DESIGN-BASED VERSUS MODEL-BASED METHODS: A COMPARATIVE STUDY USING LONGITUDINAL SURVEY DATA

S. Ghosh1[1], P. Pahwa[1, 2]

ABSTRACT

Survey data analysis using complex sampling designs ought to account for clustering, stratification and unequal probability of selection. Design-based and model-based methods are two commonly used routes taken to account for such survey designs. Several studies of cross-sectional survey designs have shown that these two approaches provide similar results when the model fits the data well. The present paper aims at comparing these two approaches using the longitudinal National Population Health Survey (NPHS) dataset. The NPHS is an ongoing longitudinal study, for which data collection was based on a stratified multi-stage sampling design. A marginal modeling approach proposed by Rao (1998) was used by way of a design-based method. The Generalized Estimating Equation (GEE) method, proposed by Liang and Zeger (1986), was used as a typical model-based approach. Results obtained from these methods were compared.

KEY WORDS: Multistage sampling, GEE, survey GEE, Taylor linearization, bootstrap, NPHS.

RÉSUMÉ

L'analyse des données d'enquêtes utilisant des plans complexes se doit de tenir compte de l'utilisation de grappes, de la stratification et des probabilités inégales de sélection. Les méthodes basées sur le plan et basées sur un modèle sont les deux principales voies utilisées pour tenir compte de tels plans de sondage. Plusieurs études des plans de sondage transversaux ont démontré que ces deux approches donnent des résultats similaires lorsque le modèle s'ajuste bien aux données. Le présent article a pour but de comparer ces deux approches en utilisant le jeu de données longitudinales de l'Enquête nationale sur la santé de la population (ENSP). L'ENSP est une enquête longitudinale continue pour laquelle la collecte de données utilise un plan de sondage stratifié et à plusieurs degrés. Une approche de modélisation marginale proposée par Rao (1998) a été utilisée comme méthode basée sur le plan de sondage. La méthode des équations d'estimation généralisées (ÉEG), proposée par Liang et Zeger (1986) a été utilisée comme approche typique basée sur un modèle. Les résultats obtenus à l'aide de ces deux méthodes ont été comparés.

KEY WORDS: Échantillonage; Enquête nationale sur la santé de la population; équations d'estimation généralisées; équations d'estimation d'enquête généralisées; Linéarisation de Taylor; Bootstrap.

1. INTRODUCTION

During the past few decades there has been a growing interest in analyzing longitudinal survey data. Data collection for these large surveys uses a complex sampling design. The growing interest in this field is mainly due to the fact that the results obtained from these surveys can be easily generalized to the target population under study, if analyzed correctly. There is a large literature available on methods for modeling longitudinal binary data when collected using simple random sampling approach. The Generalized Estimating Equations approach developed by Liang and Zeger (1986) has gained popularity during the last two decades. There are two main reasons for its popularity. First, it is an extension of the generalized linear model, and can easily handle continuous, poisson, survival as well as discrete data obtained from longitudinal studies. Second, implementation of GEE approach in the standard commercial softwares, such as SAS, STATA, and SUDAAN. However, statistical methods to analyze data from longitudinal multi-stage sampling designs are still developing. The primary objective of analyzing survey data is to make inferences about the population of interest (LaVange, Koch et al. 2001). However, to obtain correct estimates and inferences the sampling design should be taken into account (Feder, Nathan et al. 2000). Three key features of survey data are clustering, stratification and unequal selection probabilities. In longitudinal survey data one has to account for (i) clustering arising due to the fact that individuals belonging to the same cluster are correlated, (ii) within subject correlation due to repeated measurements on the same subject makes the analysis of longitudinal survey data becomes complex. Considering both the complexity of

[1]Sunita Ghosh, 11560 University Ave., Edmonton, T6J4M1, Canada Email: Sunita.ghosh@usask.ca
[2]Punam Pahwa, Community Health and Epidemiology, Saskatoon, Canada

design and longitudinal nature of the data at the analysis stage to a large extent is an ongoing research (Feder, Nathan et al. 2000). Some of the recent and important contributions are Feder et al (Feder, Nathan et al. 2000), Rao (Rao 1998), Skinner and Holmes (Skinner and Holmes 2003), Lawless (Lawless 2003). Skinner and Holmes (2003) proposed a random effect modeling for continuous data, Lawless (2003) proposed event history approach for discrete data, Feder et al (2000) used multilevel modeling with time varying random effects approach, and Rao (1998) proposed marginal modeling for discrete data approach.

To account for the complexities of the survey data, three approaches are proposed in literature: model based, design based and model assisted methods (Lehtonen and Pahkinen 2004). Each of the methods has their own advantages and disadvantages. When we use model based methods, the clustering effect reduces the statistical efficiency. For design based methods, the weighted estimators require large sample sizes to ensure sufficiently valid results (Kalton 1983). When the design is "informative", i.e., the outcome variable is correlated with the design variables as well as those not included in the model, the standard errors of the estimates can be severely biased (Feder et al, 2000). These were some of the drawbacks of using model based and design based approaches. In this paper, we primarily focus on comparing the GEE approach (model based method), proposed by Liang and Zeger (1986) and the design-based marginal modeling approach proposed by Rao (1998). We use Taylor series linearization to calculate the standard errors of regression estimates.

