

# STATISTICAL MODELING OF MENTAL DISTRESS AMONG THE NATIONAL POPULATION HEALTH SURVEY PARTICIPANTS ASSESSED LONGITUDINALLY: MISSING DATA ANALYSIS

Chandima Karunanayake<sup>1</sup>, Punam Pahwa<sup>2</sup>, Helen McDuffie<sup>3</sup>

## ABSTRACT

**BACKGROUND:** National Population Health Survey (NPHS) collects longitudinal data on the physical and mental health of Canadians. The main objective of this paper was to investigate how missing data pattern influences the results. **METHODS:** A subpopulation of Statistics Canada's longitudinal National Population Health Survey (NPHS) dataset from the first five cycles (1994/95 – 2002/03) was used to investigate the effects of missing data on the longitudinal changes of mental distress scores among the NPHS participants who reported non-malignant respiratory diseases those who are age 15 years and older. From the relatively wide range of mental health indicators available in the NPHS, the distress measure based on a subset of items of the Composite International Diagnostic Interview with six questions was chosen as outcome. The analysis was conducted using generalized estimating equation approach accounting for the complexity of multi-stage survey design using bootstrap weights available for incomplete longitudinal data. **RESULTS:** There were significant interactions between missing data pattern (completers/in-completers) and geographic area, marital status and self-reported general health index. **CONCLUSIONS:** There was a significant effect from the dropouts to the outcome distress scale and "missingness" interacts with the covariates; geographic area, marital status and self-reported general health index. Therefore it was clear that without considering missingness, loss of information occurs and our conclusions may change.

**KEY WORDS:** Longitudinal surveys, ordinal outcome, Generalized Estimating equations, Bootstrap Estimation, missing data

## RÉSUMÉ

**CONTEXTE :** L'Enquête nationale sur la santé de la population (ENSP) recueille des données longitudinales sur la santé physique et mentale des Canadiens. Le principal objectif de cet article était d'étudier de quelle façon les patrons des données manquantes affectent les résultats. **MÉTHODES :** Une sous-population tirée des jeux de données de l'Enquête nationale sur la santé de la population (ENSP), enquête longitudinale de Statistique Canada, pour les cinq premiers cycles (1994/95 - 2002/03) a été utilisée pour étudier les effets des données manquantes sur les changements longitudinaux sur les scores de détresse morale parmi les participants à l'ENSP âgés de 15 ans et plus ayant rapporté des maladies respiratoires bénignes. Parmi la variété relativement grande d'indicateurs de la santé mentale disponibles à partir de l'ENSP, la mesure de détresse basée sur un sous-ensemble des éléments du Composite International Diagnostic Interview à 6 questions a été choisi comme étant le résultat. L'analyse a été menée en utilisant l'approche des équations d'estimation généralisées tenant compte de la complexité du plan de l'enquête à plusieurs degrés en utilisant les poids bootstrap disponibles pour les données longitudinales incomplètes. **RÉSULTAT :** Il y avait des interactions significatives entre les patrons des données manquantes (complets/incomplets) et les régions géographiques, l'état matrimonial, et l'indice de santé générale autorapportée. **CONCLUSIONS:** Il y avait un effet significatif des abandons par rapport à l'échelle du résultat de détresse et aux interactions entre la disponibilité des données et les covariables : la région géographique l'état matrimonial et l'indice de santé générale autorapportée. Il était donc clair que sans considérer la disponibilité des données, une perte d'information survient et nos conclusions peuvent changer.

**KEY WORDS:** Données manquantes; enquêtes longitudinales; équations d'estimation généralisées; estimation par Bootstrap; résultat ordinal.

---

<sup>1</sup> Chandima Karunanayake, Institute of Agricultural Rural and Environmental Health, University of Saskatchewan, 103 Hospital Drive, Saskatoon, SK, Canada, S7N 0W8, cpk646@mail.usask.ca

<sup>2</sup> Punam Pahwa, Department of Community Health & Epidemiology, University of Saskatchewan, 103 Hospital Drive, Saskatoon, SK, Canada, S7N 0W8 (Corresponding Author)

<sup>3</sup> Helen McDuffie, Institute of Agricultural Rural and Environmental Health, University of Saskatchewan, 103 Hospital Drive, Saskatoon, SK, Canada, S7N 0W8, mcduffie@sask.usask.ca

# 1. INTRODUCTION

## 1.1 Importance of Handling Missing Data in Longitudinal Studies

In agricultural and medical research, data are often measured repeatedly over time, resulting in so-called longitudinal data, which usually suffer from incompleteness. That is subjects can be missed at a particular measurement time point with the result that these subjects provide data at some, but not all study points. The reasons for incompleteness are usually out of the control of the investigators. Missing data may occur, because subjects may not provide responses for all the study variables. Nonmonotone missing data arise when subjects miss one or more planned visits and return at arbitrary points in time. As a result, observed outcome data may be highly unbalanced and the availability of the data may be directly related to the outcome measure itself. Nonmonotone missing data do not have common sets of visit times or they visit at non-prescheduled times. Because the frequency and timing of subjects' visits may be useful about longitudinal outcomes, they need to be accounted for in order to make valid inference on the longitudinal outcome. Therefore it is not possible to ignore the missingness process because it related to the outcome measurement of interest. Taking proper care to analyze missing data will provide valuable information about the validity of predictive models to estimate the parameters correctly and reduce the potential bias caused by the missing data.

