HD043695, National Cancer Institute grant U54CA153460, the Barnes-Jewish Hospital Foundation, Siteman Cancer Center, and Washington University School of Medicine. The research of Dr. Li was supported by National Institute of Health grant R01CA95747.

References

1. Agresti, A. Analysis of Ordinal Categorical Data. John Wiley & Sons; 1984.
2. Agresti, A. Categorical Data Analysis. John Wiley & Sons; 1990.
3. Agresti A. Modeling ordered categorical data: Recent advances and future challenges. *Statistics in Medicine*. 1999; 18:2191–2207. [PubMed: 10474133]
4. Albert PS. A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics*. 2000; 56:602–608. [PubMed: 10877323]
5. Anderson J. Regression and ordered categorical variables. *J R Statist Soc B*. 1984; 46:1–30.
6. Anderson J, Philips P. Regression, discrimination and measurement models for ordered categorical variables. *Applied Statistics*. 1981; 30:22–31.
7. Croninger, RG.; Douglas, KM. Survey Research, Emerging Issues, New directions for institutional Research # 127. In: PD, editor. Missing data and institutional research. Vol. chap 3. San Francisco: Jossey-Bass; 2005. p. 33-50.
8. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*. 1977; 39:1–22.
9. Finkelstein DM. A proportional hazards model for interval-censored failure time data. *Biometrics*. 1986:845–854. [PubMed: 3814726]
10. Fisher LD, Lin D. Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health*. 1999; 20:145–157.
11. Hogan JW, Laird NM. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*. 1997; 16:239–257. [PubMed: 9004395]
12. Lipsitz SR. Methods for estimating the parameters of a linear model for ordered data. *Biometrics*. 1992; 48:271–281. [PubMed: 1581487]
13. Little, RJ. Encyclopaedia of Biostatistics. In: Armitage, P.; Colton, T., editors. Missing Data. Vol. 4. John Wiley & Sons; 1998. p. 2622-2635.
14. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2. New York: John Wiley & Sons; 2002.
15. McCullagh, P.; Nelder, J. Generalized Linear Models. Chapman & Hall/CRC; 1989.
16. Minard CG, Chan W, Wetter DW, Etzel CJ. Trends in smoking cessation: A markov model approach. *Journal of Applied Statistics*. 2012; 39:113–127.
17. Sorensen, G.; Barbeau, E.; Stoddard, AM.; Huntkey, M.; Goldman, R.; Smith, A.; Brennan, AA.; Wallace, L. Tools for health: The efficacy of a tailored intervention targeted for construction laborers. 2006.
18. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*. 2004; 14:809–834.

Table 1

Demographics

	Frequency	Percent
Race		
White	74	73.3
other	27	26.7
Gender		
Male	91	91.9
Female	8	8.0
Education		
Did not complete H.S.	23	22.8
H.S. diploma	45	44.6
Some post H.S. training	28	27.7
Baccalaureate or more	5	4.9
Income		
\$15K	4	4.0
\$15K–\$50K	54	53.5
\$50K	43	42.6

Table 2
Model 1: Last Call Assuming Ignorable Missing, Adjusting for Race, Income, and Education

Variable	Quit	Estimate	Standard Error	OR	Lower	Upper
<i>Model 1a: Race</i>						
Intercept	Baseline	-4.839	1.132	0.008	0.001	0.073
	0-3 months	-3.983	1.559	0.019	0.001	0.448
	4-6 months	-3.859	1.527	0.021	0.001	0.371
Non-white		0.948	0.663	2.582	0.704	9.465
Last call		0.369	0.519	1.446	0.523	3.997
<i>Model 1b: Income</i>						
Intercept	Baseline	-4.420	1.584	0.012	0.001	0.268
	0-3 months	-3.663	1.476	0.026	0.001	0.463
	4-6 months	-3.356	1.399	0.035	0.002	0.542
>\$50K		-0.646	0.868	0.524	0.096	2.871
Last call		0.390	0.466	1.477	0.593	3.679
<i>Model 1c: Education</i>						
Intercept	Baseline	-4.514	1.308	0.011	0.001	0.142
	0-3 months	-3.532	1.325	0.029	0.002	0.392
	4-6 months	-3.241	1.528	0.039	0.002	0.782
> than H.S.		-0.561	0.662	0.571	0.156	2.088
Last call		0.309	0.655	1.362	0.377	4.918