Section 2 provides an overview of the National Population Health Survey (NPHS) study design and Section 3 discusses the model based and design based methods to be used for the analysis. The application to the NPHS dataset and the results are discussed in section 4.

2. OVERVIEW OF NATIONAL POPULATION HEALTH SURVEY STUDY DESIGN

The National Population Health Survey (NPHS) household component is an ongoing longitudinal survey, designed to collect "longitudinal" information on the health of the Canadian population and related socio-demographic information. The first cycle of data collection took place in 1994-1995. The household component of the NPHS has completed five so-called Cycles: the data collection for the first Cycle started in 1994-95 and then every two years until 2002-03. This survey collects data every second year and will continue until 2014. The target population for the NPHS was initially 19,600 households, with a minimum of 1200 households per province. The population of the household component includes household residents in all provinces in 1994-95, and this does not include Indian reserves and Crown lands, health institution, Canadian Forced Bases and some remote areas in Ontario and Quebec. The survey collects data on health status, use of health services, chronic conditions, socio-demographic information such as age, sex, education, household income and labour force status and also on activity restrictions. The longitudinal sample is composed of 17,276 individuals that were selected in Cycle1, no new panel members are included in the study. In Cycle5 all longitudinal panels were 8 years old and over. A detailed description of the survey can be found elsewhere (2004).

The sampling procedure of the household component of the NPHS was based on a multi-stage sampling design. In all the provinces except Quebec this sampling scheme was maintained. In the first stage homogeneous strata were formed by dividing each province into three types of areas namely, major urban centers, urban town and rural areas and based on these separate geographic and/or socio-economic strata were formed. Independent samples of clusters (heterogeneous) were selected with *probability proportional to size* (PPS) from each stratum. PPS is a sampling technique, commonly used in multi-stage cluster sampling, in which the *probability* that a particular sampling unit will be selected in the sample is *proportional* to some known variable (e.g., in a population survey, the population *size* of the sampling unit). PPS is useful when populations of sampling units vary in size, for example population is different for clusters within the same strata or groups, and when units *do not* have the same probability of selection (unequal weights). In the next stage, a dwelling list was prepared for each cluster chosen and from this list households were selected. The country was divided into 1000 strata and approximately 3000 clusters were formed which are the primary sampling units. Within each cluster, dwelling were selected which comprised the secondary sampling units and finally one individual was selected from each household producing the tertiary sampling units. In Quebec, the households were selected based on a two-stage sampling scheme. The province was geographically subdivided into four urban density classes: Montreal Census Metropolitan Area, regional capitals, small urban agglomerations and the rural sector. Clusters were stratified based on socio-economic characteristics. Random samples of dwelling were drawn from each cluster.

The survey weights used in the longitudinal household component of the NPHS are adjusted such that these weights reflect the probability of selecting the individuals at Cycle1 (represents the population of 1994-95) and not subsequent Cycles. In addition, the weights are also adjusted for the non-response and post-stratification features. Post-stratification weights are calculated by further post-stratifying Cycle 1 stripped weight to the 1994-1995 population estimates based on 1996 Census counts by age group (0-11, 12-24, 25-44, 45-64, 65 and older) and sex within each province.
The post-stratification adjustment is given by:

$$\frac{\text{Population estimate in a province/ age/ sex category}}{\text{Sum of ``stripped'' weights of respondent household members in a province/ age/ sex category}}$$

3. MODELING METHODS

3.1 Description of the variables included from the dataset

The present analysis restricts attention to the adult female Canadian population, aged 18 to 64 years in Cycle 1 (1994-95). We have included only non-pregnant females in this study. The following variable were included for analysis: age groups (18-29 years, 30-49 years, 50-64 years and 65-72 years-reference), body mass index (underweight, normal weight-reference, over weight and obese), location (rural and urban-reference), ethnicity (white and non-white-reference), smoking status (smokers, ex-smokers and non-smokers-reference), exposure to second hand smoke (yes and no), immigration status (yes and no), socio-economic status (low-reference, middle and high income), time (Cycle 1-reference, Cycle 2, Cycle 3, Cycle 4 and Cycle 5), province regrouped into five regions (with Ontario as the reference category), physician diagnosed other allergies, food allergies, chronic bronchitis/emphysema and intestinal problems were all recoded into yes and no categories. No was the reference category for all the yes/no dichotomous variables. The prevalence rates and 95% confidence intervals for all the five Cycles were calculated using BOOTVAR program provided by Statistics Canada (2004). Variance estimation for the multi-stage sampling design is very complex, so an approximation method is required to compute variance. The bootstrap method is a computer intensive method for variance estimation and is required to calculate the standard errors in order to avoid serious underestimation of standard errors. The BOOTVAR program used 500 bootstrap replicates to calculate the standard errors and it accounts for the complexity of survey design.