There is a long history of statistical approaches to handle missing data, starting from early 1960's. Initially the researchers were mainly concerned with overcoming the lack of balance or deviations from the intended study. Later stage researchers were interested in recreating the missing data using the observed data and computer algorithms such as expectation-maximization (EM) (Dempster et al., 1977), data imputation and augmentation procedures (Rubin, 1987). In the past twenty years, many articles have been published to demonstrate methods of handling missing data in longitudinal studies (Demirtas, 2004; Molenberghs et al., 2004; Hogan et al., 1997; Gorrbein et al., 1992). They all suggest that the results of the analysis of longitudinal data severely depend on the mechanism by which data is missing.

## 1.2 Missing Data Patterns and Missing Data Mechanisms

There are two different types of missing data patterns. In arbitrary or general missing pattern we have intermittent missing observations. A data set is said to have a monotone missing pattern when a variable  $Y_j$  is missing for the  $i^{\text{th}}$  individual implies that all subsequent variables  $Y_k, k>j$  are missing for the  $i^{\text{th}}$  individual. In other words, in this pattern we have complete dropouts. According to the terminology of Little and Rubin<sup>11</sup>, three different types of missing data mechanisms are defined as follows:

Let  $\underline{Y}$ =Data matrix, with complete data;  $\underline{Y}_{\text{obs}}$ =observed part of  $\underline{Y}$ ;  $\underline{Y}_{\text{mis}}$ =Missing part of  $\underline{Y}$ ;  $\underline{R}$ = Missing data indicator matrix where  $\underline{R}_{ij}= 1$ , if  $\underline{Y}_{ij}$  missing or 0, if  $\underline{Y}_{ij}$  observed. The missing pattern type is concerned with the distribution of  $\underline{R}$  and the missing data mechanism is concerned with the distribution of  $\underline{R}$  given  $\underline{Y}$ .

### 1.2.1 Missing completely at Random (MCAR)

The dropout and the measurement process are independent, That is, cases with complete data are indistinguishable from cases with incomplete data. That is,  $P(\underline{R}|\underline{Y})=P(\underline{R})$  for all  $\underline{Y}$ .

### 1.2.2 Missing At Random (MAR)

The dropout process depends on the observed measurements, that is, cases with incomplete data differ from cases with complete data. In such cases, the missingness pattern of the incomplete data is predictable from other variables in the database rather than being due to the specific variable on which the data are missing. That is,  $P(\underline{R}|\underline{Y})=P(\underline{R}|\underline{Y}_{\text{obs}})$  for all  $\underline{Y}_{\text{mis}}$ .

### 1.2.3 Missing not at Random (MNAR)

The dropout process depends on the unobserved measurement, or in other words, the missingness pattern of the incomplete data is non-random and it is not predictable from other variables in the database, that is,  $P(\underline{R}|\underline{Y})$  depends on  $\underline{Y}_{\text{mis}}$ .

It is very important to know why the values are missing and if these missing values had any effect on the practical question or research question that we want to answer by analyzing the available data.

## **2 LITERATURE REVIEW**

### **2.1 Methods of Analyzing Incomplete Longitudinal Data**

There are several approaches for analysis of incomplete longitudinal data. For repeated categorical data with monotone missingness, a marginal proportional odds model can be fitted assuming that the dropouts are not missing completely at random (MCAR). This can be done using a likelihood-based method called n-way Dale model and the non-likelihood based method, Generalized Estimating Equation (GEE) approach (Michiels et al., 2002). Another approach for analysis of incomplete longitudinal data is pattern mixture models formulated by Little (Little, 1993,1994, 1995), which ignore the missing data mechanism. Pattern-mixture models provide a flexible class of models for incomplete data that are not MCAR. Also pattern mixture model is a solution to the non-response problem in survey data. Another approach for analysis of incomplete longitudinal data is random-effects regression models (RRM) (Hedeker et al., 1996; Hedeker et al., 1997; Stiratelli et al., 1984; Gibbons et al., 1987; Goldstein, 1991; Jansen, 1990; Ezzet et al., 1991; Hedeker et al.,1991; Hedeker et al., 1994). Random-effects model results could not presented in this paper due to convergence problems.

The first step in applying the pattern mixture model approach is to divide the subjects into groups depending on their missing data pattern. If subjects are measured at five time points, then there are 32 ( $2^5$ ) possible missing data patterns. By grouping the subjects in this way, a between-subjects variable, the missing data pattern is created, which can be used in the longitudinal data analysis as another covariate. In longitudinal studies, it is often reasonable to assume that intermittent missing observations are randomly missing. If a large percentage of the subjects complete the study, it may be reasonable to simply contrast completers versus incompleters. Another consideration is whether one is interested in estimating the main effects of the missing data patterns, or also in interactions with the missing data patterns. The applications of the pattern-mixture approach can be used with longitudinal models that allow for missing data across time (e.g., structural equation models or GEE-based models) and they are not limited to random effects modeling (Hedeker et al., 1994; Hedeker et al.,1991). In this paper pattern mixture models were examined with GEE-based models for National Population Health Survey data.