Table 3
Model 2: Cumulative Sum of Intended Calls Assuming Ignorable Missingness, Adjusting for Race, Income, and Education

Variable	Quit	Estimate	Standard Error	OR	95% CI for OR	
					Lower	Upper
<i>Model 2a: Race</i>						
Intercept	Baseline	-4.397	1.065	0.012	0.002	0.099
	0-3 months	-4.297	1.117	0.014	0.005	0.397
	4-6 months	-3.115	0.932	0.044	0.002	0.085
Non-white		0.221	0.823	1.248	0.249	6.262
Cumulative sum		-0.339	0.639	0.713	0.204	2.493
<i>Model 2b: Income</i>						
Intercept	Baseline	-4.650	1.313	0.010	0.001	0.125
	0-3 months	-3.669	1.153	0.026	0.003	0.245
	4-6 months	-3.252	0.820	0.039	0.008	0.193
>\$50K		-0.231	1.180	0.794	0.079	8.018
Cumulative sum		-0.376	1.072	0.687	0.084	5.619
<i>Model 2c: Education</i>						
Intercept	Baseline	-4.274	1.582	0.014	0.001	0.309
	0-3 months	-3.960	1.681	0.019	0.001	0.514
	4-6 months	-3.599	0.964	0.027	0.004	0.181
> than H.S.		-0.963	1.027	0.382	0.051	2.860
Cumulative sum		-0.383	1.456	0.682	0.039	11.832

Table 4
Model 3: Last Call Assuming Non-ignorable Missingness, Adjusting for Race, Income, and Education

Variable	Quit	Estimate	Standard Error	OR	95% CI for OR	
<i>Model 3a: Race</i>						
Intercept	Baseline	-4.036	1.434	0.018	0.001	0.294
	0-3 months	-3.721	1.235	0.024	0.003	0.320
	4-6 months	-3.560	1.248	0.028	0.002	0.279
Non-white						
		0.721	0.834	2.057	0.401	10.539
Last call		0.360	1.267	1.433	0.120	17.183
<i>Model 3b: Income</i>						
Intercept	Baseline	-4.561	1.312	0.010	0.001	0.137
	0-3 months	-3.210	0.903	0.040	0.007	0.237
	4-6 months	-3.133	1.154	0.044	0.005	0.418
>\$50K						
		-0.528	1.391	0.590	0.039	9.018
Last call		0.232	1.307	1.261	0.097	16.340
<i>Model 3c: Education</i>						
Intercept	Baseline	-4.853	1.438	0.008	0.000	0.131
	0-3 months	-3.871	1.060	0.021	0.003	0.166
	4-6 months	-3.623	1.735	0.027	0.001	0.801
> than H.S.						
		-0.952	0.808	0.386	0.079	1.882
Last call		0.232	0.746	1.262	0.292	5.442

Table 5
Model 4: Cumulative Sum of Intended Calls Assuming Non-ignorable Missingness, Adjusting for Race, Income, and Education

Variable	Quit	Estimate	Standard Error	OR	95% CI for OR	
					Lower	Upper
<i>Model 4a: Race</i>						
Intercept	Baseline	-5.440	0.968	0.004	0.001	0.029
	0-3 months	-4.596	1.143	0.010	0.002	0.171
	4-6 months	-4.004	0.781	0.018	0.002	0.047
Non-white		0.211	1.153	1.235	0.129	11.831
Cumulative sum		-0.781	0.952	0.458	0.071	2.956
<i>Model 4b: Income</i>						
Intercept	Baseline	-4.643	1.234	0.010	0.001	0.108
	0-3 months	-4.340	1.280	0.013	0.001	0.160
	4-6 months	-4.200	1.694	0.015	0.001	0.415
>\$50K		-0.249	1.457	0.780	0.045	13.552
Cumulative sum		-0.707	1.364	0.493	0.034	7.152
<i>Model 4c: Education</i>						
Intercept	Baseline	-4.182	1.112	0.015	0.002	0.135
	0-3 months	-3.364	1.227	0.035	0.003	0.383
	4-6 months	-3.202	0.960	0.041	0.006	0.267
> than H.S.		-0.550	1.005	0.577	0.081	4.136
Cumulative sum		-1.167	0.603	0.311	0.096	1.014