3.2 Logistic models for asthma prevalence

In the present analysis, physician- diagnosed asthma a dichotomous variable was the outcome variable. The response variable for the present analysis was presence or absence of self reported physician diagnosed asthma, which is a binary variable. Sex, age-groups (18-29, 30-49, 50-64 and 65-72 years), ten provinces of Canada regrouped into five regions (Atlantic, Quebec, Prairies, British Columbia and Ontario), location of residence (rural and urban), presence or absence of food allergies, other allergies, intestinal disorders, emphysema or chronic bronchitis (these variables were based on any long term condition diagnosed by a health professional), body mass index (under weight, normal weight, over weight, and obese), ethnicity (white and non-white), smoking status, exposure to second hand smoke, immigration status (yes and no), socio-economic status (low-reference, middle and high income), time (all five cycles of the survey) were considered as the possible covariates/risk factors for asthma prevalence. These variables were included in the final model based on standard model building strategies (chosen at $p < 0.25$ significance level). Other variables like location, ethnicity, smoking status and exposure to second hand smoke were included in the model even if not statistically significant at $p < 0.25$ level because of their clinical significance. Some of the significant interaction terms were also included in the final model. Proc GENMOD was used to obtain the model based estimates and standard errors, the BOOTVAR macro was used to compute the prevalence rate of asthma. A macro in SAS was used to obtain the estimates and standard error when accounting for the complex features of longitudinal survey design.

3.3 Statistical Methods

Model-based and design-based approaches were compared. A model-based approach does not account for the complex survey structure and first calculates the estimates under the erroneous assumption of simple random sampling. The design based approach takes into account the clustering and stratification arising from the survey design. In the present analysis

we compare the model-based and the design-based techniques. The prevalence rates and 95% confidence intervals were calculated using a macro developed by Statistics Canada for each of the five Cycles. This method is a design based technique as it accounts for all the three features of the survey design. Finally the model based and the design based methods will be compared. The generalized estimating equations approach developed by Liang and Zeger (1986) (model based methods) and marginal modelling approach proposed by Rao (1998) (design-based method) which calculates the standard error using Taylor linearization method were used. The following section discusses the methods used in this paper, and we will discuss them in turn.

3.4 Models

3.4.1 Generalized Estimating Equation (GEE)

Consider Y_{it} a dichotomous outcome variable which assumes the logit model for the first order marginal probabilities

logit $[\Pr(Y_{ij}=1)]$ = logit $\mu_{ij} = \log(\frac{\mu_{it}}{1-\mu_{it}}) = \beta_s^T x_{is} + \beta_t^T x_{it}$, …… eq (3.4.1)

$t = 1,\ldots\ldots,T$(occasion) and $i= 1,\ldots\ldots,M$ (subjects), β_s^T is a vector of stationary covariates, and β_t^T is a vector of time varying covariates

$\mu_{it} = \frac{\exp\{\beta_s' x_{is} + \beta_t' x_{it})}{1+\exp\{\beta_s' x_{is} + \beta_t' x_{it}\}}$, where x_{is} = design-matrix of time stationary covariates and x_{it} = design-matrix of time varying covariates

Liang and Zeger (1986) and Zeger and Liang (1986) proposed generalized estimating equations. A set of score equations for a marginal normal model is given by

$$U(\beta) = \sum_{i=1}^{N} D_i^T (\Delta_i^{1\backslash 2} \Re_i(\alpha) \Delta_i^{1\backslash 2})^{-1} (y_i - \mu_i) = 0, \qquad \ldots\ldots\ldots\text{eq (3.4.2)}$$

where $D_i = \partial\mu_i / \partial\beta^T$ and μ_i is the mean function, and V_i is a working covariance matrix of outcome variable $Y_i = (Y_{i1,\ldots\ldots,} Y_{ij})^T$ a t x 1 vector of i=1,……,M individuals observed at t occasions, $X_i = (X_{i1},\ldots\ldots,X_{ij})^T$ is t X P matrix of covariates for individual i. The working covariance structure can be decomposed into $V_i = \Delta_i^{1\backslash 2} \Re_i(\alpha) \Delta_i^{1\backslash 2}$, where Δ_i=diag [var(Y_{i1}),…….,var(Y_{ij})], and $\Re_i(\alpha)$=corr (Y_i) is a T X T "working" correlation matrix and α is a vector of parameters which are usually associated with a specified model for corr (Y_i) (Fitzmaurice and Laird 1993). The above equations reduce to independence equations if $\Re_i(\alpha)$ is the identity matrix.

GEE properly accounts for within-subject correlation which results in consistent estimators. Efficiency increases when the assumed correlation structure is closer to the true correlation structure. This method uses quasi-likelihood approach for discrete and continuous longitudinal data. Quasi-likelihood was proposed by Wedderburn (Wedderburn 1974) and was extended to multivariate quasi-likelihood by McCullagh and Nelder (McCullagh and Nelder 1989). Liang and Zeger (Liang and Zeger 1986)and Zeger and Liang (Zeger and Liang 1986) extended the univariate quasi-likelihood approach to include correlation arising from hierarchies in the data.