#### **2.1.1 National Population Health Survey (NPHS) –Mental Health and Its Longitudinal Nature**

Some aspects of mental health have a clear cause and effect sequence. Job dissatisfaction, low family income, being uneducated and general health conditions may result in mental health problems (Maclean et al., 2004; Buckley et al., 2003). However, in most instances, the relationship between cause and outcome is much less obvious. Most of the time, research answer to such questions is based on cross-sectional surveys, which gather information about conditions prevailing at one point in time. Instead of gathering information at various intervals from different people by means of a number of cross-sectional surveys, it is preferable to study the same individuals repeatedly and identify changes in their characteristics over time to determine whether there have been corresponding changes in mental health. With such longitudinal data, cause and effect are still difficult to separate, but the evidence is stronger because some information on the sequence of events is available. Statistics Canada conducts the National Population Health Survey (NPHS) to examine the dynamics of health. The NPHS collects both cross-sectional and longitudinal data on the physical and mental health of Canadians and their use of health services. It also collects data on the economic, social, demographic, occupational and environmental correlates of health (Swain et al., 1999).

The main objective of this paper was to investigate the results of different statistical approaches to handling missing data in the National Population Health Survey. We have selected a subpopulation of National Population Health Survey participants who reported non-malignant respiratory diseases (asthma, chronic bronchitis, emphysema) to study the longitudinal changes in mental health status among those who are 15 years and older.

### 3. OTHER INFORMATION

#### 3.1 METHODS

Data from the National Population Health Survey (NPHS) were used in this analysis. The NPHS is a longitudinal study (Statistics Canada, 2002) of a Canadian national sample. The original survey included 17626 subjects sampled in 1994-1995 (first cycle) who will be followed (or re-contacted) every 2 years for up to 20 years. To be included, respondents must have completed at least the general component of the questionnaire in 1994/95. For this analysis information from five cycles (1994/95, 1996/97, 1998/99, 2000/01, 2002/03) is available. The NPHS employed a stratified two-stage design (clusters, dwellings) based on Statistics Canada's Labour Force Survey, except in the province of Quebec. Base sample sizes for each province were determined using the Kish allocation, which balanced the reliability requirements at national and provincial levels. A minimum of 1200 households in each province was needed to ensure a specified reliability by sex and broad age groups. Populations of Indian reserves, on Canadian Forces bases, and in some remote areas of Quebec and Ontario were excluded from the household components of the Survey. Data were weighted to reflect the sample design, adjustments for non-response, and post-stratification. There are 2399 participants in the subpopulation, and in the longitudinal setting there were 6395 observations.

##### 3.1.1 Distress Scale: National Population Health Surveys (NPHS)

From the relatively wide range of mental health indicators available in the NPHS, we chose the distress measure based on a subset of items from the Composite International Diagnostic Interview (CIDI) with six questions developed by Kessler and Mroczek of the University of Michigan be the outcome of interest. The distress scale is comprised of various CIDI items that inquired into feelings of sadness, nervousness, restlessness, hopelessness, worthlessness, and feelings that "everything was an effect"<sup>4</sup>. Additional items clarified whether these symptoms occurred "a lot", "somewhat", "a little", "more than usual", "the same", or "less than usual" in the preceding month. Based on the summation of variables based on the above questions, a distress scale was derived for each of the five cycles. This derived variable determines the respondent's distress scale. Scores on the distress scale range from 0 (no distress) to 24 (highly distressed). The higher values indicate more distress. Details can be found in the NPHS derived variable directory ((Statistics Canada, 2002). Because there is no agreed-upon definition of high distress and according to Kessler's studies we identified those scoring between 0 and 5 as having no or low distress, those scoring between 6 to 12 as having moderate distress, and those scoring 13 or more as experiencing high distress. After new coding the distress scale can be considered as ordinal outcomes.

##### 3.1.2 Subpopulation

The National Population Health Survey consists of 17276 participants. This study was limited to the age group of 15 years and older. The subpopulation was the National Population Health Survey (NPHS) participants who reported non-malignant respiratory diseases (Asthma, Bronchitis, Emphysema and Pneumonia). There are 2399 participants in the subpopulation, and in the longitudinal setting there were 6395 observations. The demographic and socioeconomic variables included in these analyses are described elsewhere (Pahwa et al. 2006).

##### 3.1.3 Modeling distress as an ordinal response-Marginal models

It was of interest to investigate how the response vector evolves over time and how it relates to a set of explanatory variables. As distress was recoded to an ordinal scale, it seems natural to consider a marginal model based on cumulative probabilities. An example of such a model is the proportional odds model. Details can be found in related article (P. Pahwa, *Epidemiology*, August 2006). We fitted the GEE based proportional odds model (Diggle et al., 1994; Liang et al., 1986; Allison, 1999; Fleming et al., 2004) using the genmod procedure in SAS<sup>29</sup>. However, this procedure does not allow the specification of a working correlation matrix other than the independence matrix. When using an independent correlation structure, as the cluster sizes become larger some loss of efficiency occurs, but the estimates are consistent.

