MODELING OF LONGITUDINAL POLYTOMOUS OUTCOME FROM COMPLEX SURVEY DATA – AN APPLICATION

Punam Pahwa1[1,2], Chandima Karunanayke[2], Helen McDuffie[2]

ABSTRACT

The data from longitudinal complex surveys based on multi-stage sampling designs contain cross-sectional dependencies among units due to clustered nature of the data and within-subject dependencies due to repeated measurements. Special statistical methods are required to analyze longitudinal complex survey data. Ordered logistic regression models based on the weighted generalized estimating equations (GEE) approach were fitted to investigate the association between respiratory diseases and mental distress adjusting for other covariates of interest. Variance estimates of regression coefficients were computed by using bootstrap method and GEE approach. The GEE variance estimates were similar to those obtained from using the bootstrap method. The final model was used to predict the probabilities of prevalence of no/low, moderate or high mental distress scores.

KEY WORDS: Longitudinal, Complex surveys, Ordinal Outcome, Generalized Estimating Equations, Robust Variance Estimation, Bootstrap.

RÉSUMÉ

Les données longitudinales provenant d'enquêtes complexes utilisant des plans de sondage à plusieurs degrés comportent des dépendances transversales entre les unités en raison de la nature "en grappes" des données et de la dépendance intra-sujet due à la répétition des mesures. Des méthodes statistiques spéciales sont requises pour analyser les données longitudinales provenant d'enquêtes complexes. Des modèles de régression logistique ordonnés basés sur l'approche des équations d'estimation généralisées (ÉEG) pondérées ont été ajustés dans le but d'étudier la relation entre les maladies respiratoires et la détresse morale en ajustant pour les covariables d'intérêt. Les estimations de variance des coefficients de régression ont été calculées en utilisant la méthode du bootstrap et l'approche des ÉEG. Les estimations de variance ÉEG étaient similaires à celles obtenues à l'aide de la méthode du bootstrap. Le modèle final a été utilisé pour prédire la probabilité de prévalence des scores pour une détresse morale nulle à faible ou élevée.

MOTS CLÉS : Bootstrap; données longitudinales; enquêtes complexes; équations d'estimation généralisées; estimation de la variance robuste; résultat ordinal.

1. INTRODUCTION

Statistics Canada has engaged in conducting large scale longitudinal surveys[1-5] over long periods of time. The selection of the sample of individuals who participate in such surveys is based on complex multi-stage sampling designs. Participants in these surveys have repeated measurements on the response variables of interest and several covariables over time, which lead to the dependent observations, as encountered in standard longitudinal studies[6]. The complex multi-stage sampling designs used for these longitudinal surveys also contain cross-sectional dependencies among units (caused by inherent hierarchies in the data) in addition to the within-subject dependencies due to repeated measurements.
Many indicators of mental health status are measured on ordinal scales. Ordinal outcomes can be viewed as an extension of binary outcomes. Although models based on dichotomous ordinal outcomes do extend to polytomous ordinal outcomes, there are a number of issues specific to the ordinal case[7]. For a polytomous ordinal outcome, the most popular model is the logit model based on the concept of cumulative logits. Considerable progress in methodological development for the

[1,2] Punam Pahwa, Department of Community Health & Epidemiology, University of Saskatchewan, Institute of Agricultural Rural and Environmental Health, 03 Hospital Drive, Saskatoon, SK, Canada, S7N 0W8 (Corresponding Author)
[2] Chandima Karunanayake, Institute of Agricultural Rural and Environmental Health, University of Saskatchewan, 103 Hospital Drive, Saskatoon, SK, Canada,S7N 0W8, cpk646@mail.usask.ca
[2] Helen McDuffie, Institute of Agricultural Rural and Environmental Health, University of Saskatchewan, 103 Hospital Drive, Saskatoon, SK, Canada,S7N 0W8, mcduffie@sask.usask.ca

analysis of ordinal response data and its application to complex survey data has been made in recent years[8]. Two different sets of models for ordinal outcome were utilized in this article to analyze the longitudinal Canadian National Population Health Survey (NPHS) data. The first set of models was based on the assumption that the study design involved only subject-level clustering due to repeated measurements and was based on the Generalized Estimating Equations (GEEs) approach[9,10], thus ignoring the complexities of the survey design. The second set of models was based on the assumption that the study-design was a complex survey and incorporated complexities of the design in addition to the subject-level clustering which was an assumption for the first set of models. For the second set of models, the GEEs approach was used for the parameter coefficients estimation and complexities of the design were incorporated via appropriate variance estimation approaches. The two commonly used approaches for producing variance estimates for estimated regression coefficients are, the Taylor-linearization approach[11,12] and a replication approach[11,13]. The bootstrap method based on the replication approach was used in this article for the variance estimation.