The main inference is on the model-based coefficients, while the intra-cluster dependence is merely a nuisance characteristic, merely accounted for, but not subject to modeling in the classical sense. It can be used for Gaussian and non-Gaussian outcomes alike (Zeger and Liang 1986). GEE provides consistent regression coefficient β even under minimal assumption about the time dependence (Zeger and Liang 1986).

3.4.2 Marginal Modeling approach accounting for survey design [3]

Consider a longitudinal study with T occasion of measurement and the finite longitudinal population of size *M* is clustered into *N* primary sampling units also known as primary sampling units (psu). The subscript i in equation 3.4.1 is changed to

3 This section was modified from the notes of Suzanne Rubin-Bluer from an Internal report, January 2006, Statistics Canada

hik in survey data, where *h* represents strata, *i* represents clusters and *k* represents subject. For each stratum *h*, N_h and M_{hi} are respectively the number of clusters in stratum h and the number of secondary units in the cluster *hi*, *i = 1,……., N_h* and *h = 1,……..,L*

Assume the same logit model for the first order marginal probabilities as in equation 3.4.1

The Survey independent estimating equation (IEE) estimator are given by

$$\hat{U}_{IEE}(\beta) = \sum_{hik \in S_l} \omega_{hik} D_{hik}^{'} A_{hik}^{-1} (y_{hik} - \mu_{hik}) = 0 \qquad \text{…….. eq (3.4.3)}$$

S_l is the longitudinal sample and ω_{hik} is the longitudinal weights.

To calculate the survey IEE estimator $\hat{\beta}_{IEE}$ we do the iteration

$$\hat{\beta}_{IEE}(K) = \hat{\beta}_{IEE}(K-1) - \left(\frac{\partial \hat{U}_{IEE}}{\partial \beta} \right)^{-1} \left(\beta_{IEE}(\hat{K-1}) \right) . \hat{U}_{IEE} \left(\hat{\beta}_{IEE}(K-1) \right) \text{ eq (3.4.4)}$$

Where $\hat{U}_{IEE}$ is the survey estimate defined above and $\left(\frac{\partial \hat{U}_{IEE}}{\partial \beta} \right)$ is replaced by its expectation:

$$\left(\frac{\partial \hat{U}_{IEE}}{\partial \beta} \right) \approx \sum_{hik \in S_l} \omega_{hik} D_{hik}^{'} A_{hik}^{-1} D_{hik}$$

Where $A_{hik}^{-1} = Diag\left(\frac{1}{\mu_{hik1} - \mu_{hik1}^2}, ..., \frac{1}{\mu_{hik4} - \mu_{hik4}^2} \right)$

The Survey Generalized Estimating Equation (GEE) estimator proposed by Rao (1998) is of the form:

$$\hat{U}_{GEE}(\beta) = \sum_{hik \in S_l} \omega_{hik} D_{hik}^{'}(\beta) \Delta_{hik}^{-1/2}(\beta) \hat{R}^{-1} \Delta_{hik}^{-1/2}(\beta)(y_{hik} - \mu_{hik}(\beta)) = 0 \text{ ….eq (3.4.5)}$$

Where the matrix of "correlation" $\hat{R}$ now has the form $\hat{R} = (r_{tu})_{tu}$: with

$$r_{tu} = \sum_{hik \in S_l} \omega_{hik} \frac{\left(y_{hikt} - \mu_{hikt}\left(\hat{\beta}_{IEE} \right) \right)\left(y_{hiku} - \mu_{hiku}\left(\hat{\beta}_{IEE} \right) \right)}{\sqrt{\mu_{it}\left(\hat{\beta}_{IEE} \right) - \mu_{it}^2\left(\hat{\beta}_{IEE} \right)} \sqrt{\mu_{iu}\left(\hat{\beta}_{IEE} \right) - \mu_{iu}^2\left(\hat{\beta}_{IEE} \right)}}$$

where $\sum_{hik \in S_l} \omega_{hik}$, t and u = 1, ……..,5

The estimator $\hat{\beta}_{GEE}$ is defined as the solution of the survey GEE.

$\hat{\beta}_{GEE}$ is calculated through iteration, where the $\hat{\beta}_{GEE}$ (K-1) change at each iterations, but $\hat{R} = (r_{tu})_{tu}$ is fixed throughout the iterations to calculate $\hat{\beta}_{GEE}$

The variance matrix of $\hat{\beta}_{GEE}$ can be consistently estimated by

$$v\left(\hat{\beta}_{GEE} \right) = \hat{J}_G^{-1}\left(\hat{\beta}_{GEE} \right) . v\left(\hat{U}_{GEE} \right) . \hat{J}_G^{-1}\left(\hat{\beta}_{GEE} \right) \qquad \text{……..eq (3.4.4)}$$

evaluated at $\beta = \hat{\beta}_{GEE}$ with

$$\hat{J}_G(\beta) = -\sum_{hik \in S_l} \omega_{hik} D'_{hik}(\beta) A^{-1}_{hik}(\beta) D_{hik}(\beta) \text{ and } \nu\left(\hat{U}_{GEE}\right), \quad \text{.....eq (3.4.5)}$$

evaluated at $\beta = \hat{\beta}_{GEE}$, is the survey design variances of a survey total and can be estimated by bootstrap, calculating for each one of the 500 sets of bootstrap weights estimated.