### 3.1.4 Results

Univariate analyses were conducted to examine the relationship between distress scale and the demographic and socioeconomic variables mentioned above. To fit the related covariates to the model, the Generalized Estimating Equation (GEE) method was used. The preliminary analysis showed that the variables: sex, education level, age group, marital status, income level, general health, social involvement score, geographic area, smoking status, household smoking, were related to the mental distress scale at  $\alpha=0.20$  significance level. Ethnicity and immigration status were not significant at  $\alpha=0.20$ . Also according to the primary analysis (P. Pahwa, Epidemiology, August 2006) and results from the baseline, job, job satisfaction and social support index variables were removed from the analysis, as these variables had more than fifty percent data missing.

The next step of the analysis was to determine the effects of all the potential covariates and/or interactions on distress scale. The process by which this was determined followed a similar procedure as the univariate analysis but included all of the potential covariates and interactions terms concurrently. The variables that had been included were retained in the model as long as they were significant at  $\alpha=0.05$  level. If a variable included in the model was not significant, it was dropped from the analysis and a new model without these non-significant terms was run. This process was repeated until an appropriate model was produced. It was necessary to use generalized estimating equations for the analysis because of the longitudinal nature of the data (Diggle et al., 1994; Liang et al., 1986; Allison, 1999; Fleming et al., 2004). Standard statistical analysis assumes that all observations in a dataset are independent, but this assumption is not appropriate to apply to the longitudinal component of the National Population Health Survey because each individual respondent has more than one observation per variable (one for each cycle of the survey). But for the ordinal outcome there are limitations, which we discussed in the previous section. Because the National Population Health Survey has a complex sampling design, the bootstrap re-sampling method was used to calculate the correct variance around a given estimate. This was achieved using the 'Bootvar' SAS macro, and bootstrap weights provided by Statistics Canada. The "bootvar" macro was modified to apply to the generalized estimating equations method (Fleming et al., 2004; Statistics Canada, 2005). Table 2 (see appendix) summarizes the results from the multivariate analysis assessing the relationship between covariates and mental distress across all five cycles of NPHS using generalized estimating equations. Age, marital status, location of residence, geographic area, income level, education level, smoking status, household smoking, and general health status were included in the model for each cycle of the survey for the purpose of the generalized estimating equations procedure. Sex, ethnicity, immigration status were included into the model only at baseline (cycle I) because these variables do not change over time. The social involvement score was also included only at baseline because it was not recorded at all five cycles. The variables that were included in Table 2 (see appendix) were those significant variables that were retained from the model building process. The variables that were retained for the final model of the relationship between covariates and mental distress among respondents who reported respiratory diseases (asthma or chronic bronchitis) were: age, sex, education, marital status, income, location of residence, geographic area, general health, household smoking, cycle, and education\*income interaction.

The odds ratios reported for all covariates predicting mental distress takes into account the relationship of each of these variables with each outcome at each cycle. It is possible that at cycles I-IV those with less education were less likely to have had mental distress, but in cycle V this relationship could have changed. Taking all of this information into account, the generalized estimating equation model summarizes all these possible changes in order to report a summary measure of association (the final odds ratio that is produced). All reported odds ratios are adjusted for all other variables in the model. The odds ratios can be interpreted as the likelihood that those who were educated less than or equal to 12 years would have greater mental distress compared with those who were educated more than 12 years. This takes into account the changes in education level over the 10-year follow-up period to produce an overall estimate of the association for each relationship. A similar interpretation can be applied to each of the other variables in the model for mental distress.

Respondents in age groups 15-24 years and 25-54 years were more likely than those aged 55-69 years and 70 and older to have experienced mental distress: odds ratios [2.16 (95% CI=1.34,3.48)], [1.59 (95% CI=1.08,2.36)] respectively. Those living in Quebec reported more mental distress compared to Ontario residents [odds ratio of 1.35 (95% CI=0.97, 1.89)]. Females had nearly twice the chance of reporting high mental distress scores compared to males [OR=1.74, 95% CI=1.39, 2.19]. Fewer respondents who reported being married, living common law or with a partner reported high mental distress scores compared to single respondents [OR=0.75,( 95% CI=0.55, 1.03)].

The low and middle-income level respondents were also reported higher mental distress score compared to high-income group. Having ‘poor’ health status was very highly associated with having high mental distress [odds ratio 22.38 (95% CI=13.22, 37.88)]. Reporting ‘fair’, and ‘good’ health status were also significantly associated with having had high levels of mental distress when compared to respondents with ‘excellent’ health; odds ratios [8.25 (95% CI=5.07,13.42)] and [3.27 (95% CI=2.13,5.01)] respectively. Respondents who reported smoking in the households had a 1.72 times higher chance of having mental distress compared to the ‘no’ group.

The “cycle” variable for all cycles was also significant. In cycles I-IV the odds of having had high mental distress were lower than in cycle V.