The statistical approaches mentioned above were used to analyze longitudinal Canadian NPHS data to investigate an association between non-malignant respiratory diseases and mental distress. Mental health is as important as physical health to the overall well-being of individuals, societies and countries. Everyone suffers with mental distress at some time due to physical illness or chronic disease. Sroujian reported that mental illness accounts for 30% of disability claims, i.e. $15 to $33 billion annually in Canada[14]. Advances in neuroscience and behavioural medicine have shown that, like many physical illnesses, mental and behavioural disorders are the result of a complex interaction between biological, psychological and social factors[15].

Statistics Canada's longitudinal NPHS provided a wealth of new data, which allowed us to investigate the effects of demographic variables, social, life-style, health-related and other factors on the longitudinal changes of mental distress scores among the NPHS participants who self-reported physician diagnosed respiratory diseases, specifically asthma and chronic bronchitis. This research focused on some aspects of statistical modeling of longitudinal complex survey data, and the application of these methods to the longitudinal Canadian NPHS data to determine the probability of developing mental distress among those people who suffered with non-malignant respiratory diseases.

2. METHODS

2.1 Canadian National Population Health Survey :

The Canadian NPHS was launched in the mid 90's[1]. The longitudinal sample data used for the analyses in this report consist of 17,276 participants (Newfoundland: 1082; Prince Edward Island: 1037; Nova Scotia: 1085; New Brunswick: 1125; Québec: 3000; Ontario: 4307; Manitoba: 1205; Saskatchewan: 1168; Alberta: 1544; British Columbia: 1723) at the start of Cycle 1 (1994/95). A stratified multi-stage sampling design was used to collect the data from all the provinces except for Quebec. Details of this sampling design can be found elsewhere[16,17]. In every participating household, one person provided demographic, socio-economic and health information about each household member for the general component of the survey. One randomly selected individual was chosen to provide in-depth information about his or her own health for the health component of the survey, and was followed for the longitudinal component of the survey. This group of individuals will be surveyed every two years in future until 2014.

2.2 Statistical models:

Ordered logistic regression models[8,18] were used to predict the relationship between an ordinal mental distress outcome and a set of explanatory variables.

$$log\left(\frac{probability\ (category\ (l+1)\ or\ lower))}{(1-probability\ (category\ (l+1)\ or\ lower))}\right)=\beta_{0l}+\sum_i\sum_j\sum_k\beta_i X^t_{ijk}+\sum_i\sum_j\gamma_i X^s_i \quad \ldots (1)$$

where β_{0l} (l=1,2) are the intercepts and β_i's are regression coefficients for the covariates X^t_{ijk} and X^s_{ij}. X^t_{ijk} represents time-dependent i^{th} covariate for j^{th} subject at k^{th} cycle ($k=1,2, \ldots ,5$). X^s_{ij} represents time-independent i^{th} covariate for j^{th} subject measured at the baseline. β_{01} is the intercept for log odds of having high distress versus moderate or no/low distress and β_{02} is the intercept of log odds of having high or moderate distress vs. no/low distress. The basic assumption made to conduct this type of analysis was that the regression lines for the different outcome categories were parallel to each other but were allowed to have different intercepts. This assumption was satisfied when tested by a graphical method[19].

2.2.1 Estimation of regression coefficients:

The SAS procedure PROC GENMOD[18] was used to fit the multivariable model in order to determine the significant predictors of mental distress. The longitudinal weight variable computed by the methodologists of Statistics Canada was used in the WEIGHT statement of SAS syntax. Currently, for ordinal outcome SAS PROC GENMOD has only one option available for specifying the within subject correlation and that is 'independent". The estimates of regression coefficients for the ordinal logistic regression model given in equation (1) above were obtained by solving the set of score equations based on multivariate quasi-likelihood approach[9,10] modified for complex survey designs using the weight variable.

2.3. Variance estimation:

2.3.1. Robust Variance estimation based on the GEEs and not accounting for the design: (model-based variance estimation).

Robust variance estimation in GENMOD is based on Zeger and Liang's method[9,10] which accounts only for the within-subject dependencies due to the repeated measurements over time. The variance estimation was based on the formula given by Liang and Zeger[9,10].

2.3.2. Survey Bootstrap for Variance Estimation accounting for the design complexities: (design-based variance estimation).

Statistics Canada releases design information for variance estimation only in the form of bootstrap weights: cross-sectional weights and longitudinal weights (adjusted for non-response) that have been created from taking numerous bootstrap samples of primary sampling units from the original sample. Computation of replicate survey weights is done by the methodologists of Statistics Canada who are most familiar with the survey design and the computation of weights[20]. A Bootstrap replication method was used that made appropriate use of these longitudinal bootstrap weights for the variance estimation of regression estimates. To account for the complexities of the multi-stage stratified clustered design the BOOTVAR program which was originally developed by Statistics Canada and modified by Lam[21] was used for the variance estimation.