$$\hat{U}_{GEE}(b)\left(\hat{\beta}_{GEE}\right) = \sum_{hik \in S_l} \omega_{hik}(b) D'_{hik} \Delta^{-1/2}_{hik} \hat{R}^{-1} \Delta^{-1/2}_{hik} (y_{hik} - \mu_{hik}) \text{ ..eq (3.4.6)}$$

b= 1,.., 500

And then calculate:

$$\nu\left(\sqrt{n}\,\hat{\mathbf{U}}_{GEE}\right) = n \frac{1}{500} \sum_{b=1}^{500} \left(\hat{U}^{(b)}_{GEE} - \overline{\hat{U}}^{(b)}_{GEE}\right)\left(\hat{U}^{(b)}_{GEE} - \overline{\hat{U}}^{(b)}_{GEE}\right)'$$

4. RESULTS

4.1 Logistic modeling of asthma prevalence

This study focuses on comparing the model-based methods with the design-based methods when we have longitudinal survey data. Another objective of the study was to determine the major risk factors of asthma in adult Canadian female population.

Table 1 presents the prevalence rate and 95% confidence interval of asthma in adult females in the age range 18-64 years. Table 2 provides the estimates (standard errors) and odds ratio obtained from using the model based approach, GEE proposed by Liang and Zeger (1986). The independence and the exchangeable working correlation matrix were used and compared with the estimates (standard errors) and odds ratio obtained from Rao's (1998) marginal GEE approach (Table 3).

Table 2, shows that the estimates and standard errors when using independence and exchangeable working correlation matrix are very similar for some of the variables whereas for few variables are slightly different. Table 3, also showed similar trend and in some cases it showed very different standard errors. Next, we compared the model based independence correlation matrix with the design based methods. Both these methods provided similar estimates and odds ratio, however the standard errors were larger when using the design based methods. Only for very few variables the standard errors were similar. Similar results were obtained when we compared the exchangeable correlation matrix of design based and model based methods.

The estimates and odds ratio of the independence and exchangeable correlation matrix for model based and design based methods are different, this could be accounted to the different working correlation matrix used. The independence correlation matrix makes the working assumptions that the observations are independent, and then properly corrects the standard errors. The result shows that self reported physician diagnosed food allergy, other kinds of allergy, bronchitis and ulcer were positively associated with asthma. The odds of asthma prevalence were higher in these groups when compared to reference category. Obese females were at higher risk of developing asthma compared to normal weight females. The odds of developing asthma were lower in the immigrant population compared to the individuals born in Canada. The risk of developing asthma was higher in smokers and ex-smokers staying in rural location compared to non-smokers in urban locations, higher socio-economic status individuals staying in rural location. Asthma prevalence was higher amongst smokers and ex-smokers in all age groups compared to non-smokers females in the age group 65-72 years. However, it was negatively associated with socio-economic status amongst all three age groups.

Table 1: Self reported physician diagnosed asthma prevalence rate and 95% confidence interval of Canadian females in the age group 18-64 years

Cycles	Prevalence rate (%)	95 % Confidence Interval
1994-95 (Cycle 1)	6.2	5.5-7.0
1996-97 (Cycle 2)	7.2	6.4-8.0
1998-99 (Cycle 3)	7.3	6.5-8.2
2000-01 (Cycle 4)	7.1	6.3-7.9
2002-02 (Cycle 5)	6.9	6.1-7.7

Table 2: Estimates (Standard Errors) and Odds Ratio using Generalized Estimating equation proposed by Liang and Zeger

Covariates	Independence correlation matrix		Exchangeable Correlation matrix	
	Estimate (S.E.)	Odds Ratio	Estimate (S.E.)	Odds Ratio
BMI (Normal weight)				
Under weight	-0.46 (0.34)	0.63	-0.42 (0.30)	0.66
Over weight	0.27 (0.14)	1.30	0.19 (0.13)	1.21
Obese	0.60 (0.15)	1.82***	0.53 (0.14)	1.69***
Food Allergy (No)				
Yes	0.80 (0.12)	2.23***	0.32 (0.09)	1.38***
Other Allergy (No)				
Yes	1.38 (0.09)	3.98***	0.50 (0.06)	1.66***
Bronchitis (No)				
Yes	1.90 (0.15)	6.70***	0.69 (0.14)	2.00***
Intestinal Problem (No)				
Yes	0.59 (0.19)	1.81**	0.24 (0.12)	1.27*
Socio-economic status (Low Income)				
High Income	1.19 (0.80)	3.03	1.29 (0.58)	3.63
Middle Income	-0.85 (0.49)	0.43	-0.31 (0.27)	0.73
Location (Urban)				
Rural	-1.18 (0.30)	0.31	-0.59 (0.21)	0.55
Age Group (65-72 years)				
18-29 years	0.16 (0.46)	1.18	0.38 (0.32)	1.46
30-49 years	0.13 (0.41)	1.14	0.41 (0.28)	1.51
50-64 years	-0.51 (0.40)	0.60	0.04 (0.25)	1.04
Region (Ontario)				
Atlantic	-0.33 (0.16)	0.72*	-0.29 (0.13)	0.75*
Quebec	-0.05 (0.16)	0.95	-0.18 (0.14)	0.83
Prairies	-0.27 (0.15)	0.76	-0.24 (0.13)	0.79
British Columbia	0.28 (0.18)	1.32	0.10 (0.16)	1.11
Immigration (Citizen)				
Immigrants	-0.55 (0.22)	0.58*	-0.81 (0.19)	0.44
Ethnicity (Non-white)				
White	0.40 (0.32)	1.50	0.19 (0.30)	1.21
Smoking Status (Non-Smokers)				
Current Smokers	-1.94 (0.59)	0.14	-0.48 (0.31)	0.61
Ex-Smokers	-0.56 (0.41)	0.57	-0.34 (0.27)	0.71
Second hand smoke (No)				
Yes	0.08 (0.17)	1.08	-0.04 (0.11)	0.96
Location * Smoking				
Rural Smokers	0.87 (0.29)	2.40**	0.43 (0.16)	1.54**
Rural Ex-Smokers	0.92 (0.27)	2.51***	0.36 (0.18)	1.43*
Location * Income				