Statistically significant interactions terms were included in the final model. The odds ratio for educational level of 12 years or fewer and low income was [0.50 (95% CI=0.24, 1.06)]. The correct odds ratios for this interaction term can be obtained by multiplying the interaction term odds ratio by the odds ratios for each of the separate terms. The overall odds ratio for this interaction term is 1.42. The interpretation was that those who had 12 or fewer years of education and who had low income are 1.42 times more likely to report high distress compared with who had more than 12 years of education and high income. The overall odds ratio for the middle income and less than 12 years education is 1.21.

When conducting GEE analysis using Proc Genmod, missing data is considered missing at random and missing observations were removed from the analysis. Therefore, it was important to study the missing data patterns in the dataset. The missing data patterns for the mental distress scale based on five cycles (Cycle I-Cycle V) are given in Table 1.

It can be seen that frequency and percentage of the missing data patterns 7, 10, 11, 13, 14, 15, 19, 21, 22, 23, 24, 26, 27, 28, 30 and 31 are not presented here due to small numbers which result in confidentiality concerns. Sixty-one percent of participants completed all five cycles. Therefore we can simply contrast completers versus incompleters, creating a new variable “Drop” in the following manner.

$$\text{Drop} = \begin{cases} 0, & \text{Outcome was recorded in all five cycles} \\ 1, & \text{Outcome was not recorded in some of the five cycles} \end{cases}$$

To estimate the main effect of the missing data pattern the “Drop” variable was included in a GEE based model with other covariates. For this SAS PROC GENMOD procedure was used. The first part of the analysis was conducted without including the “Drop” variable. We called this model “Marginal proportional odds model using GEE for subpopulation” and the results of the analysis of GEE empirical parameter estimates and bootstrapped standard errors were given in Table 2 (see appendix).

The next part of the analysis estimated main effects and interactions including the missing patterns and covariates. This multivariate analysis was based on generalized estimating equations approach and results were given in Table 2. This model is called “Pattern-Mixture Marginal proportional Odds models using GEE for subpopulation”.

Incompleters were more likely to have had mental distress compared to completers [odds ratio 2.46 (95% CI=1.00, 6.02)]. Young and middle aged (ages 15-24, 25-54 years) respondents were more likely than those aged 70 years and over to have had mental distress [odds ratio of 2.10(95% CI=1.27,3.47) and 1.62(95% CI=1.07,2.44)] respectively. The youngest age group had a higher risk of reporting mental distress. Females have nearly twice the risk as males [odds ratio 1.78 (95% CI=1.41, 2.25)]. Odds ratio for education, sex, income, household smoking, cycle and education\*income interaction remained the same as the results of the marginal proportional odds model. The overall odds ratio of incompleters and Atlantic respondents was 1.78 suggesting that completers who lived in Atlantic region were 1.78 times as likely to have reported mental distress as compared with the completers who lived in Ontario. Similarly incompleters who lived in British Columbia, Prairies, and Quebec were 1.52, 2.06, and 2.41 as likely respectively to have reported mental distress compared with completers who lived in Ontario. Incompleters who were married or lived common law or with a partner and completers who were separated or widowed or divorced were 2.62 and 3.36 times respectively as likely to have reported mental distress in comparison with single completers. Incompleters who were in poor health condition were

**Table 1: Missing Data Patterns of Mental Distress in NPHS**

Pattern	Cycle					Frequency (%)
	I	II	III	IV	V	
1	x	x	x	x	x	3893 (60.88)
2	x	x	x	x	0	535 (8.37)
3	x	x	x	0	x	199 (3.11)
4	x	x	X	0	0	389 (6.08)
5	x	x	0	x	x	61 (0.95)
6	x	x	0	x	0	53 (0.83)
7	x	x	0	0	x	
8	x	x	0	0	0	281 (4.39)
9	x	0	x	x	x	79 (1.24)
10	x	0	x	x	0	
11	x	0	x	0	x	
12	x	0	x	0	0	31 (0.48)
13	x	0	0	x	x	
14	x	0	0	x	0	
15	x	0	0	0	x	
16	x	0	0	0	0	131 (2.05)
17	0	x	x	x	x	265 (4.14)
18	0	x	x	x	0	57 (0.89)
19	0	x	x	0	x	
20	0	x	x	0	0	53 (0.83)
21	0	x	0	x	x	
22	0	x	0	x	0	
23	0	x	0	0	x	
24	0	x	0	0	0	
25	0	0	x	x	x	73 (1.14)
26	0	0	x	x	0	
27	0	0	x	0	x	
28	0	0	x	0	0	
29	0	0	0	x	x	36 (0.56)
30	0	0	0	x	0	
31	0	0	0	0	x	
32	0	0	0	0	0	59 (0.92)

x - complete  
0 – missing

30.71 times more likely to have had mental distress compared with completers who were in excellent health. Similarly incompleters who had fair, good, or very good health were 11.79, 4.36, and 2.80 times respectively more likely to have had mental distress compared with completers who had excellent health.

The weight variable we used in our analysis play an important role in the complex survey data analysis. Pattern mixture marginal proportional odds model using GEE fitted to full dataset and results were presented in last two columns of Table 2. It is clearly visible that the parameter estimates of pattern mixture subpopulation model were very different from the full data model. Here we did not make an adjustment of weights for the subpopulation. Therefore further investigation is needed to find out how the weights can adjust for subpopulation.