2.3.4. Prediction of probabilities for three different mental distress categories:

Once the model was fitted, the following two predictive models were used to determine the predicted probabilities for : i) high mental distress category (p_1); ii) moderate mental distress category (p_2) and iii) no/low distress category (p_3)

$$log\left(\frac{predicted\ probability\ of\ high\ distress}{predicted\ probability\ of\ \ moderate\ or\ no/low\ distress}\right) = \hat{\beta}_{0l} + \sum_i \sum_j \sum_k \hat{\beta}_i X_{ijk}^t + \sum_i \sum_j \hat{\gamma}_i X_{ij}^s \quad ...(2)$$

$$log\left(\frac{predicted\ probability\ of\ high\ or\ moderate\ distress}{predicted\ probability\ of\ no/low\ distress}\right) = \hat{\beta}_{0l} + \sum_i \sum_j \sum_k \hat{\beta}_i X_{ijk}^t + \sum_i \sum_j \hat{\gamma}_i X_{ij}^s \quad ...(3)$$

Total probability attributable to the three distress categories is equal to 1, i.e.

$$p_1+p_2+ p_3 = 1 \quad ... (4)$$

Equations (2) to (4) were solved to estimate probabilities p_1, p_2, and p_3.

3. APPLICATION TO LONGITUDINAL NPHS DATA

The NPHS includes a set of questions designed to determine/investigate the mental health of NPHS participants. In this report, we used mental distress as a measure of mental health. The mental distress variable was derived from a set of questions designed by Kessler et al[22].

3.1. Dependent Variable:

Distress, an ordinal outcome variable was examined using a six-item scale that assessed feelings of i) sadness, ii) nervousness, iii) restlessness, iv) hopelessness , v) worthlessness and vi) the feeling that everything was an effort within

the previous month. The variable "distress scale", is based on the work of Kessler and Morczek[22] and was derived from the Composite International Diagnostic Interview. Scores on the distress scales ranged from 0 (no distress) to 24 (highly distressed). The distribution of this distress scale was highly skewed for the Canadian population. There is no agreed-upon definition of low, moderate or high distress. After personal communications with one of the psychiatrists at the University of Saskatchewan, the authors categorized the outcome variable into three categories: i) no or low distress : 0-5; ii) moderate distress: 6-12; and iii) high distress: 13-24.

3.2. Independent variables:

Mental health is an interplay among several factors, such as: demographic; socio-economic, social-support, health related, time of study and interactions between them. In this report the following variables were considered as independent variables:

Main risk factors of interest: presence or absence of asthma, presence or absence of chronic bronchitis

Demographic variables consist of age, sex, ethnicity, marital status, location of residence, geographical area, and length of stay in Canada. Age was used as a time-dependent variable with four categories: 15-24 yrs, 25-54 yrs, 55-69 yrs and 70 yrs and older (reference category: 70 yrs and older). Ethnicity[23] was a time-independent dichotomous variable with two categories: white vs. non-white (non-white as a reference category). Marital status was grouped into three categories: married/common law/partnership; separated/widowed/divorced; and single (reference category). Location of residence had two categories, rural vs. urban. Geographical area was a nominal variable with five categories: Atlantic (Halifax, New Foundland, New Brunswick, and Prince Edward Island); British Columbia; Prairies (Manitoba, Saskatoon, Alberta)); Quebec; and Ontario (reference category). The length of residence in Canada variable was dichotomized (< 25 years (reference category) or ≥ 25 years) based on the median years of stay in Canada.

Socio-economic status variables consist of education and income. Education was a dichotomous variable with two categories: education received less than or equal to12 years and education received greater than 12 years. Income was divided into three categories based on the work of Wang and EI-Gebaly[24].

Social Support variables consist of a social involvement score, which was divided into three categories: low (0-1); moderate (2-4); and high (5-8). This score was based on two questions : frequency of participation in organizations and frequency of attending religious services.

Life-style variables consist of participant's personal smoking history and household smoking status. Personal Smoking history was divided into three categories, non-smokers, ex-smokers and current smoker. Household smoking status was a dichotomous variable indicating presence or absence of smokers within a household.

Health related variable consists of a self-perceived general health status, which had five categories: poor, fair, good, vary good, and excellent (reference category: excellent).

Four dummy variables for 'Cycle' was used to study the effect of time on mental distress.

4. RESULTS

Our study population (n=17,246) consisted of the longitudinal sample of NPHS. At the baseline, 78.2% were classified with no/low distress, 19.4 % with moderate distress and 2.4% with high distress[15]. The pattern of distribution of participants in each of the five cycles is given in Table 1.

Table 1. Number of observations contributed by Canadian National Population Health Survey Participants.