Table 2 (cont'd)

High Income Rural	0.89 (0.32)	2.43**	0.56 (0.23)	1.75*
Middle Income Rural	0.36 (0.27)	1.44	0.28 (0.18)	1.32
Ethnicity * Income				
White* High SES	-0.90 (0.57)	0.40	-0.72 (0.44)	0.49
White * Middle SES	0.79 (0.39)	2.21*	0.46 (0.23)	1.59*
Second Hand Smoke * Time				
Exposure * Cycle 2	0.03 (0.21)	1.03	0.06 (0.15)	1.06
Exposure * Cycle 3	0.004 (0.19)	1.00	0.11 (0.13)	1.12
Exposure * Cycle 4	0.06 (0.17)	1.10	0.11 (0.13)	1.12
Exposure * Cycle 5	0.19 (0.14)	1.21	0.27 (0.11)	1.31*
Smoking * Age Group				
Smoker * 18-29 years	2.35 (0.64)	10.48***	0.80 (0.35)	2.23*
Ex-Smoker * 18-29 years	0.44 (0.47)	1.55	0.44 (0.31)	1.55
Smoker * 30-49 years	1.71 (0.61)	5.54**	0.29 (0.32)	1.34
Ex-Smoker * 30-49 years	0.53 (0.43)	1.71	0.37 (0.29)	1.44
Smoker * 50-64 years	1.96 (0.59)	7.14***	0.46 (0.28)	1.58
Ex-Smoker * 50-64 years	1.06 (0.42)	2.88*	0.38 (0.26)	1.46
Age Group * Income				
18-29 years * High SES	-0.91 (0.56)	0.40	-0.79 (0.34)	0.45*
18-29 years * Middle SES	-0.16 (0.41)	0.85	-0.30 (0.22)	0.74
30-49 years * High SES	-0.92 (0.53)	0.40	-0.97 (0.31)	0.38*
30-49 years * Middle SES	-0.46 (0.38)	0.63	-0.41 (0.18)	0.66*
50-64 years * High SES	-0.66 (0.51)	0.52	-0.68 (0.28)	0.50*
50-64 years * Middle SES	-0.27 (0.37)	0.76	-0.19 (0.19)	0.83

Table 3: Estimate (Standard Error) and Odds Ratio taking into account Multi-stage sampling and repeated measurements using Rao's marginal GEE method

Covariates	Independence correlation matrix		Exchangeable Correlation matrix	
	Estimate (S.E.)	Odds Ratio	Estimate (S.E.)	Odds Ratio
BMI (Normal weight)				
Under weight	-0.46 (0.35)	0.63	-0.42 (0.27)	0.66
Over weight	0.27 (0.14)	1.30	0.21 (0.13)	1.23
Obese	0.60 (0.28)	1.82*	0.55 (0.29)	1.74
Food Allergy (No)				
Yes	0.80 (0.16)	2.23***	0.39 (0.11)	1.48***
Other Allergy (No)				
Yes	1.38 (0.18)	3.98***	0.60 (0.14)	1.83***
Bronchitis (No)				
Yes	1.90 (0.19)	6.70***	0.80 (0.30)	2.23**
Intestinal Problem (No)				
Yes	0.59 (0.23)	1.81**	0.27 (0.16)	1.32
Socio-economic status (Low Income)				
High Income	1.19 (1.00)	3.30	1.31 (0.53)	3.72*
Middle Income	-0.85 (0.69)	0.43	-0.37 (0.40)	0.69
Location (Urban)				
Rural	-1.18 (0.45)	0.31	-0.66 (0.25)	0.52
Age Group (65-72 years)				
18-29 years	0.16 (0.81)	1.18	0.43 (0.44)	1.54
30-49 years	0.13 (0.62)	1.14	0.46 (0.39)	1.58
50-64 years	-0.51 (0.69)	0.60	0.03 (0.31)	1.04
Region (Ontario)				

Table 3 (cont'd)