### 3.1.5 Discussion

In this paper the main interest was to determine how missingness affects the results and interpretation. In this dataset, data were missing at intermittent waves “pattern-mixture models” were used to analyze the missing or incomplete data. We combined some of the patterns to increase interpretability. In longitudinal studies, it is often reasonable to assume that the intermittent missing observations are randomly missing. In this study, subjects were divided into groups two groups depending on their missing data pattern. As a large percentage of subjects (61%) completed the study, we simply contrasted completers versus incompleters. These groups were used to examine the effect of the missing-data pattern on the outcome of interest. For the ordinal responses, pattern-mixture proportional odds models using generalized estimation equation were considered. There was a significant effect from the dropouts to the outcome distress scale and “missingness” interacts with the following covariates: geographic area, marital status and self-reported general health index. Therefore it was clear that without considering missingness, loss of information occurs and our conclusions may change. The complete case analysis was not valid for these situations. It is clear from our results that suitable weights must be chosen for the analysis of complex datasets. There should be adjustment in weights when a subpopulation used for analysis.

## ACKNOWLEDGEMENTS

Authors would like to thank the remote data access services of Statistics Canada, providing the original master file data results.

## REFERENCES

- Allison, P.D. (1999). *Logistic Regression Using the SAS System*. Cary, North Carolina: SAS Institute.
- Buckley, N.J., Denton, F.T., Robb, A.L., and Spencer, B.G.(2003). Socio-economic influence on the health of older people: estimates based on two longitudinal surveys, *Research Institute for Quantitative Studies in Economics and Population (QSEP)*. Research Report No. 387.
- Demirtas, H. (2004). Modeling incomplete longitudinal data. *Journal of Modern Applied Statistical Methods*. **3**: 305-321.
- Dempster, A. P. , N. M. Laird, and D. B. Rubin. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, **34**, 1-38.
- Diggle, P.J., Liang, K-Y., Zeger, S.L.(1994). *Analysis of Longitudinal Data*. New York: Oxford University Press.
- Ezzet, F., and Whitehead, J. (1991). A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine*. **10**, 901-907.
- Fleming S.A., Bains N., Hunter D.J.W., Lam M.(2004). Social support and Health care use among a sample of healthy Canadian: A longitudinal analysis of the National population Health Survey, *Health information partnership Eastern Ontario Region*, Kingston, Ontario.
- Gibbons, R. D., & Bock, R. D. (1987). Trend in correlated proportions. *Psychometrika*. **52**, 113-124.
- Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*. **78**, 45-51.
- Gorrbein, J.A., Lazaro C.G., and Little R.J.A. (1992). Incomplete data in repeated measures analysis. *Statistical Methods in Medical research*. **1**:275-295.
- Hedeker, D. & Mermelstein, R.J. (1996). Application of random-effects regression models in relapse research. *Addiction*. **91** (Supplement): S211-S229.
- 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., and Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel data. *Biometrics*. **50**, 933-944.
- Hedeker, D., Gibbons R.D., and Davis J.M. (1991). Random regression models for multicenter clinical trials data. *Psychopharmacology Bulletin*. **27(1)**, 73-77.
- Hogan, J.W. and Laird N.M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*. **16**: 239-258.
- Jansen, J. (1990). On the statistical analysis of ordinal data when extravariation is present. *Applied Statistics*. **39**, 75-84.
- Liang, K-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized estimating equations. *Biometrika*. **73**, 13-22.
- 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 American Statistical Association*. **90**, 1112-1121.
- Little, R.J.A. and Rubin, D.B.(1987). *Statistical Analysis with Missing Data*. New York:John Wiley and Sons Inc.
- Maclean, H., Glynn, K., and Ansara, D. (2004) Multiple Roles and Women's Mental Health in Canada. *BioMed Central Women's Health*. 4(suppl 1): S3
- Michiels, B., Molenberghs G.M., Bijmens L., Vangeneugden T., and Thijs H.(2002). Selection models and Pattern-mixture models to analyze longitudinal quality of life data subject to dropout. *Statistics in Medicine*. **21**, 1023-1041.
- Molenberghs, G.M., Thijs H., Jansen I., Beunckens C., Kenward M.G., Mallinckrodt C., and Carroll R.J.(2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*. **5**, 445-464.
- Pahwa, P., Karunanayake, C., McDuffie H.H., (2006) Modeling of Longitudinal Polytomous Outcome From Complex Survey Data – An Application. (*To appear in SSC -2006 proceedings*)
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Statistics Canada. (2005). Estimation of the variance using the Bootstrap Weights. User's Guide for the *BOOTVARE\_V21.SPS Program* (Version 2.1).
- Statistics Canada.(2002).Documentation of the Longitudinal Component of the National Population Health Survey.
- Stiratelli, R., Laird N. M. and Ware J. (1984). Random-effects models for serial observations with binary response. *Biometrics*. **40**, 961-971.
- Swain, L., Catlin, G., Beaudet, M.P.(1999). The National Population Health Survey--its longitudinal nature. *Health Reports*. **10(4)** 69-82.