	Cycle I	Cycle II	Cycle III	Cycle IV	Cycle V	Frequency	Percent (%)
Participants with no missing values	x	x	x	x	x	8210	56.75
Participants with one missing values	x	x	x	x	.		
	x	x	x	.	x		
	x	x	.	x	x	2383	16.47
	x	.	x	x	x		
	.	x	x	x	x		
Participants with two missing values	x	x	x	.	.		
	x	x	.	x	.		
	x	x	.	.	x		
	x	.	x	.	x		
	x	.	.	x	x	1541	10.65
	.	x	x	x	.		
	.	x	x	.	x		
	.	x	.	x	x		
	.	.	x	x	x		
	x	.	x	x	.		
Participants with three missing values	x	x	.	.	.		
	x	.	.	.	x		
	.	.	.	x	x		
	x	.	x	.	.		
	.	x	x	.	.	1188	8.21
	.	.	x	x	.		
	x	.	.	x	.		
	.	x	.	.	x		
	.	.	x	.	x		
Participants with four missing values	.	.	.	.	x		
	x	.	.	.	.		
	.	x	.	.	.	1144	7.91
	.	.	x	.	.		
	.	.	.	x	.		

4.1. Multivariable Model:

Table 2 summarizes the results from the multivariable model to assess the relationship between non-malignant respiratory diseases and mental health adjusting for important covariates: demographic, socio-economic, social-support, lifestyle, self-perceived general health status and time (cycle) and the effects of interactions using the generalized estimating equations approach. These covariates for the multivariable model were selected based on standard model building strategies[25]. The standard errors of regression coefficients ignoring the design complexities (based on Zeger and Liang's formula) and accounting for the complexities of stratified multi-stage design (based on bootstrap methods) were computed. The standard errors obtained by the two methods were very similar. The results based on bootstrap variance estimation were used to interpret the effect of each independent variable adjusting for other covariates as described below:
The main risk factor of interest was non-malignant respiratory diseases (asthma or chronic bronchitis). The NPHS participants who said 'yes' to physician-diagnosed asthma were not at a high risk of mental distress when adjusted for important covariates. Participants suffering with chronic bronchitis were significantly at a higher risk (OR_{adj}=1.37; 95% CI.: 1.12-1.66) of reporting high levels of mental distress compared to those who did not have chronic bronchitis. Participants in the younger age groups (15-24, 25-54, and 55-69) were significantly more likely than those 70+ to report high levels of distress with odds ratio of 3.63 (95% CI; 2.03-4.36), 2.47 (95% CI: 2.15-2.84) and 1.23 (95% CI: 1.06-1.43) respectively. White people were less likely to have high mental distress [OR_{adj}=0.97, (95% CI.: 0.82-1.15)] compared to non-white people but this difference was not statistically significant. Rural participants were less likely [OR_{adj}=0.83; (95% CI: 0.75-0.93)] to report high level of mental distress compared to urban participants. Participants from Quebec were significantly at a higher risk [OR_{adj}=1.54; (95% CI: 1.37-1.74)] to report high level of distress compared to Ontario participants. Immigrant participants were at a higher risk [OR_{adj}=1.12; (95% C.I: 0.99-1.27)] of reporting high level of distress compared to non-immigrants with a borderline significance. Participants with low [OR_{adj}=1.13; (95% C.I: 1.00-1.28)] or moderate [OR_{adj}=1.20; (95% CI: 1.07-1.35)] social involvement scores were significantly at a higher risk of reporting high level of distress compared to those participants who had high social involvement score. Current smokers [OR_{adj}=1.39; (95% CI: 1.23-1.57)] and ex-smokers (OR_{adj}=1.13; 95% C.I. 1.02-1.24) were significantly more likely to have high level of distress compared to the non-smokers.

Various interaction terms were tested in the multivariable model for statistical significance. The following interaction terms: education*income ($p<0.1$), general health-status*sex ($p<0.05$), and general health-status*household smoking ($p<0.1$) were retained in the final model. The interactions education*income and general health-status*household were considered scientifically important and were kept in the model. The overall odds ratios for educational level $\leq$ 12 years*low income, educational level $\leq$ 12 years *middle income indicate that participants in these two categories were more likely to have had high distress compared to those who had high income and more than 12 years of education. Female participants with self-perceived 'poor' health were at the highest risk (overall OR_{adj}=43.91) to have had high distress, followed by female participants with self-perceived 'fair' (overall OR_{adj}=11.85), 'good' (overall OR_{adj}=4.96) and 'very good' (overall OR_{adj}=2.36) general health status compared to the male participants with 'excellent' self-perceived general health status. Participants who were exposed to smoking within their household and had self-perceived 'poor' health status were at the highest risk (overall OR_{adj}=22.22) followed by those who were exposed to cigarette smoke at home and had 'fair' (overall OR_{adj}=7.3), 'good' (overall OR_{adj}=2.75) and 'very good' (overall OR_{adj}=1.71) health compared to males with 'excellent' general health status.
The predicted probability of developing no/low, moderate or high distress adjusting for other covariates is shown in figure 1. The risk of developing any level of distress was higher in those participants who self-reported physician diagnosed asthma or chronic bronchitis.