Atlantic	-0.33 (0.27)	0.72	-0.31 (0.32)	0.73
Quebec	-0.05 (0.25)	0.95	-0.16 (0.18)	0.85
Prairies	-0.27 (0.21)	0.76	-0.24 (0.20)	0.79
British Columbia	0.28 (0.24)	1.32	0.15 (0.18)	1.16
Immigration (Citizen)				
Immigrants	-0.55 (0.30)	0.58	-0.76 (0.30)	0.47*
Ethnicity (Non-white)				
White	0.40 (0.57)	1.50	0.22 (0.36)	1.25
Smoking Status (Non-Smokers)				
Current Smokers	-1.94 (2.26)	0.14	-0.57 (0.69)	0.57
Ex-Smokers	-0.56 (0.53)	0.57	-0.33 (0.36)	0.72
Second hand smoke (No)				
Yes	0.08 (0.18)	1.08	-0.03 (0.20)	0.97
Time (Cycle 1)				
Cycle 5	0.26 (0.14)	1.30	0.37 (0.14)	1.45
Cycle 4	0.31 (0.15	1.37	0.39 (0.15)	1.48
Cycle 3	0.28 (0.17)	1.32	0.32 (0.14)	1.38
Cycle 2	0.07 (0.12)	1.07	0.10 (0.10)	1.11
Location * Smoking				
Rural Smokers	0.87 (0.40)	2.40	0.48 (0.22)	1.61*
Rural Ex-Smokers	0.92 (0.43)	2.51	0.42 (0.19)	1.52*
Location * Income				
High Income Rural	0.89 (0.45)	2.43	0.60 (0.27)	1.82*
Middle Income Rural	0.36 (0.36)	1.44	0.30 (0.23)	1.34
Ethnicity * Income				
White* High SES	-0.90 (0.88)	0.40	-0.73 (0.42)	0.48
White * Middle SES	0.79 (0.57)	2.21	0.51 (0.26)	1.67*
Second Hand Smoke * Time				
Exposure * Cycle 2	0.03 (0.33)	1.03	0.06 (0.34)	1.06
Exposure * Cycle 3	0.004 (0.33)	1.00	0.11 (0.30)	1.11
Exposure * Cycle 4	0.06 (0.32)	1.06	0.10 (0.25)	1.11
Exposure * Cycle 5	0.19 (0.20)	1.21	0.27 (0.16)	1.31
Smoking * Age Group				
Smoker * 18-29 years	2.35 (2.45)	10.48	0.92 (0.83)	2.51
Ex-Smoker * 18-29 years	0.44 (0.61)	1.55	0.42 (0.41)	1.52
Smoker * 30-49 years	1.71 (2.43)	5.54	0.40 (0.69)	1.50
Ex-Smoker * 30-49 years	0.53 (0.56)	1.71	0.36 (0.38)	1.44
Smoker * 50-64 years	1.96 (2.22)	7.17	0.57 (0.71)	1.77
Ex-Smoker * 50-64 years	1.06 (0.59)	2.88	0.41 (0.39)	1.51
Age Group * Income				
18-29 years * High SES	-0.91 (0.74)	0.40	-0.85 (0.38)	0.43*
18-29 years * Middle SES	-0.16 (0.77)	0.85	-0.31 (0.43)	0.73
30-49 years * High SES	-0.92 (0.67)	0.40	-1.02 (0.34)	0.36**
30-49 years * Middle SES	-0.46 (0.55)	0.63	-0.44 (0.36)	0.64
50-64 years * High SES	-0.66 (0.71)	0.52	-0.72 (0.36)	0.49*
50-64 years * Middle SES	-0.27 (0.57)	0.76	-0.22 (0.30)	0.80

5 CONCLUSION

In this paper we compared the model based and design based methods. Data collected using multi-stage sampling and unequal probability of selection is common for large national databases. The analysis of data arising under such schemes are consequently less easy, even though expanding commercial software abilities have somewhat alleviated the complexity of this task. It is important to consider the design while analyzing the data as if not accounted for can result in biased and incorrect estimates. Research is needed towards survey designs that can accommodate both traditional complex elements (unequal selection probabilities, stratification, multi-stage sampling) and longitudinal collection of data.

The results shows that estimates obtained from both methods and using exchangeable correlation matrix are very close to each other. When we did not include sampling weights for model based and design based methods, the results obtained were too conservative for model based methods (table not provided). The weight variables used for purpose of analysis were specifically created for longitudinal survey design. Including the weight variables in the model produced correct estimates but the standard error obtained was incorrect. In order to calculate correct standard error we used linearized estimating function bootstrap technique (Binder, Kovacevic et al. 2004). Both the methods produced similar estimates but the standard errors were larger using design based methods. The method proposed by Rao (1998) does account for the complex sampling design as well as the longitudinal nature. Rao (1998) suggests that if the correlation matrix is not correctly specified, GEE method proposed by Liang and Zeger (1986) can result in inefficient estimates of β. The reason for such differences in the standard errors can be due to the reason that the proposed model doesn't fit the data well. When there is larger variation of weights, it can result in larger standard errors for the weighted estimators (Reiter, Zanutto et al. 2005). The larger standard errors can also be due to smaller sample size (as we are considering only female population ages 18-64 years) and larger variability in weights (Pfeffermann 1996).