## Appendix

**Table 2: Results- Marginal Proportional Odds Model using GEE and Pattern-Mixture Marginal Proportional Odds Model using GEE**

Parameter/ Categories	Marginal Proportional Odds Model using GEE (subpopulation)		Pattern-Mixture Marginal Proportional Odds Model using GEE (subpopulation)		Pattern-Mixture Marginal Proportional Odds Model using GEE (full data)	
	Estimate (Standard error)	Odds Ratio [95% CI]	Estimate (Standard error)	Odds Ratio [95% CI]	Estimate (Standard error)	Odds Ratio [95% CI]
Intercept1	-5.31 (0.40)	0.0049 [0.002,0.011]	-5.72 (0.47)	0.0033 [0.001,0.008]	-5.95 (0.14)	0.0026 [0.002,0.003]
Intercept2	-2.99(0.39)	0.05 [0.02,0.11]	-3.38 (0.46)	0.03 [0.01,0.08]	-3.40 (0.14)	0.03 [0.02,0.04]
Drop/Incompleters			0.90 (0.46)	2.46 [1.00,6.02]	0.02 (0.14)	1.02 [0.77,1.35]
Drop/Completers			Reference	1.00	Reference	1.00
Age group	0.77 (0.24)	2.16 [1.34,3.48]	0.74 (0.25)	2.10 [1.27,3.47]	1.33 (0.09)	3.78 [3.19,4.47]
15-24years		1.59 [1.08,2.36]	0.48 (0.21)	1.62 [1.07,2.44]	0.96 (0.07)	2.62 [2.28,3.01]
25-54 years	0.47 (0.20)	0.94 [0.63,1.40]	-0.06 (0.21)	0.94 [0.62,1.43]	0.25 (0.08)	1.28 [1.09,1.49]
55-69 years	-0.06 (0.20)	1.00	Reference	1.00	Reference	1.00
70 years/over	Reference	1.00	Reference	1.00	Reference	1.00
Education Level						
≤ 12 years	0.28 (0.31)	1.32 [0.72,2.42]	0.29 (0.31)	1.33 [0.72,2.45]	0.13 (0.11)	1.14 [0.92,1.41]
> 12 years	Reference	1.00	Reference	1.00	Reference	1.00
Marital Status						
Married/ Common law/living with a partner	-0.28 (0.16)	0.75 [0.55,1.03]	-0.46 (0.18)	0.63 [0.44,0.90]	-0.44 (0.07)	0.64 [0.56,0.74]
Separated/widowed/divorced	-0.08 (0.21)	0.92 [0.61,1.40]	-0.40 (0.23)	0.67 [0.42,1.05]	-0.07 (0.08)	0.93 [0.79,1.09]
Single	Reference	1.00	Reference	1.00	Reference	1.00
Total household income						
Lowest income	0.77 (0.23)	2.15 [1.36,3.40]	0.77 (0.23)	2.15 [1.36,3.41]	0.56 (0.08)	1.74 [1.47,2.06]
Middle income	0.25 (0.19)	1.29 [0.89,1.87]	0.26 (0.19)	1.30 [0.89,1.89]	0.12 (0.07)	1.12 [0.98,1.28]
High income	Reference	1.00	Reference	1.00	Reference	1.00

Cont...

Parameter/ Categories	Marginal Proportional Odds Model using GEE (subpopulation)		Pattern-Mixture Marginal Proportional Odds Model using GEE (subpopulation)		Pattern-Mixture Marginal Proportional Odds Model using GEE (full data)	
	Estimate (Standard error)	Odds Ratio [95% CI]	Estimate (Standard error)	Odds Ratio [95% CI]	Estimate (Standard error)	Odds Ratio [95% CI]
General Health Index						
Poor	3.11 (0.27)	22.38 [13.22,37.88]	3.47 (0.40)	32.01 [14.57,70.34]	3.31 (0.15)	27.44 [20.42,36.89]
Fair	2.11 (0.25)	8.25 [5.07,13.42]	2.48 (0.34)	11.98 [6.12,23.44]	2.18 (0.09)	8.86 [7.41,10.58]
Good	1.18 (0.22)	3.27 [2.13,5.01]	1.59 (0.32)	4.92 [2.64,9.16]	1.21 (0.07)	3.36 [2.95,3.84]
Very Good	0.43 (0.22)	1.53 [1.00,2.36]	0.66 (0.32)	1.93 [1.03,3.61]	0.47 (0.06)	1.59 [1.41,1.80]
Excellent	Reference	1.00	Reference	1.00	Reference	1.00
Sex						
Female	0.56 (0.11)	1.74 [1.39,2.19]	0.58 (0.12)	1.78 [1.41,2.25]	0.45 (0.04)	1.57 [1.44,1.71]
Male	Reference	1.00	Reference	1.00	Reference	1.00
Location of Residence						
Rural	-0.46 (0.16)	0.63 [0.46,0.85]	-0.49 (0.16)	0.61 [0.44,0.83]	-0.19 (0.06)	0.82 [0.74,0.92]
Urban	Reference	1.00	Reference	1.00	Reference	1.00
Household Smoking						
Yes	0.54 (0.11)	1.72 [1.38,2.14]	0.53 (0.11)	1.71 [1.36,2.14]	0.27 (0.04)	1.31 [1.21,1.42]
No	Reference	1.00	Reference	1.00	Reference	1.00
Geographical area						
Atlantic	0.25 (0.19)	1.28 [0.87,1.88]	0.59 (0.28)	1.81 [1.05,3.11]	-0.10 (0.08)	0.90 [0.77, 1.06]
British Columbia	0.02 (0.17)	1.02 [0.73,1.44]	0.44 (0.26)	1.55 [0.94,2.56]	-0.05 (0.09)	0.94 [0.80,1.12]
Prairies	-0.12 (0.17)	0.88 [0.64,1.23]	-0.10 (0.23)	0.90 [0.57,1.43]	-0.09 (0.07)	0.91 [0.79,1.05]
Quebec	0.30 (0.17)	1.35 [0.97,1.89]	0.54 (0.21)	1.72 [1.13,2.60]	0.40 (0.07)	1.49 [1.30,1.72]
Ontario	Reference	1.00	Reference	1.00	Reference	1.00
Cycle						
5- 2002/03	-0.53 (0.15)	0.59 [0.44,0.79]	-0.49 (0.15)	0.61 [0.45,0.83]	-0.36 (0.05)	0.70 [0.63,0.77]
4-2000/01	-0.59 (0.15)	0.55 [0.41,0.74]	-0.56 (0.16)	0.57 [0.42,0.77]	-0.47 (0.05)	0.62 [0.57,0.68]
3-1998/99	-0.50 (0.15)	0.61 [0.45,0.82]	-0.49 (0.16)	0.61 [0.45,0.84]	-0.25 (0.05)	0.77 [0.71,0.85]
2-1996/97	-0.48 (0.12)	0.62 [0.48,0.79]	-0.48 (0.12)	0.62 [0.48,0.79]	-0.35 (0.04)	0.70 [0.65,0.77]
1-1994/95	Reference	1.00	Reference	1.00	Reference	1.00