Table 2. Regression estimates ($\hat{\beta}$); GEE-based [$s.e.(\hat{\beta})]_{Robust}$] and bootstrapped standard errors [$s.e.$ $(\hat{\beta})_{Bootstrap}$] and odds ratio (OR) and their 95% confidence interval (95% C.I.) based on ordinal logistics regression of the prevalence of mental distress (Modeling probability of high distress)

	$\hat{\beta}$ $[s.e(\hat{\beta})]_{Robust}$	$[s.e\,(\hat{\beta})]_{Bootstrap}$	$OR(95\%\ C.I.)_{GEE}$	$OR(95\%\ C.I.)_{Bootstrap}$
Intercept1	-6.07 [0.18]	-6.07 [0.19]	0.0023(0.0016,0.0033)	0.0023(0.0016,0.0034)
Intercept2	-3.52 [0.18]	-3.52 [0.19]	0.03(0.02,0.04)	0.03(0.02,0.04)
Non-Malignant Respiratory Diseases				
Asthma				
Yes	-0.06 [0.08]	-0.06 [0.08]	0.94(0.81,1.09)	0.94(0.80,1.10)
No			1.00	1.00
Chronic Bronchitis				
Yes	0.31 [0.10]	0.31 [0.10]	1.37 (1.12,1.67)	1.37 (1.12,1.66)
No			1.00	1.00
Demographic Information				
Age Group				
15-24 years	1.29[0.09]	1.29[0.09]	3.63(3.03,4.36)	3.63(3.06,4.30)
25-54 years	0.90[0.07]	0.90[0.07]	2.47(2.15,2.84)	2.47(2.15,2.85)
55-69 years	0.21 [0.08]	0.21 [0.08]	1.23(1.06,1.43)	1.23(1.05,1.43)
70 years and over			1.00	1.00
Sex				
Female	0.31 [0.11]	0.31 [0.10]	1.36(1.10,1.68)	1.36(1.11,1.66)
Male			1.00	1.00
Ethnicity				
White	-0.03 [0.08]	-0.03 [0.09]	0.97(0.82,1.15)	0.97(0.81,1.16)
Non-White			1.00	1.00
Marital Status				
Married/Common law/ Partnership	-0.37 [0.06]	-0.37 [0.06]	0.69(0.62,0.78)	0.69(0.61,0.78)
Separated/ Widowed/ Divorced	-0.01 [0.07]	-0.01 [0.07]	0.98(0.85,1.13)	0.98(0.86,1.13)
Single			1.00	1.00
Location of residence				
Rural	-0.19 [0.05]	-0.19 [0.06]	0.83(0.75,0.91)	0.83(0.74,0.93)
Urban			1.00	1.00
Geographical area				
Atlantic	-0.08 [0.06]	-0.08 [0.07]	0.92(0.81,1.05)	0.92(0.81,1.06)
British Columbia	-0.01 [0.07]	-0.01 [0.07]	0.99(0.86,1.13)	0.99(0.85,1.14)
Prairies	-0.04 [0.06]	-0.04 [0.06]	0.96(0.85,1.08)	0.96(0.86,1.08)
Quebec	0.43 [0.06]	0.43 [0.06]	1.54(1.37,1.73)	1.54(1.37,1.74)
Ontario			1.00	1.00
Immigration status				
Yes	0.12 [0.06]	0.12 [0.06]	1.12(0.99,1.28)	1.12(0.99,1.27)
No			1.00	1.00

Table 2(Cont'd)