The design based methods should be preferred as argued by Pfefferman (1996) as bias is the main issue in large sample surveys, and use of these methods removes bias so they should be preferred even if at the cost of large variances. The design-based estimators using the weighted estimates allows to obtain unbiased coefficient estimates of the independent variables in the regression model (Reiter, Zanutto et al. 2005). Binder (1983) and Kott (1991) suggest the use of design-based approaches as the weighted estimates provide unbiased estimates of the coefficients of the independent variables in the model, even when the other relevant independent variables are excluded from the model (Binder 1983; Kott 1991). The uses of design-based methods are limited. Large samples are required for hypothesis tests to be valid (Pfeffermann 1993). The generalization of the result obtained from design-based approach are not readily made to different populations as the inferences are specific to a particular finite population (Kalton 1983). The model-based methods have gained popularity over the design-based methods as these methods can be readily implemented using standard commercial software. Model-based approaches are more valid and powerful than design-based approaches when the model assumption is reasonable. The standard errors obtained using model based methods are smaller when compared to design based methods. This is because unweighted estimates have smaller variances, as the variation in magnitude of weights are not included (Reiter, Zanutto, 2005). However, if the model does not fit the data well, biased estimation can result. Model-based standards errors have smaller standard errors than design-based estimators (Pfeffermann 1993).

Another approach to obtain valid inferences with model-based methods, the design variables like age groups, race, sex , socio-economic status should also be included in the model (Reiter, Zanutto et al. 2005). In this paper we had included the design variables (province and age) in the model, with the exception of sex as we were considering only the female gender. The result of both the methods used in this study should have been closer but there were differences and it could be attributed to smaller sample size (as discussed above) we were studying a real smaller subpopulation. Studies will be conducted with a larger sample size, to check the hypothesis that model based and design based methods provide similar results. It is important to consider the relative advantages of these two approaches carefully.

ACKNOWLEDGEMENT

The authors of the manuscript would like to thank Susana Rubin-Bluer and Abdul Nasser Saidi of Statistics Canada for their help with macro to compute variance using Rao's method.

REFERENCES

Statistics Canada, (2004). National Population Health Survey Household Component, Cycle 5 (2002-2003) Longitudinal Component Documentation, Statistics Canada.

Binder, D. (1983). "On the variances of Asymptotically Normal estimators from Complex Surveys." *International Statistical Review* **51**(2), 279-292.

Binder, D., M. Kovacevic, et al. (2004). *Design Based Methods for Survey Data: Alternative uses of estimating Functions*. Proceeding of the Section on Survey Resrearch Methods of the American Statistical Association.

Feder, M., G. Nathan, et al. (2000). "Multilevel Modeling of Complex Survey Longitudinal Data with Time Varying Random Effects." *Survey Methodology* **26**(1), 53-65.

Fitzmaurice, G. M. and N. M. Laird (1993). "A likelihood based method for analysing longitudinal binary data." *Biometrika* **80,** 141-151.

Kalton, G. (1983). "Model in the Practice of Survey Sampling." *International Statistical Review* **51**(1), 175-188.

Kott, P. S. (1991). "A model based look at the Linear Regression with Survey Data." *The American Statistician*. **45**(2), 107-112.

LaVange, L. M., G. G. Koch, et al. (2001). "Applying sample survey methods to clinical trials data." St*atistics in Medicine* **20**: 2609-2623.

Lawless, J. F. (2003). Event history analysis and longitudinal surveys. *Analysis of Survey data*. R. L. Chambers and C. J. Skinner, John Wiley and Sons.

Lehtonen, R. and E. Pahkinen (2004). *Practical Methods for Design and Analysis of Complex Surveys*, John Wiley & Sons.

Liang, K. Y. and S. L. Zeger (1986). "Longitudinal data analysis using generalized linear models." *Biometrics* **73**: 13-22.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. London, Chapman & Hall.

Pfeffermann, D. (1993). "The role of sampling weights when Modelling Survey Data." *International Statistical Review* **61**(2), 317-337.

Pfeffermann, D. (1996). "The use of sampling weights for Survey Data Analysis." *Statistical Methods in Medical Research* **5**(1), 239-261.

Rao, J. N. K. (1998). Marginal Models for reapeated observations: Inference with survey data. *Proceedings of the Survey Research Methods Section*.

Reiter, J. P., E. L. Zanutto, et al. (2005). "Analytical Modeling in Complex Surveys of Work Practices." *Industrial and Labor Relations Review* **59**(1), 82-100.

Skinner, C. J. and D. J. Holmes (2003). Random effects models for longitudinal survey data. *Analysis of survey data*. R. L. Chambers and C. J. Skinner, John Wiley and Sons.

Wedderburn, R. W. M. (1974). "Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method." *Biometrika* **61,** 439-447.

Zeger, S. L. and K. Y. Liang (1986). "Longitudinal data analysis for discrete and continuous outcomes." *Biometrics* **42**(1) 121-30.