Cont....

Parameter/ Categories	Marginal Proportional Odds Model using GEE (subpopulation)		Pattern-Mixture Marginal Proportional Odds Model using GEE (subpopulation)		Pattern-Mixture Marginal Proportional Odds Model using GEE (full data)	
	Estimate (Standard error)	Odds Ratio [95% CI]	Estimate (Standard error)	Odds Ratio [95% CI]	Estimate (Standard error)	Odds Ratio [95% CI]
<b>Education*income</b>						
≤ 12 years*low income	-0.68 (0.38)	0.50 [0.24,1.06]	-0.70 (0.38)	0.49 [0.23,1.05]	-0.22 (0.13)	0.80 [0.62,1.03]
Overall ≤ 12 years*low income		1.42		1.40		1.59
≤ 12 years*middle income	-0.34 (0.34)	0.71 [0.36,1.38]	-0.36 (0.35)	0.69 [0.35,1.38]	-0.07 (0.12)	0.93 [0.74,1.18]
Overall ≤ 12 years* middle income		1.21		1.19		1.19
<b>Drop* geographical area</b>						
Incompleters*Atlantic			-0.92 (0.37)	0.40 [0.19,0.83]	-0.04 (0.11)	0.96 [0.77,1.20]
Overall Incompleters*Atlantic				1.78		0.88
Incompleters*British Columbia			-0.91 (0.38)	0.40 [0.19,0.84]	0.09 (0.12)	1.10 [0.86,1.40]
Overall Incompleters*British Columbia				1.52		1.05
Incompleters*Prairies			-0.07 (0.35)	0.93 [0.46,1.87]	0.08 (0.11)	1.08 [0.88,1.33]
Overall Incompleters*Prairies				2.06		1.00
Incompleters* Quebec			-0.57 (0.38)	0.57 [0.27,1.20]	0.01 (0.10)	1.01 [0.83,1.24]
Overall Incompleters* Quebec				2.41		1.53
<b>Drop*marital status</b>						
Incompleters*married/ common law/ living with a partner			0.52 (0.26)	1.69 [1.02,2.79]	0.28 (0.10)	1.32 [1.08,1.61]
Overall Incompleters*married/ common law/ living with a partner				2.62		0.86
Incompleters*separated/ widowed/divorced			0.71 (0.33)	2.04 [1.05,3.94]	0.33 (0.12)	1.39 [1.10,1.75]
Overall Incompleters*separated/ widowed/divorced				3.36		1.32
<b>Drop* General Health Index</b>						
Incompleters*poor			-0.94 (0.50)	0.39 [0.14,1.04]	-0.22 (0.22)	0.80 [0.52,1.23]
Overall Incompleters*poor				30.71		22.39
Incompleters*fair			-0.92 (0.48)	0.40 [0.15,1.02]	-0.25 (0.14)	0.78 [0.59,1.02]
Overall Incompleters*fair				11.79		7.05
Incompleters * good			-1.02 (0.43)	0.36 [0.15,0.85]	-0.12 (0.11)	0.88 [0.71,1.10]
Overall Incompleters * good				4.36		3.01
Incompleters*very good			-0.52 (0.45)	0.59 [0.24,1.43]	-0.06 (0.11)	0.94 [0.75,1.18]
OverallIncompleters*very good				2.80		1.52