Socio-economic status				
Education level				
Less or equal to 12 years	0.15 [0.11]	0.15 [0.12]	1.17(0.94,1.45)	1.17(0.92,1.48)
Greater than 12 years			1.00	1.00
Income level				
Low	0.56 [0.09]	0.56 [0.09]	1.74(1.46,2.08)	1.74(1.46,2.08)
Middle	0.11 [0.07]	0.11 [0.07]	1.11(0.98,1.27)	1.11(0.97,1.28)
High			1.00	1.00
Social Support				
Social Involvement Score				
Low	0.12 [0.06]	0.12 [0.06]	1.13(1.00,1.28)	1.13(1.00,1.28)
Moderate	0.18 [0.06]	0.18 [0.06]	1.20(1.06,1.35)	1.20(1.07,1.35)
High			1.00	1.00
Life-style				
Smoking Status				
Current smoker	0.33 [0.07]	0.33 [0.06]	1.39(1.22,1.58)	1.39(1.23,1.57)
Ex-Smoker	0.12 [0.05]	0.12 [0.05]	1.13(1.02,1.24)	1.13(1.02,1.24)
Non-Smoker			1.00	1.00
Household Smoking				
Yes	0.22 [0.10]	0.22 [0.11]	1.25(1.01,1.53)	1.25(1.01,1.54)
No			1.00	1.00
Health- Related:				
General Health status				
Poor	2.90 [0.18]	2.90 [0.18]	18.14(12.85,25.62)	18.14(12.71,25.89)
Fair	1.90 [0.14]	1.90 [0.13]	6.70(5.13,8.75)	6.70(5.20,8.62)
Good	1.14 [0.11]	1.14 [0.10]	3.12(2.51,3.89)	3.12(2.55,3.82)
Very Good	0.50 [0.11]	0.50 [0.11]	1.65(1.33,2.03)	1.65(1.34,2.03)
Excellent			1.00	1.00
Time point				
Cycle 5	-0.40 [0.05]	-0.40 [0.05]	0.67(0.60,0.74)	0.67(0.60,0.74)
Cycle 4	-0.53 [0.05]	-0.53 [0.05]	0.59(0.53,0.65)	0.59(0.53,0.65)
Cycle 3	-0.28 [0.04]	-0.28 [0.05]	0.76(0.69,0.83)	0.76(0.68,0.84)
Cycle 2	-0.36 [0.04]	-0.36 [0.04]	0.69(0.64,0.75)	0.69(0.63,0.76)
Cycle 1			1.00	1.00
Education*income				
12 or < 12 years*low	-0.24 [0.14]	-0.24 [0.14]	0.78(0.60,1.03)	0.78(0.60,1.03)
Overall 12 or <12 years*low income			1.03	
12 or < 12 years*middle income	-0.07 [0.12]	-0.07 [0.13]	0.93(0.74,1.17)	0.93(0.71,1.21)
Overall 12 or < 12 years*middle income			1.21	
General Health*Sex				
Poor*female	0.58 [0.21]	0.58 [0.22]	1.78(1.19,2.68)	1.78(1.15,2.76)
Overall Poor*female			43.91	
Fair*female	0.26 [0.15]	0.26 [0.16]	1.30(0.96,1.76)	1.30(0.95,1.77)
Overall Fair*female			11.85	
Good*female	0.16 [0.12]	0.16 [0.12]	1.17(0.92,1.50)	1.17(0.92,1.50)
Overall Good*female			4.96	
Very good*female	0.05 [0.12]	0.05 [0.11]	1.05(0.83,1.33)	1.05(0.84,1.31)
Overall Very Good*female			2.36	

Table 2(Cont'd)

General Health*household smoking				
Poor*yes	-0.01 [0.21]	-0.01 [0.23]	0.98(0.65,1.49)	0.98(0.63,1.54)
Overall Poor*yes			22.22	
Fair*yes	0.08 [0.14]	0.08 [0.16]	1.09(0.82,1.44)	1.09(0.80,1.45)
Overall Fair*yes			7.30	
Good*yes	-0.13 [0.11]	-0.13 [0.12]	0.88(0.70,1.10)	0.88(0.70,1.10)
Overall Good*yes			2.75	
Very good*yes	-0.18 [0.11]	-0.18 [0.11]	0.83(0.67,1.04)	0.83(0.67,1.04)
Overall Vary Good*yes			1.71	

Overall Odds Ratio= (Interaction term OR*Main effect 1 OR * Main effect 2 OR)

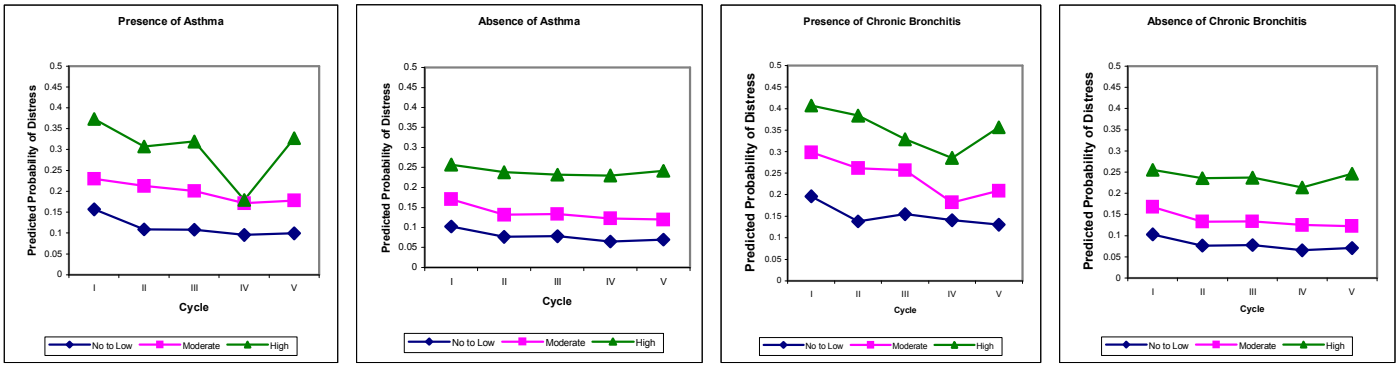

Figure 1: Predicted probability of developing mental distress over time among NPHS participants with presence/absence Asthm

5. DISCUSSION

5.1. Issues related to longitudinal complex survey data sets:

There are several issues related to complex survey data analysis. First, participants in these surveys have repeated measurements of the response variables of interest and several covariables over time, which lead to the dependent observations and similar challenges of analyzing these data as encountered in standard longitudinal studies[5]. Second, The complex multi-stage sampling designs used for these longitudinal surveys also contain cross-sectional dependencies among units (caused by inherent hierarchies in the data) in addition to the within-subject dependencies due to repeated measurements, which make the statistical analyses of these data sets intricate. Third, it is very common to have missing values in longitudinal surveys. In the NPHS, several methods were used by interviewers to trace non-respondents. Non-response was mainly due to no contact or refusal by the participant. Letters were sent, second calls were made and refusals were followed up by senior interviewers to try to convince non-responders to participate. A large number of non-responders were followed up in subsequent collection periods. A detailed description can be found in Statistics Canada documentation for longitudinal surveys[5]. Missingness is an important characteristic of longitudinal studies . In this article, statistical methods used were based on the assumption that all observations were missing at random. In an accompanied paper[26] statistical techniques which account for missing values were utilized to analyse these data. In complex surveys it is possible to have clusters of missing data, and accounting for such clusters is a complicated and an entirely different issue, which will be attempted in another manuscript. Fourth, issues related to stratification and clustering, which are characteristics of multi-stage complex surveys. Stratification decreases the variability and thus provides more precise variance estimates, while clustering increases variability and thus variance estimates are less precise. Overall, the multi-stage design has the effect of increasing variability, thus variance estimates (if not adjusted for design complexities) are less precise compared to simple random sampling. Even though there are problems for variance estimates, the main two reasons for the popularity or acceptance of complex survey designs are that these surveys are efficient for interviewing and have better coverage of the entire region of interest[27].

Around 1970, investigators started accounting for design effects in the statistical modeling[28]. Regression parameters are affected by the weights, therefore the weight variable was used to obtain consistent and valid estimates of regression parameters. Variance estimates are affected by clustering and stratification. Exact formulas for variance estimation are very complex, so approximate method (bootstrap) was used. The approximate methods are gaining popularity for cross-sectional and longitudinal complex survey data analysis, and their properties have been investigated theoretically and empirically[29]. In our report, the GEE variance estimates were similar to those obtained from using the bootstrap method, which supports the following statement: '*Design-based approach reduces to the Liang-Zeger* "*Sandwich" estimator for longitudinal samples when the longitudinal units are independent*'[29]. As suggested by Binder and Roberts: even though these two sets of variance estimates are similar, it is more appropriate to use bootstrap variance estimates, bootstrap estimates were used for the purposes of formulating inferences.

5.2. Association between respiratory diseases and mental distress:

In the present paper, we evaluated the longitudinal relationship between the presence of respiratory diseases (asthma and chronic bronchitis) and mental distress among Canadian NPHS participants who self-reported physician diagnosed asthma or chronic bronchitis, studied over a 10 year period (1994/1995-2001/2003). Our analysis showed that there is a positive association between the physician diagnosed asthma or chronic bronchitis and an increased prevalence of mental distress. In 1999, one study projected that depression will be the second leading contributor to of the overall burden of illness in 2020, following ischaemic heart disease[15]. Stephens et al. published a comprehensive report on the mental health of the Canadian population[16]. Several studies[13,14,30-32] have reported that patients with bronchial asthma have higher than expected levels of psychiatric morbidity and our results support these findings. Dales et al[33] found a strong positive association between respiratory symptoms and psychological status indicators. In 1994, a study by Janson et al[12] supported the findings by Dales et al with respect to respiratory symptoms but did not find an association between a diagnosis of asthma or objective asthma-related measurements and anxiety and depression. Of those who self-reported physician diagnosed asthma were significantly at a risk of having high-level mental distress compared to those who reported no asthma .This difference was not significant when adjusted for other covariates. Similarly, NPHS participants who self-reported a physician diagnosed chronic bronchitis were significantly at a high risk of having high-level of mental distress compared to those who reported no chronic bronchitis and this difference remained significant after adjusting for the other covariates. As reported by the World Health Organization, the relationship between poverty (defined as: lack of money or material possessions) and mental health is multifaceted[15]. Our data showed that subjects

with low income and low education were at higher risk compared to those who had high education and high income. It is hard to explain the decreasing trend over time (figure 1) in the predicted probabilities of developing distress for those who self- reported a physician diagnosed asthma or chronic bronchitis. This needs further investigation of comparison between participants who stayed-in to those who dropped out from the survey.

ACKNOWLEDGEMENTS

Authors would like to thank the Remote Data Access (RDA) services of Statistics Canada. After successful execution of SAS syntax on the dummy data provided by Statistics Canada, we sent our syntax to RDA office to run it on the original master data file and they provided us with the results.

REFERENCES

Beland, Y. and MacNabb, H. Population Health Surveys Bootstrap Hands-on Workshop. data.library.ubc.ca/rdc/other/0702Hands_on.ppt

Binder, D.A. and Roberts, G.A. (2003). Statistical inference in survey data analysis: where does the sample design fit in? http://socserv.socsci.mcmaster.ca/rdc2003/binderoberts.pdf

Dales, R.E., Spitzer, W.O., Schecter, M.T. and Suissa, S.(1989) The influence of psychological status on respiratory symptoms reporting. *Am Rev. Respir Dis*. **139**, 1459-1463.

Demnati, A. and Rao, J.N.K (2004). Linearization Variance Estimators for Survey Data. *Survey Methodology*, **2004**; 138-143

Diggle, P., Lang, K., Zeger, S. (1995). *Analysis of Longitudinal Data.* Oxford Science Publication.

Hosmer, D.W and Lemshow, S. (1989) Applied Logistic Regression. A Wiley-Interscience Publication, John Wiley and Sons Inc., Canada, 1989; 82-134.

Karunanayke, C., Pahwa, P., McDuffie, H.H. (2006). Statistical Modeling of Mental Distress among the National Population Health Survey Participants Assessed Longitudinally: Missing Data Analysis. *(To appear in the SSC-2006 proceedings)*

Kessler, R.C. and Morcyzk, . http://www.mentalhealth.org/publications/allpubs/SMA04-3938/Chapter12.asp.

Kim, Ji-H http://stat.soongsil.ac.kr/~jhkim/Publication/stat&prob_2003.pdf. Assessing Practical Signifcance of the Proportional Odds Assumption.

Kish, L. Multipurpose Sample Design. *Survey Methdology*. **14**, 19-32.

Lam M. Personal Communication.

Liang, K-Y and Zeger Sl. (1986). Longitudinal data analysis using generalized linear models. *Biometrika.* 1986; 73:13-22.

Liu, I and Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. Socieddad de Estadistica e Investigacion Operativa Test, 14(1): 1-73.

Lyketsos, C.G., Lyketsos, G.C., Richardson, S.C. and Beis, A. (1986) Dsythymic states and depression syndrome in physical conditions of presumably psychogenic origin. *Acta Psychiatr Scand*. **76**: 529-534.

Molenberghs G and Verbeke G. Models for Discrete Longitudinal Data. 2005. Springer Science and Business Media Inc.

Ng, E., Altman, B. and Berthelot, J-M. (2006). Racial differences in HUI-based disability using the 2003 Joint Canada/United States Survey of Health: a cross-national comparison. http://paa2006.princeton.edu/download.aspx?submissionId=60679.

Oswald, N.C., Waller, R.E. and Drinkwater, J.(1970) Relationship between breathless and anxiety in asthma and bronchitis; a comparative study. *British Med. Jou*. **2,** 14-17.

Rao, J.N.K. (2005) Interplay between sample survey theory and practice: An appraisal. Survey Methodology, **31**(2):117-138.

Rust, K., and Rao, J.N.K. Variance Estimation for Complex Estimators in Sample. Surveys. Statistics in Medical Research, 1996; **5**, 381-397.

Sroujian C. (2003). Mental Health is the number one cause of disability in Canada. The Insurance Journal, 2003; August: 8.

Statistics Canada. 1994-1995 National Population Health Survey Public Use Microdata Documentation. 1995

Statistics Canada. 1996-1997 National Population Health Survey Public Use Microdata Documentation. 1997

Statistics Canada. 1998-1999 National Population Health Survey Public Use Microdata Documentation. 1999

Statistics Canada. 2000-2001 National Population Health Survey Public Use Microdata Documentation. 2001.

Statistics Canada (2002). Documentation of the Longitudinal Component of the National Population Health Survey.

Stokes, E.M., Davis, C.S. and Koch, G.G. *Categorical Data Analysis Using The SAS System (2000).*, SAS Inst. Inc.

Tambay, J-L and Catlin, G. (1995) Sample design of the National Population Health Survey. *Health Reports*. **1995**;7:29-38.

Wang, J. and Nady, E-G.(2004) Socio-demographic factors associated with co-morbid major depressive episodes and alcohol dependence in the general population. *Canadian J of Psychiatry.* **49**(1):37-44.

World Health Organization [WHO]. (2001). The World Health Report 2001 Mental Health: New Understanding, New Hope. Geneva: WHO.

Yeo, D., Mantel, H. and Liu T-P. (1999) Bootstrap variance estimation for the National Population Health Survey., http://www.amstat.org/Sections/Srms/Proceedings/papers/1999_136.pdf

Yellowless, P.M., Haynes, S., Potts, N. and Ruffin, R.E. (1983) Psychiatric morbidity in patients with life-threatening asthma: Initial report of a controlled study. *Med J. Aust*. **67**: 361-370.

Zeger Sl and Liang K-Y (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics,* **42**:121-130.